James E. Gentle

Theory of Statistics

Preface: Mathematical Statistics

After teaching mathematical statistics for several years using chalk on a blackboard (and, later, smelly "dry erase markers" on a whiteboard) mostly doing proofs of theorems, I decided to lecture from computer slides that provide an outline of the "big picture". Rather than spend class time "doing" proofs that are given in standard texts, I decided that time would be better spent discussing the material from a different, higher-level perspective.

While lecturing from canned slides, I cannot, however, ignore certain details of proofs and minutiae of examples. But what details and which minutiae? To be effective, my approach depends on an understanding between students and the instructor, an understanding that is possibly implicit. I lecture; but I ask "what is ... ?" and "why is ... ?"; and I encourage students to ask "what is ... ?" and "why is ... ?". I adopt the attitude that there are many things that I don't know, but if there's something that I wonder about, I'll admit ignorance and pursue the problem until I've attained some resolution. I encourage my students to adopt a similar attitude.

I am completely dissatisfied with a class that sits like stumps on a log when I ask "what is ... ?" or "why is ... ?" during a lecture. What can I say?

After writing class slides (in LaTeX 2_ε, of course), mostly in bullet form, I began writing text around the bullets, and I put the notes on the class website. Later I decided that a single document with a fairly extensive subject index (see pages **??** through **??**) would be useful to serve as a companion for the study of mathematical statistics. The other big deal about this document is the internal links, which of course is not something that can be done with a hardcopy book. (One thing I must warn you about is that there is a (known) bug in the LaTeX package hyperref; if the referenced point happens to occur at the top of a page, the link takes you to the previous page – so if you don't see what you expect, try going to the next page.)

Much of the present document reflects its origin as classroom notes; it contains incomplete sentences or sentence fragments, and it lacks connective material in some places. (The connective material was (probably!) supplied orally during the lectures.)

Another characteristic of this document that results from the nature of its origin, as well as from the medium itself (electronic) is its length. A long book doesn't use any more paper or "kill any more trees" than a short book. Usually, however, we can expect the density of "importance" in a short document to be greater than that in a long document. If I had more time I could make this book shorter by prioritizing its content, and I may do that someday.

Several sections are incomplete and several proofs are omitted. Also, I plan to add more examples. *I just have not had time to type up the material.*

I do not recommend that you print these notes. First of all, they are evolving, so a printed version is just a snapshot. Secondly, the PDF file contains active internal links, so navigation is easy. (For versions without active links, I try to be friendly to the reader by providing page numbers with most internal references.)

This document is directed toward students for whom the theory of statistics is or will become an important part of their lives. Obviously, such students should be able to work through the details of "hard" proofs and derivations; that is, students should master the fundamentals of mathematical statistics. In addition, students at this level should acquire, or begin acquiring, a deep appreciation for the field, including its historical development and its relation to other areas of mathematics and science generally; that is, students should master the fundamentals of the broader theory of statistics. Some of the chapter endnotes are intended to help students gain such an appreciation by leading them to background sources and also by making more subjective statements than might be made in the main body.

It is important to understand the intellectual underpinnings of our science. There are certain *principles* (such as sufficiency, for example) that guide our approaches to statistical inference. There are various general approaches (see page 239) that we follow. Within these general approaches there are a number of specific methods (see page 240). The student should develop an appreciation for the relations between principles, approaches, and methods.

This book on mathematical statistics assumes a certain amount of background in mathematics. Following the final chapter on mathematical statistics Chapter 8, there is Chapter 0 on "statistical mathematics" (that is, mathematics with strong relevance to statistical theory) that provides much of the general mathematical background for probability and statistics. The mathematics in this chapter is prerequisite for the main part of the book, and it is hoped that the reader already has command of the material; otherwise, Chapter 0 can be viewed as providing "just in time" mathematics. Chapter 0 grew (and is growing) recursively. Every time I add material, it seems that I need to add some background for the new material. This is obviously a game one cannot win.

Probability theory is the most directly relevant mathematical background, and it is assumed that the reader has a working knowledge of measure-theory-based probability theory. Chapter 1 covers this theory at a fairly rapid pace.

The objective in the discussion of probability theory in Chapter 1, as in that of the other mathematical background, is to provide some of the most relevant material for statistics, which is the real topic of this text. Chapter 2 is also on probability, but the focus is on the applications in statistics. In that chapter, I address some important properties of probability distributions that determine properties of statistical methods when applied to observations from those distributions.

Chapter 3 covers many of the fundamentals of statistics. It provides an overview of the topics that will be addressed in more detail in Chapters 4 through 8.

This document is organized in the order in which I cover the topics (more-or-less!). Material from Chapter 0 may be covered from time to time during the course, but I generally expect this chapter to be used for reference as needed for the statistical topics being covered.

The primary textbooks I have used in the past few years are Shao (2003), Lehmann and Casella (1998), and Lehmann (1986) (the precursor to Lehmann and Romano (2005)). At various places in this document, references are given to the related sections of Shao (2003) ("MS2"), Lehmann and Casella (1998) ("TPE2"), and Lehmann and Romano (2005) ("TSH3"). These texts state all of the important theorems, and in most cases, provide the proofs. They are also replete with examples. Full bibliographic citations for these references, as well as several other resources are given in the bibliography beginning on page 873.

It is of course expected that the student will read the primary textbook, as well as various other texts, and to work through all proofs and examples in the primary textbook. As a practical matter, obviously, even if I attempted to cover all of these in class, there just is not enough class time to do it.

The purpose of this evolving document is not just to repeat all of the material in those other texts. Its purpose, rather, is to provide some additional background material, and to serve as an outline and a handy reference of terms and concepts. The nature of the propositions vary considerably; in some cases, a fairly trivial statement will be followed by a proof, and in other cases, a rather obtuse statement will not be supported by proof. In all cases, the student should understand why the statement is true (or, if it's not, immediately send me email to let me know of the error!). More importantly, the student should understand why it's relevant.

Each student should read this and other texts and work through the proofs and examples at a rate that matches the individual student's understanding of the individual problem. What one student thinks is rather obtuse, another student comprehends quickly, and then the tables are turned when a different problem is encountered. There is a lot of lonely work required, and this is why lectures that just go through the details are often not useful.

It is commonplace for textbooks in mathematics to include examples and exercises without reference to the source of the examples or exercises

and yet without implying any claim of originality. (A notable exception is Graham et al. (1994).) My book is not intended to present new and original work, and it follows the time-honored tradition of reusing examples and exercises from long-forgotten sources.

Notation

Adoption of notation is an overhead in communication. I try to minimize that overhead by using notation that is "standard", and using it locally consistently.

Examples of sloppy notation abound in mathematical statistics. Functions seem particularly susceptible to abusive notation. It is common to see "$f(x)$" and "$f(y)$" used in the same sentence to represent two different functions. (These often represent two different PDFs, one for a random variable X and the other for a random variable Y. When I want to talk about two different things, I denote them by different symbols, so when I want to talk about two different PDFs, I often use notation such as "$f_X(\cdot)$" and "$f_Y(\cdot)$". If $x = y$, which is of course very different from saying $X = Y$, then $f_X(x) = f_X(y)$; however, $f_X(x) \neq f_Y(x)$ in general.)

For a function and a value of a function, there is a certain amount of ambiguity that is almost necessary. I generally try to use notation such as "$f(x)$" or "$Y(\omega)$" to denote the value of the function f at x or Y at ω, and I use "f", "$f(\cdot)$", or "Y" to denote the function itself (although occasionally, I do use "$f(x)$" to represent the function — notice the word "try" in this discussion).

If in the notation "$f(x)$", "x" denotes an element of a set A, and $B \subseteq A$ (that is, B is a set of the same kinds of elements as A), then "$f(B)$" does not make much sense. For the image of B under f, I use "$f[B]$".

I also freely use the notation $f^{-1}(y)$ or $f^{-1}[B]$ to denote the preimage, whether or not f^{-1} is actually a function; that is, whether or not f is invertible.

There are two other areas in which my notation may be slightly different from common notation. First, to designate the open interval between the real numbers $a < b$, I use the Bourbaki notation "$]a, b[$". (I eschew most of the weird Bourbaki notation, however. This particular notation is also used in my favorite book on analysis, Hewitt and Stromberg (1965).) Second, I do not use any special notation, such as boldface, for vectors; thus, x may represent a scalar or a vector.

All vectors are "column vectors". This is only relevant in vector-vector or vector-matrix operations, and has nothing to do with the way we represent a vector. It is far more convenient to represent the vector x as

$$x = (x_1, \ldots, x_d)$$

than as

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix},$$

and there is certainly no need to use the silly notation

$$x = (x_1, \ldots, x_d)^{\mathrm{T}}.$$

A vector is not a matrix.

There are times, however, when a vector may be treated like a matrix in certain operations. In such cases, the vector is treated as a matrix with one column.

Appendix C provides a list of the common notation that I use. The reader is encouraged to look over that list both to see the notation itself and to get some idea of the objects that I discuss.

Solving Problems

The main ingredient for success in a course in mathematical statistics is the ability to work problems. The only way to enhance one's ability to work problems is to *work problems. It is not sufficient to read, to watch, or to hear solutions to problems.* One of the most serious mistakes students make in courses in mathematical statistics is to work through a solution that somebody else has done and to think they have worked the problem.

While sometimes it may not be possible to solve a given problem, rather than looking for a solution that someone else has come up with, it is much better to stop with a partial solution or a hint and then sometime later return to the effort of completing the solution. *Studying a problem without its solution is much more worthwhile than studying the solution to the problem.*

Do you need to see a solution to a problem that you have solved? Except in rare cases, if you have solved a problem, you know whether or not your purported solution is correct. It is like a Sudoku puzzle; although solutions to these are always published in the back of the puzzle book or in a later edition of the medium, I don't know what these are for. If you have solved the puzzle you know that your solution is correct. If you cannot solve it, I don't see any value in looking at the solution. It's not going to make you a better Sudoku solver. (Sudoku is different from crossword puzzles, another of my pastimes. Seeing the solution or partial solution to a crossword puzzle can make you a better crossword solver.) There is an important difference in Sudoku puzzles and mathematical problems. In Sudoku puzzles, there is only one correct solution. In mathematical problems, there may be more than one way to solve a problem, so occasionally it is worthwhile to see someone else's solution.

The common wisdom (or cliché, depending on your viewpoint) that it takes 10000 hours to master a field or a skill is probably applicable to statistics.

This means working on this stuff for about 40 hours a week for 5 years. This is approximately the amount of work that a student should do for receipt of a PhD degree (preferably in less than 5 years).

Many problems serve as models of "standard operations" for solving other problems. Some problems should become "easy pieces".

Standard Operations

There are a number of operations and mathematical objects that occur over and over in deriving results or in proving propositions. These operations are sometimes pejoratively called "tricks". In some sense, perhaps they are; but it is useful to identify these operations outside of the context of a specific application. Some of these standard operations and useful objects are listed in Section 0.0.9 on page 676.

Easy Pieces

I recommend that all students develop a list of easy pieces. These are propositions or examples and counterexamples that the student can state and prove or describe and work through *without resort to notes*. An easy piece is something that is important in its own right, but also may serve as a model or template for many other problems. A student should attempt to accumulate a large bag of easy pieces. If development of this bag involves some memorization, that is OK, but things should just naturally get into the bag in the process of working problems and observing similarities among problems — and by seeing the same problem over and over.

Some examples of easy pieces are

- State and prove the information inequality (CRLB) for a d-vector parameter. (Get the regularity conditions correct.)
- Give an example to distinguish the asymptotic bias from the limiting bias.
- State and prove Basu's theorem.
- Give an example of a function of some parameter in some family of distributions that is not U-estimable.
- A statement of the Neyman-Pearson Lemma (with or without the randomization) and its proof.

Some easy pieces in the background area of "statistical mathematics" are

- Let $\mathcal{C}$ be the class of all closed intervals in $\mathbb{R}$. Show that $\sigma(\mathcal{C}) = \mathcal{B}(\mathbb{R})$ (the real Borel σ-field).
- Define induced measure and prove that it is a measure. That is, prove: If $(\Omega, \mathcal{F}, \nu)$ is a measure space and $(\Lambda, \mathcal{G})$ is a measurable space, and f is a function from Ω to Λ that is measurable with respect to $\mathcal{F}/\mathcal{G}$, then $\nu \circ f^{-1}$ is a measure with domain $\mathcal{G}$.
- Define the Lebesgue integral for a general Borel function.

- State and prove Fatou's lemma conditional on a sub-σ-field.

Make your own list of easy pieces.

Relevance and Boundaries

For any proposition or example, you should have a clear understanding of *why* the proposition or example is important. Where is it subsequently used? Is it used to prove something else important, or does it justify some statistical procedure?

Propositions and definitions have boundaries; that is, they apply to a specific situation. You should look at the "edges" or "boundaries" of the hypotheses. What would happen if you were to remove one or more assumptions? (This is the idea behind counterexamples.) What would happen if you make stronger assumptions?

"It is clear" and "It can be shown"

I tend to use the phrase "it is clear ..." often. (I only realized this recently, because someone pointed it out to me.) When I say "it is clear ...", I expect the reader to agree with me actively, not passively.

I use this phrase only when the statement is "clearly" true to me. I must admit, however, sometimes when I read the statement a few weeks later, it's not very clear! It may require many minutes of difficult reasoning. In any event, the reader should attempt to supply the reasoning for everything that I say is clear.

I also use the phrase "it can be shown ..." in connection with a fact (theorem) whose proof at that point would be distracting, or else whose proof I just don't want to write out. (In later iterations of this document, however, I may decide to give the proof.) A statement of fact preceded by the phrase "it can be shown", is likely to require more thought or background information than a statement preceded by the phrase "it is clear", although this may be a matter of judgement.

Study of mathematical statistics at the level appropriate for this document is generally facilitated by reference to a number of texts and journal articles; and I assume that the student does refer to various sources.

My Courses

The courses in mathematical statistics at George Mason University are CSI/STAT 972 and CSI/STAT 973. The prerequisites for these courses include measure-theoretic-based probability theory, such as is covered in CSI/STAT 971. Chapters 0 and 1 address the prerequisite material briefly, and in CSI/STAT 972 some class time is devoted to this material. Although

Chapters 1 and 2 are on "probability", some of their focus is more on what is usually covered in "statistics" courses, such as families of distributions, in particular, the exponential class of families.

My notes on these courses are available at

`http://mason.gmu.edu/~jgentle/csi9723/`

Acknowledgements

A work of this size is bound to have some errors (at least if I have anything to do with it!). Errors must first be detected and then corrected.

I would appreciate any feedback – errors, comments, or suggestions. Email me at `jgentle@gmu.edu`

Fairfax County, Virginia James E. Gentle
March 20, 2020

Contents

Appendices

1

Probability Theory

Probability theory provides the basis for mathematical statistics.

Probability theory has two distinct elements. One is just a special case of measure theory and can be approached in that way. For this aspect, **the presentation in this chapter assumes familiarity with the material in Section 0.1** beginning on page 692. This aspect is "pure" mathematics. The other aspect of probability theory is essentially built on a gedanken experiment involving drawing balls from an urn that contains balls of different colors, and noting the colors of the balls drawn. In this aspect of probability theory, we may treat "probability" as a primitive (that is, undefined) concept. In this line of development, we relate "probability" informally to some notion of long-term frequency or to expectations or beliefs relating to the types of balls that will be drawn. Following some additional axiomatic developments, however, this aspect of probability theory is also essentially "pure" mathematics.

Because it is just mathematics, in probability theory *per se*, we do not ask "what do you think is the probability that ...?" Given an axiomatic framework, one's beliefs are irrelevant, whether probability is a measure or is a primitive concept. In statistical theory or applications, however, we may ask questions about "beliefs", and the answer(s) may depend on deep philosophical considerations in connecting the mathematical concepts of probability theory with decisions about the "real world". This may lead to a different definition of *probability*. For example, Lindley and Phillips (1976), page 115, state "Probability is a relation between you and the external world, expressing your opinion of some aspect of that world..." I am sure that an intellectually interesting theory could be developed based on ways of "expressing your opinion[s]", but I will not use "probability" in this way; rather, throughout this book, I will use the term *probability* as a *measure* (see Definition 0.1.10, page 704). For specific events in a given application, of course, certain values of the probability measure may be assigned based on "opinions", "beliefs", or whatever.

Another useful view of "probability" is expressed by Gnedenko and Kolmogorov (1954) (page 1): "The very concept of mathematical *probability* [their emphasis] would be fruitless if it did not find its realization in the *frequency* [their

emphasis] of occurrence of events under large-scale repetition of uniform conditions." (See a more complete quotation on page 143.)

Ranks of Mathematical Objects

In probability theory we deal with various types of mathematical objects. We would like to develop concepts and identify properties that are independent of the type of the underlying objects, but that is not always possible. Occasionally we will find it necessary to discuss scalar objects, rank one objects (vectors), and rank two objects (matrices) separately. In general, most degree-one properties, such as expectations of linear functions, can be considered uniformly across the different types of mathematical objects. Degree-two properties, such as variances, however, must usually be considered separately for scalars, vectors, and matrices.

Overview of Chapter

This chapter covers important topics in probability theory at a fairly fast pace. Some of the material in this chapter, such as the properties of certain families of distributions, is often considered part of "mathematical statistics", rather than a part of probability theory. Unless the interest is in use of *data* for describing a distribution or for making inferences about the distribution, however, the study of properties of the distribution is part of probability theory, rather than statistics.

We begin in Section 1.1 with statements of definitions and some basic properties. The initial development of this section parallels the first few subsections of Section 0.1 for more general measures, and then the development of expectations depends on the results of Section 0.1.6 for integration.

Sections 1.3 and 1.4 are concerned with sequences of independent random variables. The limiting properties of such sequences are important. Many of the limiting properties can be studied using expansions in power series, which is the topic of Section 1.2.

In Section 1.5 we do a fairly careful development of the concept of conditioning. We do not take conditional probability to be a fundamental concept, as we take (unconditional) probability itself to be. Conditional probability, rather, is based on conditional expectation as the fundamental concept, so we begin that discussion by considering conditional expectation. This provides a more general foundation for conditional probability than we would have if we defined it more directly in terms of a measurable space. Conditional probability plays an important role in sequences that lack a simplifying assumption of independence. We discuss sequences that lack independence in Section 1.6. Many interesting sequences also do not have identical marginal distributions, but rather follow some kind of evolving model whose form depends on, but is not necessarily determined by, previous variates in the sequence.

In the next chapter, beginning on page 155, I identify and describe useful classes of probability distributions. These classes are important because they are good models of observable random phenomena, and because they are easy to work with. The properties of various statistical methods discussed in subsequent chapters depend on the underlying probability model, and some of the properties of the statistical methods can be worked out easily for particular models discussed in Chapter 2.

1.1 Some Important Probability Definitions and Facts

A probability distribution is built from a measure space in which the measure is a probability measure.

Definition 1.1 (probability measure)
A measure ν whose domain is a σ-field defined on the sample space Ω with the property that $\nu(\Omega) = 1$ is called a *probability measure*. ∎

We often use P to denote a probability measure, just as we often use λ, μ, or ν to denote a general measure.

Properties of the distribution and statistical inferences regarding it are derived and evaluated in the context of the "probability triple",

$$(\Omega, \mathcal{F}, P). \tag{1.1}$$

Definition 1.2 (probability space)
If P in the measure space $(\Omega, \mathcal{F}, P)$ is a probability measure, the triple $(\Omega, \mathcal{F}, P)$ is called a *probability space*. ∎

Probability spaces are the basic structures we will consider in this chapter. In a probability space $(\Omega, \mathcal{F}, P)$, a set $A \in \mathcal{F}$ is called an "event".

The full σ-field $\mathcal{F}$ in the probability space $(\Omega, \mathcal{F}, P)$ may not be necessary to define the space.

Definition 1.3 (determining class)
If P and Q are probability measures defined on the measurable space $(\Omega, \mathcal{F})$, a collection of sets $\mathcal{C} \subseteq \mathcal{F}$ is called a *determining class* of P and Q, iff

$$P(A) = Q(A)\ \forall A \in \mathcal{C} \Longrightarrow P(B) = Q(B)\ \forall B \in \mathcal{F}.$$

∎

For measures P and Q defined on the measurable space $(\Omega, \mathcal{F})$, the condition $P(B) = Q(B)\,\forall B \in \mathcal{F}$, of course, is the same as the condition $P = Q$.

Notice that the determining class is not necessarily a sub-σ-field. If it is, however, a probability measure on the measurable space of the determining class results in a probability space that is essentially the same as that formed

by the probability measure on the original measurable space. That is, in the notation of Definition 1.3, if $\mathcal{C}$ is a determining class, then the probability space $(\Omega, \sigma(\mathcal{C}), P)$ is essentially equivalent to $(\Omega, \mathcal{F}, P)$ in so far as properties of the probability measure are concerned.

We now define complete probability measures and spaces as special cases of complete measures and complete measure spaces (Definitions 0.1.16 and 0.1.21 on pages 707 and 709). Completeness is often necessary in order to ensure convergence of sequences of probabilities.

Definition 1.4 (complete probability space)
A probability measure P defined on the σ-field $\mathcal{F}$ is said to be *complete* if $A_1 \subseteq A \in \mathcal{F}$ and $P(A) = 0$ implies $A_1 \in \mathcal{F}$. If the probability measure P in the probability space $(\Omega, \mathcal{F}, P)$ is complete, we also say that the probability space is a *complete probability space*. ∎

An event A such that $P(A) = 0$ is called a negligible event or negligible set. For a set A_1 that is a subset of a negligible set, as in Definition 1.4, it is clear that A_1 is also negligible.

Definition 1.5 (almost surely (a.s.))
Given a probability space $(\Omega, \mathcal{F}, P)$, a property that holds for all elements of $\mathcal{F}$ with positive probability is said to hold *almost surely*, or a.s. ∎

This is the same as almost everywhere for general measures, and there is no essential difference in "almost everywhere" and "almost surely".

1.1.1 Probability and Probability Distributions

The elements in the probability space can be any kind of objects. They do not need to be numbers. Later, on page 9, we will define a real-valued measurable function (to be called a "random variable"), and consider the measure on $\mathbb{R}$ induced or "pushed forward" by this function. (See page 712 for definition of an induced measure.) This induced measure, which is usually based either on the counting measure (defined on countable sets as their cardinality) or on the Lebesgue measure (the length of intervals), is also a probability measure.

First, however, we continue with some definitions that do not involve random variables.

Probability Measures on Events; Independence and Exchangeability

Definition 1.6 (probability of an event)
In a probability space $(\Omega, \mathcal{F}, P)$, the *probability of the event* A is $P(A)$. This is also written as $\Pr(A)$. ∎

In the probability space $(\Omega, \mathcal{F}, P)$, for $A \in \mathcal{F}$, we have

$$\Pr(A) = P(A) = \int_A \mathrm{d}P. \tag{1.2}$$

We use notation such as "$\Pr(\cdot)$", "$\mathrm{E}(\cdot)$", "$\mathrm{V}(\cdot)$", and so on (to be introduced later) as generic symbols to represent specific quantities within the context of a given probability space. Whenever we discuss more than one probability space, it may be necessary to qualify the generic notation or else use an alternative notation for the same concept. For example, when dealing with the probability spaces $(\Omega, \mathcal{F}, P)$ and $(\Lambda, \mathcal{G}, Q)$, we may use notation of the form "$\Pr_P(\cdot)$" or "$\Pr_Q(\cdot)$"; but in this case, of course, the notation "$P(\cdot)$" or "$Q(\cdot)$"is simpler.

One of the most important properties that involves more than one event or more than one function or more than one measurable function is *independence*. We define independence in a probability space in three steps.

Definition 1.7 (independence)
Let $(\Omega, \mathcal{F}, P)$ be a probability space.

1. **Independence of events** (within a collection of events).
   Let $\mathcal{C}$ be a collection of events; that is, a collection of subsets of $\mathcal{F}$. The *events in* $\mathcal{C}$ *are independent* iff for a positive integer n and distinct events $A_1, \ldots, A_n$ in $\mathcal{C}$,
   $$P(A_1 \cap \cdots \cap A_n) = P(A_1) \cdots P(A_n). \tag{1.3}$$
2. **Independence of collections of events** (and, hence, of σ-fields).
   For any index set $\mathcal{I}$, let $\mathcal{C}_i$ be a collection of sets with $\mathcal{C}_i \subseteq \mathcal{F}$. The *collections* $\mathcal{C}_i$ *are independent* iff the events in any union of the $\mathcal{C}_i$ are independent; that is, $\{A_i \in \mathcal{C}_i \, : \, i \in \mathcal{I}\}$ are independent events.
3. **Independence of Borel functions** (and, hence, of random variables, which are special functions defined below).
   For i in some index set $\mathcal{I}$, the *Borel-measurable functions* X_i *are independent* iff $\sigma(X_i)$ for $i \in \mathcal{I}$ are independent.

∎

While we have defined independence in terms of a single probability measure (which gives meaning to the left side of equation (1.3)), we could define the concept over different probability spaces in the obvious way that requires the probability of all events simultaneously to be the product of the probabilities of the individual events.

Notice that Definition 1.7 provides meaning to mixed phrases such as "the event A is independent of the σ-field $\mathcal{F}$" or "the random variable X (defined below) is independent of the event A".

We will often consider a *sequence* or a *process* of events, σ-fields, and so on. In this case, the collection $\mathcal{C}$ in Definition 1.7 is a sequence. For events $\mathcal{C} =$

$A_1, A_2, \ldots$, which we may write as $\{A_n\}$, we say the sequence is a sequence of independent events. We also may abuse the terminology slightly and say that "the sequence is independent". Similarly, we speak of independent sequences of collections of events or of Borel functions.

Notice that each pair of events within a collection of events may be independent, but the collection itself is not independent.

Example 1.1 pairwise independence
Consider an experiment of tossing a coin twice. Let

A be "heads on the first toss"

B be "heads on the second toss"

C be "exactly one head and one tail on the two tosses"

We see immediately that any pair is independent, but that the three events are not independent; in fact, the intersection is $\emptyset$. ∎

We refer to this situation as "pairwise independent". The phrase "mutually independent", is ambiguous, and hence, is best avoided. Sometimes people use the phrase "mutually independent" to try to emphasize that we are referring to independence of *all* events, but the phrase can also be interpreted as "pairwise independent".

Notice that an event is independent of itself if its probability is 0 or 1.

If collections of sets that are independent are closed wrt intersection, then the σ-fields generated by those collections are independent, as the following theorem asserts.

Theorem 1.1
Let $(\Omega, \mathcal{F}, P)$ be a probability space and suppose $\mathcal{C}_i \subseteq \mathcal{F}$ for $i \in \mathcal{I}$ are independent collections of events. If $\forall i \in \mathcal{I}, A, B \in \mathcal{C}_i \Rightarrow A \cap B \in \mathcal{C}_i$, then $\sigma(\mathcal{C}_i)$ for $i \in \mathcal{I}$ are independent.

Proof. Exercise. ∎

Independence also applies to the complement of a set, as we see next.

Theorem 1.2
Let $(\Omega, \mathcal{F}, P)$ be a probability space. Suppose $A, B \in \mathcal{F}$ are independent. Then A and B^c are independent.

Proof. We have

$$P(A) = P(A \cap B) + P(A \cap B^c),$$

hence,

$$\begin{aligned} P(A \cap B^c) &= P(A)(1 - P(B)) \\ &= P(A)P(B^c), \end{aligned}$$

and so A and B^c are independent. ∎

In the interesting cases in which the events have equal probability, a concept closely related to independence is *exchangeability*. We define exchangeability in a probability space in three steps, similar to those in the definition of independence.

Definition 1.8 (exchangeability)
Let $(\Omega, \mathcal{F}, P)$ be a probability space.

1. **Exchangeability of events** within a collection of events.
   Let $\mathcal{C} = \{A_i : i \in \mathcal{I}\}$ for some index set $\mathcal{I}$ be a collection of events; that is, a collection of subsets of $\mathcal{F}$. Let n be any positive integer (less than or equal to $\#(\mathcal{C})$ if $\#(\mathcal{C}) < \infty$) and let $\{i_1, \ldots, i_n\}$ and $\{j_1, \ldots, j_n\}$ each be sets of distinct positive integers in $\mathcal{I}$. The *events in* $\mathcal{C}$ *are exchangeable* iff for any positive integer n and distinct events $A_{i_1}, \ldots, A_{i_n}$ and distinct events $A_{j_1}, \ldots, A_{j_n}$ in $\mathcal{C}$,
   $$P(\cup_{k=1}^{n} A_{i_k}) = P(\cup_{k=1}^{n} A_{j_k}). \tag{1.4}$$
2. **Exchangeability of collections of events** (and, hence, of σ-fields).
   For any index set $\mathcal{I}$, let $\mathcal{C}_i$ be a collection of sets with $\mathcal{C}_i \subseteq \mathcal{F}$. The *collections* $\mathcal{C}_i$ *are exchangeable* iff the events in any collection of the form $\{A_i \in \mathcal{C}_i : i \in \mathcal{I}\}$ are exchangeable.
3. **Exchangeability of Borel functions** (and, hence, of random variables, which are special functions defined below).
   For i in some index set $\mathcal{I}$, the *Borel-measurable functions* X_i *are exchangeable* iff $\sigma(X_i)$ for $i \in \mathcal{I}$ are exchangeable.
   (This also defines exchangeability of any generators of σ-fields.)

∎

Notice that events being exchangeable requires that they have equal probabilities.

As mentioned following Definition 1.7, we will often consider a sequence or a process of events, σ-fields, and so on. In this case, the collection $\mathcal{C}$ in Definition 1.8 is a sequence, and we may say the sequence $\{A_n\}$ is a sequence of exchangeable events. Similarly, we speak of exchangeable sequences of collections of events or of Borel functions. As with independence, we also may abuse the terminology slightly and say that "the sequence is exchangeable".

For events with equal probabilities, independence implies exchangeability, but exchangeability does not imply independence.

Theorem 1.3
Let $(\Omega, \mathcal{F}, P)$ be a probability space and suppose $\mathcal{C} \subseteq \mathcal{F}$ is a collection of independent events with equal probabilities. Then $\mathcal{C}$ is an exchangeable collection of events.

Proof. Exercise. ∎

The next example shows that exchangeability does not imply independence.

Example 1.2 independence and exchangeability
A simple urn example may illustrate the difference in exchangeability and independence. Suppose an urn contains 3 balls, 2 of which are red and 1 of

which is not red. We "randomly" draw balls from the urn without replacing them (that is, if there are n balls to draw from, the probability that any specific one is drawn is $1/n$).

Let R_i be the event that a red ball is drawn on the i^{th} draw, and R_i^{c} be the event that a non-red ball is drawn. We see the following

$$\Pr(R_1) = \Pr(R_2) = \Pr(R_3) = 2/3$$

and

$$\Pr(R_1^{\text{c}}) = \Pr(R_2^{\text{c}}) = \Pr(R_3^{\text{c}}) = 1/3.$$

Now

$$\Pr(R_1 \cap R_2) = 1/3;$$

hence, R_1 and R_2 are not independent. Similarly, we can see that R_1 and R_3 are not independent and that R_2 and R_3 are not independent. Hence, the collection $\{R_1, R_2, R_3\}$ is certainly not independent (in fact, $\Pr(R_1 \cap R_2 \cap R_3) = 0$). The events R_1, R_2, and R_3 are exchangeable, however. The probabilities of singletons are equal and of course the probability of the full set is equal to itself however it is ordered, so all we need to check are the probabilities of the doubletons:

$$\Pr(R_1 \cup R_2) = \Pr(R_1 \cup R_3) = \Pr(R_2 \cup R_3) = 1.$$

∎

Using a binomial tree, we could extend the computations in the preceding example for an urn containing n balls m of which are red, with the events R_i defined as before, to see that the elements of any subset of the m R_is is exchangeable, but that they are not independent.

While, of course, checking definitions explicitly is necessary, it is useful to develop an intuition for such properties as independence and exchangeability. A little simple reasoning about the urn problem of Example 1.2 should provide heuristic justification for exchangeability as well as for the lack of independence.

1.1.2 Random Variables

In many applications of probability concepts, we define a measurable function X, called a random variable, from $(\Omega, \mathcal{F})$ to $(\mathbb{R}^d, \mathcal{B}^d)$:

$$X : (\Omega, \mathcal{F}) \mapsto \left(\mathbb{R}^d, \mathcal{B}^d\right). \tag{1.5}$$

The random variable, together with a probability measure, P, on the measurable space $(\Omega, \mathcal{F})$ determines a new probability space $(\mathbb{R}^d, \mathcal{B}^d, P \circ X^{-1})$.

We can study the properties of the probability space $(\Omega, \mathcal{F}, P)$ through the random variable and the probability space $(\mathbb{R}^d, \mathcal{B}^d, P \circ X^{-1})$, which is

easier to work with because the sample space, $X[\Omega]$, is $\mathbb{R}^d$ or some subset of it, rather than some abstract set Ω. In most applications, it is more natural to begin with Ω as some subset, $\mathcal{X}$, of $\mathbb{R}^d$, to develop a vague notion of some σ-field on $\mathcal{X}$, and to define a random variable that relates in a meaningful way to the problem being studied.

The mapping of the random variable allows us to assign meaning to the elements of Ω consistent with the application of interest. The properties of one space carry over to the other one, subject to the random variable, and we may refer to objects of either space equivalently. Random variables allow us to develop a theory of probability that is useful in statistical applications.

Definition 1.9 (random variable)
A measurable function, $X(\omega)$ or just X, from a measurable space $(\Omega, \mathcal{F})$ to the measurable space $(\mathbb{R}^d, \mathcal{B}^d)$ is called a *random variable*, or, to be more specific, a *d-variate random variable*. ∎

This definition means that "Borel function" (see page 719) and "random variable" are synonymous. Notice that the words "random" and "variable" do not carry any separate meaning.

Many authors define a random variable only for the case $d = 1$, and for the case of $d \geq 1$, call the Borel function a "random vector". I see no reason for this distinction. Recall that I use "real" to refer to an element of $\mathbb{R}^d$ for any positive integer d. My usage is different from an alternate usage in which "real" means what I call a "real scalar"; in that alternate usage, a random variable takes values only in $\mathbb{R}$.

We often denote the image of X, that is, $X[\Omega]$, as $\mathcal{X}$. If $B \in \mathcal{B}(\mathcal{X})$, then $X^{-1}[B] \in \mathcal{F}$.

Although we define the random variable X to be real, we could form a theory of probability and statistics that allowed X to be a function into a general field. Complex-valued random variables are often useful, especially, for example, in harmonic analysis of such things as electrical signals, but we will not consider them in any detail in this text.

Notice that a random variable is finite a.s. If this were not the case, certain problems would arise in the definitions of some useful functions of the random variable that we will discuss below.

A random variable could perhaps more appropriately be defined as an equivalence class of real-valued measurable functions that are equal almost surely; that is, a class in which if $X \overset{\text{a.s.}}{=} Y$, then X and Y are the same random variable.

Note that a real constant is a random variable. If c is a real constant and if $X \overset{\text{a.s.}}{=} c$, then we call X a *degenerate* random variable; that is, any constant c is a degenerate random variable. We call a random variable that is not a degenerate random variable a *nondegenerate random variable*.

Another comment on a.s. may be in order here. The expression "$X \neq c$ a.s." means the measure of $\Omega_c = \{\omega : X(\omega) = c\}$ is 0. (That is, the expression

does not mean the there is some set with positive measure on which $X(\omega) \neq c$.) Similar interpretations apply to other expressions such as "$X > c$ a.s.".

Simple Random Variables

Some useful random variables assume only a finite number of different values; these are called simple random variables because they are simple functions.

Definition 1.10 (simple random variable)
A random variable that has a finitely-countable range is called a simple random variable. ∎

This is just the same as a simple function, Definition 0.1.28 on page 719.

We will also speak of "discrete" random variables. A discrete random variable has a countable range. A simple random variable is discrete, but a discrete random variable is not necessarily simple.

σ-Fields Generated by Random Variables

A random variable defined on $(\Omega, \mathcal{F})$ determines a useful sub-σ-field of $\mathcal{F}$. First, we establish that a certain collection of sets related to a measurable function is a σ-field.

Theorem 1.4
If $X : \Omega \mapsto \mathcal{X} \subseteq \mathbb{R}^d$ is a random variable, then $\sigma(X^{-1}[\mathcal{B}(\mathcal{X})])$ is a sub-σ-field of $\mathcal{F}$.

Proof. Exercise. (Note that instead of $\mathcal{B}(\mathcal{X})$ we could write $\mathcal{B}^d$.) ∎

Now we give a name to that collection of sets.

Definition 1.11 (σ-field generated by a random variable)
Let $X : \Omega \mapsto \mathbb{R}^d$ be a random variable. We call $\sigma(X^{-1}[\mathcal{B}^d])$ *the σ-field generated by X* and denote it as $\sigma(X)$. ∎

Theorem 1.4 ensures that $\sigma(X)$ is a σ-field and in fact a sub-σ-field of $\mathcal{F}$.

If X and Y are random variables defined on the same measurable space, we may write $\sigma(X, Y)$, with the obvious meaning (see equation (0.1.5) on page 704). As with σ-fields generated by sets or functions discussed in Sections 0.1.1 and 0.1.2, it is clear that $\sigma(X) \subseteq \sigma(X, Y)$. This idea of sub-$\sigma$-fields generated by random variables is important in the analysis of a sequence of random variables. (It leads to the ideas of a filtration; see page 125.)

Random Variables and Probability Distributions

Notice that a random variable is defined in terms only of a measurable space $(\Omega, \mathcal{F})$ and a measurable space defined on the reals $(\mathcal{X}, \mathcal{B}^d)$. No associated probability measure is necessary for the definition, but for meaningful applications of a random variable, we need some probability measure. For a random

variable X defined on $(\Omega, \mathcal{F})$ in the probability space $(\Omega, \mathcal{F}, P)$, the *probability measure* of X is $P \circ X^{-1}$. (This is a pushforward measure; see page 712. In Exercise 1.9, you are asked to show that it is a probability measure.)

A probability space is also called a *population*, a *probability distribution*, a *distribution*, or a *law*. The probability measure itself is the final component that makes a measurable space a probability space, so we associate the distribution most closely with the measure. Thus, "P" may be used to denote both a population and the associated probability measure. We use this notational convention throughout this book.

For a given random variable X, a probability distribution determines $\Pr(X \in B)$ for $B \subseteq \mathcal{X}$. The underlying probability measure P of course determines $\Pr(X^{-1} \in A)$ for $A \in \mathcal{F}$.

Quantiles

Because the values of random variables are real, we can define various special values that would have no meaning in an abstract sample space. As we develop more structure on a probability space characterized by a random variable, we will define a number of special values relating to the random variable. Without any further structure, at this point we can define a useful value of a random variable that just relies on the ordering of the real numbers.

For the random variable $X \in \mathbb{R}$ and given $\pi \in]0, 1[$, the quantity x_π defined as

$$x_\pi = \inf\{x, \text{ s.t. } \Pr(X \leq x) \geq \pi\} \tag{1.6}$$

is called the π *quantile* of X.

For the random variable $X \in \mathbb{R}^d$, there are two ways we can interpret the quantiles. If the probability associated with the quantile, π, is a scalar, then the quantile is a level curve or contour in $X \in \mathbb{R}^d$. Such a quantile is obviously much more complicated, and hence, less useful, than a quantile in a univariate distribution. If π is a d-vector, then the definition in equation (1.6) applies to each element of X and the quantile is a point in $\mathbb{R}^d$.

Multiple Random Variables on the Same Probability Space

If two random variables X and Y have the same distribution, we write

$$X \overset{\mathrm{d}}{=} Y. \tag{1.7}$$

We say that they are *identically distributed*. Note the difference in this and the case in which we say X and Y are the *same random variable*. If X and Y are the same random variable, then $X \overset{\text{a.s.}}{=} Y$. It is clear that

$$X \overset{\text{a.s.}}{=} Y \Longrightarrow X \overset{\mathrm{d}}{=} Y, \tag{1.8}$$

but the implication does not go the other way. (A simple example, using notation to be developed later, is the following. Let $X \sim \mathrm{U}(0, 1)$, and let $Y = 1 - X$. Then $X \overset{\mathrm{d}}{=} Y$ but clearly it is not the case that $X \overset{\text{a.s.}}{=} Y$.)

Support of a Random Variable

Definition 1.12 (support of a distribution or of a random variable)
The *support of the distribution* (or of the random variable) is the smallest closed set $\mathcal{X}_S$ in the image of X such that $P(X^{-1}[\mathcal{X}_S]) = 1$. ∎

We have seen that a useful definition of the support of a general measure requires some structure on the measure space (see page 710). Because the range of a random variable has sufficient structure (it is a metric space), in Definition 1.12, we arrive at a useful concept, while avoiding the ambiguities of a general probability space.

Product Distribution

If X_1 and X_2 are independent random variables with distributions P_1 and P_2, we call the joint distribution of (X_1, X_2) the *product distribution* of P_1 and P_2.

Parameters, Parameter Spaces, and Parametric Families

We often restrict our attention to a *probability family* or a *family of distributions*, $\mathcal{P} = \{P_\theta\}$, where θ is some convenient index.

Definition 1.13 (parametric family of probability distributions)
A family of distributions on a measurable space $(\Omega, \mathcal{F})$ with probability measures P_θ for $\theta \in \Theta$ is called a *parametric family* if $\Theta \subseteq \mathbb{R}^k$ for some fixed positive integer k and θ fully determines the measure. We call θ the *parameter* and Θ the *parameter space*. ∎

If the dimension of Θ is large (there is no precise meaning of "large" here), we may refrain from calling θ a parameter, because we want to refer to some statistical methods as "nonparametric". (In nonparametric methods, our analysis usually results in some general description of the distribution, rather than in a specification of the distribution.)

We assume that every parametric family is *identifiable*; that is, $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ is an identifiable parametric family if it is a parametric family and for $\theta_1, \theta_2 \in \Theta$ if $\theta_1 \neq \theta_2$ then $P_{\theta_1} \neq P_{\theta_2}$.

A family that cannot be indexed in this way might be called a nonparametric family. The term "nonparametric" is most commonly used to refer to a statistical procedure, rather than to a family, however. In general terms, a nonparametric procedure is one that does not depend on strict assumptions about a parametric family.

Example 1.3 a parametric family
An example of a parametric family of distributions for the measurable space $(\Omega = \{0, 1\}, \mathcal{F} = 2^\Omega)$ is that formed from the class of the probability measures

$P_\pi(\{1\}) = \pi$ and $P_\pi(\{0\}) = 1 - \pi$. This is a parametric family, namely, the Bernoulli distributions. The index of the family, π, is called the parameter of the distribution. The measures are dominated by the counting measure. ∎

Example 1.4 a nonparametric family
An example of a nonparametric family over a measurable space $(\mathbb{R}, \mathcal{B})$ is $\mathcal{P}_c = \{P : P \ll \nu\}$, where ν is the Lebesgue measure. This family contains all of the parametric families of Tables A.2 through A.6 of Appendix A as well as many other families. ∎

There are a number of useful parametric distributions to which we give names. For example, the normal or Gaussian distribution, the binomial distribution, the chi-squared, and so on. Each of these distributions is actually a family of distributions. A specific member of the family is specified by specifying the value of each parameter associated with the family of distributions.

For a few distributions, we introduce special symbols to denote the distribution. We use $\mathrm{N}(\mu, \sigma^2)$ to denote a univariate normal distribution with parameters μ and σ^2 (the mean and variance). To indicate that a random variable has a normal distribution, we use notation of the form

$$X \sim \mathrm{N}(\mu, \sigma^2),$$

which here means that the random variable X has a normal distribution with parameters μ and σ^2. We use

$$\mathrm{N}_d(\mu, \Sigma)$$

to denote a d-variate normal distribution with parameters μ and Σ

We use

$$\mathrm{U}(\theta_1, \theta_2)$$

to denote a uniform distribution with support $[\theta_1, \theta_2]$. The most common uniform distribution that we will use is $\mathrm{U}(0, 1)$.

In some cases, we also use special symbols to denote random variables with particular distributions. For example, we often use χ^2_ν to denote a random variable with a chi-squared distribution with ν degrees of freedom.

In Chapter 2 I discuss types of families of probability distributions, and in Tables A.1 through A.6 beginning on page 838 of Appendix A we give descriptions of some parametric families.

The Cumulative Distribution Function (CDF)

The cumulative distribution function provides an alternative expression of a probability measure on $\mathbb{R}^d$. This function gives a clearer picture of the probability distribution, and also provides the basis for defining other useful functions for studying a distribution.

Definition 1.14 (cumulative distribution function (CDF))
If $(\mathbb{R}^d, \mathcal{B}^d, P)$ is a probability space, and F is defined by

$$F(x) = P(]-\infty, x]) \quad \forall\, x \in \mathbb{R}^d, \tag{1.9}$$

then F is called a *cumulative distribution function*, or *CDF*. ∎

The CDF is also called the distribution function, or DF.

There are various forms of notation used for CDFs. The CDF of a given random variable X is often denoted as F_X. A CDF in a parametric family P_θ is often denoted as F_θ, or as $F(x;\theta)$.

If the probability measure P is dominated by the measure ν, then we also say that the associated CDF F is dominated by ν.

The probability space completely determines F, and likewise, F completely determines P a.s.; hence, we often use the CDF and the probability measure interchangeably. More generally, given the probability space $(\Omega, \mathcal{F}, P)$ and the random variable X defined on that space, if F is the CDF of X, the basic probability statement for an event $A \in \mathcal{F}$ given in equation (1.2) can be written as

$$P(A) = \int_A \mathrm{d}P = \int_{X[A]} \mathrm{d}F. \tag{1.10}$$

If the random variable is assumed to be in a family of distributions indexed by θ, we may use the notation $F_\theta(x)$ or $F(x;\theta)$.

For a given random variable X, $F(x) = \Pr(X \leq x)$. We sometimes use the notation $F_X(x)$ to refer to the CDF of the random variable X.

For a given CDF F, we define $\overline{F}$ called the tail CDF by

$$\overline{F}(x) = 1 - F(x). \tag{1.11}$$

This function, which is also denoted by F^{C}, is particularly interesting for random variables whose support is $\mathbb{R}_+$.

The CDF is particularly useful in the case $d = 1$. (If X is a vector-valued random variable, and x is a vector of the same order, $X \leq x$ is interpreted to mean that $X_i \leq x_i$ for each respective element.)

Theorem 1.5 (properties of a CDF)
If F is a CDF then

1. $\lim_{x\downarrow-\infty} F(x) = 0$.
2. $\lim_{x\uparrow\infty} F(x) = 1$.
3. $F(x_1) \leq F(x_2)$ *if* $x_1 < x_2$.
4. $\lim_{\epsilon\downarrow 0} F(x+\epsilon) = F(x)$. *(A CDF is continuous from the right.)*

Proof. Each property is an immediate consequence of the definition. ∎

These four properties characterize a CDF, as we see in the next theorem.

Theorem 1.6
If F is a function defined on $\mathbb{R}^d$ that satisfies the properties of Theorem 1.5, then F is a CDF (for some probability space).

Proof. Exercise. (*Hint:* Given $(\mathbb{R}^d, \mathcal{B}^d)$ and a function F defined on $\mathbb{R}^d$ satisfying the properties of Theorem 1.5, define P as

$$P(]-\infty, x]) = F(x) \quad \forall\, x \in \mathbb{R}^d$$

and show that P is a probability measure.) ∎

Because the four properties of Theorem 1.5 characterize a CDF, they can serve as an alternate definition of a CDF, without reference to a probability distribution. Notice, for example, that the Cantor function (see Section 0.1.5) is a CDF if we extend its definition to be 0 on $]-\infty, 0[$ and to be 1 on $]1, \infty[$. The distribution associated with this CDF has some interesting properties; see Exercise 1.12.

One of the most useful facts about an absolutely continuous CDF is its relation to a $\mathrm{U}(0,1)$ distribution.

Theorem 1.7
If X is a random variable with absolutely continuous CDF F then $F(X) \sim \mathrm{U}(0,1)$.

Proof. If X is a random variable with CDF F then

$$\Pr(F(X) \le t) = \begin{cases} 0 & t < 0 \\ t & 0 \le t < 1 \\ 1 & 1 \le t. \end{cases}$$

This set of probabilities characterize the $\mathrm{U}(0,1)$ distribution. ∎

Although Theorem 1.7 applies to continuous random variables, a discrete random variable has a similar property when we "spread out" the probability between the mass points.

The Quantile Function: The Inverse of the CDF

Although as I indicated above, quantiles can be defined for random variables in $\mathbb{R}^d$ for general positive integer d, they are more useful for $d = 1$. I now define a useful function for that case. (The function could be generalized, but, again, the generalizations are not as useful.)

Definition 1.15 (quantile function)
If $(\mathbb{R}, \mathcal{B}, P)$ is a probability space with CDF F, and F^{-1} is defined on $]0,1[$ by

$$F^{-1}(p) = \inf\{x, \text{ s.t. } F(x) \ge p\}, \tag{1.12}$$

then F^{-1} is called a *quantile function*. ∎

Notice that if F is strictly increasing, the quantile function is the ordinary inverse of the cumulative distribution function. If F is not strictly increasing, the quantile function can be interpreted as a *generalized inverse* of the cumulative distribution function. This definition is reasonable (at the expense of

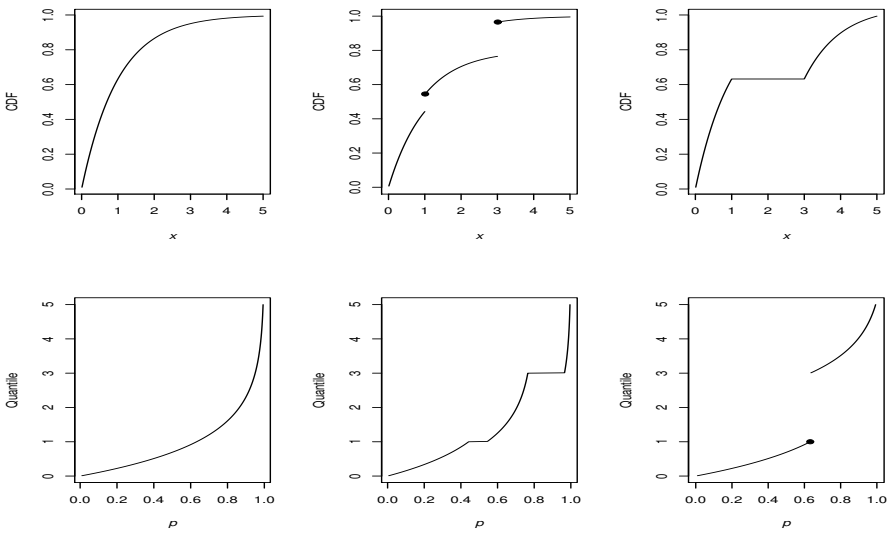


Figure 1.1. CDFs and Quantile Functions

overloading the notation "$\cdot^{-1}$") because, while a CDF may not be an invertible function, it is monotonic nondecreasing.

Notice that for the random variable X with CDF F, if

$$x_\pi = F^{-1}(\pi), \tag{1.13}$$

then x_π is the π quantile of X as defined in equation (1.6). Equation (1.13) is usually taken as the definition of the π quantile.

The quantile function, just as the CDF, fully determines a probability distribution.

Theorem 1.8 (properties of a quantile function)
If F^{-1} is a quantile function and F is the associated CDF,

1. $F^{-1}(F(x)) \leq x$.
2. $F(F^{-1}(p)) \geq p$.
3. $F^{-1}(p) \leq x \iff p \leq F(x)$.
4. $F^{-1}(p_1) \leq F^{-1}(p_2)$ *if* $p_1 \leq p_2$.
5. $\lim_{\epsilon \downarrow 0} F^{-1}(p - \epsilon) = F^{-1}(p)$.
   (A quantile function is continuous from the left.)
6. *If U is a random variable distributed uniformly over* $]0, 1[$, *then* $X = F^{-1}(U)$ *has CDF F.*

Proof. Exercise. ∎

The first five properties of a quantile function given in Theorem 1.8 characterize a quantile function, as stated in the following theorem.

Theorem 1.9
Let F be a CDF and let G be function such that

1. $G(F(x)) \leq x$,
2. $F(G(p)) \geq p$,
3. $G(p) \leq x \iff p \leq F(x)$,
4. $G(p_1) \leq G(p_2)$ *if* $p_1 \leq p_2$, *and*
5. $\lim_{\epsilon \downarrow 0} G(p - \epsilon) = G(p)$.

Then G is the quantile function associated with F, that is, $G = F^{-1}$.

Proof. Exercise. (The definitions of a CDF and a quantile function are sufficient.) ∎

As we might expect, the quantile function has many applications that parallel those of the CDF. For example, we have an immediate corollary to Theorem 1.7.

Corollary 1.7.1
If F is a CDF and $U \sim \mathrm{U}(0,1)$, then $F^{-1}(U)$ is a random variable with CDF F.

Corollary 1.7.1 is actually somewhat stronger than Theorem 1.7 because no modification is needed for discrete distributions. One of the most common applications of this fact is in random number generation, because the basic pseudorandom variable that we can simulate has a $\mathrm{U}(0,1)$ distribution.

The Probability Density Function: The Derivative of the CDF

Another function that may be very useful in describing a probability distribution is the probability density function. This function also provides a basis for straightforward definitions of meaningful characteristics of the distribution.

Definition 1.16 (probability density function (PDF))
The derivative of a CDF (or, equivalently, of the probability measure) with respect to an appropriate measure, if it exists, is called the *probability density function*, *PDF*. ∎

The PDF is also called the density function.

There are various forms of notation used for PDFs. As I mentioned on page 14, common forms of notation for CDFs are F_X, F_θ, and $F(x;\theta)$. Of course, instead of "F", other upper case letters such as G and H are often used similarly. The common notation for the associated PDF parallels that of the CDF and uses the corresponding lower case letter, for example, f_X, f_θ, and $f(x;\theta)$. I will use the standard forms of mathematical notation for functions for denoting CDFs and PDFs. (Other statisticians often use a sloppy notation, called "generic notation"; see page 21.)

Theorem 1.10 (properties of a PDF)
Let F be a CDF defined on $\mathbb{R}^d$ dominated by the measure ν. Let f be the PDF defined as

$$f(x) = \frac{\mathrm{d}F(x)}{\mathrm{d}\nu}.$$

Then over $\mathbb{R}^d$

$$f(x) \geq 0 \quad \text{a.e. } \nu, \tag{1.14}$$

$$f(x) < \infty \quad \text{a.e. } \nu, \tag{1.15}$$

and

$$\int_{\mathbb{R}^d} f \mathrm{d}\nu = 1. \tag{1.16}$$

If $\mathcal{X}_S$ is the support of the distribution, then

$$0 < f(x) < \infty \quad \forall x \in \mathcal{X}_S.$$

Proof. Exercise. ∎

A characteristic of some distributions that is easily defined in terms of the PDF is the mode.

Definition 1.17 (mode of a probability distribution)
If x_0 is a point in the support $\mathcal{X}_S$ of a distribution with PDF f such that

$$f(x_0) \geq f(x), \quad \forall x \in \mathcal{X}_S,$$

then x_0 is called a *mode* of the distribution. ∎

If the mode exists it may or may not be unique.

Dominating Measures

Although we use the term "PDF" and its synonyms for either discrete random variables and the counting measure or for absolutely continuous random variables and Lebesgue measure, there are some differences in the interpretation of a PDF of a discrete random variable and a continuous random variable. In the case of a discrete random variable X, the value of the PDF at the point x is the probability that $X = x$; but this interpretation does not hold for a continuous random variable. For this reason, the PDF of a discrete random variable is often called a probability mass function, or just probability function. There are some concepts defined in terms of a PDF, such as self-information, that depend on the PDF being a probability, as it would be in the case of discrete random variables.

The general meaning of the term "discrete random variable" is that the probability measure is dominated by the counting measure; and likewise for a "absolutely continuous random variable" the general meaning is that the probability measure is dominated by Lebesgue measure. Any simple CDF

has a PDF wrt the counting measure, but not every continuous CDF has a PDF wrt Lebesgue measure (the Cantor function, see page 723, is a classic counterexample – see Exercise 1.12b), but every absolutely continuous CDF does have a PDF wrt Lebesgue measure.

The "appropriate measure" in the definition of PDF above must be σ-finite and must dominate the CDF. For a random variable $X = (X_1, \ldots, X_d)$ with CDF $F_X(x)$ dominated by Lebesgue measure, the PDF, if it exists, is $\partial^d F_X(x)/\partial x_1 \cdots \partial x_d$.

In most distributions of interest there will be a PDF wrt a σ-finite measure. Many of our definitions and theorems will begin with a statement that includes a phrase similar to "a distribution with a PDF wrt a σ-finite measure". In the case of a discrete random variable, that σ-finite measure is the counting measure (Definition 0.1.20 on page 708), and in the case of an absolutely continuous random variable, it is the Lebesgue measure (see page 717).

Parameters and PDFs

In addition to being functions of points x in the support, the PDFs of the distributions within a given parametric family P_θ, are also functions of θ. There may also be other constants in the PDF that can be separated from the functional dependence on x. It is often of interest to focus on the PDF solely as a function of x. (This may be because in applications, the "inverse" problem of deciding on the form of the PDF is simpler if we consider only the role of the observable x.) Given a PDF $f_\theta(x)$, a useful decomposition is

$$f_\theta(x) = g(\theta)k(x), \tag{1.17}$$

where $0 < g$ and $0 \leq k$, and $k(x)$ encodes all of the dependence of $f_\theta(x)$ on x. In this decomposition, we call $k(x)$ the "kernel".

From equation (1.16), we have

$$\frac{1}{g(\theta)} = \int_{\mathbb{R}^d} k(x) \mathrm{d}\nu(x). \tag{1.18}$$

The function $(g(\theta))^{-1}$ is called the "normalizing function" or the "partition function". (The latter name comes from statistical physics, and in many areas of application, the partition function is a meaningful characteristic of the problem. Both or either of these names is sometimes applied to $g(\theta)$.)

The Likelihood; A Function of the Parameters

It is often useful to consider the PDF (or CDF) as a function of the parameter instead of the range of the random variable. We call such a function, a *likelihood function*, and for the parametric PDF $f_\theta(x)$ with support $\mathcal{X}_S$ and parameter space Θ, we denote the corresponding likelihood function as $L(\theta; x)$:

$$L(\theta; x) = f_\theta(x), \quad \forall \theta \in \Theta, \ \forall x \in \mathcal{X}_S. \tag{1.19}$$

(Actually, any positive scalar multiple of $L(\theta; x)$ in (1.19), is called a likelihood function corresponding to $f_\theta(x)$. I will discuss likelihood functions later in more detail, particularly in Chapter 6, where their use in statistical inference is discussed.)

While the likelihood and the PDF are the same, at any two fixed points x and θ, they differ fundamentally as functions.

Example 1.5 likelihood in the exponential family
Consider the exponential family of distributions with parameter θ. The PDF is

$$p_X(x\,;\theta) = \theta^{-1}\mathrm{e}^{-x/\theta}\mathrm{I}_{\mathbb{R}_+}(x), \tag{1.20}$$

for $\theta \in \mathbb{R}_+$. Plots of this PDF for $\theta = 1$ and $\theta = 5$ are shown on the left side of Figure 1.2.

Given a single observation x, the likelihood is

$$L(\theta\,;x) = \theta^{-1}\mathrm{e}^{-x/\theta}\mathrm{I}_{\mathbb{R}_+}(\theta). \tag{1.21}$$

Plots of this likelihood for $x = 1$ and $x = 5$ are shown on the right side of Figure 1.2. ∎

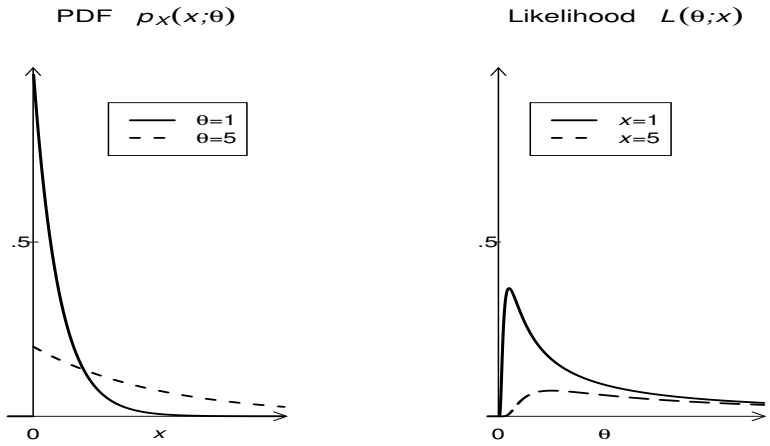


Figure 1.2. PDF and Likelihood for Exponential Distribution

The log of the likelihood, called the *log-likelihood function*,

$$l_L(\theta\,;\,x) = \log L(\theta\,;\,x), \tag{1.22}$$

is also useful. We often denote the log-likelihood without the "L" subscript. (The notation for the likelihood and the log-likelihood varies with authors. My own choice of an uppercase "L" for the likelihood and a lowercase "l" for the log-likelihood is long-standing, and not based on any notational optimality consideration. Because of the variation in the notation for the log-likelihood, I will often use the "l_L" notation because this expression is suggestive of the meaning.)

In cases when the likelihood or the log-likelihood is differentiable wrt the parameter, the derivative is of interest because it indicates the sensitivity of the parametric family to the parameter when it is considered to be a variable. The most useful derivative is that of the log-likelihood, $\partial l_L(\theta;x)/\partial\theta$. The expectation of the square of this appears in a useful quantity in statistics, the Cramér-Rao lower bound, inequality (3.39) on page 234.

Parametric Families

As mentioned above, if a specific CDF is F_θ, we often write the corresponding PDF as f_θ:

$$f_\theta = \frac{\mathrm{d}F_\theta}{\mathrm{d}\nu}. \tag{1.23}$$

There may be some ambiguity in the use of such subscripts, however, because when we have defined a specific random variable, we may use the symbol for the random variable as the identifier of the CDF or PDF. The CDF and the PDF corresponding to a given random variable X are often denoted, respectively, as F_X and f_X. Adding to this confusion is the common usage by statisticians of the "generic notation", that is, for given random variables X and Y, the notation $f(x)$ may refer to a different function than the notation $f(y)$. I will not use the "generic" notation for CDFs and PDFs.

We assume that every parametric family is *identifiable*; that is, if F_θ and f_θ are the CDF and PDF for distributions within the family and these are dominated by the measure ν, then over the parameter space Θ for $\theta_1 \neq \theta_2$,

$$\nu(\{x : F_{\theta_1}(x) \neq F_{\theta_2}(x)\}) > 0 \tag{1.24}$$

and

$$\nu(\{x : f_{\theta_1}(x) \neq f_{\theta_2}(x)\}) > 0. \tag{1.25}$$

Dominating Measures

The dominating measure for a given probability distribution is not unique, but use of a different dominating measure may change the representation of the distribution. For example, suppose that the support of a distribution is S, and so we write the PDF as

$$\frac{\mathrm{d}F_\theta(x)}{\mathrm{d}\nu} = g_\theta(x)\mathrm{I}_S(x). \tag{1.26}$$

If we define a measure λ by

$$\lambda(A) = \int_A \mathrm{I}_S \mathrm{d}\nu \quad \forall A \in \mathcal{F}, \tag{1.27}$$

then we could write the PDF as

$$\frac{\mathrm{d}F_\theta}{\mathrm{d}\lambda} = g_\theta. \tag{1.28}$$

Mixtures of Distributions

If $F_1, F_2, \ldots$ are CDFs and $\pi_1, \pi_2, \ldots \geq 0$ are real constants such that $\sum_i \pi_i = 1$, then

$$F = \sum_i \pi_i F_i \tag{1.29}$$

is a CDF (exercise). If each F_i in equation (1.29) is dominated by Lebesgue measure, then F is dominated by Lebesgue measure. Likewise, if each F_i is dominated by the counting measure, then F is dominated by the counting measure.

The PDF corresponding to F is also the same linear combination of the corresponding PDFs.

It is often useful to form mixtures of distributions of continuous random variables with distributions of discrete random variables. The ϵ-mixture distribution, whose CDF is given in equation (2.45) on page 194, is an example.

A mixture distribution can also be thought of as a random variable X whose distribution, randomly, is the same as that of some other random variable $X_1, X_2, \ldots$; that is,

$$X \overset{\mathrm{d}}{=} X_I,$$

where I is a random variable taking values in the index set of $X_1, X_2, \ldots$. In terms of the π_i in equation (1.29), we can define the random variable of the mixture X by

$$X \overset{\mathrm{d}}{=} X_i \quad \text{with probability } \pi_i.$$

We must be careful not to think of the linear combination of the CDFs or PDFs as applying to the random variables; the random variable of the mixture is not a linear combination of the constituent random variables.

Joint and Marginal Distributions

For a random variable consisting of two components, (X_1, X_2), we speak of its distribution as a "joint distribution", and we call the separate distributions the "marginal distributions". We might denote the PDF of the joint distribution

as f_{X_1,X_2} and the PDF for X_1 alone as f_{X_1}. We call f_{X_1,X_2} a joint PDF and f_{X_1} a marginal PDF.

We have the simple relationship

$$f_{X_1}(x_1) = \int f_{X_1,X_2}(x_1, x_2)\,\mathrm{d}x_2. \tag{1.30}$$

Independence and Exchangeability of Random Variables

We have defined independence and exchangeability in general. We now give equivalent definitions for random variables.

Definition 1.18 (independence of random variables)
The random variables $X_1, \ldots, X_k$ on $(\Omega, \mathcal{F}, P)$ are said to be *independent* iff for any sets $B_1, \ldots, B_k$ in the σ-field of the image space,

$$P(X_1 \in B_1, \ldots, X_k \in B_k) = \prod_{i=1}^{k} P(X_i \in B_i).$$

∎

Notice that this definition is essentially the same as Definition 1.7.3 on page 5, and it corresponds to the definition of independence of events; that is, independence of $X_i^{-1}[B_i]$ (Definition 1.7.1). These are not just separate definitions of independence of various objects. The following theorem (which we could have stated analogously following Definition 1.7) relates Definition 1.18 above to Definition 1.7.2.

Theorem 1.11
The random variables $X_1, \ldots, X_k$ on $(\Omega, \mathcal{F}, P)$ are independent iff the σ-fields $\sigma(X_1), \ldots, \sigma(X_k)$ are independent.

Proof. Exercise. ∎

The following factorization theorem is often useful in establishing independence of random variables.

Theorem 1.12
If X_1 and X_2 are random variables with joint PDF f_{X_1,X_2} and marginal PDFs f_{X_1} and f_{X_2}, then X_1 and X_2 are independent iff

$$f_{X_1,X_2}(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2). \tag{1.31}$$

Proof. Exercise. ∎

We are often interested in random variables that have the same distributions. In that case, exchangeability of the random variables is of interest.

The following definition of exchangeability is essentially the same as Definition 1.8.3, and similar comments relating to exchangeability of random variables, sets, and σ-fields as made above relating to independence hold.

Definition 1.19 (exchangeability of random variables)
The random variables $X_1, \ldots, X_k$ on $(\Omega, \mathcal{F}, P)$ are said to be *exchangeable* iff the joint distribution of $X_1, \ldots, X_k$ is the same as the joint distribution of $\Pi(\{X_1, \ldots, X_k\})$, for any Π, where $\Pi(A)$ denotes a permutation of the elements of the set A. ∎

As we have seen, exchangeability requires identical distributions, but, given that, it is a weaker property than independence.

Example 1.6 Polya's urn process
Consider an urn that initially contains r red and b blue balls. One ball is chosen randomly from the urn, and its color noted. The ball is then put back into the urn together with c balls of the same color. (Hence, the number of total balls in the urn changes. We can allow $c = -1$, in which case, the drawn ball is not returned to the urn.) Now define a binary random variable $R_i = 1$ if a red ball is drawn and $R_i = 0$ if a blue ball is drawn. (The random variable R_i in this example is the indicator function for the event R_i in Example 1.2.) The sequence $R_1, R_2, \ldots$ is exchangeable, but not independent. ∎

An interesting fact about infinite sequences of exchangeable binary random variables, such as those in Example 1.6 with $c \geq 0$, is that they are mixtures of independent Bernoulli sequences; see page 75. This provides a link between exchangeability and independence. This connection between exchangeability and independence does not necessarily hold in finite sequences, as in the urn process of Example 1.2.

Random Samples

If the random variables $X_1, \ldots, X_k$ are *independent* and

$$X_1 \overset{\mathrm{d}}{=} \cdots \overset{\mathrm{d}}{=} X_k,$$

we say $X_1, \ldots, X_k$ are identically and independently distributed, which we denote as iid. A set of iid random variables is called a *simple random sample*, and the cardinality of the set is called the "size" of the simple random sample. The common distribution of the variables is called the *parent distribution* of the simple random sample. We also often use the phrase "excangeable random sample", with the obvious meaning.

There are many situations in which a sample is generated randomly, but the sample is not a simple random sample or even an exchangeable random sample. Two of the most common such situations are in finite population sampling (see Section 5.5.2) and a process with a stopping rule that depends on the realizations of the random variable, such as sampling from an urn until a ball of a certain color is drawn or sampling from a binary (Bernoulli) process until a specified number of 1s have occurred (see Example 3.12).

Despite the more general meaning of random sample, we often call a simple random sample just a "random sample".

In statistical applications we often form functions of a random sample, and use known or assumed distributions of those functions to make inferences abput the parent distribution. Two of the most common functions of a random sample are the sample mean and the sample variance. Given a random sample $X_1, \ldots, X_n$, the *sample mean* is defined as

$$\overline{X} = \sum_{i=1}^{n} X_i/n, \tag{1.32}$$

and if $n \geq 2$, the *sample variance* is defined as

$$S^2 = \sum_{i=1}^{n} (X_i - \overline{X})^2/(n-1). \tag{1.33}$$

(Notice that the divisor in the sample variance is $n-1$.)

If the parent distribution is $\mathrm{N}(\mu, \sigma^2)$, the sample mean and sample variance have simple and useful distributions (see page 187).

For functions of a random sample such as the sample mean and variance, we often include the sample size in the notation for the function, as $\overline{X}_n$ or S_n^2, and we may be interested in the properties of these functions as n gets larger (see Example 1.23).

The Empirical Distribution Function

Given a random sample $X_1, \ldots, X_n$, we can form a conditional discrete distribution, dominated by a counting measure on $\{1, \ldots, n\}$, with CDF

$$F_n(x) = \#\{X_i \,|\, X_i \leq x\}/n, \quad \text{for } i = 1, \ldots, n. \tag{1.34}$$

The simple CDF F_n is called the empirical cumulative distribution function (ECDF) for the sample. It has at most $n+1$ distinct values.

The ECDF has wide-ranging applications in statistics.

1.1.3 Definitions and Properties of Expected Values

First we define the expected value of a random variable.

Definition 1.20 (expected value of a random variable)
Given a probability space $(\Omega, \mathcal{F}, P)$ and a d-variate random variable X defined on $\mathcal{F}$, we define the *expected value* of X with respect to P, which we denote by $\mathrm{E}(X)$ or for clarity by $\mathrm{E}_P(X)$, as

$$\mathrm{E}(X) = \int_\Omega X \, \mathrm{d}P, \tag{1.35}$$

if this integral exists. ∎

For the random variable X, $\mathrm{E}(X)$, if it exists, is called the *first moment of* X.

Although by properties following immediately from the definition, $\Pr(-\infty < X < \infty) = 1$, it could happen that $\int_{\mathbb{R}^d} X \, \mathrm{d}P = \infty$. In that case we say the expected value, or first moment, is infinite. It could also happen that $\int_{\mathbb{R}^d} X \, \mathrm{d}P$ does not exist (see Definition 0.1.41 on page 728), and in that case we say the expected value does not exist. Recall, however, what we mean by the expression "integrable function". This carries over to a random variable; we say the random variable X is *integrable* (or L_1 *integrable*) iff $-\infty < \int_{\mathbb{R}^d} X \, \mathrm{d}P < \infty$.

Example 1.7 existence and finiteness of expectations
Let the random variable X have the PDF

$$f_X(x) = \frac{1}{(1+x)^2} \mathrm{I}_{\mathbb{R}_+}(x). \tag{1.36}$$

We have

$$\begin{aligned} \mathrm{E}(X) &= \int_0^\infty \frac{x}{(1+x)^2} \mathrm{d}x \\ &= \left(\log(1+x) + \frac{1}{1+x} \right) \Big|_0^\infty \\ &= \infty. \end{aligned}$$

That is, $\mathrm{E}(X)$ is infinite. (Although $\mathrm{E}(X) \notin \mathbb{R}$, and some people would say that it does not exist in this case, we consider $\mathrm{E}(X)$ "to exist", just as we speak of the existence of infinite integrals (page 727). Just as we identify conditions in which integrals do not exist because of indeterminancy (page 728), however, we likewise will identify situations in which the expectations do not exist.)

Let the random variable Y have the PDF

$$f_Y(y) = \frac{1}{\pi(1+y^2)}. \tag{1.37}$$

We have

$$\begin{aligned} \mathrm{E}(Y) &= \int_{-\infty}^\infty \frac{x}{\pi(1+x^2)} \mathrm{d}x \\ &= \frac{1}{2\pi} \log(1+x^2) \Big|_{-\infty}^\infty . \end{aligned}$$

That is, $\mathrm{E}(Y)$ is not infinite; it does not exist.

The random variable X has the same distribution as the ratio of two standard exponential random variables, and the random variable Y has the same distribution as the ratio of two standard normal random variables, called a Cauchy distribution. (It is an exercise to show these facts.) ∎

It is clear that E is a linear operator; that is, for random variables X and Y defined on the same probability space, and constant a,

$$\mathrm{E}(aX + Y) = a\mathrm{E}(X) + \mathrm{E}(Y), \tag{1.38}$$

if $\mathrm{E}(X)$ and $\mathrm{E}(Y)$ are finite.

Look carefully at the integral (1.35). It is the integral of a function, X, over Ω with respect to a measure, P, over the σ-field that together with Ω forms the measurable space. To emphasize the meaning more precisely, we could write the integral in the definition as

$$\mathrm{E}(X) = \int_{\Omega} X(\omega)\,\mathrm{d}P(\omega).$$

The integral (1.35) is over an abstract domain Ω. We can also write the expectation over the real range of the random variable and an equivalent measure on that range. If the CDF of the random variable is F, we have, in the abbreviated form of the first expression given in the definition,

$$\mathrm{E}(X) = \int_{\mathbb{R}^d} x\,\mathrm{d}F, \tag{1.39}$$

or in the more precise form,

$$\mathrm{E}(X) = \int_{\mathbb{R}^d} x\,\mathrm{d}F(x).$$

If the PDF exists and is f, we also have

$$\mathrm{E}(X) = \int_{\mathbb{R}^d} xf(x)\,\mathrm{d}x.$$

An important and useful fact about expected values is given in the next theorem.

Theorem 1.13
Let X be a random variable in $\mathbb{R}^d$ such that $\mathrm{E}(\|X\|_2) < \infty$. Then

$$\mathrm{E}(X) = \arg\min_{a \in \mathbb{R}^d} \mathrm{E}(\|X - a\|_2). \tag{1.40}$$

Proof. Exercise. Also, see equation (0.0.101). ∎

In statistical applications this result states that $\mathrm{E}(X)$ is the best prediction of X given a quadratic loss function in the absence of additional information about X.

We define the expected value of a Borel function of a random variable in the same way as above for a random variable.

Definition 1.21 (expected value of a Borel function)
If g is a Borel function of the random variable X with CDF F, then the expected value of $g(X)$ is defined as

$$\mathrm{E}(g(X)) = \int_{\mathbb{R}^d} g(x)\,\mathrm{d}F(x). \tag{1.41}$$

∎

Theorem 1.14
If X and Y are random variables defined over the same probability space and g is a Borel function of the random variable X, then

$$X \geq Y \text{ a.s.} \quad \Rightarrow \quad \mathrm{E}(X) \geq \mathrm{E}(Y), \tag{1.42}$$

and

$$g(X) \geq 0 \text{ a.s.} \quad \Rightarrow \quad \mathrm{E}(g(X)) \geq 0. \tag{1.43}$$

Proof. Each property is an immediate consequence of the definition. ∎

Expected Values of Probability Density Functions

Expected values of PDFs or of functionals of PDFs are useful in applications of probability theory. The expectation of the negative of the log of a PDF, $-\log(f(X))$, is the entropy (Definition 1.26 on page 41), which is related to the concept of "information". We have also mentioned the expectation of the square of $\partial \log(f(X;\theta))/\partial\theta$, which appears in the Cramér-Rao lower bound. The expectation of $\partial \log(f(X;\theta))/\partial\theta$ is zero (see page 230).

There are several interesting expectations of the PDFs of two different distributions, some of which we will mention on page 36.

Expected Values and Quantile Functions of Univariate Random Variables

As we mentioned earlier, the quantile function has many applications that parallel those of the CDF.

If X is a univariate random variable with CDF F, then the expected value of X is

$$\mathrm{E}(X) = \int_0^1 F^{-1}(x)\mathrm{d}x. \tag{1.44}$$

See Exercise 1.25.

Expected Value and Probability

There are many interesting relationships between expected values and probabilities, such as the following.

Theorem 1.15
If X is a random variable such that $X > 0$ a.s., then

$$\mathrm{E}(X) = \int_0^\infty \Pr(X > t)\mathrm{d}t. \tag{1.45}$$

Proof. This is a simple application of Fubini's theorem, using the CDF F of X:

$$\begin{aligned}
\mathrm{E}(X) &= \int_0^\infty x\,\mathrm{d}F(x) \\
&= \int_0^\infty \int_{]0,x[} \mathrm{d}t\,\mathrm{d}F(x) \\
&= \int_0^\infty \int_{]t,\infty[} \mathrm{d}F(x)\mathrm{d}t \\
&= \int_0^\infty (1 - F(t))\mathrm{d}t \\
&= \int_0^\infty \Pr(X > t)\mathrm{d}t
\end{aligned}$$

∎

Theorem 1.15 leads in general to the following useful property for any given random variable X for which $\mathrm{E}(X)$ exists:

$$\mathrm{E}(X) = \int_0^\infty (1 - F(t))\mathrm{d}t - \int_{-\infty}^0 F(t)\mathrm{d}t. \tag{1.46}$$

It is an exercise to write a proof of this statement.

Another useful fact in applications involving the Bernoulli distribution with parameter π is the relationship

$$\mathrm{E}(X) = \Pr(X = 1) = \pi.$$

Expected Value of the Indicator Function

We define the indicator function, $\mathrm{I}_A(x)$, as

$$\mathrm{I}_A(x) = \begin{cases} 1 \text{ if } x \in A \\ 0 \text{ otherwise.} \end{cases} \tag{1.47}$$

(This is also sometimes called the "characteristic function", but we use that term to refer to something else.) If X is an integrable random variable over A, then $\mathrm{I}_A(X)$ is an integrable random variable, and

$$\Pr(A) = \mathrm{E}(\mathrm{I}_A(X)). \tag{1.48}$$

It is an exercise to write a proof of this statement. When it is clear from the context, we may omit the X, and merely write $\mathrm{E}(\mathrm{I}_A)$.

Expected Value over a Measurable Set

The expected value of an integrable random variable over a measurable set $A \subseteq \mathbb{R}^d$ is

$$\mathrm{E}(X\mathrm{I}_A(X)) = \int_A X \, \mathrm{d}P. \tag{1.49}$$

It is an exercise to write a proof of this statement. We often denote this as $\mathrm{E}(X\mathrm{I}_A)$.

Expected Value of General Measurable Functions

A real-valued measurable function g of a random variable X is itself a random variable, possibly with a different probability measure. Its expected value is defined in exactly the same way as above. If the probability triple associated with the random variable X is $(\Omega, \mathcal{F}, P)$ and $Y = g(X)$, we could identify a probability triple associated with Y. Being measurable, the relevant measurable space of $g(X)$ is $(\Omega, \mathcal{F})$, but the probability measure is not necessarily P. If we denote the probability triple associated with the random variable Y is $(\Omega, \mathcal{F}, Q)$, we may distinguish the defining integrals with respect to $\mathrm{d}P$ and $\mathrm{d}Q$ by E_P and E_Q.

We can also write the expected value of Y in terms of the CDF of the original random variable. The expected value of a real-valued measurable function g of a random variable X with CDF F is $\mathrm{E}(g(X)) = \int g(x)\mathrm{d}F(x)$.

Moments of Scalar Random Variables

The higher-order moments are the expected values of positive integral powers of the random variable. If X is a scalar-valued random variable, the r^{th} *raw moment* of X, if it exists, is $\mathrm{E}(X^r)$. We often denote the r^{th} raw moment as μ'_r. There is no requirement, except notational convenience, to require that r be an integer, and we will often allow it non-integral values, although "first moment", "second moment", and so on refer to the integral values.

For $r \geq 2$, *central moments* or *moments about* $\mathrm{E}(X)$ are often more useful. The r^{th} *central moment* of the univariate random variable X, denoted as μ_r, is $\mathrm{E}((X - \mathrm{E}(X))^r)$:

$$\mu_r = \int (x - \mu)^r \mathrm{d}F(x). \tag{1.50}$$

if it exists. We also take μ_1 to be μ, which is also μ'_1. For a discrete distribution, this expression can be interpreted as a sum of the values at the mass points times their associated probabilities.

Notice that the moment may or may not exist, and if it exists, it may or may not be finite. (As usual, whenever I speak of some property of a moment, such as that it is finite, I am assuming that the moment exists, even though I may not make a statement to that effect.)

The first two central moments are usually the most important; μ_1 is called the *mean* and μ_2 is called the *variance*. The variance of X is denoted by $\mathrm{V}(\cdot)$. Because $(X - \mathrm{E}(X))^2 \geq 0$ a.s., we see that the variance is nonnegative. Further, unless $X \overset{\text{a.s.}}{=} \mathrm{E}(X)$, the variance is positive.

The square root of the variance is called the *standard deviation*.

$\mathrm{E}(|X|)$ is called the *first absolute moment* of X; and generally, $\mathrm{E}(|X|^r)$ is called the r^{th} absolute moment.

Theorem 1.16
If the r^{th} absolute moment of a scalar random variable is finite, then the absolute moments of order $1, 2, \ldots, r-1$ are finite.

Proof. For $s \leq r$, $|x|^s \leq 1 + |x|^r$. ∎

Example 1.8 moments of random variables in the t family
Consider the t family of distributions (page 841) with PDF

$$f(y) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\nu\pi}}(1 + y^2/\nu)^{-(\nu+1)/2},$$

in which the parameter ν is called the degrees of freedom. By direct integration, it is easy to see that the r^{th} absolute moment moment exists iff $r \leq \nu - 1$. ∎

We define the r^{th} *standardized moment* as

$$\eta_r = \mu_r / \mu_2^{r/2}. \tag{1.51}$$

The first raw moment or the *mean*, is an indicator of the general "location" of the distribution. The second central moment or the *variance*, denoted as μ_2 or σ^2 is a measure of the "spread" of the distribution. The nonnegative square root, σ, is sometimes called the "scale" of the distribution. The third standardized moment, η_3, is an indicator of whether the distribution is skewed; it is called the *skewness coefficient*. If $\eta_3 \neq 0$, the distribution is asymmetric, but $\eta_3 = 0$ does not mean that the distribution is symmetric.

The fourth standardized moment, η_4 is called the *kurtosis coefficient*. It is an indicator of how "peaked" the distribution is, or how "heavy" the tails of the distribution are. (Actually, exactly what this standardized moment measures cannot be described simply. Because, for the random variable X, we have

$$\eta_4 = \mathrm{V}\left(\frac{(X-\mu)^2}{\sigma^2}\right) + 1,$$

it can be seen that the minimum value for η_4 is attained for a discrete distribution with mass points $-\sigma$ and σ. We might therefore interpret η_4 as a

measure of variation about the two points $-\sigma$ and σ. This, of course, leads to the two characteristics mentioned above: peakedness and heaviness in the tails.)

The sequence of (raw) moments is very useful in characterizing a distribution of a scalar random variable, but often the central moments are to be preferred because the second and higher central moments are invariant to change in the first moment (the "location").

Uniqueness of Moments of Scalar Random Variables

An interesting question is whether the full set of moments fully determines a distribution of a scalar random variable.

It seems reasonable to guess that this would not be the case if not all moments of all orders exist; and, indeed, it is a simple matter to construct two different distributions whose moments beyond the k^{th} are infinite but whose first k moments are the same. The question of interest, therefore, is whether the full set of moments fully determine a distribution, given that moments of all orders are finite. In general, they do not.

Example 1.9 different distributions with equal finite moments of all orders
Consider the complex integral related to the gamma function,

$$\int_0^\infty t^{\alpha-1}\mathrm{e}^{-t/\beta}\mathrm{d}t = \beta^\alpha\Gamma(\alpha),$$

where $\alpha > 0$ and $\beta = 1/(\gamma + \mathrm{i}\xi)$ with $\gamma > 0$. Now, make a change of variable, $x^\rho = t$ for $0 < \rho < 1/2$; and for a nonnegative integer k, choose $\alpha = (k+1)/\rho$, and $\xi/\gamma = \tan(\rho\pi)$. Noting that $(1 + \mathrm{i}\tan(\rho\pi))^{(k+1)/\rho}$ is real, we have that the imaginary part of the integral after substitution is 0:

$$\int_0^\infty x^k\mathrm{e}^{-\gamma x^\rho}\sin(\xi x^\rho)\mathrm{d}x = 0. \tag{1.52}$$

Hence, for all $|\alpha| \leq 1$, the distributions over $\bar{\mathbb{R}}_+$ with PDF

$$p(x) = c\mathrm{e}^{-\gamma x^\rho}(1 + \alpha\sin(\xi x^\rho))\mathrm{I}_{\bar{\mathbb{R}}_+}(x), \tag{1.53}$$

where $\gamma > 0$ and $0 < \rho < 1/2$ have the same moments of all orders $k = 0, 1, 2, \ldots$. ∎

We could in a similar way develop a family of distributions over $\mathbb{R}$ that have the same moments. (In that case, the ρ is required to be in the range $]-1, 1[$.) The essential characteristic of these examples is that there exists x_0, such that for $\gamma > 0$,

$$p(x) > \mathrm{e}^{-\gamma|x|^\rho} \quad \text{for } x > x_0, \tag{1.54}$$

because we can add a function all of whose moments are 0. (Such a function would necessarily take on negative values, but not sufficiently small to make $p(x)$ plus that function be negative.)

A distribution all of whose moments are the same as some other distribution is called a *moment-indeterminant distribution*. The distributions within a family of distributions all of which have the same moments are called *moment-equivalent distributions*. In Exercise 1.28, you are asked to show that there are other distributions moment-equivalent to the lognormal family of distributions.

So can we identify conditions that are sufficient for the moments to characterize a probability distribution? The answer is yes; in fact, there are several criteria that ensure that the moments uniquely determine a distribution. For scalar random variables, one criterion is given in the following theorem.

Theorem 1.17
Let $\nu_0, \nu_1, \ldots \in \mathbb{R}$ be the moments of some probability distribution. (The moments can be about any origin.) The probability distribution is uniquely determined by those moments if

$$\sum_{j=0}^{\infty} \frac{\nu_j t^j}{j!} \tag{1.55}$$

converges for some real nonzero t.

A simple proof of this theorem is based on the uniqueness of the characteristic function, so we defer its proof to page 49, after we have discussed the characteristic function.

Corollary 1.17.1 *The moments of a probability distribution with finite support uniquely determine the distribution.*

The following theorem, which we state without proof, tells us that the moments determine the distribution if the central moments are close enough to zero.

Theorem 1.18
Let $\mu_0, \mu_1, \ldots$ be the central moments of some probability distribution. If the support of the probability distribution is $\mathbb{R}$, it is uniquely determined by those moments if

$$\sum_{j=0}^{\infty} \frac{1}{\mu_{2j}^{1/2j}} \tag{1.56}$$

diverges.

If the support of the probability distribution is $\overline{\mathbb{R}}_+$, it is uniquely determined by those moments if

$$\sum_{j=0}^{\infty} \frac{1}{\mu_j^{1/2j}} \tag{1.57}$$

diverges.

Corollary 1.18.1
If a probability distribution has a PDF $p(x)$ with support $\mathbb{R}$, then the moments uniquely determine $p(x)$ if

$$p(x) < M|x|^{\lambda-1}e^{-\gamma|x|^{\rho}} \quad \text{for } x > x_0, \tag{1.58}$$

where $M, \lambda, \gamma > 0$ and $\rho \geq 1$.

If a probability distribution has a PDF $p(x)$ with support $\overline{\mathbb{R}}_+$, then the moments uniquely determine $p(x)$ if

$$p(x) < Mx^{\lambda-1}e^{-\gamma x^{\rho}} \quad \text{for } x > 0, \tag{1.59}$$

where $M, \lambda, \gamma > 0$ and $\rho \geq 1/2$.

The conditions in Theorem 1.18 are called the Carleman criteria (after Torsten Carleman). Compare the conditions in the corollary with inequality (1.54).

Within a specific family of distributions, the full set of moments can usually be expected to identify the distribution. For narrowly-defined families of distributions, such as the normal or gamma families, often only one or two moments completely identify the family.

Cumulants of Scalar Random Variables

Another useful sequence of constants for describing the distribution of a scalar random variables are called cumulants, if they exist. Cumulants, except for the first, are also invariant to change in the first moment. The cumulants and the moments are closely related, and cumulants can be defined in terms of raw moments, if they exist. For the first few cumulants, $\kappa_1, \kappa_2, \ldots$, and raw moments, $\mu'_1, \mu'_2, \ldots$, of a scalar random variable, for example,

$$\begin{aligned} \mu'_1 &= \kappa_1 \\ \mu'_2 &= \kappa_2 + \kappa_1^2 \\ \mu'_3 &= \kappa_3 + 3\kappa_2\kappa_1 + \kappa_1^3. \end{aligned} \tag{1.60}$$

The expressions for the cumulants are a little more complicated, but can be obtained easily from the triangular system (1.60).

Factorial Moments of Discrete Scalar Random Variables

A discrete distribution with support $x_1, x_2, \ldots \in \mathbb{R}$ is equivalent to a discrete distribution with support $0, 1, \ldots$, and for such a distribution, another kind of moment is sometimes useful. It is the factorial moment, related to the r^{th} *factorial* of the real number y:

$$y^{[r]} = y(y-1)\cdots(y-(r-1)). \tag{1.61}$$

(We see that $y^{[y]} = y!$. It is, of course, not necessary that y be an integer, but factorials are generally more useful in the context of nonnegative integers.)

The r^{th} *factorial moment* of the random variable X above is

$$\mu'_{[r]} = \sum_{i=0}^{\infty} x_i^{[r]} p_i. \tag{1.62}$$

We see that $\mu'_{[1]} = \mu'_1 = \mu_1$.

The r^{th} central factorial moment, denoted $\mu_{[r]}$ is the r^{th} factorial moment about μ.

1.1.4 Relations among Random Variables

In many applications there are two or more random variables of interest. For a given probability space $(\Omega, \mathcal{F}, P)$, there may be a collection of random variables, $\mathcal{W}$. If the random variables have some common properties, for example, if either all are discrete or all are continuous and if all have the same structure, we may identify the collection as a "space".

If the random variables have some relationship to each other, that is, if they are not independent, we seek useful measures of their dependence. The appropriate measure depends on the nature of their dependence. If they are quadratically related, a measure that is appropriate for linear relationships may be inappropriate, and vice versa.

We will consider two ways of studying the relationships among random variables. The first is based on second-degree moments, called covariances, between *pairs* of variables, and the other is based on functions that relate the CDF of one variable or one set of variables to the CDFs of other variables. These functions, called copulas, can involve more than single pairs of variables.

Random Variable Spaces

As with any function space, $\mathcal{W}$ may have interesting and useful properties. For example, $\mathcal{W}$ may be a linear space; that is, for $X, Y \in \mathcal{W}$ and $a \in \mathbb{R}$, $aX + Y \in \mathcal{W}$.

The concept of $\mathcal{L}^p$ random variable spaces follows immediately from the general property of function spaces, discussed on page 741. for random variables in an $\mathcal{L}^p$ random variable space, the p^{th} absolute is finite.

The closure of random variable spaces is often of interest. We define various forms of closure depending on types of convergence of a sequence X_n in the space. For example, given any sequence $X_n \in \mathcal{W}$ if $X_n \xrightarrow{\text{L}_r} X$ implies $X \in \mathcal{W}$ then $\mathcal{W}$ is closed for the r^{th} moment.

Expectations

We may take expectations of functions of random variables in terms of their joint distribution or in terms of their marginal distributions. To indicate the

distribution used in an expectation, we may use notation for the expectation operator similar to that we use on the individual distribution, as described on page 23. Given the random variables X_1 and X_2, we use the notation E_{X_1} to indicate an expectation taken with respect to the marginal distribution of X_1.

We often denote the expectation taken with respect to the joint distribution as simply E, but for emphasis, we may use the notation E_{X_1,X_2}.

We also use notation of the form E_P, where P denotes the relevant probability distribution of whatever form, or E_θ in a parametric family of probability distributions.

Expectations of PDFs and of Likelihoods

If the marginal PDFs of the random variables X_1 and X_2 are f_{X_1} and f_{X_2}, we have the equalities

$$\mathrm{E}_{X_1}\left(\frac{f_{X_2}(X_1)}{f_{X_1}(X_1)}\right) = \mathrm{E}_{X_2}\left(\frac{f_{X_1}(X_2)}{f_{X_2}(X_2)}\right) = 1. \tag{1.63}$$

On the other hand,

$$\mathrm{E}_{X_1}(-\log(f_{X_1}(X_1))) \leq \mathrm{E}_{X_1}(-\log(f_{X_2}(X_1))), \tag{1.64}$$

with equality only if $f_{X_1}(x) = f_{X_2}(x)$ a.e. (see page 41).

When the distributions are in the same parametric family, we may write f_θ with different values of θ instead of f_{X_1} and f_{X_2}. In that case, it is more natural to think of the functions as likelihoods since the parameter is the variable. From equation (1.63), for example, we have for the likelihood ratio,

$$\mathrm{E}_{\theta_1}\left(\frac{L(\theta_2; X)}{L(\theta_1; X)}\right) = 1. \tag{1.65}$$

Covariance and Correlation

Expectations are also used to define relationships among random variables. We will first consider expectations of scalar random variables, and then discuss expectations of vector and matrix random variables.

For two scalar random variables, X and Y, useful measures of a *linear* relationship between them are the covariance and correlation. The covariance of X and X, if it exists, is denoted by $\mathrm{Cov}(X, Y)$, and is defined as

$$\mathrm{Cov}(X, Y) = \mathrm{E}\left((X - \mathrm{E}(X))(Y - \mathrm{E}(Y))\right) \tag{1.66}$$

From the Cauchy-Schwarz inequality (B.21) (see page 853), we see that

$$(\mathrm{Cov}(X, Y))^2 \leq \mathrm{V}(X)\mathrm{V}(Y). \tag{1.67}$$

The *correlation* of X and Y, written $\mathrm{Cor}(X,Y)$, is defined as

$$\mathrm{Cor}(X,Y) = \mathrm{Cov}(X,Y) \Big/ \sqrt{\mathrm{V}(X)\mathrm{V}(Y)}\,. \tag{1.68}$$

The correlation is also called the *correlation coefficient* and is often written as $\rho_{X,Y}$.

From inequality (1.67), we see that the correlation coefficient is in $[-1,1]$.

If X and Y are independent, then $\mathrm{Cov}(X,Y) = \mathrm{Cor}(X,Y) = 0$ (exercise).

Structure of Random Variables

Random variables may consist of individual $\mathbb{R}$ elements arrayed in some structure, such as a vector so that the random variable itself is in $\mathbb{R}^d$ or as a matrix so that the random variable is in $\mathbb{R}^{d\times m}$. Many of the properties of random variables are essentially the same whatever their structure, except of course those properties may have structures dependent on that of the random variable.

Multiplication is an operation that depends very strongly on the structure of the operand. If x is a scalar, x^2 is a scalar. If x is a is vector, however, there are various operations that could be interpreted as extensions of a squaring operation. First, of course, is elementwise squaring. In this interpretation x^2 has the same structure as x. Salient relationships among the individual elements of x may be lost by this operation, however. Other interpretations are $x^{\mathrm{T}}x$, which preserves none of the structure of x, and xx^{T}, which is in $\mathbb{R}^{d\times d}$. The point of this is that what can reasonably be done in the analysis of random variables depends on the structure of the random variables, and such relatively simple concepts as moments require some careful consideration. In many cases, a third-order or higher-order moment is not useful because of its complexity.

Structural Moments

For random variables that have a structure such as a vector or matrix, the elementwise moments are the same as those for a scalar-valued random variable as described above, and hence, the first moment, the mean, has the same structure as the random variable itself.

Higher order moments of vectors and matrices present some problems because the number of individual scalar moments is greater than the number of elements in the random object itself. For multivariate distributions, the higher-order marginal moments are generally more useful than the higher-order joint moments. We define the second-order moments (variances and covariances) for random vectors and for random matrices below.

Definition 1.22 (variance-covariance of a random vector)
The *variance-covariance of a vector-valued random variable* X is the expectation of the outer product,

$$\mathrm{V}(X) = \mathrm{E}\left((X - \mathrm{E}(X))(X - \mathrm{E}(X))^{\mathrm{T}}\right), \tag{1.69}$$

if it exists.

For a constant vector, the rank of an outer product is no greater than 1, but unless $X \overset{\text{a.s.}}{=} \mathrm{E}(X)$, $\mathrm{V}(X)$ is nonnegative definite. We see this by forming the scalar random variable $Y = c^{\mathrm{T}}X$ for any $c \neq 0$, and writing

$$\begin{aligned} 0 &\leq \mathrm{V}(Y) \\ &= \mathrm{E}((c^{\mathrm{T}}X - c^{\mathrm{T}}\mathrm{E}(X))^2) \\ &= \mathrm{E}((c^{\mathrm{T}}(X - \mathrm{E}(X))(X - \mathrm{E}(X))c) \\ &= c^{\mathrm{T}}\mathrm{V}(X)c. \end{aligned}$$

(If $X \overset{\text{a.s.}}{=} \mathrm{E}(X)$, then $\mathrm{V}(X) = 0$, and while it is true that $c^{\mathrm{T}}0c = 0 \geq 0$, we do not say that the 0 matrix is nonnegative definite. Recall further that whenever I write a term such as $\mathrm{V}(X)$, I am implicitly assuming its existence.)

Furthermore, if it is not the case that $X \overset{\text{a.s.}}{=} \mathrm{E}(X)$, unless some element X_i of a vector X is such that

$$X_i \overset{\text{a.s.}}{=} \sum_{j \neq i} (a_j + b_j X_j),$$

then $\mathrm{V}(X)$ is positive definite a.s. To show this, we show that $\mathrm{V}(X)$ is full rank a.s. (exercise).

The elements of $\mathrm{V}(X)$ are the bivariate moments of the respective elements of X; the (i, j) element of $\mathrm{V}(X)$ is the covariance of X_i and X_j, $\mathrm{Cov}(X_i, X_j)$.

If $\mathrm{V}(X)$ is nonsingular, then the correlation matrix of X, written $\mathrm{Cor}(X)$ is

$$\mathrm{Cor}(X) = \left(\mathrm{E}(X - \mathrm{E}(X))^{\mathrm{T}}\right)(\mathrm{V}(X))^{-1}\mathrm{E}(X - \mathrm{E}(X)). \tag{1.70}$$

The (i, j) element of $\mathrm{Cor}(X)$ is the correlation of X_i and X_j, and so the diagonal elements are all 1.

Definition 1.23 (variance-covariance of a random matrix)
The *variance-covariance of a matrix random variable* X is defined as the variance-covariance of $\mathrm{vec}(X)$:

$$\mathrm{V}(X) = \mathrm{V}(\mathrm{vec}(X)) = \mathrm{E}\left(\mathrm{vec}(X - \mathrm{E}(X))(\mathrm{vec}(X - \mathrm{E}(X)))^{\mathrm{T}}\right), \tag{1.71}$$

if it exists.

Linearity of Moments

The linearity property of the expectation operator yields some simple linearity properties for moments of first or second degree of random variables over the same probability space.

For random variables X, Y, and Z with finite variances and constants a, b, and c, we have

$$\mathrm{V}(aX + Y + c) = a^2\mathrm{V}(X) + \mathrm{V}(Y) + 2a\mathrm{Cov}(X, Y); \tag{1.72}$$

that is, $\mathrm{V}(\cdot)$ is not a linear operator (but it simplifies nicely), and

$$\mathrm{Cov}(aX + bY + c, X + Z) = a\mathrm{V}(X) + a\mathrm{Cov}(X, Z) + b\mathrm{Cov}(X, Y) + b\mathrm{Cov}(Y, Z); \tag{1.73}$$

that is, $\mathrm{Cov}(\cdot, \cdot)$ is a bilinear operator. Proofs of these two facts are left as exercises.

Copulas

A copula is a function that relates a multivariate CDF to lower dimensional marginal CDFs. The basic ideas of copulas can all be explored in the context of a bivariate distribution and the two associated univariate distributions, and the ideas extend in a natural way to higher dimensions.

Definition 1.24 (two-dimensional copula)
A *two-dimensional copula* is a function C that maps $[0, 1]^2$ onto $[0, 1]$ with the following properties:

1. for every $u \in [0, 1]$,
$$C(0, u) = C(u, 0) = 0, \tag{1.74}$$
and
$$C(1, u) = C(u, 1) = u, \tag{1.75}$$
2. for every $(u_1, u_2), (v_1, v_2) \in [0, 1]^2$ with $u_1 \leq v_1$ and $u_2 \leq v_2$,
$$C(u_1, u_2) - C(u_1, v_2) - C(v_1, u_2) + C(v_1, v_2) \geq 0. \tag{1.76}$$

∎

A two-dimensional copula is also called a *2-copula*.

The arguments to a copula C are often taken to be CDFs, which of course take values in $[0, 1]$.

The usefulness of copulas derive from *Sklar's theorem*, which we state without proof.

Theorem 1.19 (Sklar's theorem)
Let P_{XY} be a bivariate CDF with marginal CDFs P_X and P_Y. Then there exists a copula C such that for every $x, y \in \mathbb{R}$,

$$P_{XY}(x, y) = C(P_X(x), P_Y(y)). \tag{1.77}$$

If P_X and P_Y are continuous everywhere, then C is unique; otherwise C is unique over the support of the distributions defined by P_X and P_Y.

Conversely, if C is a copula and P_X and P_Y are CDFs, then the function $P_{XY}(x, y)$ defined by equation (1.77) is a CDF with marginal CDFs $P_X(x)$ and $P_Y(y)$.

Thus, a copula is a joint CDF of random variables with $\mathrm{U}(0,1)$ marginals. The proof of the first part of the theorem is given in Nelsen (2006), among other places. The proof of the converse portion is straightforward and is left as an exercise.

For many bivariate distributions the copula is the most useful way to relate the joint distribution to the marginals, because it provides a separate description of the individual distributions and their association with each other.

One of the most important uses of copulas is to combine two marginal distributions to form a joint distribution with known bivariate characteristics.

Certain standard copulas are useful in specific applications. The copula that corresponds to a bivariate normal distribution with correlation coefficient ρ is

$$C_{\mathrm{N}\rho}(u,v) = \int_{-\infty}^{\Phi^{-1}(u)} \int_{-\infty}^{\Phi^{-1}(v)} \phi_\rho(t_1,t_2)\,\mathrm{d}t_2\mathrm{d}t_1, \tag{1.78}$$

where $\Phi(\cdot)$ is the standard (univariate) normal CDF, and $\phi_\rho(\cdot,\cdot)$ is the bivariate normal PDF with means 0, variances 1, and correlation coefficient ρ. This copula is usually called the Gaussian copula and has been widely used in financial applications.

The association determined by a copula is not the same as that determined by a correlation; that is, two pairs of random variables may have the same copula but different correlations.

1.1.5 Entropy

Probability theory is developed from models that characterize uncertainty inherent in random events. Information theory is developed in terms of the information revealed by random events. The premise is that the occurrence of an event with low probability is more informative than the occurrence of an event of high probability. For a discrete random variable we can effectively associate a value of the random variable with an event, and we quantify information in such a way that the information revealed by a particular outcome decreases as the probability increases. Thus, there is more information revealed by a rare event than by a common event.

Definition 1.25 (self-information)
Let X be a discrete random variable with probability mass function p_X. The *self-information* of $X = x$ is $-\log_2(p_X(x))$. ∎

Self-information is also called Shannon information. The logarithm to the base 2 comes from the basic representation of information in base 2, but we can equivalently use any base, and it is common to use the natural log in the definition of self-information.

The logarithm of the PDF is an important function in studying random variables. It is used to define logconcave families (see page 165). The negative

of its expected value is called "entropy", or sometimes, "Shannon entropy" to distinguish it from other measures of entropy.

Definition 1.26 (entropy)
Let X be a random variable with PDF p_X with respect to a σ-finite measure. The *entropy* of the random variable X is

$$\mathrm{E}(-\log(p_X(X))). \tag{1.79}$$

∎

The expected value of the derivative of the logarithm of the PDF with respect to a parameter of a distributional family yields another definition of "information" (see Section 1.1.6), and it appears in an important inequality in statistical theory (see pages 399 and 854).

The expected value in expression (1.79) is smallest when the it is taken wrt the distribution with PDF p_X, as we see in the following theorem, known as the Gibbs lemma.

Theorem 1.20 (Gibbs lemma)
Let P_1 and P_2 be probability distributions with PDFs p_1 and p_2 respectively. Then

$$\mathrm{E}_{P_1}(-\log(p_1(X))) \le \mathrm{E}_{P_1}(-\log(p_2(X))), \tag{1.80}$$

with equality only if $p_1(x) = p_2(x)$ a.e.

Proof. I give a proof for distributions with PDFs dominated by Lebesgue measure, but it is clear that a similar argument would hold for other measures.

Let p_1 and p_2 be PDFs dominated by Lebesgue measure. Let $\mathcal{X}$ be the set of x such that $p_1(x) > 0$. Over $\mathcal{X}$

$$\log\left(\frac{p_2(x)}{p_1(x)}\right) \le \frac{p_2(x)}{p_1(x)} - 1,$$

with inequality iff $p_2(x) = p_1(x)$. So we have

$$p_1(x)\log\left(\frac{p_2(x)}{p_1(x)}\right) \le p_2(x) - p_1(x).$$

Now

$$\begin{aligned}
\mathrm{E}_{P_1}(-\log(p_1(X))) - \mathrm{E}_{P_1}(-\log(p_2(X))) &= -\int_{\mathcal{X}} p_1(x)\log(p_1(x))\mathrm{d}x \\
&\quad + \int_{\mathcal{X}} p_1(x)\log(p_2(x))\mathrm{d}x \\
&\le \int_{\mathcal{X}} p_2(x)\mathrm{d}x - \int_{\mathcal{X}} p_1(x)\mathrm{d}x \\
&\le \int_{\mathbb{R}^d} p_2(x)\mathrm{d}x - \int_{\mathbb{R}^d} p_1(x)\mathrm{d}x \\
&= 0.
\end{aligned}$$

The definitions of information theory are generally given in the context of a countable sample space, and in that case, we can see that the entropy is the expected value of the self-information, and equation (1.79) becomes

$$H(X) = -\sum_x p_X(x) \log(p_X(x)), \tag{1.81}$$

which is the more familiar form in information theory.

We can likewise define the *joint entropy* $H(X, Y)$ in terms of the joint PDF $p_{X,Y}$.

We can see that the entropy is maximized if all outcomes are equally probable. In the case of a discrete random variable with two outcomes with probabilities π and $1-\pi$ (a Bernoulli random variable with parameter π), the entropy is

$$-\pi \log(\pi) - (1-\pi) \log(1-\pi).$$

We note that it is maximized when $\pi = 1/2$.

1.1.6 Fisher Information

The concept of "information" is important in applications of probability theory. We have given one formal meaning of the term in Section 1.1.5. We will give another formal definition in this section, and in Section 1.6.1, we will give an informal meaning of the term in the context of evolution of σ-Fields.

If a random variable X has a PDF $f(x;\theta)$ wrt a σ-finite measure that is differentiable in θ, the rate of change of the PDF at a given x with respect to different values of θ intuitively is an indication of the amount of information x provides. For such distributions, we define the "information" (or "Fisher information") that X contains about θ as

$$I(\theta) = \mathrm{E}_\theta \left(\left(\frac{\partial \log f(X;\theta)}{\partial \theta} \right) \left(\frac{\partial \log f(X;\theta)}{\partial \theta} \right)^{\mathrm{T}} \right). \tag{1.82}$$

1.1.7 Generating Functions

There are various functionals of the PDF or CDF that are useful for determining properties of a distribution. One important type are "generating functions". These are functions whose derivatives evaluated at specific points yield important quantities that describe the distribution, such as moments, or as in the case of discrete distributions, the probabilities of the points in the support.

We will describe these functions in terms of d-variate random variables, although, as we discussed on pages 30 through 39, the higher-order moments of d-variate random variables have structures that depends on d.

Moment-Generating Functions

Definition 1.27 (moment-generating function)
For the random d-variate X, the moment-generating function (MGF) is

$$\psi_X(t) = \mathrm{E}\left(\mathrm{e}^{t^{\mathrm{T}}X}\right), \quad t \in \mathcal{N}_\epsilon(0) \subseteq \mathbb{R}^d, \tag{1.83}$$

if this expectation is finite for some $\epsilon > 0$. ∎

The moment-generating function is obviously nonnegative.

Many common distributions do not have moment-generating functions. Two examples are the Cauchy distribution (exercise) and the lognormal distribution (Example 1.10).

For a scalar random variable X, the MGF yields the (raw) moments directly (assuming existence of all quantities):

$$\left.\frac{\mathrm{d}^k \psi_X(t)}{\mathrm{d}t^k}\right|_{t=0} = \mathrm{E}(X^k). \tag{1.84}$$

For vector-valued random variables, the moments become tensors, but the first two moments are very simple:

$$\nabla \psi_X(t)|_{t=0} = \mathrm{E}(X) \tag{1.85}$$

and

$$\nabla\nabla \psi_X(t)|_{t=0} = \mathrm{E}(X^{\mathrm{T}}X). \tag{1.86}$$

Because of the foregoing and the fact that for $\forall k \geq 1, |t| > 0, \exists x_0 \ni |x| \geq x_0 \Rightarrow \mathrm{e}^{tx} > x^k$, we have the following fact.

Theorem 1.21
If for the scalar random variable X the MGF $\psi_X(t)$ exists, then $\mathrm{E}(|X|^k) < \infty$ and $\mathrm{E}(X^k) = \psi_X^{(k)}(0)$.

The theorem also holds for vector-valued random variables, with the appropriate definition of X^k.

The converse does not hold, as the following example shows.

Example 1.10 a distribution whose moments exist but whose moment-generating function does not exist
The lognormal distribution provides some interesting examples with regard to moments. First, recall from page 33 that the moments do not uniquely determine a lognormal distribution (see also Exercise 1.28).

The standard lognormal distribution is related to the standard normal distribution. If Y has a standard normal distribution, then $X = \mathrm{e}^Y$ has a lognormal distribution (that is, the log of a lognormal random variable has a normal distribution).

Let Y be a standard normal variable and let $X = \mathrm{e}^Y$. We see that the moments of the lognormal random variable X exist for all orders $k = 1, 2, \ldots$:

$$\begin{aligned} \mathrm{E}\left(X^k\right) &= \mathrm{E}\left(\mathrm{e}^{kY}\right) \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathrm{e}^{ky} \mathrm{e}^{-y^2/2} \mathrm{d}y \\ &= \mathrm{e}^{k^2/2}. \end{aligned}$$

However, for $t > 0$,

$$\begin{aligned} \mathrm{E}\left(\mathrm{e}^{tX}\right) &= \mathrm{E}\left(\mathrm{e}^{t\mathrm{e}^Y}\right) \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathrm{e}^{t\mathrm{e}^y - y^2/2} \mathrm{d}y. \end{aligned}$$

By expanding e^y in the exponent of the integrand as $1+y+y^2/2+y^3/3!+\cdots$, we see that the integrand is greater than $\exp(t(1 + y + y^3/3!))$. However,

$$\int_{-\infty}^{\infty} \exp(t(1 + y + y^3/3!))\mathrm{d}y = \infty;$$

hence, the MGF does not exist. ∎

When it exists, the MGF is very similar to the characteristic function, which we will define and discuss in Section 1.1.8 beginning on page 45.

Generating Functions for Discrete Distributions

Frequency-generating functions or probability-generating functions (the terms are synonymous) are useful for discrete random variables.

Definition 1.28 (probability-generating function)
For the discrete random variable X taking values $x_1, x_2, \ldots$ with probabilities $0 < p_1, p_2, \ldots$, the *frequency-generating function* or *probability-generating function* is the polynomial

$$P(t) = \sum_{i=0}^{\infty} p_{i+1} t^i. \tag{1.87}$$

∎

The probability of x_r is

$$\frac{\mathrm{d}^{r+1}}{\mathrm{d}t^{r+1}} P(t) \,|_{t=0} \tag{1.88}$$

The probability-generating function for the binomial distribution with parameters π and n, for example, is

$$P(t) = (\pi t + (1-\pi))^n.$$

For a discrete distribution, there is also a generating function for the factorial moments. We see immediately that the *factorial-moment-generating function* is the same as the probability-generating function evaluated at $t+1$:

$$\begin{aligned} P(t+1) &= \sum_{j=0}^{\infty} p_{j+1}(t+1)^j \\ &= \sum_{j=0}^{\infty} p_{j+1} \sum_{i=1}^{j} \binom{j}{i} t^i \\ &= \sum_{i=0}^{\infty} \frac{t^i}{i!} \sum_{j=0}^{\infty} (p_{j+1} j(j-1)\cdots(j-i+1)) \\ &= \sum_{i=0}^{\infty} \frac{t^i}{i!} \mu'_{[i]}. \end{aligned} \tag{1.89}$$

1.1.8 Characteristic Functions

One of the most useful functions determining a probability distribution is the characteristic function, or CF. The CF is also a generating function for the moments.

Definition and Properties

Definition 1.29 (characteristic function)
For the random d-variate variable X, the characteristic function (CF), with $t \in \mathbb{R}^d$, is

$$\varphi_X(t) = \mathrm{E}\left(\mathrm{e}^{\mathrm{i}t^{\mathrm{T}}X}\right) \tag{1.90}$$

where $\mathrm{i} = \sqrt{-1}$. ∎

The characteristic function is the Fourier transform of the density with argument $-t/(2\pi)$.

The function e^z has many useful properties. There are several relationships to trigonometric functions, series expansions, limits, and inequalities involving this function that are useful in working with characteristic functions.

We see that the integral in equation (1.83) exists (as opposed to the integral in equation (1.90) defining the MGF) by use of Euler's formula (equation (0.0.67)) to observe that

$$\left|\mathrm{e}^{\mathrm{i}t^{\mathrm{T}}x}\right| = \left|\cos\left(t^{\mathrm{T}}x\right) + \mathrm{i}\sin\left(t^{\mathrm{T}}x\right)\right| = \sqrt{\cos^2\left(t^{\mathrm{T}}x\right) + \sin^2\left(t^{\mathrm{T}}x\right)} = 1.$$

Euler's formula also provides an alternate expression for the CF that is sometimes useful:

$$\varphi_X(t) = \mathrm{E}\left(\cos\left(t^{\mathrm{T}}X\right)\right) + \mathrm{i}\mathrm{E}\left(\sin\left(t^{\mathrm{T}}X\right)\right). \tag{1.91}$$

Although the CF always exists, it may not have an explicit representation. The CF for the lognormal distribution, for example, cannot be represented explicitly, but can be approximated to any tolerance by a divergent series (exercise).

Note that the integration in the expectation operator defining the CF is not complex integration; we interpret it as ordinary Lebesgue integration in the real dummy variable in the PDF (or $\mathrm{d}F$). Hence, if the MGF is finite for all t such that $|t| < \epsilon$ for some $\epsilon > 0$, then the CF can be obtained by replacing t in $\psi_X(t)$ by $\mathrm{i}t$. Note also that the MGF may be defined only in a neighborhood of 0, but the CF is defined over $\mathbb{R}$.

There are some properties of the characteristic function that are immediate from the definition:

$$\varphi_X(-t) = \overline{\varphi_X(t)} \tag{1.92}$$

and

$$\varphi_X(0) = 1. \tag{1.93}$$

The CF is real if the distribution is symmetric about 0. We see this from equation (1.91).

The CF is bounded:

$$|\varphi_X(t)| \leq 1. \tag{1.94}$$

We see this property by first observing that

$$\left|\mathrm{E}\left(\mathrm{e}^{\mathrm{i}t^{\mathrm{T}}X}\right)\right| \leq \mathrm{E}\left(\left|\mathrm{e}^{\mathrm{i}t^{\mathrm{T}}X}\right|\right),$$

and then by using Euler's formula on $\left|\mathrm{e}^{\mathrm{i}t^{\mathrm{T}}x}\right|$.

The CF and the moments of the distribution if they exist are closely related, as we will see below. Another useful property provides a bound on the difference of the CF at any point and a partial series in the moments in terms of expected absolute moments.

$$\left|\varphi_X(t) - \sum_{k=0}^{n} \frac{(\mathrm{i}t)^k}{k!}\mathrm{E}(X^k)\right| \leq \mathrm{E}\left(\min\left(\frac{2|tX|^n}{n!}, \frac{|tX|^{n+1}}{(n+1)!}\right)\right), \tag{1.95}$$

This property follows immediately from inequality 0.0.71 on page 663.

Another slightly less obvious fact is

Theorem 1.22
The CF is uniformly continuous on $\mathbb{R}$.

Proof. We want to show that for any $t \in \mathbb{R}$, $|\varphi_X(t+h) - \varphi_X(t)| \to 0$ as $h \to 0$.

$$\begin{aligned}|\varphi_X(t+h) - \varphi_X(t)| &= \left| \int \mathrm{e}^{\mathrm{i}t^{\mathrm{T}}X} \left(\mathrm{e}^{\mathrm{i}h^{\mathrm{T}}X} - 1\right) \mathrm{d}F \right| \\ &\leq \int \left| \mathrm{e}^{\mathrm{i}h^{\mathrm{T}}X} - 1 \right| \mathrm{d}F\end{aligned}$$

By the bounded convergence theorem (Corollary 0.1.25.1 on page 734), the last integral goes to 0 as $h \to 0$, so we have the desired result. ∎

Moments and the CF

As with the MGF, the (raw) moments of X, if they exist, can be obtained from the derivatives of the CF evaluated at 0. For a scalar random variable X, if the k^{th} derivative of the CF exists in a neighborhood of 0, then

$$\left. \frac{\mathrm{d}^k \varphi_X(t)}{\mathrm{d}t^k} \right|_{t=0} = (-1)^{k/2} \mathrm{E}(X^k). \tag{1.96}$$

For vector-valued random variables, the moments become tensors of course, but the first two moments are very simple:

$$\nabla \varphi_X(t)|_{t=0} = \mathrm{i}\mathrm{E}(X) \tag{1.97}$$

and

$$\nabla\nabla \varphi_X(t)|_{t=0} = -\mathrm{E}(X^{\mathrm{T}}X). \tag{1.98}$$

Note that these derivatives are wrt a real variable. This means, for example, that existence of the first derivative in a neighborhood does not imply the existence of all derivatives in that neighborhood.

There is a type of converse to the statement that includes equation (1.96).

Theorem 1.23
Let X be a scalar random variable. If $\mathrm{E}(X^k)$ exists and is finite, then the k^{th} derivative of the CF of X exists and satisfies equation (1.96).

We will not give a formal proof here; But note that we have the existence of the derivative of the CF because we can interchange the differentiation and integration operators. For example, for the first moment (and the first derivative) we have

$$\infty > \left| \int x \mathrm{d}F \right| = \left| \int \frac{\mathrm{d}}{\mathrm{d}t} \mathrm{e}^{\mathrm{i}tx}\big|_{t=0} \mathrm{d}F \right| = \left| \frac{\mathrm{d}}{\mathrm{d}t} \int \mathrm{e}^{\mathrm{i}tx} \mathrm{d}F\big|_{t=0} \right|.$$

This is very different from the situation for MGFs; we saw in Example 1.10 that the moment may exist but the MGF, let alone its derivative, may not exist.

The converse of Theorem 1.23 is true for even moments (see Chung (2000) or Gut (2005), for example); that is, if the CF has a finite even derivative of even order k at 0, then the random variable has a finite moment of order k.

Perhaps surprisingly, however, the converse does not hold for odd moments (see Exercise 1.36).

CF and MGF of Multiples and Sums

The CF or MGF for scalar multiples of a random variable or for sums of iid random variables is easily obtained from the CF or MGF of the underlying random variable(s). It is easy to see from the definition that if the random variable X has CF $\varphi_X(t)$ and $Z = aX$, for a fixed scalar a, then

$$\varphi_Z(t) = \varphi_X(at). \tag{1.99}$$

Likewise, if $X_1, \ldots, X_n$ are iid with CF $\varphi_X(t)$, and $Y = X_1 + \cdots + X_n$, then the CF or MGF is just the n^{th} power of the CF or MGF of X_i:

$$\varphi_Y(t) = (\varphi_X(t))^n. \tag{1.100}$$

Combining these and generalizing (for independent but not necessarily identically distributed X_i), for $W = \sum_i a_i X_i$, we have

$$\varphi_W(t) = \prod_i \varphi_{X_i}(a_i t). \tag{1.101}$$

On page 59, we discuss the use of CFs and MGFs in studying general transformations of random variables.

Uniqueness

The importance of the characteristic function or of the moment-generating function if it exists is that it is unique to a given distribution. This fact is asserted formally in the following theorem.

Theorem 1.24 (inversion theorem)
The CF (or MGF if it exists) completely determines the distribution.

We will not prove this theorem here; its proof can be found in a number of places, for example, Billingsley (1995), page 346. It is essentially the same theorem as the one often used in working with Fourier transforms.

The limit of a sequence of CFs or MGFs also determines the limiting distribution if it exists, as we will see in Theorem 1.37. This fact, of course, also depends on the uniqueness of CFs or MGFs, if they exist.

We now illustrate an application of the CF by proving Theorem 1.17 stated on page 33. This is a standard result, and its method of proof using analytic continuation is standard.

Proof. (Theorem 1.17)
We are given the finite scalar moments $\nu_0, \nu_1, \ldots$ (about any origin) of some probability distribution, and the condition that

$$\sum_{j=0}^{\infty} \frac{\nu_j t^j}{j!}$$

converges for some real nonzero t. We want to show that the moments uniquely determine the distribution. We will do this by showing that the moments uniquely determine the CF.

Because the moments exist, the characteristic function $\varphi(t)$ is continuous and its derivatives exist at $t = 0$. We have for t in a neighborhood of 0,

$$\varphi(t) = \sum_{j=0}^{r} \frac{(\mathrm{i}t)^j \mu_j}{j!} + R_r, \tag{1.102}$$

where $|R_r| < \nu_{r+1}|t|^{r+1}/(r+1)!$.

Now because $\sum \nu_j t^j/j!$ converges, $\nu_j t^j/j!$ goes to 0 and hence the right hand side of equation (1.102) is the infinite series $\sum_{j=0}^{\infty} \frac{(\mathrm{i}t)^j \mu_j}{j!}$ if it converges. This series does indeed converge because it is dominated termwise by $\sum \nu_j t^j/j!$ which converges. Thus, $\varphi(t)$ is uniquely determined within a neighborhood of $t = 0$. This is not sufficient, however, to say that $\varphi(t)$ is uniquely determined a.e.

We must extend the region of convergence to $\mathbb{R}$. We do this by analytic continuation. Let t_0 be arbitrary, and consider a neighborhood of $t = t_0$. Analogous to equation (1.102), we have

$$\varphi(t) = \sum_{j=0}^{r} \frac{\mathrm{i}^j (t - t_0)^j}{j!} \int_{-\infty}^{\infty} x^j \mathrm{e}^{\mathrm{i}t_0 x} \mathrm{d}F + \widetilde{R}_r.$$

Now, the modulus of the coefficient of $(t - t_0)^j/j!$ is less than or equal to ν_j; hence, in a neighborhood of t_0, $\varphi(t)$ can be represented as a convergent Taylor series. But t_0 was arbitrary, and so $\varphi(t)$ can be extended through any finite interval by taking it as the convergent series. Hence, $\varphi(t)$ is uniquely defined over $\mathbb{R}$ in terms of the moments; and therefore the distribution function is uniquely determined. ∎

Characteristic Functions for Functions of Random Variables; Joint and Marginal Distributions

If $X \in \mathbb{R}^d$ is a random variable and for a Borel function g, $g(X) \in \mathbb{R}^m$, the characteristic function of $g(X)$ is

$$\varphi_{g(X)}(t) = \mathrm{E}_X\left(\mathrm{e}^{\mathrm{i}t^{\mathrm{T}} g(X)}\right), \quad t \in \mathbb{R}^m, \tag{1.103}$$

where E_X represents expectation wrt the distribution of X. Other generating functions for $g(X)$ are defined similarly.

In some cases of interest, the Borel function may just be a projection. For a random variable X consisting of two components, (X_1, X_2), either component is just a projection of X. In this case, we can factor the generating functions to correspond to the two components.

If $\varphi_X(t)$ is the CF of X, and if we decompose the vector t to be conformable to $X = (X_1, X_2)$, then we have the CF of X_1 as $\varphi_{X_1}(t_1) = \varphi_X(t_1, 0)$.

Note that $\varphi_{X_1}(t_1)$ is not (necessarily) the CF of the marginal distribution of X_1. The expectation is taken with respect to the joint distribution.

Following equation (1.31), we see immediately that X_1 and X_2 are independent iff

$$\varphi_X(t) = \varphi_{X_1}(t_1)\varphi_{X_2}(t_2). \tag{1.104}$$

Cumulant-Generating Function

The cumulant-generating function, defined in terms of the characteristic function, can be used to generate the cumulants if they exist.

Definition 1.30 (cumulant-generating function)
For the random variable X with characteristic function $\varphi(t)$ the *cumulant-generating function* is

$$K(t) = \log(\varphi(t)). \tag{1.105}$$

∎

(The "K" in the notation for the cumulant-generating function is the Greek letter kappa.) The cumulant-generating function is often called the "second characteristic function".

The derivatives of the cumulant-generating function can be used to evaluate the cumulants, similarly to the use of the CF to generate the raw moments, as in equation (1.96).

If $Z = X + Y$, given the random variables X and Y, we see that

$$K_Z(t) = K_X(t) + K_Y(t). \tag{1.106}$$

The cumulant-generating function has useful properties for working with random variables of a form such as

$$Y_n = \left(\sum X_i - n\mu\right) / \sqrt{n}\sigma,$$

that appeared in the central limit theorem above. If $X_1, \ldots, X_n$ are iid with cumulant-generating function $K_X(t)$, mean μ, and variance $0 < \sigma^2 < \infty$, then the cumulant-generating function of Y_n is

$$K_{Y_n}(t) = -\frac{\sqrt{n}\mu t}{\sigma} + nK_X\left(\frac{t}{\sqrt{n}\sigma}\right). \tag{1.107}$$

A Taylor series expansion of this gives

$$K_{Y_n}(t) = -\frac{1}{2}t^2 + \frac{K'''(0)}{6\sqrt{n}\sigma^3}t^3 + \cdots, \tag{1.108}$$

from which we see, as $n \to \infty$, that $K_{Y_n}(t)$ converges to $-t^2/2$, which is the cumulant-generating function of the standard normal distribution.

1.1.9 Functionals of the CDF; Distribution "Measures"

We often use the term "functional" to mean a function whose arguments are functions. The value of a functional may be any kind of object, a real number or another function, for example. The domain of a functional is a set of functions. I will use notation of the following form: for the functional, a capital Greek or Latin letter, Υ, M, etc.; for the domain, a calligraphic Latin letter, $\mathcal{F}$, $\mathcal{G}$, etc.; for a function, an italic letter, g, F, G, etc.; and for the value, the usual notation for functions, $\Upsilon(G)$ where $G \in \mathcal{G}$, for example.

Parameters of distributions as well as other interesting characteristics of distributions can often be defined in terms of functionals of the CDF. For example, the mean of a distribution, if it exists, may be written as the functional M of the CDF F:

$$M(F) = \int y \, \mathrm{d}F(y). \tag{1.109}$$

Viewing this mean functional as a Riemann–Stieltjes integral, for a discrete distribution, it reduces to a sum of the mass points times their associated probabilities.

A functional operating on a CDF is called a *statistical functional* or *statistical function*. (This is because they are often applied to the ECDF, and in that case are "statistics".) I will refer to the values of such functionals as *distributional measures*. (Although the distinction is not important, "M" in equation (1.109) is a capital Greek letter mu. I usually—but not always—will use upper-case Greek letters to denote functionals, especially functionals of CDFs and in those cases, I usually will use the corresponding lower-case letters to represent the measures defined by the functionals.)

Many statistical functions, such as $M(F)$ above, are expectations; but not all are expectations. For example, the quantile functional in equation (1.113) below cannot be written as an expectation. (This was shown by Bickel and Lehmann (1969).)

Linear functionals are often of interest. The statistical function M in equation (1.109), for example, is linear over the distribution function space of CDFs for which the integral exists.

It is important to recognize that a given functional may not exist at a given CDF. For example, if

$$F(y) = 1/2 + \tan^{-1}((y-\alpha)/\beta))/\pi \tag{1.110}$$

(that is, the distribution is Cauchy), then $M(F)$ does not exist. (Recall that I follow the convention that when I write an expression such as $M(F)$ or $\Upsilon(F)$, I generally imply the existence of the functional for the given F. That is, I do not always use a phrase about existence of something that I implicitly assume exists.)

Also, for some parametric distributions, such as the family of beta distributions, there may not be a "nice" functional that yields the parameter.

A functional of a CDF is generally a function of any parameters associated with the distribution, and in fact we often define a parameter as a statistical function. For example, if μ and σ are parameters of a distribution with CDF $F(y\,;\,\mu,\sigma)$ and Υ is some functional, we have

$$\Upsilon(F(y\,;\,\mu,\sigma)) = g(\mu,\sigma),$$

for some function g. If, for example, the M in equation (1.109) above is Υ and the F is the normal CDF $F(y\,;\,\mu,\sigma)$, then $\Upsilon(F(y\,;\,\mu,\sigma)) = \mu$.

Moments

For a univariate distribution with CDF F, the r^{th} *central moment* from equation (1.50), if it exists, is the functional

$$\begin{aligned}\mu_r &= M_r(F)\\ &= \int (y-\mu)^r \mathrm{d}F(y).\end{aligned} \tag{1.111}$$

For general random vectors or random variables with more complicated structures, this expression may be rather complicated. For $r = 2$, the matrix of joint moments for a random vector, as given in equation (1.69), is the functional

$$\Sigma(F) = \int (y-\mu)(y-\mu)^{\mathrm{T}}\,\mathrm{d}F(y). \tag{1.112}$$

Quantiles

Another set of useful distributional measures for describing a univariate distribution with CDF F are the quantiles. For $\pi \in]0,1[$, the π *quantile* is given by the functional $\Xi_\pi(F)$:

$$\Xi_\pi(F) = \inf\{y,\ \text{s.t. } F(y) \geq \pi\}. \tag{1.113}$$

This functional is the same as the quantile function or the generalized inverse CDF,

$$\Xi_\pi(F) = F^{-1}(\pi), \tag{1.114}$$

as given in Definition 1.15.

The 0.5 quantile is an important one; it is called the *median*. For the Cauchy distribution, for example, the moment functionals do not exist, but

the median does. An important functional for the Cauchy distribution is, therefore, $\Xi_{0.5}(F)$ because that is the location of the "middle" of the distribution.

Quantiles can be used for measures of scale and of characteristics of the shape of a distribution. A measure of the scale of a distribution, for example, is the *interquartile range*:

$$\Xi_{0.75} - \Xi_{0.25}. \tag{1.115}$$

Various measures of skewness can be defined as

$$\frac{(\Xi_{1-\pi} - \Xi_{0.5}) - (\Xi_{0.5} - \Xi_{\pi})}{\Xi_{1-\pi} - \Xi_{\pi}}, \tag{1.116}$$

for $0 < \pi < 0.5$. For $\pi = 0.25$, this is called the *quartile skewness* or the *Bowley coefficient*. For $\pi = 0.125$, it is called the *octile skewness*. These can be especially useful with the measures based on moments do not exist. The extent of the peakedness and tail weight can be indicated by the ratio of interquantile ranges:

$$\frac{\Xi_{1-\pi_1} - \Xi_{\pi_1}}{\Xi_{1-\pi_2} - \Xi_{\pi_2}}. \tag{1.117}$$

These measures can be more useful than the kurtosis coefficient based on the fourth moment, because different choices of π_1 and π_2 emphasize different aspects of the distribution. In expression (1.117), $\pi_1 = 0.025$ and $\pi_2 = 0.125$ yield a good measure of tail weight, and $\pi_1 = 0.125$ and $\pi_2 = 0.25$ in expression (1.117) yield a good measure of peakedness.

L_J Functionals

Various modifications of the mean functional M in equation (1.109) are often useful, especially in robust statistics. A functional of the form

$$L_J(F) = \int y J(y) \, \mathrm{d}F(y), \tag{1.118}$$

for some given function J, is called an L_J *functional*. If $J \equiv 1$, this is the mean functional. Often J is defined as a function of $F(y)$.

A "trimmed mean", for example, is defined by an L_J functional with $J(y) = (\beta - \alpha)^{-1} \mathrm{I}_{]\alpha,\beta[}(F(y))$, for constants $0 \leq \alpha < \beta \leq 1$ and where I is the indicator function. In this case, the L_J functional is often denoted as $T_{\alpha,\beta}$. Often β is taken to be $1-\alpha$, so the trimming is symmetric in probability content.

M_ρ Functionals

Another family of functionals that generalize the mean functional are defined as a solution to the minimization problem

$$\int \rho(y, M_\rho(F))\,\mathrm{d}F(y) = \min_{\theta\in\Theta} \int \rho(y,\theta)\,\mathrm{d}F(y), \tag{1.119}$$

for some function ρ and where Θ is some open subset of $\mathbb{R}^d$. A functional defined as the solution to this optimization problem is called an M_ρ *functional*. (Note the similarity in names and notation: we call the M in equation (1.109) the mean functional; and we call the M_ρ in equation (1.119) the M_ρ functional.)

Two related functions that play important roles in the analysis of M_ρ functionals are

$$\psi(y,t) = \frac{\partial \rho(y,t)}{\partial t}, \tag{1.120}$$

and

$$\lambda_F(t) = \int \psi(y,t)\mathrm{d}F(y) = \frac{\partial}{\partial t}\int \rho(y,t)\mathrm{d}F(y) \tag{1.121}$$

If y is a scalar and $\rho(y,\theta) = (y-\theta)^2$ then $M_\rho(F)$ is the mean functional from equation (1.109). Other common functionals also yield solutions to the optimization problem (1.119); for example, for $\rho(y,\theta) = |y-\theta|$, $\Xi_{0.5}(F)$ from equation (1.113) is an M_ρ functional (possibly nonunique).

We often choose the ρ in an M_ρ functional to be a function of $y-\theta$, and to be convex and differentiable. In this case, the M_ρ functional is the solution to

$$\mathrm{E}(\psi(Y-\theta)) = 0, \tag{1.122}$$

where

$$\psi(y-\theta) = \mathrm{d}\rho(y-\theta)/\mathrm{d}\theta,$$

if that solution is in the interior of Θ.

1.1.10 Transformations of Random Variables

We often need to determine the distribution of some transformation of a given random variable or a set of random variables. In the simplest case, we have a random variable X with known distribution and we want to determine the distribution of $Y = h(X)$, where h is a full-rank transformation; that is, there is a function h^{-1} such that $X = h^{-1}(Y)$. In other cases, the function may not be full-rank, for example, X may be an n-vector, and $Y = \sum_{i=1}^n X_i$.

There are three general approaches to the problem: the method of CDFs; the method of direct change of variables, including convolutions; and the method of CFs or MGFs. Sometimes one method works best, and other times some other method works best.

Method of CDFs

Given X with known CDF F_X and $Y = h(X)$ is invertible as above, we can write the CDF F_Y of Y as

$$\begin{aligned} F_Y(y) &= \Pr(Y \le y) \\ &= \Pr(h(X) \le y) \\ &= \Pr(X \le h^{-1}(y)) \\ &= F_X(h^{-1}(y)). \end{aligned}$$

Example 1.11 distribution of the minimum order statistic in a two-parameter exponential distribution
Consider a shifted version of the exponential family of distributions, called the two-parameter exponential with parameter (α, θ). Suppose the random variables $X_1, \ldots, X_n$ are iid with Lebesgue PDF

$$p_{\alpha,\theta}(x) = \theta^{-1} \mathrm{e}^{-(x-\alpha)/\theta} \mathrm{I}_{]\alpha,\infty[}(x).$$

We want to find the distribution of the minimum of $\{X_1, \ldots, X_n\}$. Let us denote that minimum by Y. (This is an order statistic, for which we use a general notation, $Y = X_{(1)}$. We discuss distributions of order statistics more fully in Section 1.1.12.) We have

$$\begin{aligned} \Pr(Y \le t) &= 1 - \Pr(Y > t) \\ &= 1 - \Pr(X_i > t \text{ for } i = 1, \ldots, n) \\ &= 1 - (\Pr(X_i > t \; \forall X_i))^n \\ &= 1 - (1 - \Pr(X_i \le t \; \forall X_i))^n \\ &= 1 - (\mathrm{e}^{-(t-\alpha)/\theta})^n \\ &= 1 - \mathrm{e}^{-n(t-\alpha)/\theta}. \end{aligned}$$

This is the CDF for a two-parameter exponential distribution with parameters α and θ/n. If instead of a two-parameter exponential distribution, we began with the more common one-parameter exponential distribution with parameter θ, the distribution of Y would be the one-parameter exponential distribution with parameter θ/n. ∎

Example 1.12 distribution of the square of a continuous random variable
Given X with CDF F_X and Lebesgue PDF f_X. Let $Y = X^2$. For $x < 0$, $Y^{-1}[]-\infty, x]] = \emptyset$, and $Y^{-1}[]-\infty, x]] = X^{-1}[[-\sqrt{x}, \sqrt{x}]]$, otherwise. Therefore the CDF F_Y of Y is

$$\begin{aligned} F_Y(x) &= \Pr(Y^{-1}[] - \infty, x]]) \\ &= \Pr(X^{-1}[[-\sqrt{x}, \sqrt{x}]]) \\ &= (F_X(\sqrt{x}) - F_X(-\sqrt{x}))\mathrm{I}_{\overline{\mathbb{R}}_+}(x). \end{aligned}$$

Differentiating, we have the Lebesgue PDF of Y:

$$f_Y(x) = \frac{1}{2\sqrt{x}}(f_X(\sqrt{x}) + f_X(-\sqrt{x}))\mathrm{I}_{\overline{\mathbb{R}}_+}(x). \tag{1.123}$$

∎

Method of Change of Variables

If X has density $p_X(x|\alpha,\theta)$ and $Y = h(X)$, where h is a full-rank transformation (that is, there is a function h^{-1} such that $X = h^{-1}(Y)$), then the density of Y is

$$p_Y(y|\alpha,\theta) = p_X\left(h^{-1}(y)|\alpha,\theta\right)|\mathrm{J}_{h^{-1}}(y)|, \tag{1.124}$$

where $\mathrm{J}_{h^{-1}}(y)$ is the Jacobian of the inverse transformation, and $|\cdot|$ is the determinant.

Constant linear transformations are particularly simple. If X is an n-vector random variable with PDF f_X and A is an $n \times n$ constant matrix of full rank, the PDF of $Y = AX$ is $f_X|\det(A^{-1})|$.

In the change of variable method, we think of h as a mapping of the range $\mathcal{X}$ of the random variable X to the range $\mathcal{Y}$ of the random variable Y, and the method works by expressing the probability content of small regions in $\mathcal{Y}$ in terms of the probability content of the pre-image of those regions in $\mathcal{X}$.

For a given function h, we often must decompose $\mathcal{X}$ into disjoint sets over each of which h is one-to-one.

Example 1.13 distribution of the square of a standard normal random variable
Suppose $X \sim \mathrm{N}(0,1)$, and let $Y = h(X) = X^2$. The function h is one-to-one over $]-\infty, 0[$ and, separately, one-to-one over $[0, -\infty[$. We could, of course, determine the PDF of Y using equation (1.123), but we will use the change-of-variables technique.

The absolute value of the Jacobian of the inverse over both regions is $x^{-1/2}$. The Lebesgue PDF of X is

$$f_X(x) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-x^2/2};$$

hence, the Lebesgue PDF of Y is

$$f_Y(y) = \frac{1}{\sqrt{2\pi}}y^{-1/2}\mathrm{e}^{-y/2}\mathrm{I}_{\mathbb{R}_+}(y). \tag{1.125}$$

This is the PDF of a chi-squared random variable with one degree of freedom χ^2_1 (see Table A.3). ∎

Sums

A simple application of the change of variables method is in the common situation of finding the distribution of the sum of two scalar random variables that are independent but not necessarily identically distributed.

Suppose X is a d-variate random variable with PDF f_X, Y is a d-variate random variable with PDF f_Y, and X and Y are independent. We want the density of $U = X + Y$. We form another variable $V = Y$ and the matrix

$$A = \begin{pmatrix} I_d & I_d \\ 0 & I_d \end{pmatrix},$$

so that we have a full-rank transformation, $(U,V)^{\mathrm{T}} = A(X,Y)^{\mathrm{T}}$ The inverse of the transformation matrix is

$$A^{-1} = \begin{pmatrix} I_d & -I_d \\ 0 & I_d \end{pmatrix},$$

and the Jacobian is 1. Because X and Y are independent, their joint PDF is $f_{XY}(x,y) = f_X(x)f_Y(y)$, and the joint PDF of U and V is $f_{UV}(u,v) = f_X(u-v)f_Y(v)$; hence, the PDF of U is

$$\begin{aligned} f_U(u) &= \int_{\mathbb{R}^d} f_X(u-v)f_Y(v)\mathrm{d}v \\ &= \int_{\mathbb{R}^d} f_Y(u-v)f_X(v)\mathrm{d}v. \end{aligned} \tag{1.126}$$

We call f_U the *convolution* of f_X and f_Y. The commutative operation of convolution occurs often in applied mathematics, and we denote it by $f_U = f_X \star f_Y$. We often denote the convolution of a function f with itself by $f^{(2)}$; hence, the PDF of $X_1 + X_2$ where X_1, X_2 are iid with PDF f_X is $f_X^{(2)}$. From equation (1.126), we see that the CDF of U is the convolution of the CDF of one of the summands with the PDF of the other:

$$F_U = F_X \star f_Y = F_Y \star f_X. \tag{1.127}$$

In the literature, this operation is often referred to as the convolution of the two CDFs, and instead of as in equation (1.127), may be written as

$$F_U = F_X \star F_Y.$$

Note the inconsistency in notation. The symbol "$\star$" is overloaded. Following the latter notation, we also denote the convolution of the CDF F with itself as $F^{(2)}$.

Example 1.14 sum of two independent Poissons
Suppose X_1 is distributed as Poisson(θ_1) and X_2 is distributed independently as Poisson(θ_2). By equation (1.126), we have the probability function of the sum $U = X_1 + X_2$ to be

$$\begin{aligned} f_U(u) &= \sum_{v=0}^{u} \frac{\theta_1^{u-v}\mathrm{e}^{\theta_1}}{(u-v)!}\frac{\theta_2^{v}\mathrm{e}^{\theta_2}}{v!} \\ &= \frac{1}{u!}\mathrm{e}^{\theta_1+\theta_2}(\theta_1+\theta_2)^u. \end{aligned}$$

The sum of the two independent Poissons is distributed as Poisson($\theta_1+\theta_2$). ∎

Table 1.1. Distributions of the Sums of Independent Random Variables

Distributions of X_i for $i = 1, \ldots, k$	Distribution of $\sum X_i$
Poisson(θ_i)	Poisson$(\sum \theta_i)$
Bernoulli(π)	binomial(k, π)
binomial(n_i, π)	binomial$(\sum n_i, \pi)$
geometric(π)	negative binomial(k, π)
negative binomial(n_i, π)	negative binomial$(\sum n_i, \pi)$
normal(μ_i, σ_i^2)	normal$(\sum \mu_i, \sum \sigma_i^2)$
exponential(β)	gamma(k, β)
gamma(α_i, β)	gamma$(\sum \alpha_i, \beta)$

The property shown in Example 1.14 obviously extends to k independent Poissons. Other common distributions also have this kind of property for sums, as shown in Table 1.1. For some families of distributions such as binomial, negative binomial, and gamma, the general case is the sum of special cases.

The additive property of the gamma distribution carries over to the special cases: the sum of k iid exponentials with parameter θ is gamma(k, θ) and the sum of independent chi-squared variates with $\nu_1, \ldots, \nu_k$ degrees of freedom is distributed as $\chi^2_{\sum \nu_i}$.

The are other distributions that could be included in Table 1.1 if the parameters met certain restrictions, such as being equal; that is, the random variables in the sum are iid.

In the case of the inverse Gaussian(μ, λ) distribution, a slightly weaker restriction than iid allows a useful result on the distribution of the sum so long as the parameters have a fixed relationship. If $X_1, \ldots, X_k$ are independent and X_i is distributed as inverse Gaussian$(\mu_0 \alpha_i, \lambda_0 \alpha_i^2)$, then $\sum X_i$ is distributed as inverse Gaussian$(\mu_0 \sum \alpha_i, \lambda_0 (\sum \alpha_i)^2)$.

Products

Another simple application of the change of variables method is for finding the distribution of the product or the quotient of two scalar random variables that are independent but not necessarily identically distributed.

Suppose X is a random variable with PDF f_X and Y is a random variable with PDF f_Y and X and Y are independent, and we want the density of the product $U = XY$. As for the case with sums, we form another variable $V = Y$, form the joint distribution of U and V using the Jacobian of the inverse transformation, and finally integrate out V. Analogous to equation (1.126), we have

$$f_U(u) = \int_{-\infty}^{\infty} f_X(u/v) f_Y(v) v^{-1} \mathrm{d}v, \tag{1.128}$$

and for the quotient $W = X/Y$, we have

$$f_W(w) = \int_{-\infty}^{\infty} f_X(wv) f_Y(v) v \mathrm{d}v. \tag{1.129}$$

Example 1.15 the F family of distributions
Suppose Y_1 and Y_2 are independent chi-squared random variables with ν_1 and ν_2 degrees of freedom respectively. We want to find the distribution of $W = Y_1/Y_2$. Along with the PDFs of chi-squared random variables, equation (1.129) yields

$$f_W(w) \propto w^{\nu_1/2-1} \int_0^{\infty} v^{(\nu_1+\nu_2)/2-1} \mathrm{e}^{-(w+1)v/2} \mathrm{d}v.$$

This integral can be evaluated by making the change of variables $z = (w+1)v$. After separating out the factors involving w, the remaining integrand is the PDF of a chi-squared random variable with $\nu_1 + \nu_2 - 1$ degrees of freedom. Finally, we make one more change of variables: $F = W\nu_2/\nu_1$. This yields

$$f_F(f) \propto \frac{f^{\nu_1/2-1}}{(\nu_2 + \nu_1 f)^{(\nu_1+\nu_2)/2}}. \tag{1.130}$$

This is the PDF of an F random variable with ν_1 and ν_2 degrees of freedom (see Table A.3). It is interesting to note that the mean of such a random variable depends only on ν_2. ∎

Method of MGFs or CFs

In this method, for the transformation $Y = h(X)$ we write the MGF of Y as $\mathrm{E}\left(\mathrm{e}^{t^{\mathrm{T}}Y}\right) = \mathrm{E}\left(\mathrm{e}^{t^{\mathrm{T}}h(X)}\right)$, or we write the CF in a similar way. If we can work out the expectation (with respect to the known distribution of X), we have the MGF or CF of Y, which determines its distribution.

The MGF or CF technique is particularly useful in the case when Y is the sum from a simple random sample. If

$$Y = X_1 + \cdots + X_n,$$

where $X_1, \ldots, X_n$ are iid with CF $\varphi_X(t)$, we see from the linearity of the expectation operator that the CF of Y is

$$\varphi_Y(t) = (\varphi_X(t))^n. \tag{1.131}$$

We use this approach in the proof of the simple CLT, Theorem 1.38 on page 87.

The MGF or CF for a linear transformation of a random variable has a simple relationship to the MGF or CF of the random variable itself, as we can easily see from the definition. Let X be a random variable in $\mathbb{R}^d$, A be a $d \times m$ matrix of constants, and b be a constant m-vector. Now let

$$Y = A^{\mathrm{T}}X + b.$$

Then

$$\varphi_Y(t) = \mathrm{e}^{\mathrm{i}b^{\mathrm{T}}t} \varphi_X(At), \quad t \in \mathbb{R}^m. \tag{1.132}$$

Example 1.16 distribution of the sum of squares of independent standard normal random variables
Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{N}(0,1)$, and let $Y = \sum X_i^2$. In Example 1.13, we saw that $Y_i = X_i^2 \overset{\mathrm{d}}{=} \chi^2_1$. Because the X_i are iid, the Y_i are iid. Now the MGF of a χ^2_1 is

$$\begin{aligned}\mathrm{E}\left(\mathrm{e}^{tY_i}\right) &= \int_0^\infty \frac{1}{\sqrt{2\pi}} y^{-1/2} \mathrm{e}^{-y(1-2t)/2} \mathrm{d}y \\ &= (1-2t)^{-1/2} \quad \text{for} \quad t < \frac{1}{2}.\end{aligned}$$

Hence, the MGF of Y is $(1-2t)^{n/2}$ for $t < 1/2$, which is seen to be the MGF of a chi-squared random variable with n degrees of freedom.

This is a very important result for applications in statistics. ∎

1.1.11 Decomposition of Random Variables

We are often interested in the sum of random numbers,

$$S_k = X_1 + \cdots + X_k. \tag{1.133}$$

Because the sum may grow unwieldy as k increases, we may work with normed sums of the form S_k/a_k.

In order to develop interesting properties of S_k, there must be some commonality among the individual X_i. The most restrictive condition is that the X_i be iid. Another condition for which we can develop meaningful results is that the X_i be independent, but different subsequences of them may have different distributions.

The finite sums that we consider in this section have relevance in the limit theorems discussed in Sections 1.4.1 and 1.4.2.

Infinitely Divisible Distributions

Instead of beginning with the X_i and forming their sum, we can think of the problem as beginning with a random variable X and decomposing it into additive components. A property of a random variable that allows a particular kind of additive decomposition is called *divisibility*.

Definition 1.31 (n-divisibility of random variables)
Given a random variable X and an integer $n \geq 2$, we say X is *n-divisible* if there exist iid random variables $X_1, \ldots, X_n$ such that

$$X \overset{\mathrm{d}}{=} X_1 + \cdots + X_n.$$

∎

Notice that $(n+1)$-divisibility does not imply n-divisibility. We can see this in the example of a random variable X with a binomial $(3, \pi)$ distribution, which is clearly 3-divisible (as the sum of three Bernoullis). We see, however, that X is not 2-divisible; that is, there cannot be iid X_1 and X_2 such that $X = X_1 + X_2$, because if there were, each X_i would take values in $[0, 3/2]$ with $\Pr(X_i = 0) > 0$ and $\Pr(X_i = 0) > 0$, but in that case $\Pr(X = 3/2) > 0$.

Rather than divisibility for some fixed n, a more useful form of divisibility is infinite divisibility; that is, divisibility for any $n \geq 2$.

Definition 1.32 (infinite divisibility of random variables)
A nondegenerate random variable X is said to be *infinitely divisible* if for every positive integer n, there are iid random variables $X_{n1}, \ldots, X_{nn}$, such that

$$X \stackrel{\mathrm{d}}{=} X_{n1} + \cdots + X_{nn}. \tag{1.134}$$

∎

Note that the distributions of the X_ns may be different from each other, but for a given n, $X_{n1}, \ldots, X_{nn}$ are identically distributed.

The random variables $X_{n1}, \ldots, X_{nn}$ in the definition above are a special case of a *triangular array* in which for given n, the $X_{n1}, \ldots, X_{nn}$ are iid. We encounter similar, more general triangular arrays in some of the limit theorems of Section 1.4.2 (on page 106).

Because infinite divisibility is associated with the distribution of the random variable, we will refer to other characteristics of an infinitely divisible random variable, such as the PDF or CDF, as being infinitely divisible. In the same form as equation (1.131), the characteristic function of an infinitely divisible random variable for any positive integer n can be written as

$$\varphi_X(t) = (\varphi_{X_n}(t))^n,$$

for some characteristic function $\varphi_{X_n}(t)$. Furthermore, if the characteristic function of a random variable X can be expressed in this way for any n, then X is infinitely divisible.

The normal, the Cauchy, and the Poisson families of distributions are all infinitely divisible (exercise).

Divisibility properties are particularly useful in stochastic processes.

Stable Distributions

Another type of decomposition leads to the concept of *stability*. In this setup we require that the full set of random variables be iid.

Definition 1.33 (stable random variables) ˜
Let $X, X_1, \ldots, X_n$ be iid nondegenerate random variables. If for each n there exist numbers d_n and positive numbers c_n, such that

$$X_1 + \cdots + X_n \overset{\mathrm{d}}{=} c_n X + d_n, \tag{1.135}$$

then X is said to have a *stable distribution* (or to be a *stable random variable*). ∎

This family of distributions is symmetric. We see this by noting that for the variables X_1 and X_2 in Definition 1.33 or in equations (1.135) or (1.136), $Y = X_1 - X_2$ has a stable distribution that is symmetric about 0 (exercise). Because there is a generalization of the stable family of distributions that is not symmetric (see page 184), the family defined here is sometimes called the *symmetric stable family*.

Definition 1.33 is equivalent to the requirement that there be three iid nondegenerate random variables, X, X_1, and X_2, such that for arbitrary constants a_1 and a_2, there exists constants c and d such that

$$a_1 X_1 + a_2 X_2 \overset{\mathrm{d}}{=} cX + d. \tag{1.136}$$

In this case, X is has a stable distribution (or is a stable random variable). It is an exercise to show that this is the case.

If X is a stable random variable as defined in equation (1.135), then there exists a number $\alpha \in]0, 2]$ such that

$$c^\alpha = n^{1/\alpha}. \tag{1.137}$$

(See Feller (1971), Section VI.1, for a proof.) The number α is called the *index of stability* or the *characteristic exponent*, and the random variable with that index is said to be α-stable. The normal family of distributions is stable with characteristic exponent of 2 and the Cauchy family is stable with characteristic exponent of 1 (exercise).

An infinitely divisible family of distributions is stable (exercise). The Poisson family is an example of a family of distributions that is infinitely divisible, but not stable (exercise).

1.1.12 Order Statistics

In a set of iid random variables $X_1, \ldots, X_n$, it is often of interest to consider the ranked values $X_{i_1} \leq \cdots \leq X_{i_n}$. These are called the *order statistics* and are denoted as $X_{(1:n)}, \ldots, X_{(n:n)}$. For $1 \leq k \leq n$, we refer to $X_{(k:n)}$ as the k^{th} order statistic. We often use the simpler notation $X_{(k)}$, assuming that n is some fixed and known value. Also, many other authors drop the parentheses in the other representation, $X_{k:n}$.

Theorem 1.25
In a random sample of size n from a distribution with PDF f dominated by a σ-finite measure and CDF F given the k^{th} order statistic $X_{(k)} = a$, the $k-1$ random variables less than $X_{(k)}$ are iid with PDF

$$f_{\text{right}}(x) = \frac{1}{F(a)} f(x) \mathrm{I}_{]-\infty,a[}(x),$$

and the $n-k$ random variables greater than $X_{(k)}$ are iid with PDF

$$f_{\text{left}}(x) = \frac{1}{1-F(a)} f(x) \mathrm{I}_{]a,\infty[}(x).$$

Proof. Exercise. ∎

Using Theorem 1.25, forming the joint density, and integrating out all variables except the k^{th} order statistic, we can easily work out the density of the k^{th} order statistic as

$$f_{X_{(k)}}(x) = \binom{n}{k} \Big(F(x)\Big)^{k-1} \Big(1-F(x)\Big)^{n-k} f(x). \tag{1.138}$$

Example 1.17 distribution of order statistics from $\mathrm{U}(0,1)$
The distribution of the k^{th} order statistic from a $\mathrm{U}(0,1)$ is the beta distribution with parameters k and $n-k+1$, as we see from equation (1.138). ∎

From equation (1.138), the CDF of the k^{th} order statistic in a sample from a distribution with CDF F can be expressed in terms of the regularized incomplete beta function (equation (C.10) on page 866) as

$$F_{X_{(k)}}(x) = I_{F(x)}(k, n-k+1). \tag{1.139}$$

The joint density of all order statistics is

$$n! \prod_{i=1}^{n} f(x_{(i)}) \mathrm{I}_{x_{(1)} \le \cdots \le x_{(n)}}(x_{(1)}, \ldots, x_{(n)}) \tag{1.140}$$

The joint density of the i^{th} and j^{th} $(i < j)$ order statistics is

$$\frac{n!}{(i-1)!(j-i-1)!(n-j)!} \cdot$$

$$\Big(F(x_{(i)})\Big)^{i-1} \Big(F(x_{(j)}) - F(x_{(i)})\Big)^{j-i-1} \Big(1-F(x_{(j)})\Big)^{n-j} f(x_{(i)}) f(x_{(j)}). \tag{1.141}$$

Although order statistics are obviously not independent, differences of order statistics or functions of those differences are sometimes independent, as we see in the case of the exponential family a uniform family in the following examples. These functions of differences of order statistics often are useful in statistical applications.

Example 1.18 another distribution from a two-parameter exponential distribution (continuation of Example 1.11)

Suppose the random variables $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{exponential}(\alpha, \theta)$. (Note that $n \geq 2$.) Let $Y = X_{(1)}$ as before and let $Y_1 = \sum(X_i - X_{(1)})$. We want to find the distribution of Y_1.

Note that

$$\begin{aligned} Y_1 &= \sum \left(X_i - X_{(1)}\right) \\ &= \sum X_i - nX_{(1)} \\ &= \sum \left(X_{(i)} - X_{(1)}\right). \end{aligned}$$

Now, for $k = 2, \ldots, n$, let

$$Y_k = (n - k + 1)\left(X_{(k)} - X_{(k-1)}\right).$$

Using the change-of-variables technique, we see that $Y_k \overset{\text{iid}}{\sim} \text{exponential}(0, \theta)$, and are independent of $X_{(i)}$. We have independence because the resulting joint density function factorizes. (This is the well-known exponential spacings property.)

Now, $\sum_{k=2}^{n} Y_k = \sum(X_{(i)} - X_{(1)}) = Y_1$, and the distribution of the sum of $n - 1$ iid exponentials with parameters 0 and θ, multiplied by θ, is a gamma with parameters $n - 1$ and 1. ∎

Example 1.19 distribution of order statistics from $\mathrm{U}(\theta_1, \theta_2)$
Let $X_{(1)}, \ldots, X_{(n)}$ be the order statistics in a sample of size n from the $\mathrm{U}(\theta_1, \theta_2)$ distribution. While it is clear that $X_{(1)}$ and $X_{(n)}$ are not independent, the random variables

$$Y_i = \frac{X_{(i)} - X_{(1)}}{X_{(n)} - X_{(1)}}, \quad \text{for } i = 2, \ldots, n-1,$$

are independent of both $X_{(1)}$ and $X_{(n)}$. (Proving this is Exercise 1.47.) ∎

Quantiles and Expected Values of Order Statistics

We would expect the k^{th} order statistic in an iid sample of size n to have some relationship to the k/n quantile of the underlying distribution. Because of the limited set of values that k/n can take on for any given n, there are obvious problems in determining a direct relationship between quantiles and order statistics. For given $0 < \pi < 1$, we call the $(\lceil n\pi \rceil)^{\text{th}}$ order statistic, $X_{(\lceil n\pi \rceil:n)}$, the *sample quantile* of order π.

The expected value of the k^{th} order statistic in a sample of size n, if it exists, should be approximately the same as the k/n quantile of the underlying distribution; and indeed, this is the case. We will consider this issue for large n in Theorem 1.48, but for now, the first question is whether $\mathrm{E}(X_{(k:n)})$ exists.

For simplicity in the following, let $\mu_{(k:n)} = \mathrm{E}(X_{(k:n)})$. It is easy to see that $\mu_{(k:n)}$ exists and is finite if $\mathrm{E}(X)$ exists and is finite, where the distribution of

X is the underlying distribution of the sample (exercise). The converse of this statement is not true, and there are important cases in which $\mu_{(k:n)}$ exists and is finite, but the $\mathrm{E}(X)$ does not exist. For example, in the Cauchy distribution, $\mu_{(k:n)}$ exists and is finite unless $k = 1$ or n (see Exercise 1.48b).

For the iid random variables $X_1, X_2, \ldots$ if the expectations $\mu_{(k:n)}$ exist, we have the following relationship:

$$(n-k)\mu_{(k:n)} + k\mu_{(k+1:n)} = n\mu_{(k:n-1)}. \tag{1.142}$$

(See Exercise 1.49.)

1.2 Series Expansions

Series expansions are useful in studying properties of probability distributions. We may represent a CDF or PDF as a series of some basis functions. For example, the density may be written as

$$f(x) = \sum_{r=0}^{\infty} c_r g_r(x).$$

In such an expansion, the basis functions are often chosen as derivatives of the normal distribution function, as in the Edgeworth expansions discussed in Section 1.2.4.

1.2.1 Asymptotic Properties of Functions

***The development of the series representation allows us to investigate the rate of convergence of the moments.

*** Consider a set of iid random variables, $X_1, \ldots, X_n$, and a function T_n of those random variables. Suppose T_n converges in distribution to the distribution of the normal variate Z. A simple example of such a T_n is the standardized sample mean $n^{1/2}(\overline{X}_n - \mu)/\sigma$ from a sample of size n. If

$$T_n \xrightarrow{d} Z,$$

then the characteristic function converges:

$$\varphi_{T_n}(t) = \mathrm{E}(e^{itT_n}) \to \mathrm{E}(e^{itZ}).$$

***The question is how fast does it converge. At what rate (in n) do the moments of T_n approach those of Z?

***fix Taylor series expansion *** equation (1.108) $\kappa_3 = K'''(0)$

$$K_{Y_n}(t) = \frac{1}{2}t^2 + \frac{\kappa_3}{6\sqrt{n}\sigma^3}t^3 + \cdots,$$

Thus, if the characteristic function for a random variable T_n, $\varphi_{T_n}(t)$, can be written as a series in terms involving n, it may be possible to determine the order of convergence of the moments (because the derivatives of the cf yield the moments).

expansion of statistical functionals

1.2.2 Expansion of the Characteristic Function

The characteristic function

$$\varphi_X(t) = \mathrm{E}(e^{itX})$$

is useful for determining moments or cumulants of the random variable X, for example, $\varphi'(0) = i\mathrm{E}(X)$ and $\varphi''(0) = -\mathrm{E}(X^2)$.

If $T_n = n^{1/2}(\overline{X}_n - \mu)/\sigma$, it has an asymptotic standard normal distribution, and

$$\begin{aligned}\varphi_{T_n}(t) &= \mathrm{E}(e^{itT_n}) \\ &= \mathrm{E}(e^{itn^{1/2}Z}) \\ &= \left(\varphi_Z(t/n^{1/2})\right)^n,\end{aligned}$$

where Z is standard normal.

Now the j^{th} cumulant, κ_j, of Z is the coefficient of $(it)^j/j!$ in a power series expansion of the log of the characteristic function, so

$$\varphi_Z(t) = \exp\left(\kappa_1 it + \frac{1}{2}\kappa_2(it)^2 + \ldots + \frac{1}{j!}\kappa_j(it)^j + \ldots\right),$$

but also

$$\varphi_Z(t) = 1 + \mathrm{E}(Z)it + \frac{1}{2}\mathrm{E}(Z^2)(it)^2 + \ldots + \frac{1}{j!}\mathrm{E}(Z^j)(it)^j + \ldots.$$

1.2.3 Cumulants and Expected Values

With some algebra, we can write the cumulants in terms of the expected values as

$$\begin{aligned}\sum_{j\geq 1}\frac{1}{j!}\kappa_j(it)^j &= \log\left(1 + \sum_{j\geq 1}\frac{1}{j!}\mathrm{E}(Z^j)(it)^j\right) \\ &= \sum_{k\geq 1}(-1)^{k+1}\frac{1}{k}\left(\sum_{j\geq 1}\frac{1}{j!}\mathrm{E}(Z^j)(it)^j\right)^k.\end{aligned}$$

Equating coefficients, we get

$$\begin{aligned}
\kappa_1 &= \mathrm{E}(Z) \\
\kappa_2 &= \mathrm{E}(Z^2) - (\mathrm{E}(Z))^2 = \mathrm{V}(Z) \\
\kappa_3 &= \mathrm{E}((Z - \mathrm{E}(Z))^3) \\
\kappa_4 &= \mathrm{E}((Z - \mathrm{E}(Z))^4) - 3(\mathrm{V}(Z))^2
\end{aligned}$$

and so on. (There's a lot of algebra here!)

One thing to notice is that the cumulant κ_j is a homogeneous polynomial of degree j in the moments (i.e., each term is of degree j).

Now, back to the characteristic function of T_n: $\varphi_{T_n}(t) = (\varphi_Z(t/n^{1/2}))^n$. Using the fact that Z is standard normal (so $\kappa_1 = 0$ and $\kappa_2 = 1$), we can write (using the series expansion of the exponential function the last step),

$$\begin{aligned}
\varphi_{T_n}(t) &= \exp\left(\frac{1}{2}(it)^2 + n^{-1/2}\frac{1}{3!}\kappa_3(it)^3 + \ldots + \right. \\
&\qquad \left. n^{-(j-2)/2}\frac{1}{j!}\kappa_j(it)^j + \ldots\right), \\
&= e^{-t^2/2}\exp\left(n^{-1/2}\frac{1}{3!}\kappa_3(it)^3 + \ldots + \right. \\
&\qquad \left. n^{-(j-2)/2}\frac{1}{j!}\kappa_j(it)^j + \ldots\right), \\
&= e^{-t^2/2}\left(1 + n^{-1/2}r_1(it)n^{-1}r_2(it) + \ldots + \right. \\
&\qquad \left. n^{-j/2}r_j(it) + \ldots\right),
\end{aligned}$$

where r_j is a polynomial of degree $3j$, with real coefficients, and depends on the cumulants $\kappa_3, \ldots, \kappa_{j+2}$, but does not depend on n.

$$r_1(x) = \frac{1}{6}\kappa_3 x^3$$

and

$$r_2(x) = \frac{1}{24}\kappa_4 x^4 + \frac{1}{72}\kappa_3^2 x^6$$

for example. Note that r_j is an even polynomial when j is even and is odd when j is odd.

The relevance of this is that it indicates the rate of convergence of the moments.

1.2.4 Edgeworth Expansions in Hermite Polynomials

An Edgeworth expansion represents a distribution function as a series in derivatives of the normal distribution function, or of the density, $\phi(x)$. Taking successive derivatives of the normal distribution function yields a series of polynomials that are orthogonal with respect to the normal density.

If $D = \frac{d}{dx}$, we have,

$$\begin{aligned} D\phi(x) &= -x\phi(x), \\ D^2\phi(x) &= (x^2 - 1)\phi(x), \\ D^3\phi(x) &= (-x^3 + 3x)\phi(x) \\ &\cdots \vdots \cdots \end{aligned}$$

Letting $p_i(x) = D^{(i)}\phi(x)$, we see that the $p_i(x)$ are orthogonal with respect to the normal density, that is,

$$\int_{-\infty}^{\infty} p_i(x)p_j(x)\phi(x)\,dx = 0,$$

when $i \neq j$.

From the polynomial factors of **********, we identify the Hermite polynomials

$$\begin{aligned} H_0 &= 1 \\ H_1 &= x \\ H_2 &= x^2 - 1 \\ H_3 &= x^3 - 3x \\ &\cdots \vdots \cdots \end{aligned}$$

which as we discuss on page 753, is a series in Hermite polynomials times $\phi(x)$ and in the cumulants (or moments). A series using these Hermite polynomials is often called a Gram-Charlier series Edgeworth series.

The Edgeworth expansion for a given CDF F_X is

$$F_X(x) = \sum c_r H_r(x)\phi(x).$$

The series representation is developed by equating terms in the characteristic functions.

The adequacy of the series representation depends on how "close" the distribution of the random variable is to normal. If the random variable is asymptotically normal, the adequacy of the series representation of course depends on the rate of convergence of the distribution of the random variable to a normality.

1.2.5 The Edgeworth Expansion

Inverting the characteristic function transform yields a power series for the distribution function, F_{T_n}. After some more algebra, we can get this series in the form

$$F_{T_n}(x) = \Phi(x) + n^{-1/2}p_1(x)\phi(x) + n^{-1}p_2(x)\phi(x) + \ldots,$$

where the p_j's are polynomials containing the cumulants and simple combinations of Hermite polynomials.

This is the Edgeworth expansion of the distribution function of T_n.

The degree of p_j is $3j-1$, and is even for odd j, and odd for even j. This is one of the most important properties of this representation.

In this expansion, heuristically, the term of order $n^{-1/2}$ corrects the approximation for the effect of skewness, and the term of order n^{-1} corrects the approximation for the effect of kurtosis.

The first few Hermite polynomials are shown in equation (0.1.98) on page 753.

This is an instance of the more general method of representing a given function in terms of basis functions, as we discuss beginning on page 749.

1.3 Sequences of Spaces, Events, and Random Variables

Countably infinite sequences play the main role in the definition of the basic concept of a σ-field, and consequently, in the development of a theory of probability. Sequences of sets correspond to sequences of events and, consequently, to sequences of random variables. Unions, intersections, and complements of sequences of sets are important for studying sequences of random variables. The material in this section depends heavily on the properties of sequences of sets discussed on page 626 and the following pages.

At the most general level, we could consider a sequence of sample spaces, $\{(\Omega_n, \mathcal{F}_n, P_n)\}$, but this level of generality is not often useful. Sequences of σ-fields and probability measures over a fixed sample space, or sequences of probability measures over a fixed sample space and fixed σ-field are often of interest, however.

Sequences of σ-Fields and Associated Probability Measures

Given a sample space Ω, we may be interested in a sequence of probability spaces, $\{(\Omega, \mathcal{F}_n, P_n)\}$. Such a sequence could arise naturally from a sequence of events $\{A_n\}$ that generates a sequence of σ-fields. Beginning with some base σ-field $\mathcal{F}_0$, we have the increasing sequence

$$\mathcal{F}_0 \subseteq \mathcal{F}_1 \;=\; \sigma(\mathcal{F}_0 \cup \{A_1\}) \subseteq \mathcal{F}_2 \;=\; \sigma(\mathcal{F}_1 \cup \{A_2\}) \subseteq \cdots. \tag{1.143}$$

We have a sequence of probability spaces,

$$\{(\Omega, \mathcal{F}_n, P_n) \mid \mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}\}, \tag{1.144}$$

where the domains of the measures P_n are evolving, but otherwise the sample space and the other characteristics of P are not changing. This evolving sequence is the underlying paradigm of some stochastic processes, particularly martingales, that we discuss in Section 1.6.

Such a sequence of σ-fields could of course equivalently be generated by a sequence of random variables, instead of by a sequence of sets.

The sequence of probability measures exists and the measures are unique, as the Carathéodory extension theorem ensures (see page 712).

Sequences of Probability Measures

It is also of interest to consider a sequence of probability measures $\{P_n\}$ over a fixed measurable space $(\Omega, \mathcal{F})$. Such sequences have important applications in statistics. Convergent sequences of probability measures form the basis for asymptotic inference. We will consider the basic properties beginning on page 78, and then consider asymptotic inference in Section 3.8 beginning on page 306.

Sequences of Events; lim sup and lim inf

In most of our discussion of sequences of events and of random variables, we will assume that we have a single probability space, $(\Omega, \mathcal{F}, P)$. This assumption is implicit in a phrase such as "a sequence of events and an associated probability measure".

We begin by recalling a basic fact about a sequence of events and an associated probability measure:

$$\lim P(A_i) \leq P(\lim A_i), \tag{1.145}$$

and it is possible that

$$\lim P(A_i) \neq P(\lim A_i).$$

Compare this with the related fact about a useful sequence of intervals (page 647):

$$\lim_{n\to\infty} \bigcup_{i=1}^{n} \left[a + \frac{1}{i},\ b - \frac{1}{i}\right] \neq \bigcup \lim_{i\to\infty} \left[a + \frac{1}{i},\ b - \frac{1}{i}\right].$$

Consider a sequence of probabilities defined in terms of a sequence of events, $\{P(A_n)\}$. Two important types of limits of such sequences of probabilities are, similar to the analogous limits for sets defined on page 626,

$$\limsup_n P(A_n) \stackrel{\text{def}}{=} \inf_n \sup_{i\geq n} P(A_i) \tag{1.146}$$

$$\liminf_n P(A_n) \stackrel{\text{def}}{=} \sup_n \inf_{i\geq n} P(A_i). \tag{1.147}$$

Similarly to the corresponding relationship between unions and intersections of sequences of sets, we have the relationships in the following theorem.

Theorem 1.26
Let $\{A_n\}$ be a sequence of events in a probability space. Then

$$P(\limsup_n A_n) \leq \limsup_n P(A_n) \tag{1.148}$$

and

$$P(\liminf_n A_n) \geq \liminf_n P(A_n). \tag{1.149}$$

Proof.
Consider

$$B_n = \cup_{i=n}^{\infty} A_i$$

and

$$C_n = \cap_{i=n}^{\infty} A_i.$$

We see

$$B_n \searrow \limsup_n A_n,$$

and likewise

$$C_n \nearrow \liminf_n A_n.$$

Now we use the continuity of the measure to get

$$P(A_n) \leq P(B_n) \to P(\limsup_n A_n)$$

and

$$P(A_n) \geq P(C_n) \to P(\liminf_n A_n).$$

∎

For a sequence of sets $\{A_n\}$, we recall the intuitive interpretations of $\limsup_n A_n$ and $\liminf_n A_n$:

- An element ω is in $\limsup_n A_n$ iff for each n, there is some $i \geq n$ for which $\omega \in A_i$. This means that ω must lie in infinitely many of the A_n.
- An element ω is in $\liminf_n A_n$ iff there is some n such that for all $i \geq n$, $\omega \in A_i$. This means that ω must lie in all but finitely many of the A_n.

In applications of probability theory, the sets correspond to events, and generally we are more interested in those events that occur infinitely often; that is, we are more interested in $\limsup_n A_n$. We often denote this as "i.o.", and we define

$$\{A_n \text{ i.o.}\} = \limsup_n A_n. \tag{1.150}$$

Sequences of Random Variables; lim sup and lim inf

The lim sup and lim inf of a sequence of random variables $\{X_n\}$ mean the lim sup and lim inf of a sequence of functions, which we defined on page 725:

$$\limsup_n X_n \stackrel{\text{def}}{=} \inf_n \sup_{i \geq n} X_i \tag{1.151}$$

and

$$\liminf_n X_n \stackrel{\text{def}}{=} \sup_n \inf_{i \geq n} X_i. \tag{1.152}$$

1.3.1 The Borel-Cantelli Lemmas

The analysis of any finite sequence is straightforward, so the interesting behavior of a sequence is determined by what happens as n gets large.

Tail Events and the Kolmogorov Zero-One Law

As n gets large, our interest will be in the "tail" of the sequence. In the following, we consider sequences of σ-fields that are not necessarily increasing or increasing, as were the collections of events used in discussing inequalities (1.148) and (1.149).

Definition 1.34 (tail σ-field; tail event)
Let $\{\mathcal{F}_n\}$ be a sequence of σ-fields. The σ-field

$$\mathcal{T} = \cap_{n=1}^{\infty} \mathcal{F}_n$$

is called the *tail σ-field of the sequence*.
An event $A \in \mathcal{T}$ is called a *tail event of the sequence*. ∎

A tail event occurs infinitely often (exercise).

We are often interested in tail σ-fields of sequences of σ-fields generated by given sequences of events or sequences of random variables. Given the sequence of events $\{A_n\}$, we are interested in the σ-fields $\mathcal{F}_i = \sigma(A_i, A_{i+1}, \ldots)$. A tail event in the sequence $\{\mathcal{F}_i\}$ is also called a tail event of the sequence of events $\{A_n\}$.

Given the sequence of random variables $\{X_n\}$, we are also often interested in the σ-fields $\sigma(X_i, X_{i+1}, \ldots)$.

If the events, σ-fields, or random variables in the sequence are independent, the independence carries over to subsequences and sub-σ-fields in a useful way. We will focus on sequences of random variables that define tail events, but the generating sequence could also be of events or of σ-fields.

Lemma 1.27.1
Let $\{X_n\}$ be a sequence of independent random variables and let $\mathcal{T}$ be the tail σ-field, $\cap_{i=1}^{\infty} \sigma(X_i, X_{i+2}, \ldots)$. Then the events $A \in \mathcal{T}$ and $B \in \sigma(X_1, \ldots, X_{i-1})$ are independent for each i. Furthermore, A is independent of $\sigma(X_1, X_2, \ldots)$.

Proof.
Because the random variables $X_1, X_2, \ldots$ are independent, $\mathcal{T}$ and $\sigma(X_1, \ldots, X_{i-1})$ are independent and hence the events $A \in \mathcal{T}$ and $B \in \sigma(X_1, \ldots, X_{i-1})$ are independent. (This is the same reasoning as in the proof of Theorem 1.11, which applies to random.) Therefore, A is independent of $\sigma(X_1, \ldots, X_{i-1})$ and hence, also independent of $\mathcal{F}_0 = \cup_{i=1}^{\infty} \sigma(X_1, \ldots, X_i)$. By Theorem 1.1 then, A is independent of $\sigma(X_1, X_2, \ldots)$. ∎

Lemma 1.27.1 has an interesting implication. Tail events in sequences generated by independent events or random variables are independent of themselves.

Theorem 1.27 (Kolmogorov's Zero-One Law)
Let $\{X_n\}$ be a sequence of independent random variables and A be an event in the tail σ-field of the sequence of σ-fields generated by $\{X_n\}$. Then $P(A)$ is either zero or one.

Proof. An event A be an event in the tail σ-field is independent of itself; hence

$$P(A) = P(A \cup A) = P(A)P(A),$$

and so $P(A)$ must have a probability of 0 or 1. ∎

For a sequence of events in a given probability space, the Borel-Cantelli lemmas address the question implied by the Kolmogorov zero-one law. These lemmas tell us the probability of the lim sup. Under one condition, we get a probability of 0 without requiring independence. Under the other condition, with a requirement of independence, we get a probability of 1.

Theorem 1.28 (Borel-Cantelli Lemma I)
Let $\{A_n\}$ be a sequence of events and P be a probability measure. Then

$$\sum_{n=1}^{\infty} P(A_n) < \infty \quad \Longrightarrow \quad P(\limsup_n A_n) = 0. \tag{1.153}$$

Proof. First, notice that $P(\cup_{i=n}^{\infty} A_i)$ can be arbitrarily small if n is large enough. From $\limsup_n A_n \subseteq \cup_{i=n}^{\infty} A_i$, we have

$$\begin{aligned} P(\limsup_n A_n) &\leq P(\cup_{i=n}^{\infty} A_i) \\ &\leq \sum_{i=n}^{\infty} P(A_i) \\ &\to 0 \quad \text{as } n \to \infty \quad \text{because } \sum_{n=1}^{\infty} P(A_n) < \infty. \end{aligned}$$

∎

The requirement that $\sum_{n=1}^{\infty} P(A_n) < \infty$ in Theorem 1.28 means that A_n must be approaching sets with ever-smaller probability. If that is not

the case, say for example, if A_n is a constant set with positive probability, then of course $\sum_{n=1}^{\infty} P(A_n) = \infty$, but we cannot say much about about $P(\limsup_n A_n)$. However, if the A_n are disjoint, then $\sum_{n=1}^{\infty} P(A_n) = \infty$ may have a meaningful implication. The second Borel-Cantelli requires that the sets be independent.

Theorem 1.29 (Borel-Cantelli Lemma II)
Let $\{A_n\}$ be a sequence of independent events and P be a probability measure. Then

$$\sum_{n=1}^{\infty} P(A_n) = \infty \quad \Longrightarrow \quad P(\limsup_n A_n) = 1. \tag{1.154}$$

Proof. Applying de Morgan's law (equation (0.0.21)), we just need to show that $P(\liminf_n A_n^{\mathrm{c}}) = 0$. We use the fact that for $x \geq 0$

$$1 - x \leq \mathrm{e}^{-x}$$

to get, for any n and j,

$$\begin{aligned} P\left(\cap_{k=n}^{n+j} A_k^{\mathrm{c}}\right) &= \prod_{k=n}^{n+j} (1 - P(A_k)) \\ &\leq \exp\left(-\sum_{k=n}^{n+j} P(A_k)\right). \end{aligned}$$

Since $\sum_{n=1}^{\infty} P(A_n)$ diverges, the last expression goes to 0 as $j \to \infty$, and so

$$\begin{aligned} P\left(\cap_{k=n}^{\infty} A_k^{\mathrm{c}}\right) &= \lim_{j \to \infty} P\left(\cap_{k=n}^{n+j} A_k^{\mathrm{c}}\right) \\ &= 0. \end{aligned}$$

∎

1.3.2 Exchangeability and Independence of Sequences

The sequences that we have discussed may or may not be exchangeable or independent. The sequences in Theorem 1.29 are independent, for example, but most sequences that we consider are not necessarily independent. When we discuss limit theorems in Section 1.4, we will generally require independence. In Section 1.6, we will relax the requirement of independence, but will require some common properties of the elements of the sequence. Before proceeding, however, in this general section on sequences of random variables, we will briefly consider exchangeable sequences and state without proof *de Finetti's representation theorem*, which provides a certain connection between exchangeability and independence. This theorem tells us that infinite sequences of exchangeable binary random variables are mixtures of independent Bernoulli sequences.

Theorem 1.30 (de Finetti's representation theorem)
Let $\{X_i\}_{i=1}^{\infty}$ be an infinite sequence of binary random variables such that for any n, $\{X_i\}_{i=1}^{n}$ is exchangeable. Then there is a unique probability measure P on $[0,1]$ such that for each fixed sequence of zeros and ones $\{e_i\}_{i=1}^{n}$,

$$\Pr(X_1 = e_1, \ldots, X_n = e_n) = \int_0^1 \pi^k (1-\pi)^{n-k} \mathrm{d}\mu(\pi),$$

where $k = \sum_{i=1}^{n} e_i$.

The converse clearly holds; that is, if a P as specified in the theorem exists, then the sequence is exchangeable.

A proof of a more general version of de Finetti's representation theorem (for random variables that are not necessarily binary) is given by Hewitt and Stromberg (1965) and in Schervish (1995).

1.3.3 Types of Convergence

The first important point to understand about asymptotic theory is that there are different kinds of convergence of a sequence of random variables, $\{X_n\}$. Three of these kinds of convergence have analogues in convergence of general measurable functions (see Appendix 0.1) and a fourth type applies to convergence of the measures themselves. Different types of convergence apply to

- a function, that is, directly to the random variable (Definition 1.35). This is the convergence that is ordinarily called "strong convergence".
- expected values of powers of the random variable (Definition 1.36). This is also a type of strong convergence.
- probabilities of the random variable being within a range of another random variable (Definition 1.37). This is a weak convergence.
- the distribution of the random variable (Definition 1.39, stated in terms of weak convergence of probability measures, Definition 1.38). This is the convergence that is ordinarily called "weak convergence".

In statistics, we are interested in various types of convergence of procedures of statistical inference. Depending on the kind of inference, one type of convergence may be more relevant than another. We will discuss these in later chapters. At this point, however, it is appropriate to point out that an important property of *point estimators* is *consistency*, and the various types of consistency of point estimators, which we will discuss in Section 3.8.1, correspond directly to the types of convergence of sequences of random variables we discuss below.

Almost Sure Convergence

Definition 1.35 (almost sure (a.s.) convergence)
We say that $\{X_n\}$ *converges almost surely* to X if

$$\lim_{n\to\infty} X_n = X \text{ a.s.} \tag{1.155}$$

We write

$$X_n \xrightarrow{\text{a.s.}} X.$$

∎

Writing this definition in the form of Definition 0.1.38 on page 726, with X_n and X defined on the probability space $(\Omega, \mathcal{F}, P)$, we have

$$P(\{\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\}) = 1. \tag{1.156}$$

This expression provides a very useful heuristic for distinguishing a.s. convergence from other types of convergence.

Almost sure convergence is equivalent to

$$\lim_{n\to\infty} \Pr\left(\cup_{m=n}^{\infty} \|X_m - X\| > \epsilon\right) = 0, \tag{1.157}$$

for every $\epsilon > 0$ (exercise).

Almost sure convergence is also called "almost certain" convergence, and written as $X_n \xrightarrow{\text{a.c.}} X$.

The condition (1.155) can also be written as

$$\Pr\left(\lim_{n\to\infty} \|X_n - X\| < \epsilon\right) = 1, \tag{1.158}$$

for every $\epsilon > 0$. For this reason, almost sure convergence is also called *convergence with probability 1*, and may be indicated by writing $X_n \xrightarrow{\text{wp1}} X$. Hence, we may encounter three equivalent expressions:

$$\xrightarrow{\text{a.s.}} \equiv \xrightarrow{\text{a.c.}} \equiv \xrightarrow{\text{wp1}}.$$

Almost sure convergence of a sequence of random variables $\{X_n\}$ to a constant c implies $\limsup_n X_n = \liminf_n X_n = c$, and implies $\{X_n = c \text{ i.o.}\}$; by itself, however, $\{X_n = c \text{ i.o.}\}$ does not imply any kind of convergence of $\{X_n\}$.

Convergence in r^{th} Moment

Definition 1.36 (convergence in r^{th} moment (convergence in L_r))
For fixed $r > 0$, we say that $\{X_n\}$ *converges in* r^{th} *moment* to X if

$$\lim_{n\to\infty} \mathrm{E}(\|X_n - X\|_r^r) = 0. \tag{1.159}$$

We write

$$X_n \xrightarrow{\mathrm{L}_r} X.$$

∎

(Compare Definition 0.1.50 on page 748.)

Convergence in r^{th} moment requires that $\mathrm{E}(\|X_n\|_r^r) < \infty$ for each n. Convergence in r^{th} moment implies convergence in s^{th} moment for $s \leq r$ (and, of course, it implies that $\mathrm{E}(\|X_n\|_s^s) < \infty$ for each n). (See Theorem 1.16, which was stated only for scalar random variables.)

For $r = 1$, convergence in r^{th} moment is called *convergence in absolute mean*. For $r = 2$, it is called *convergence in mean square* or *convergence in second moment*, and of course, it implies convergence in mean. (Recall our notational convention: $\|X_n - X\| = \|X_n - X\|_2$.)

The Cauchy criterion (see Exercise 0.0.6d on page 689) is often useful for proving convergence in mean or convergence in mean square, without specifying the limit of the sequence. The sequence $\{X_n\}$ converges in mean square (to some real number) iff

$$\lim_{n,m\to\infty} \mathrm{E}(\|X_n - X_m\|) = 0. \tag{1.160}$$

Convergence in Probability

Definition 1.37 (convergence in probability)
We say that $\{X_n\}$ *converges in probability* to X if for every $\epsilon > 0$,

$$\lim_{n\to\infty} \Pr(\|X_n - X\| > \epsilon) = 0. \tag{1.161}$$

We write

$$X_n \xrightarrow{\mathrm{p}} X.$$

∎

(Compare Definition 0.1.51 on page 748 for general measures.)

Notice the difference in convergence in probability and convergence in r^{th} moment. Convergence in probability together with uniform integrability implies convergence in mean, but not in higher r^{th} moments. It is easy to construct examples of sequences that converge in probability but that do not converge in second moment (exercise).

Notice the difference in convergence in probability and almost sure convergence; in the former case the limit of probabilities is taken, in the latter the case a probability of a limit is evaluated; compare equations (1.157) and (1.161). It is easy to construct examples of sequences that converge in probability but that do not converge almost surely (exercise).

Although convergence in probability does not imply almost sure converge, it does imply the existence of a subsequence that does converge almost surely, as stated in the following theorem.

Theorem 1.31
Suppose $\{X_n\}$ converges in probability to X. Then there exists a subsequence $\{X_{n_i}\}$ that converges almost surely to X.

Stated another way, this theorem says that if $\{X_n\}$ converges in probability to X, then there is an increasing sequence $\{n_i\}$ of positive integers such that

$$\lim_{i\to\infty} X_{n_i} \overset{\text{a.s.}}{=} X.$$

Proof. The proof is an exercise. You could first show that there is an increasing sequence $\{n_i\}$ such that

$$\sum_{i=1}^{\infty} \Pr(|X_{n_i} - X| > 1/i) < \infty,$$

and from this conclude that $X_{n_i} \overset{\text{a.s.}}{\to} X$. ∎

Weak Convergence

There is another type of convergence that is very important in statistical applications; in fact, it is the basis for asymptotic statistical inference. This convergence is defined in terms of pointwise convergence of the sequence of CDFs; hence it is a *weak* convergence. We will give the definition in terms of the sequence of CDFs or, equivalently, of probability measures, and then state the definition in terms of a sequence of random variables.

Definition 1.38 (weak convergence of probability measures)
Let $\{P_n\}$ be a sequence of probability measures and $\{F_n\}$ be the sequence of corresponding CDFs, and let F be a CDF with corresponding probability measure P. If at each point of continuity t of F,

$$\lim_{n\to\infty} F_n(t) = F(t), \tag{1.162}$$

we say that the sequence of CDFs $\{F_n\}$ *converges weakly* to F, and, equivalently, we say that the sequence of probability measures $\{P_n\}$ *converges weakly* to P. We write

$$F_n \overset{\text{w}}{\to} F$$

or

$$P_n \overset{\text{w}}{\to} P$$

∎

Definition 1.39 (convergence in distribution (in law))
If $\{X_n\}$ have CDFs $\{F_n\}$ and X has CDF F, we say that $\{X_n\}$ *converges in distribution* or *in law* to X iff $F_n \overset{\text{w}}{\to} F$. We write

$$X_n \overset{\text{d}}{\to} X.$$

∎

Because convergence in distribution is not precisely a convergence of the random variables themselves, it may be preferable to use a notation of the form

$$\mathcal{L}(X_n) \to \mathcal{L}(X),$$

where the symbol $\mathcal{L}(\cdot)$ refers to the distribution or the "law" of the random variable.

When a random variable converges in distribution to a distribution for which we have adopted a symbol such as $\mathrm{N}(\mu, \sigma^2)$, for example, we may use notation of the form

$$X_n \xrightarrow{\sim} \mathrm{N}(\mu, \sigma^2).$$

Because this notation only applies in this kind of situation, we often write it more simply as just

$$X_n \to \mathrm{N}(\mu, \sigma^2),$$

or in the "law" notation, $\mathcal{L}(X_n) \to \mathrm{N}(\mu, \sigma^2)$

For certain distributions we have special symbols to represent a random variable. In such cases, we may use notation of the form

$$X_n \xrightarrow{\mathrm{d}} \chi^2_\nu,$$

which in this case indicates that the sequence $\{X_n\}$ converges in distribution to a random variable with a chi-squared distribution with ν degrees of freedom. The "law" notation for this would be $\mathcal{L}(X_n) \to \mathcal{L}(\chi^2_\nu)$.

Determining Classes

In the case of multiple probability measures over a measurable space, we may be interested in how these measures behave over different sub-σ-fields, in particular, whether there is a determining class smaller than the σ-field of the given measurable space. For convergent sequences of probability measures, the determining classes of interest are those that preserve convergence of the measures for all sets in the σ-field of the given measurable space.

Definition 1.40 (convergence-determining class)
Let $\{P_n\}$ be a sequence of probability measures defined on the measurable space $(\Omega, \mathcal{F})$ that converges (weakly) to P, also a probability measure defined on $(\Omega, \mathcal{F})$. A collection of subsets $\mathcal{C} \subseteq \mathcal{F}$ is called a *convergence-determining class* of the sequence, iff

$$P_n(A) \to P(A)\ \forall A \in \mathcal{C} \ni P(\partial A) = 0 \Longrightarrow P_n(B) \to P(B)\ \forall B \in \mathcal{F}.$$

∎

It is easy to see that a convergence-determining class is a determining class (exercise), but the converse is not true, as the following example from Romano and Siegel (1986) shows.

Example 1.20 a determining class that is not a convergence-determining class
For this example, we use the familiar measurable space $(\mathbb{R}, \mathcal{B})$, and construct a determining class $\mathcal{C}$ whose sets exclude exactly one point, and then define a probability measure P that puts mass one at that point. All that is then required is to define a sequence $\{P_n\}$ that converges to P. The example given by Romano and Siegel (1986) is the collection $\mathcal{C}$ of all finite open intervals that do not include the single mass point of P. (It is an exercise to show that this is a determining class.) For definiteness, let that special point be 0, and let P_n be the probability measure that puts mass one at n. Then, for any $A \in \mathcal{C}$, $P_n(A) \to 0 = P(A)$, but for any interval (a, b) where $a < 0$ and $0 < b < 1$, $P_n((a, b)) = 0$ but $P((a, b)) = 1$. ∎

Both convergence in probability and convergence in distribution are weak types of convergence. Convergence in probability, however, means that the probability is high that the two random variables are close to each other, while convergence in distribution means that two random variables have the same distribution. That does not mean that they are very close to each other.

The term "weak convergence" is often used specifically for convergence in distribution because this type of convergence has so many applications in asymptotic statistical inference. In many interesting cases the limiting distribution of a sequence $\{X_n\}$ may be degenerate, but for some sequence of constants a_n, the limiting distribution of $\{a_n X_n\}$ may not be degenerate and in fact may be very useful in statistical applications. The limiting distribution of $\{a_n X_n\}$ for a reasonable choice of a sequence of normalizing constants $\{a_n\}$ is called the asymptotic distribution of $\{X_n\}$. After some consideration of the relationships among the various types of convergence, in Section 1.3.7, we will consider the "reasonable" choice of normalizing constants and other properties of weak convergence in distribution in more detail. The relevance of the limiting distribution of $\{a_n X_n\}$ will become more apparent in the statistical applications in Section 3.8.2 and later sections.

Relationships among Types of Convergence

Almost sure convergence and convergence in r^{th} moment are both strong types of convergence, but they are not closely related to each other. We have the logical relations shown in Figure 1.3.

The directions of the arrows in Figure 1.3 correspond to theorems with straightforward proofs. Where there are no arrows, as between L_r and a.s., we can find examples that satisfy one condition but not the other (see Examples 1.21 and 1.22 below). For relations in the opposite direction of the arrows, we can construct counterexamples, as for example, the reader is asked to do in Exercises 1.54a and 1.54b.

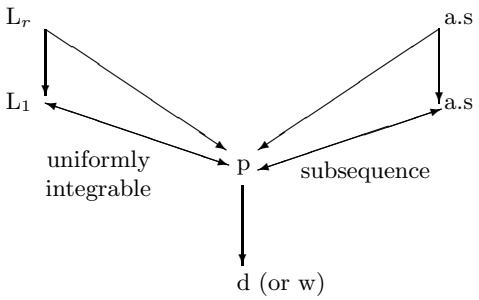


Figure 1.3. Relationships of Convergence Types

Useful Sequences for Studying Types of Convergence

Just as for working with limits of unions and intersections of sets where we find it useful to identify sequences of sets that behave in some simple way (such as the intervals $[a + 1/n, b - 1/n]$ on page 646), it is also useful to identify sequences of random variables that behave in interesting but simple ways.

One useful sequence begins with $\{U_n\}$, where $U_n \sim \mathrm{U}(0, 1/n)$. We define

$$X_n = nU_n. \tag{1.163}$$

This sequence can be used to show that an a.s. convergent sequence may not converge in L_1.

Example 1.21 converges a.s. but not in mean
Let $\{X_n\}$ be the sequence defined in equation (1.163). Since $\Pr(\lim_{n\to\infty} X_n = 0) = 1$, $X_n \xrightarrow{\text{a.s.}} 0$. The mean and in fact the r^{th} moment (for $r > 0$) is 0. However,

$$\mathrm{E}(|X_n - 0|^r) = \int_0^{1/n} n^r \mathrm{d}u = n^{r-1}.$$

For $r = 1$, this does not converge to the mean of 0, and for $r > 1$, it diverges; hence $\{X_n\}$ does not converge to 0 in r^{th} moment for any $r \geq 1$. (It does converge to the correct r^{th} moment for $0 < r < 1$, however.) ∎

This example is also an example of a sequence that converges in probability (since a.s. convergence implies that), but does not converge in r^{th} moment.

Other kinds of interesting sequences can be constructed as indicators of events; that is, 0-1 random variables. One such simple sequence is the Bernoulli random variables $\{X_n\}$ with probability that $X_n = 1$ being $1/n$. This sequence can be used to show that a sequence that converges to X in probability does not necessarily converge to X a.s.

Other ways of defining 0-1 random variables involve breaking a $\mathrm{U}(0,1)$ distribution into uniform distributions on partitions of $]0,1[$. For example, for a positive integer k, we may form 2^k subintervals of $]0,1[$ for $j = 1, \ldots, 2^k$ as

$$\left]\frac{j-1}{2^k}, \frac{j}{2^k}\right[.$$

As k gets larger, the Lebesgue measure of these subintervals approaches 0 rapidly. Romano and Siegel (1986) build an indicator sequence using random variables on these subintervals for various counterexamples. This sequence can be used to show that an L_2 convergent sequence may not converge a.s., as in the following example.

Example 1.22 converges in second moment but not a.s.
Let $U \sim \mathrm{U}(0,1)$ and define

$$X_n = \begin{cases} 1 & \text{if } \dfrac{j_n - 1}{2^{k_n}} < U < \dfrac{j_n}{2^{k_n}} \\ 0 & \text{otherwise,} \end{cases}$$

where $j_n = 1, \ldots, 2^{k_n}$ and $k_n \to \infty$ as $n \to \infty$. We see that

$$\mathrm{E}((X_n - 0)^2) = 1/(2^{k_n}),$$

hence $\{X_n\}$ converges in quadratic mean (or in mean square) to 0. We see, however, that $\lim_{n\to\infty} X_n$ does not exist (since for any value of U, X_n takes on each of the values 0 and 1 infinitely often). Therefore, $\{X_n\}$ cannot converge a.s. (to anything!). ∎

This is another example of a sequence that converges in probability (since convergence in r^{th} moment implies that), but does not converge a.s.

Convergence of PDFs

The weak convergence of a sequence of CDFs $\{F_n\}$ is the basis for most asymptotic statistical inference. The convergence of a sequence of PDFs $\{f_n\}$ is a stronger form of convergence because it implies uniform convergence of probability on any given Borel set.

Theorem 1.32 (Scheffé)
Let $\{f_n\}$ be a sequence of PDFs that converge pointwise to a PDF f; that is, at each x

$$\lim_{n\to\infty} f_n(x) = f(x).$$

Then

$$\lim_{n\to\infty} \int_B |f_n(x) - f(x)| \mathrm{d}x = 0 \tag{1.164}$$

uniformly for any Borel set B.

For a proof see Scheffé (1947).

Hettmansperger and Klimko (1974) showed that if a weakly convergent sequence of CDFs $\{F_n\}$ has an associated sequence of PDFs $\{f_n\}$, and if these PDFs are unimodal at a given point, then on any closed interval that does not contain the modal point the sequence of PDFs converge uniformly to a PDF.

Big O and Little o Almost Surely

We are often interested in nature of the convergence or the rate of convergence of a sequence of random variables to another sequence of random variables. As in general spaces of real numbers that we consider in Section 0.0.5 on page 652, we distinguish two types of limiting behavior by big O and little o. These are involve the asymptotic ratio of the elements of one sequence to the elements of a given sequence $\{a_n\}$. We defined two order classes, $\mathrm{O}(a_n)$ and $\mathrm{o}(a_n)$. In this section we begin with a given sequence of random variables $\{Y_n\}$ and define four different order classes, $\mathrm{O}(Y_n)$ a.s., $\mathrm{o}(Y_n)$ a.s., $\mathrm{O_P}(Y_n)$, and $\mathrm{o_P}(Y_n)$, based on whether or not the ratio is approaching 0 (that is, big O or little o) and on whether the converge is almost sure or in probability.

For sequences of random variables $\{X_n\}$ and $\{Y_n\}$ defined on a common probability space, we identify different types of convergence, either almost sure or in probability.

- Big O almost surely, written $\mathrm{O}(Y_n)$ a.s.

$$X_n \in \mathrm{O}(Y_n) \text{ a.s. iff } \Pr\left(\|X_n\| \in \mathrm{O}(\|Y_n\|)\right) = 1$$

- Little o almost surely, written $\mathrm{o}(Y_n)$ a.s.

$$X_n \in \mathrm{o}(Y_n) \text{ a.s. iff } \|X_n\|/\|Y_n\| \xrightarrow{\text{a.s.}} 0.$$

Compare $X_n/Y_n \xrightarrow{\text{a.s.}} 0$ for $X_n \in \mathbb{R}^m$ and $Y_n \in \mathbb{R}$.

Big O and Little o Weakly

We also have relationships in which one sequence converges to another in probability.

- Big O in probability, written $\mathrm{O_P}(Y_n)$.

$$X_n \in \mathrm{O_P}(Y_n) \text{ iff } \forall \epsilon > 0\ \exists \text{ constant } C_\epsilon > 0 \ni \sup_n \Pr(\|X_n\| \geq C_\epsilon \|Y_n\|) < \epsilon.$$

If $X_n \in \mathrm{O_P}(1)$, X_n is said to be *bounded in probability*.

If $X_n \xrightarrow{\text{d}} X$ for any random variable X, then $X_n \in \mathrm{O_P}(1)$. (Exercise.)

- Little o in probability, written $o_P(Y_n)$.

$$X_n \in o_P(Y_n) \text{ iff } \|X_n\|/\|Y_n\| \xrightarrow{p} 0.$$

If $X_n \in o_P(1)$, then X_n converges in probability to 0, and conversely.
If $X_n \in o_P(1)$, then also $X_n \in O_P(1)$. (Exercise.)

Instead of a defining sequence $\{Y_n\}$ of random variables, the sequence of interest may be a sequence of constants $\{a_n\}$.

Some useful properties are the following, in which $\{X_n\}$, $\{Y_n\}$, and $\{Z_n\}$ are random variables defined on a common probability space, and $\{a_n\}$ and $\{b_n\}$ are sequences of constants.

$$X_n \in o_p(a_n) \Longrightarrow X_n \in O_p(a_n) \tag{1.165}$$

$$X_n \in o_p(1) \Longleftrightarrow X_n \to 0. \tag{1.166}$$

$$X_n \in O_p(1/a_n), \quad \lim b_n/a_n < \infty \Longrightarrow X_n \in O_p(m_n). \tag{1.167}$$

$$X_n \in O_p(a_n) \Longrightarrow X_n Y_n \in O_p(a_n Y_n). \tag{1.168}$$

$$X_n \in O_p(a_n),\ Y_n \in O_p(b_n) \Longrightarrow X_n Y_n \in O_p(a_n b_n). \tag{1.169}$$

$$X_n \in O_p(a_n),\ Y_n \in O_p(b_n) \Longrightarrow X_n + Y_n \in O_p(\|a_n\| + \|b_n\|). \tag{1.170}$$

$$X_n \in O_p(Z_n),\ Y_n \in O_p(Z_n) \Longrightarrow X_n + Y_n \in O_p(Z_n). \tag{1.171}$$

$$X_n \in o_p(a_n),\ Y_n \in o_p(b_n) \Longrightarrow X_n Y_n \in o_p(a_n b_n). \tag{1.172}$$

$$X_n \in o_p(a_n),\ Y_n \in o_p(b_n) \Longrightarrow X_n + Y_n \in o_p(\|a_n\| + \|b_n\|). \tag{1.173}$$

$$X_n \in o_p(a_n),\ Y_n \in O_p(b_n) \Longrightarrow X_n Y_n \in o_p(a_n b_n). \tag{1.174}$$

You are asked to prove these statements in Exercise 1.61. There are, of course, other variations on these relationships. The order of convergence of sequence of absolute expectations can be related to order of convergence in probability:

$$a_n \in \mathbb{R}_+,\ \mathrm{E}(|X_n|) \in O(a_n) \Longrightarrow X_n \in O_p(a_n). \tag{1.175}$$

Almost sure convergence implies that the sup is bounded in probability. For any random variable X (recall that a random variable is finite a.s.),

$$X_n \xrightarrow{a.s.} X \Longrightarrow \sup |X_n| \in O_p(1). \tag{1.176}$$

You are asked to prove these statements in Exercise 1.62.

The defining sequence of interest is often an expression in n; for examples, $O_P(n^{-1})$, $O_P(n^{-1/2})$, and so on. For such orders of convergence, we have relationships similar to those given in statement (0.0.59) for nonstochastic convergence.

$$O_P(n^{-1}) \subseteq O_P(n^{-1/2}) \quad \text{etc.} \tag{1.177}$$

Example 1.23 order in probability of the sample mean
Suppose $\overline{X}_n$ is the sample mean (equation (1.32)) from a random sample $X_1, \ldots, X_n$ from a distribution with finite mean, μ, and finite variance, σ^2. We first note that $\mathrm{E}(\overline{X}_n) = \mu$ and $\mathrm{V}(\overline{X}_n) = \sigma^2/n$. By Chebyshev's inequality (page 848), we have for $\epsilon > 0$,

$$\Pr(|\overline{X}_n - \mu| \geq \epsilon) \leq \frac{\sigma^2/n}{\epsilon^2},$$

which goes to 0 as $n \to \infty$. Hence, $\overline{X}_n \xrightarrow{\mathrm{p}} \mu$ or $\overline{X}_n - \mu \in \mathrm{o_P}(1)$.

Now, rewriting the inequality above, we have, for $\delta(\epsilon) > 0$,

$$\Pr(\sqrt{n}|\overline{X}_n - \mu| \geq \delta(\epsilon)) \leq \frac{\sigma^2/n}{\delta(\epsilon)^2/n}.$$

Now letting $\delta(\epsilon) = \sigma/\sqrt{\epsilon}$, we have $\overline{X}_n - \mu \in \mathrm{O_P}(n^{-1/2})$. ∎

1.3.4 Weak Convergence in Distribution

Convergence in distribution, sometimes just called "weak convergence", plays a fundamental role in statistical inference. It is the type of convergence in the central limits (see Section 1.4.2) and it is the basis for the definition of asymptotic expectation (see Section 1.3.8), which, in turn is the basis for most of the concepts of asymptotic inference. (Asymptotic inference is not based on the limits of the properties of the statistics in a sequence, and in Section 3.8.3, beginning on page 311, we will consider some differences between "aysmptotic" properties and "limiting" properties.)

In studying the properties of a sequence of random variables $\{X_n\}$, the holy grail often is to establish that $a_n X_n \to \mathrm{N}(\mu, \sigma^2)$ for some sequence $\{a_n\}$, and to determine reasonable estimates of μ and σ^2. In this section we will show how this is sometimes possible, and we will consider it further in Section 1.3.7, and later in Section 3.8, where we will emphasize the statistical applications. Weak convergence to normality under less rigorous assumptions will be discussed in Section 1.4.

Convergence in distribution of a sequence of random variables is defined in terms of convergence of a sequence of CDFs. For a sequence that converges to a continuous CDF F, the Chebyshev norm of the difference between a function in the sequence and F goes to zero, as stated in the following theorem.

Theorem 1.33 (Polya's theorem)
If $F_n \xrightarrow{\mathrm{w}} F$ and F is continuous in $\mathbb{R}^k$, then

$$\lim_{n\to\infty} \sup_{t\in\mathbb{R}^k} |F_n(t) - F(t)| = 0.$$

Proof. The proof proceeds directly by use of the δ-ϵ definition of continuity. ∎

Theorem 1.34
Let $\{F_n\}$ be a sequence of CDFs on $\mathbb{R}$. Let

$$G_n(x) = F_n(b_{gn}x + a_{gn})$$

and

$$H_n(x) = F_n(b_{hn}x + a_{hn}),$$

where $\{b_{dn}\}$ and $\{b_{hn}\}$ are sequences of positive real numbers and $\{a_{gn}\}$ and $\{a_{hn}\}$ are sequences of real numbers. Suppose

$$G_n \xrightarrow{\text{w}} G$$

and

$$H_n \xrightarrow{\text{w}} H,$$

where G and H are nondegenerate CDFs. Then

$$b_{gn}/b_{hn} \to b > 0,$$

$$(a_{gn} - a_{hn})/b_{gn} \to a \in \mathbb{R},$$

and

$$H(bx + a) = G(x) \quad \forall x \in \mathbb{R}.$$

Proof. ** fix ∎

The distributions in Theorem 1.34 are in a location-scale family (see Section 2.6, beginning on page 178).

There are several necessary and sufficient conditions for convergence in distribution. A set of such conditions is given in the following "portmanteau" theorem.

Theorem 1.35 (characterizations of convergence in distribution; "portmanteau" theorem)
Given the sequence of random variables X_n and the random variable X, all defined on a common probability space, then each of the following is a necessary and sufficient condition that $X_n \xrightarrow{\text{d}} X$.

(i) $\mathrm{E}(g(X_n)) \to \mathrm{E}(g(X))$ *for all real bounded continuous functions g.*
(ii) $\mathrm{E}(g(X_n)) \to \mathrm{E}(g(X))$ *for all real functions g such that $g(x) \to 0$ as $|x| \to \infty$.*
(iii) $\Pr(X_n \in B) \to \Pr(X \in B)$ *for all Borel sets B such that $\Pr(X \in \partial B) = 0$.*
(iv) $\liminf \Pr(X_n \in S) \geq \Pr(X \in S)$ *for all open sets S.*
(v) $\limsup \Pr(X_n \in T) \leq \Pr(X \in T)$ *for all closed sets T.*

Proof. The proofs of the various parts of this theorem are in Billingsley (1995), among other resources. ∎

Although convergence in distribution does not imply a.s. convergence, convergence in distribution does allow us to construct an a.s. convergent sequence. This is stated in Skorokhod's representation theorem.

Theorem 1.36 (Skorokhod's representation theorem)
If for the random variables (vectors!) $X_1, X_2, \ldots$, we have $X_n \xrightarrow{\mathrm{d}} X$, then there exist random variables $Y_1 \overset{\mathrm{d}}{=} X_1, Y_2 \overset{\mathrm{d}}{=} X_2, \ldots$, and $Y \overset{\mathrm{d}}{=} X$, such that $Y_n \xrightarrow{\text{a.s.}} Y$.

Proof. Exercise. ∎

Theorem 1.37 (continuity theorem)
Let $X_1, X_2, \cdots$ be a sequence of random variables (not necessarily independent) with characteristic functions $\varphi_{X_1}, \varphi_{X_2}, \cdots$ and let X be a random variable with characteristic function φ_X. Then

$$X_n \xrightarrow{\mathrm{d}} X \quad \Longleftrightarrow \quad \varphi_{X_n}(t) \to \varphi_X(t) \; \forall t.$$

Proof. Exercise. ∎

The $\Longleftarrow$ part of the continuity theorem is called the Lévy-Cramér theorem and the $\Longrightarrow$ part is sometimes called the *first limit theorem*.

The continuity theorem also applies to MGFs if they exist for all X_n.

A nice use of the continuity theorem is in the proof of a simple form of the central limit theorem, or CLT. Here I will give the proof for scalar random variables. There are other forms of the CLT, and other important limit theorems, which will be the topic of Section 1.4. Another reason for introducing this simple CLT now is so we can use it for some other results that we discuss before Section 1.4.

Theorem 1.38 (central limit theorem)
If $X_1, \ldots, X_n$ are iid with mean μ and variance $0 < \sigma^2 < \infty$, then $Y_n = (\sum X_i - n\mu)/\sqrt{n}\sigma$ has limiting distribution $\mathrm{N}(0,1)$.

Proof. It will be convenient to define a function related to the CF: let $h(t) = \mathrm{e}^{\mu t}\varphi_X(t)$; hence $h(0) = 1$, $h'(0) = 0$, and $h''(0) = \sigma^2$. Now expand h in a Taylor series about 0:

$$h(t) = h(0) + h'(0)\mathrm{i}t - \frac{1}{2}h''(\xi)t^2,$$

for some ξ between 0 and t. Substituting for $h(0)$ and $h'(0)$, and adding and subtracting $\sigma^2 t/2$ to this, we have

$$h(t) = 1 - \frac{\sigma^2 t^2}{2} - \frac{(h''(\xi) - \sigma^2)t^2}{2}.$$

This is the form we will find useful. Now, consider the CF of Y_n:

$$\begin{aligned}
\varphi_{Y_n}(t) &= \mathrm{E}\left(\exp\left(\mathrm{i}t\left(\frac{\sum X_i - n\mu)}{\sqrt{n}\sigma}\right)\right)\right) \\
&= \left(\mathrm{E}\left(\exp\left(\mathrm{i}t\left(\frac{X - \mu)}{\sqrt{n}\sigma}\right)\right)\right)\right)^n \\
&= \left(h\left(\frac{\mathrm{i}t}{\sqrt{n}\sigma}\right)\right)^n.
\end{aligned}$$

From the expansion of h, we have

$$h\left(\frac{\mathrm{i}t}{\sqrt{n}\sigma}\right) = 1 - \frac{t^2}{2n} - \frac{(h''(\xi) - \sigma^2)t^2}{2n\sigma^2}.$$

So,

$$\varphi_{Y_n}(t) = \left(1 - \frac{t^2}{2n} - \frac{(h''(\xi) - \sigma^2)t^2}{2n\sigma^2}\right)^n.$$

Now we need a well-known (but maybe forgotten) result (see page 652): If $\lim_{n\to\infty} f(n) = 0$, then

$$\lim_{n\to\infty}\left(1 + \frac{a}{n} + \frac{f(n)}{n}\right)^b n = \mathrm{e}^{ab}.$$

Therefore, because $\lim_{n\to\infty} h''(\xi) = h''(0) = \sigma^2$, $\lim_{n\to\infty}\varphi_{Y_n}(t) = \mathrm{e}^{-t^2/2}$, which is the CF of the $\mathrm{N}(0,1)$ distribution. (Actually, the conclusion relies on the Lévy-Cramér theorem, the $\Longleftarrow$ part of the continuity theorem, Theorem 1.37 on page 87; that is, while we know that the CF determines the distribution, we must also know that the convergent of a sequence of CFs determines a convergent distribution.) ∎

An important CLT has a weaker hypothesis than the simple one above; instead of iid random variables, we only require that they be independent (and have finite first and second moments, of course). In Section 1.6, we relax the hypothesis in the other direction; that is, we allow dependence in the random variables. (In that case, we must impose some conditions of similarity of the distributions of the random variables.)

Tightness of Sequences

In a convergent sequence of probability measures on a metric space, we may be interested in how concentrated the measures in the sequence are. (If the space does not have a metric, this question would not make sense.) We refer to this as "tightness" of the sequence, and we will define it only on the metric space $\mathbb{R}^d$.

Definition 1.41 (tightness of a sequence of probability measures)
Let $\{P_n\}$ be a sequence of probability measures on $(\mathbb{R}^d, \mathcal{B}^d)$. The sequence is said to be *tight* iff for every $\epsilon > 0$, there is a compact (bounded and closed) set $C \in \mathcal{B}^d$ such that

$$\inf_n P_n(C) > 1 - \epsilon.$$

∎

Notice that this definition does not require that $\{P_n\}$ be convergent, but of course, we are interested primarily in sequences that converge. The following theorem, whose proof can be found in Billingsley (1995) on page 336, among other places, connects tightness to convergence.

Theorem 1.39
Let $\{P_n\}$ be a sequence of probability measures on $(\mathbb{R}^d, \mathcal{B}^d)$.
(i) The sequence $\{P_n\}$ is tight iff for every subsequence $\{P_{n_i}\}$ there exists a further subsequence $\{P_{n_j}\} \subseteq \{P_{n_i}\}$ and a probability measure P on $(\mathbb{R}^d, \mathcal{B}^d)$ such that

$$P_{n_j} \xrightarrow{\text{w}} P, \text{ as } j \to \infty.$$

(ii) If $\{P_n\}$ is tight and each weakly convergent subsequence converges to the same measure P, then $P_n \xrightarrow{\text{w}} P$.

Tightness of a sequence of random variables is defined in terms of tightness of their associated probability measures.

Definition 1.42 (tightness of a sequence of random variables)
Let $\{X_n\}$ be a sequence of random variables, with associated probability measures $\{P_n\}$. The sequence $\{X_n\}$ is said to be *tight* iff

$$\forall \epsilon > 0 \; \exists M < \infty \ni \sup_n P_n(|X_n| > M) < \epsilon.$$

∎

1.3.5 Expectations of Sequences; Sequences of Expectations

The monotonicity of the expectation operator (1.42) of course carries over to sequences.

The three theorems that relate to the interchange of a Lebesgue integration operation and a limit operation stated on page 733 (monotone convergence, Fatou's lemma, and Lebesgue's dominated convergence) apply immediately to expectations:

- monotone convergence
  For $0 \leq X_1 \leq X_2 \cdots$ a.s.

$$X_n \xrightarrow{\text{a.s.}} X \quad \Rightarrow \quad \mathrm{E}(X_n) \to \mathrm{E}(X) \tag{1.178}$$

- Fatou's lemma

$$0 \leq X_n \text{ a.s. } \forall\, n \quad \Rightarrow \quad \mathrm{E}(\liminf_n X_n) \leq \liminf_n \mathrm{E}(X_n) \tag{1.179}$$

- dominated convergence
  Given a fixed Y with $\mathrm{E}(Y) < \infty$,

$$|X_n| \leq Y \; \forall\, n \text{ and } X_n \xrightarrow{\text{a.s.}} X \quad \Rightarrow \quad \mathrm{E}(X_n) \to \mathrm{E}(X). \tag{1.180}$$

These results require a.s. properties. Skorokhod's Theorem 1.36, however, often allows us to extend results based on a.s. convergence to sequences that converge in distribution. Skorokhod's theorem is the main tool used in the proofs of the following theorems, which we state without proof.

Theorem 1.40
If for the random variables $X_1, X_2, \ldots$, we have $X_n \xrightarrow{\mathrm{d}} X$ and for each $k > 0$ $\mathrm{E}(|X_n|^k) < \infty$ and $\mathrm{E}(|X|^k) < \infty$, then

$$\mathrm{E}(|X_n|^k) \to \mathrm{E}(|X|^k). \tag{1.181}$$

With additional conditions, we have a useful converse. It requires a limiting distribution that is not moment-indeterminant (see page 33). In that case, the converse says that the moments determine the limiting distribution.

Theorem 1.41 *Let X be a random variable that does not have a moment-indeterminant distribution, and let $X_1, X_2, \ldots$ be random variables. If for each $k > 0$ $\mathrm{E}(|X_n|^k) < \infty$ and $\mathrm{E}(|X|^k) < \infty$, and if $\mathrm{E}(|X_n|^k) \to \mathrm{E}(|X|^k)$, then $X_n \xrightarrow{\mathrm{d}} X$.*

Another useful convergence result for expectations is the Helly-Bray theorem (or just the Helly theorem):

Theorem 1.42 (Helly-Bray theorem)
If g is a bounded and continuous Borel function over the support of $\{X_n\}$, then

$$X_n \xrightarrow{\mathrm{d}} X \;\Leftrightarrow\; \mathrm{E}(g(X_n)) \to \mathrm{E}(g(X)). \tag{1.182}$$

With additional conditions there is also a converse of Theorem 1.42.

The properties we have considered so far are all "nice", "positive" results. We now consider an unhappy fact: in general,

$$\lim_{n\to\infty} \mathrm{E}(X_n) \neq \mathrm{E}(\lim_{n\to\infty} X_n), \tag{1.183}$$

as we see in the following example.

Example 1.24 gambler's ruin
Let $Y_1, Y_2, \ldots$ be a sequence of iid random variables with

$$\Pr(Y_i = 0) = \Pr(Y_i = 2) = \frac{1}{2} \quad \forall i = 1, 2, \ldots.$$

Now, let

$$X_n = \prod_{i=1}^{n} Y_i. \tag{1.184}$$

It is intuitive that some Y_k will eventually be 0, and in that case $X_n = 0$ for any $n \geq k$.

*** finish: show that $\mathrm{E}(X_n) = 1$ and $\lim_{n\to\infty} X_n = 0$ a.s.; hence, $\lim_{n\to\infty} \mathrm{E}(X_n) = 1$ and $\mathrm{E}(\lim_{n\to\infty} X_n) = 0$. ∎

1.3.6 Convergence of Functions

In working with sequences of random variables we often encounter a situation in which members of the sequence may be represented as sums or products of elements one of which converges to a constant. Slutsky's theorem provides very useful results concerning the convergence of such sequences.

Theorem 1.43 (Slutsky's theorem)
Let X, $\{X_n\}$, $\{B_n\}$, and $\{C_n\}$ be random variables on a common probability space, and let $b, c \in \mathbb{R}^k$. Suppose

$$X_n \xrightarrow{\mathrm{d}} X$$

and

$$B_n \xrightarrow{\mathrm{p}} b \quad \text{and} \quad C_n \xrightarrow{\mathrm{p}} c.$$

Then

$$B_n^{\mathrm{T}} X_n + C_n \xrightarrow{\mathrm{d}} b^{\mathrm{T}} X + c \tag{1.185}$$

Proof. Exercise. ∎

Slutsky's theorem is one of the most useful results for showing convergence in distribution, and you should quickly recognize some special cases of Slutsky's theorem. If $X_n \xrightarrow{\mathrm{d}} X$, and $Y_n \xrightarrow{\mathrm{p}} c$, then

$$X_n + Y_n \xrightarrow{\mathrm{d}} X + c \tag{1.186}$$

$$Y_n^{\mathrm{T}} X_n \xrightarrow{\mathrm{d}} c^{\mathrm{T}} X \tag{1.187}$$

and, if $Y_n \in \mathbb{R}$ (that is, Y_n is a scalar), then

$$X_n / Y_n \xrightarrow{\mathrm{d}} X/c \quad \text{if} \quad c \neq 0. \tag{1.188}$$

More General Functions

The next issue has to do with functions of convergent sequences. We consider a sequence $X_1, X_2, \ldots$ in $\mathbb{R}^k$. The first function we consider is a simple linear projection, $t^{\mathrm{T}} X_n$ for $t \in \mathbb{R}^k$.

Theorem 1.44 (Cramér-Wold "device")
Let $X_1, X_2, \cdots$ be a sequence of random variables in $\mathbb{R}^k$ and let X be a random variable in $\mathbb{R}^k$.

$$X_n \xrightarrow{\mathrm{d}} X \iff t^{\mathrm{T}} X_n \xrightarrow{\mathrm{d}} t^{\mathrm{T}} X \ \forall t \in \mathbb{R}^k.$$

Proof. Follows from the continuity theorem and equation (1.132). ∎

Now consider a general function g from $(\mathbb{R}^k, \mathcal{B}^k)$ to $(\mathbb{R}^m, \mathcal{B}^m)$. Given convergence of $\{X_n\}$, we consider the convergence of $\{g(X_n)\}$. Given the limiting distribution of a sequence $\{X_n\}$, the convergence of $\{g(X_n)\}$ for a

general function g is not assured. In the following we will consider the sequence $\{g(X_n)\}$ for the case that g is a continuous Borel function. (To speak about continuity of a function of random variables, we must add some kind of qualifier, such as a.s., which, of course, assumes a probability measure.) The simple facts are given in Theorem 1.45.

Theorem 1.45
Let X and $\{X_n\}$ be random variables (k-vectors) and let g be a continuous Borel function from $\mathbb{R}^k$ to $\mathbb{R}^k$.

$$X_n \xrightarrow{\text{a.s.}} X \Rightarrow g(X_n) \xrightarrow{\text{a.s.}} g(X) \tag{1.189}$$

$$X_n \xrightarrow{\text{p}} X \Rightarrow g(X_n) \xrightarrow{\text{p}} g(X) \tag{1.190}$$

$$X_n \xrightarrow{\text{d}} X \Rightarrow g(X_n) \xrightarrow{\text{d}} g(X) \tag{1.191}$$

Proof. Exercise. ∎

Theorem 1.45 together with Slutsky's theorem provide conditions under which we may say that $g(X_n, Y_n)$ converges to $g(X, c)$.

In the following we will consider normalizing constants that may make a sequence more useful. We may wish consider, for example, the asymptotic variance of a sequence whose limiting variance is zero.

1.3.7 Asymptotic Distributions

We will now resume the consideration of weak convergence of distributions that we began in Section 1.3.4. Asymptotic distributions are the basis for the concept of asymptotic expectation, discussed in Section 1.3.8 below.

In many interesting cases, the limiting distribution of a sequence $\{X_n\}$ is degenerate. The fact that $\{X_n\}$ converges in probability to some given constant may be of interest, but in statistical applications, we are likely to be interested in how fast it converges, and what are the characteristics the sequence of the probability distribution that can be used for "large samples".

In this section we discuss how to modify a sequence so that the convergence is not degenerate. Statistical applications are discussed in Section 3.8.

Normalizing Constants

Three common types of sequences $\{X_n\}$ of interest are iid sequences, sequences of partial sums, and sequences of order statistics. Rather than focusing on the sequence $\{X_n\}$, it may be more useful to to consider a sequence of linear transformations of X_n, $\{X_n - b_n\}$, where the form of b_n is generally different for iid sequences, sequences of partial sums, and sequences of order statistics.

Given a sequence of constants b_n, if $X_n - b_n \xrightarrow{\text{d}} 0$, we may be interested in the rate of convergence, or other properties of the sequence as n becomes

large. It may be useful to magnify the difference $X_n - b_n$ by use of some normalizing sequence of constants a_n:

$$Y_n = a_n(X_n - b_n). \tag{1.192}$$

While the distribution of the sequence $\{X_n - b_n\}$ may be degenerate, the sequence $\{a_n(X_n - b_n)\}$ may have a distribution that is nondegenerate, and this asymptotic distribution may be useful in statistical inference. (This approach is called "asymptotic inference".) We may note that even though we are using the asymptotic distribution of $\{a_n(X_n - b_n)\}$, for a reasonable choice of a sequence of normalizing constants $\{a_n\}$, we sometimes refer to it as the asymptotic distribution of $\{X_n\}$ itself, but we must remember that it is the distribution of the normalized sequence, $\{a_n(X_n - b_n)\}$.

The shift constants generally serve to center the distribution, especially if the limiting distribution is symmetric. Although linear transformations are often most useful, we could consider sequences of more general transformations of X_n; instead of $\{a_n(X_n - b_n)\}$, we might consider $\{h_n(X_n)\}$, for some sequence of functions $\{h_n\}$.

The Asymptotic Distribution of $\{g(X_n)\}$

Applications often involve a differentiable Borel scalar function g, and we may be interested in the convergence of $\{g(X_n)\}$. (The same general ideas apply when g is a vector function, but the higher-order derivatives quickly become almost unmanageable.) When we have $\{X_n\}$ converging in distribution to $X + b$, what we can say about the convergence of $\{g(X_n)\}$ depends on the differentiability of g at b.

Theorem 1.46
Let X and $\{X_n\}$ be random variables (k-vectors) such that

$$a_n(X_n - b_n) \xrightarrow{\mathrm{d}} X, \tag{1.193}$$

where $b_1, b_2, \ldots$ is a sequence of constants such that $\lim_{n\to\infty} b_n = b < \infty$, and $a_1, a_2, \ldots$ is a sequence of constant scalars such that $\lim_{n\to\infty} a_n = \infty$ or such that $\lim_{n\to\infty} a_n = a > 0$. Now let g be a Borel function from $\mathbb{R}^k$ to $\mathbb{R}$ that is continuously differentiable at each b_n. Then

$$a_n(g(X_n) - g(b_n)) \xrightarrow{\mathrm{d}} (\nabla g(b))^{\mathrm{T}} X. \tag{1.194}$$

Proof. This follows from a Taylor series expansion of $g(X_n)$ and Slutsky's theorem. ∎

A common application of Theorem 1.46 arises from the simple corollary for the case when X in expression (1.193) has the multivariate normal distribution $\mathrm{N}_k(0, \Sigma)$ and $\nabla g(b) \neq 0$:

$$a_n(g(X_n) - g(b_n)) \xrightarrow{\mathrm{d}} Y, \tag{1.195}$$

where $Y \sim \mathrm{N}_k(0, (\nabla g(b))^{\mathrm{T}} \Sigma \nabla g(b))$.

One reason limit theorems such as Theorem 1.46 are important is that they can provide approximations useful in statistical inference. For example, we often get the convergence of expression (1.193) from the central limit theorem, and then the convergence of the sequence $\{g(X_n)\}$ provides a method for determining approximate confidence sets using the normal distribution, so long as $\nabla g(b) \neq 0$. This method in asymptotic inference is called the *delta method*, and is illustrated in Example 1.25 below. It is particularly applicable when the asymptotic distribution is normal.

The Case of $\nabla g(b) = 0$

Suppose $\nabla g(b) = 0$ in equation (1.194). In this case the convergence in distribution is to a degenerate random variable, which may not be very useful. If, however, $\mathrm{H}_g(b) \neq 0$ (where H_g is the Hessian of g), then we can use a second order the Taylor series expansion and get something useful:

$$2a_n^2(g(X_n) - g(b_n)) \xrightarrow{\mathrm{d}} X^{\mathrm{T}} \mathrm{H}_g(b) X, \tag{1.196}$$

where we are using the notation and assuming the conditions of Theorem 1.46. Note that while $a_n(g(X_n) - g(b_n))$ may have a degenerate limiting distribution at 0, $a_n^2(g(X_n) - g(b_n))$ may have a nondegenerate distribution. (Recalling that $\lim_{n\to\infty} a_n = \infty$, we see that this is plausible.) Equation (1.196) allows us also to get the asymptotic covariance for the pairs of individual elements of X_n.

Use of expression (1.196) is called a *second order delta method*, and is illustrated in Example 1.25.

Example 1.25 an asymptotic distribution in a Bernoulli family
Consider the Bernoulli family of distributions with parameter π. The variance of a random variable distributed as Bernoulli(π) is $g(\pi) = \pi(1 - \pi)$. Now, suppose $X_1, X_2, \ldots \overset{\mathrm{iid}}{\sim}$ Bernoulli(π). Since $\mathrm{E}(\overline{X}_n) = \pi$, we may be interested in the distribution of $T_n = g(\overline{X}_n) = \overline{X}_n(1 - \overline{X}_n)$.

From the central limit theorem (Theorem 1.38),

$$\sqrt{n}(\overline{X}_n - \pi) \to \mathrm{N}(0, \pi(1 - \pi)), \tag{1.197}$$

and so if $\pi \neq 1/2$, $g'(\pi) \neq 0$, we can use the delta method from expression (1.194) to get

$$\sqrt{n}(T_n - g(\pi)) \to \mathrm{N}(0, \pi(1 - \pi)(1 - 2\pi)^2). \tag{1.198}$$

If $\pi = 1/2$, $g'(\pi) = 0$ and this is a degenerate distribution, so we cannot use the delta method. Let's use expression (1.196). The Hessian is particularly simple.

First, we note that in this case, the CLT yields $\sqrt{n}(\overline{X} - 1/2) \to \mathrm{N}\left(0, \frac{1}{4}\right)$. Hence, if we scale and square, we get $4n(\overline{X} - \frac{1}{2})^2 \xrightarrow{\mathrm{d}} \chi^2_1$, or

$$4n(T_n - g(\pi)) \xrightarrow{\mathrm{d}} \chi^2_1.$$

∎

We can summarize the previous discussion and the special results of Example 1.25 as follows (assuming all of the conditions on the objects involved),

$$\left.\begin{array}{c} \sqrt{n}(T_n - b_n) \to \mathrm{N}(0, \sigma^2) \\ g'(b) = 0 \\ g''(b) \neq 0 \end{array}\right\} \Longrightarrow 2n\frac{(g(T_n) - g(b_n))^2}{\sigma^2 g''(b)} \xrightarrow{\mathrm{d}} \chi^2_1. \tag{1.199}$$

Higher Order Expansions

Suppose the second derivatives of $g(b)$ are zero. We can easily extend this to higher order Taylor expansions in Theorem 1.47 below. (Note that because higher order Taylor expansions of vector expressions can become quite messy, in Theorem 1.47 we use $Y = (Y_1, \ldots, Y_k)$ in place of X as the limiting random variable.)

Theorem 1.47
Let Y and $\{X_n\}$ be random variables (k-vectors) such that

$$a_n(X_n - b_n) \xrightarrow{\mathrm{d}} Y,$$

where b_n is a constant sequence and $a_1, a_2, \ldots$ is a sequence of constant scalars such that $\lim_{n\to\infty} a_n = \infty$. Now let g be a Borel function from $\mathbb{R}^k$ to $\mathbb{R}$ whose m^{th} order partial derivatives exist and are continuous in a neighborhood of b_n, and whose j^{th}, for $1 \leq j \leq m-1$, order partial derivatives vanish at b. Then

$$m! a_n^m (g(X_n) - g(b_n)) \xrightarrow{\mathrm{d}} \sum_{i_1=1}^{k} \cdots \sum_{i_m=1}^{k} \left.\frac{\partial^m g}{\partial x_{i_1} \cdots \partial x_{i_m}}\right|_{x=b} Y_{i_1} \cdots Y_{i_m}. \tag{1.200}$$

Expansion of Statistical Functions

*** refer to functional derivatives, Sections 0.1.13 and 0.1.13.

Variance Stabilizing Transformations

The fact that the variance in the asymptotic distribution in expression (1.198) depends on π may complicate our study of T_n and its relationship to π. Of course, this dependence results initially from the variance $\pi(1-\pi)$ in the asymptotic distribution in expression (1.197). If $g(\pi)$ were chosen so that

$(g(\pi)')^2 = (\pi(1-\pi)^{-1})$, the variance in an expression similar to (1.198) would be constant (in fact, it would be 1).

Instead of $g(\pi) = \pi(1-\pi)$ as in Example 1.25, we can use a solution to the differential equation

$$g'(\pi) = \pi(1-\pi)^{-1/2}.$$

One solution is $g(t) = 2\arcsin(\sqrt{t})$, and following the same procedure in Example 1.25 but using this function for the transformations, we have $2\sqrt{n}\left(\arcsin(\sqrt{X_n}) - \arcsin(\sqrt{\pi})\right) \xrightarrow{\mathrm{d}} \mathrm{N}(0,1)$.

A transformation such as this is called a *variance stabilizing transformation* for obvious reasons.

Example 1.26 variance stabilizing transformation in a normal family

Consider the normal family of distributions with known mean 0 and variance σ^2, and suppose $X_1, X_2, \ldots \overset{\mathrm{iid}}{\sim} \mathrm{N}(0,\sigma^2)$. Since $\mathrm{E}(X_i^2) = \sigma^2$, we may be interested in the distribution of $T_n = \sum X_i^2/n$. We note that $\mathrm{V}(X_i^2) = 2\sigma^4$, hence, the central limit theorem gives

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \sigma^2\right) \to \mathrm{N}(0, 2\sigma^4).$$

Following the ideas above, we seek a transformation $g(\sigma^2)$ such that $(g'(\sigma^2))^2\sigma^4$ is constant wrt σ^2. A solution to the differential equation that expresses this relationship is $g(t) = \log(t)$, and as above, we have

$$\sqrt{n}\left(\log\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) - \log\left(\sigma^2\right)\right) \to \mathrm{N}(0,2). \tag{1.201}$$

∎

Order Statistics and Quantiles

The asymptotic distributions of order statistics $X_{k_n:n}$ are often of interest. The asymptotic properties of "central" order statistics are different from those of "extreme" order statistics.

A sequence of central order statistics $\{X_{(k:n)}\}$ is one such that for given $\pi \in]0,1[$, $k/n \to \pi$ as $n \to \infty$. (Notice that k depends on n, but we will generally not use the notation k_n.) As we suggested on page 64, the expected value of the k^{th} order statistic in a sample of size n, if it exists, should be approximately the same as the k/n quantile of the underlying distribution. Under mild regularity conditions, a sequence of asymptotic central order statistics can be shown to converge in expectation to x_π, the π quantile.

Sample quantiles (defined on page 64) are the ordinary quantiles in the sense of equation (1.13) of the discrete distribution defined by the sample, $X_1, \ldots, X_n$, which has CDF $F_n(x)$, the ECDF, as defined in equation (1.34); that is, the π sample quantile is

$$x_\pi = F_n^{-1}(\pi). \tag{1.202}$$

Properties of quantiles, of course, are different for discrete and continuous distributions. In the following, for $0 < \pi < 1$ we will assume that $F(x_\pi)$ is twice differentiable in some neighborhood of x_π and F'' is bounded and $F'(x_\pi) > 0$ in that neighborhood. Denote $F'(x)$ as $f(x)$, and let $F_n(x)$ be the ECDF. Now, write the k^{th} order statistic as

$$X_{(k:n)} = x_\pi - \frac{F_n(x_\pi) - \pi}{f(x_\pi)} + R_n(\pi). \tag{1.203}$$

This is called the Bahadur representation, after Bahadur (1966), who showed that $R_n(\pi) \to 0$ as $n \to \infty$. Kiefer (1967) determined the exact order of $R_n(\pi)$, so equation (1.203) is sometimes called the Bahadur-Kiefer representation. The Bahadur representation is useful in studying asymptotic properties of central order statistics.

There is some indeterminacy in relating order statistics to quantiles. In the Bahadur representation, for example, the details are slightly different if $n\pi$ happens to be an integer. (The results are the same, however.) Consider a slightly different formulation for a set of m order statistics. The following result is due to Ghosh (1971).

Theorem 1.48
Let $X_1, \ldots, X_n$ be iid random variables with PDF f. For $k = n_1, \ldots, n_m \leq n$, let $\lambda_k \in]0, 1[$ be such that $n_k = \lceil n\lambda_k \rceil + 1$. Now suppose $0 < \lambda_1 < \cdots < \lambda_m < 1$ and for each k, $f(x_{\lambda_k}) > 0$. Then the asymptotic distribution of the random m-vector

$$\left(n^{1/2}(X_{(n_1:n)} - x_{\lambda_1}),\ \ldots,\ n^{1/2}(X_{(n_m:n)} - x_{\lambda_m}) \right)$$

is m-variate normal with mean of 0, and covariance matrix whose i, j element is

$$\left(\frac{\lambda_i(1 - \lambda_j)}{f(x_{\lambda_i}) f(x_{\lambda_j})} \right).$$

For a proof of this theorem, see David and Nagaraja (2003).

A sequence of extreme order statistics $\{X_{(k:n)}\}$ is one such that $k/n \to 0$ or $k/n \to 1$ as $n \to \infty$. Sequences of extreme order statistics from a distribution with bounded support generally converge to a degenerate distribution, while those from a distribution with unbounded support do not have a meaningful distribution unless the sequence is normalized in some way. We will consider asymptotic distributions of extreme order statistics in Section 1.4.3.

We now consider some examples of sequences of order statistics. In Examples 1.27 and 1.28 below, we obtain degenerate distributions unless we introduce a normalizing factor. In Example 1.29, it is necessary to introduce a sequence of constant shifts.

Example 1.27 asymptotic distribution of min or max order statistics from $\mathrm{U}(0,1)$
Suppose $X_1, \ldots, X_n$ are iid $\mathrm{U}(0,1)$. The CDFs of the min and max, $X_{(1:n)}$ and $X_{(n:n)}$, are easy to work out. For $x \in [0,1]$,

$$\begin{aligned} F_{X_{(1:n)}}(x) &= 1 - \Pr(X_1 > x, \ldots, X_n > x) \\ &= 1 - (1-x)^n \end{aligned}$$

and

$$F_{X_{(n:n)}}(x) = x^n.$$

Notice that these are beta distributions, as we saw in Example 1.17.

Both of these extreme order statistics have degenerate distributions. For $X_{(1:n)}$, we have $X_{(1:n)} \xrightarrow{\mathrm{d}} 0$ and

$$\mathrm{E}\left(X_{(1:n)}\right) = \frac{1}{n+1}$$

and so

$$\lim_{n\to\infty} \mathrm{E}\left(X_{(1:n)}\right) = 0.$$

This suggests the normalization $nX_{(1:n)}$. We have

$$\begin{aligned} \Pr\left(nX_{(1:n)} \le x\right) &= 1 - \left(1 - \frac{x}{n}\right)^n \\ &\to 1 - \mathrm{e}^{-x} \quad x > 0. \end{aligned} \tag{1.204}$$

This is the CDF of a standard exponential distribution. The distribution of $nX_{(1:n)}$ is more interesting than that of $X_{(1:n)}$.

For $X_{(n:n)}$, we have $X_{(n:n)} \xrightarrow{\mathrm{d}} 1$ and

$$\mathrm{E}\left(X_{(n:n)}\right) = \frac{n}{n+1}$$

and so

$$\lim_{n\to\infty} \mathrm{E}\left(X_{(n:n)}\right) = 1,$$

and there is no normalization to yield a nondegenerate distribution. ∎

Now consider the asymptotic distribution of central order statistics from $\mathrm{U}(0,1)$.

Example 1.28 asymptotic distribution of a central order statistic from U(0, 1)
Let $X_{(k:n)}$ be the k^{th} order statistic from a random sample of size n from U(0, 1) and let

$$Y = nX_{(k:n)}.$$

Using Equation (1.138) with the CDF of U(0, 1), we have

$$f_Y(y) = \binom{n}{k}\left(\frac{y}{n}\right)^{k-1}\left(1-\frac{y}{n}\right)^{n-k}\mathrm{I}_{[0,1]}(y).$$

Observing that

$$\lim_{n\to\infty}\binom{n}{k} = \frac{n^{k-1}}{k!},$$

we have for fixed k,

$$\lim_{n\to\infty} f_Y(y) = \frac{1}{\Gamma(k)}y^{k-1}\mathrm{e}^{-y}\mathrm{I}_{[0,\infty[}(y), \tag{1.205}$$

that is, the limiting distribution of $\{nX_{(k:n)}\}$ is gamma with scale parameter 1 and shape parameter k. (Note, of course, $k! = k\Gamma(k)$.) For finite values of n, the asymptotic distribution provides better approximations when k/n is relatively small. When k is large, n must be much larger in order for the asymptotic distribution to approximate the true distribution closely.

If $k = 1$, the PDF in equation (1.205) is the exponential distribution, as shown in Example 1.27. For $k \to n$, however, we must apply a limit similar to what is done in equation (1.204). ∎

While the min and max of the uniform distribution considered in Example 1.27 are "extreme" values, the more interesting extremes are those from distributions with infinite support. In the next example, we consider an extreme value that has no bound. In such a case, in addition to any normalization, we must do a shift.

Example 1.29 extreme value distribution from an exponential distribution
Let $X_{(n:n)}$ be the largest order statistic from a random sample of size n from an exponential distribution with PDF $\mathrm{e}^{-x}\mathrm{I}_{\mathbb{R}_+}(x)$ and let

$$Y = X_{(n:n)} - \log(n).$$

We have

$$\lim_{n\to\infty}\Pr(Y \le y) = \mathrm{e}^{-\mathrm{e}^{-y}} \tag{1.206}$$

(Exercise 1.65). The distribution with CDF given in equation (1.206) is called an *extreme value distribution*. There are two other classes of "extreme value distributions", which we will discuss in Section 1.4.3. The one in this example, which is the most common one, is called a type 1 extreme value distribution or a Gumbel distribution. ∎

1.3.8 Asymptotic Expectation

The properties of the asymptotic distribution, such as its mean or variance, are the asymptotic values of the corresponding properties of T_n. Let $\{T_n\}$ be a sequence of random variables with $\mathrm{E}(|T_n|) < \infty$ and $T_n \xrightarrow{\mathrm{d}} T$, with $\mathrm{E}(|T|) < \infty$. Theorem 1.40 (on page 90) tells us that

$$\mathrm{E}(|T_n|^k) \to \mathrm{E}(|T|^k).$$

When T_n is a normalized statistic, such as $\overline{X}$, with variance of the form σ^2/n, the limiting value of some properties of T_n may not be very useful in statistical inference. We need an "asymptotic variance" different from $\lim_{n\to\infty} \sigma^2/n$.

Because $\{T_n\}$ may converge to a degenerate random variable, it may be more useful to generalize the definition of asymptotic expectation slightly. We will define "an asymptotic expectation", and distinguish it from the "limiting expectation". We will consider a sequence of the form $\{a_n T_n\}$.

Definition 1.43 (asymptotic expectation)
Let $\{T_n\}$ be a sequence of random variables, and let $\{a_n\}$ be a sequence of positive constants with $\lim_{n\to\infty} a_n = \infty$ or with $\lim_{n\to\infty} a_n = a > 0$, and such that $a_n T_n \xrightarrow{\mathrm{d}} T$, with $\mathrm{E}(|T|) < \infty$. An *asymptotic expectation* of $\{T_n\}$ is $\mathrm{E}(T/a_n)$. ∎

Notice that an asymptotic expectation many include an n; that is, the order of an asymptotic expression may be expressed in the asymptotic expectation. For example, the asymptotic variance of a sequence of estimators $\sqrt{n}T_n(X)$ may be of the form $\mathrm{V}(T/n)$; that is, the order of the asymptotic variance is n^{-1}.

We refer to $\lim_{n\to\infty} \mathrm{E}(S_n)$ as the *limiting expectation*. It is important to recognize the difference in limiting expectation and asymptotic expectation. The limiting variance of a sequence of estimators $\sqrt{n}T_n(X)$ may be 0, while the asymptotic variance is of the form $\mathrm{V}(T/n)$.

Asymptotic expectation has a certain arbitrariness associated with the choice of $\{a_n\}$. The range of possibilities for "an" asymptotic expectation, however, is limited, as the following theorem shows.

Theorem 1.49
Let $\{T_n\}$ be a sequence of random variables, and let $\{c_n\}$ be a sequence of positive constants with $\lim_{n\to\infty} c_n = \infty$ or with $\lim_{n\to\infty} c_n = c > 0$, and such that $c_n T_n \xrightarrow{\mathrm{d}} R$, with $\mathrm{E}(|R|) < \infty$. Likewise, let $\{d_n\}$ be a sequence of positive constants with $\lim_{n\to\infty} d_n = \infty$ or with $\lim_{n\to\infty} d_n = d > 0$, and such that $d_n T_n \xrightarrow{\mathrm{d}} S$, with $\mathrm{E}(|S|) < \infty$. (This means that both $\mathrm{E}(R/c_n)$ and $\mathrm{E}(S/d_n)$ are asymptotic expectations of T_n.) Then it must be the case that either

(i) $\mathrm{E}(R) = \mathrm{E}(S) = 0$,

(ii) either $\mathrm{E}(R) \neq 0$, $\mathrm{E}(S) = 0$, *and* $d_n/c_n \to 0$ *or* $\mathrm{E}(R) = 0$, $\mathrm{E}(S) \neq 0$, *and* $c_n/d_n \to 0$,
or
(iii) $\mathrm{E}(R) \neq 0$, $\mathrm{E}(S) \neq 0$, *and* $\mathrm{E}(R/c_n)/\mathrm{E}(S/d_n) \to 1$.

Proof. Exercise. (*Use Theorem 1.34 on page 86.*) ∎

Multivariate Asymptotic Expectation

The multivariate generalization of asymptotic expectation is straightforward: Let $\{X_n\}$ be a sequence of random k-vectors, and let $\{A_n\}$ be a sequence of $k \times k$ positive definite matrices such that either $\lim_{n\to\infty} A_n$ diverges (that is, in the limit has no negative diagonal elements and some diagonal elements that are positively infinite) or else $\lim_{n\to\infty} A_n = A$, where A is positive definite and such that $A_n X_n \xrightarrow{\mathrm{d}} X$, with $\mathrm{E}(|X|) < \infty$. Then *an asymptotic expectation* of $\{X_n\}$ is $\mathrm{E}(A_n^{-1} X)$.

If the asymptotic expectation of $\{X_n\}$ is $B(n)\mu$ for some matrix $B(n)$, and g is a Borel function from $\mathbb{R}^k$ to $\mathbb{R}^k$ that is differentiable at μ, then by Theorem 1.46 on page 93 the asymptotic expectation of $\{g(X_n)\}$ is $B(n)\mathrm{J}_g(\mu))^{\mathrm{T}}\mu$.

1.4 Limit Theorems

We are interested in functions of a sequence of random variables $\{X_i \,|\, i = 1, \ldots, n\}$, as n increases without bound. The functions of interest involve either sums or extreme order statistics. There are three general types of important limit theorems: *laws of large numbers*, *central limit theorems*, and *extreme value theorems*.

Laws of large numbers give limits for probabilities or for expectations of sequences of random variables. The convergence to the limits may be weak or strong.

Historically, the first versions of both laws of large numbers and central limit theorems applied to sequences of binomial random variables.

Central limit theorems and extreme value theorems provide weak convergence results, but they do even more; they specify a limiting distribution. Central limit theorems specify a limiting *infinitely divisible distribution*, often a normal distribution; and extreme value theorems specify a limiting *extreme value distribution*, one of which we encountered in Example 1.29.

The functions of the sequence of interest are of the form

$$a_n \left(\sum_{i=1}^{n} X_i - b_n \right) \tag{1.207}$$

or

$$a_n \left(X_{(n:n)} - b_n \right), \tag{1.208}$$

where $\{a_n\}$ is a sequence of positive real constants and $\{b_n\}$ is a sequence of real constants. The sequence of normalizing constants $\{a_n\}$ for either case often have the form $a_n = n^{-p}$ for some fixed $p > 0$.

For both laws of large numbers and central limit theorems, we will be interested in a function of the form of expression (1.207), whereas for the extreme value theorems, we will be interested in a function of the form of expression (1.208). An extreme value theorem, of course, may involve $X_{(1:n)}$ instead of $X_{(n:n)}$. The simplest version of a central limit theorem applies to sequences of iid random variables with finite variance, as in Theorem 1.38. The simplest version of the extreme value theorem applies to sequences of exponential random variables, as in Example 1.29.

For the laws of large numbers and the central limit theorems, we will find it convenient to define

$$S_n = \sum_{i=1}^{n} X_i. \tag{1.209}$$

We distinguish different types of sequences of random variables based on the distributions of the individual terms in the sequence and on the correlational structure of the terms. Most of the results discussed in Section 1.3 did not place any restrictions on the distributions or on their correlational structure. The limit theorems often require identical distributions (or at least distributions within the same family and which vary in a systematic way). Even when different distributions are allowed, the limit theorems that we discuss in this section require that the terms in the sequence be independent. We will consider sequences of correlated random variables in Section 1.6.

1.4.1 Laws of Large Numbers

The first law of large numbers was *Bernoulli's* (Jakob Bernoulli's) *theorem*. In this case S_n is the sum of n iid Bernoullis, so it has a binomial distribution.

Theorem 1.50 (Bernoulli's theorem (binomial random variables))
If S_n has a binomial distribution with parameters n and π, then

$$\frac{1}{n} S_n \xrightarrow{\text{p}} \pi. \tag{1.210}$$

Proof. This follows from $\int_\Omega (S_n/n - \pi)^2 \mathrm{d}P = \pi(1-\pi)/n$, which means S_n/n converges in mean square to π, which in turn means that it converges in probability to π. ∎

This is a *weak* law because the convergence is in probability.

Bernoulli's theorem applies to binomial random variables. We now state without proof four theorems about large numbers. Proofs can be found in Petrov (1995), for example. The first two apply to iid random numbers and the second two require only independence and finite expectations. Two are weak laws (WLLN) and two are strong (SLLN). For applications in statistics, the weak laws are generally more useful than the strong laws.

A generalization of Bernoulli's theorem is the *weak law of large numbers (WLLN) for iid random variables*:

Theorem 1.51 (WLLN for iid random variables)
Let $X_1, X_2, \ldots$ be a sequence of iid random variables (and $S_n = \sum_{i=1}^n X_i$). There exists a sequence of real numbers $a_1, a_2, \ldots$ such that $\forall i$

$$n\Pr(|X_i| > n) \to 0 \iff \frac{1}{n}S_n - b_n \xrightarrow{\text{p}} 0 \tag{1.211}$$

The b_n can be chosen so that $b_n \leq n$, as $b_n = \mathrm{E}\left(X_i \mathrm{I}_{\{|X_i| \leq n\}}\right)$, for example.

Theorem 1.52 (SLLN for iid random variables)
Let $X_1, X_2, \ldots$ be a sequence of iid random variables such that $\forall i$ $\mathrm{E}(|X_i|) = \mu < \infty$. Then

$$\frac{1}{n}S_n \xrightarrow{\text{a.s.}} \mu. \tag{1.212}$$

A slight generalization is the alternate conclusion

$$\frac{1}{n}\sum_{i=1}^n a_i(X_i - \mathrm{E}(X_1)) \xrightarrow{\text{a.s.}} 0,$$

for any bounded sequence of real numbers $a_1, a_2, \ldots$.

We can generalize these two limit theorems to the case of independence but not necessarily identical distributions, by putting limits on normalized p^{th} moments.

Theorem 1.53 (WLLN for independent random variables)
Let $X_1, X_2, \ldots$ be a sequence of independent random variables such for some constant $p \in [1, 2]$,

$$\lim_{n\to\infty} \frac{1}{n^p}\sum_{i=1}^n \mathrm{E}(|X_i|^p) = 0.$$

Then

$$\frac{1}{n}\left(S_n - \sum_{i=1}^n \mathrm{E}(X_i)\right) \xrightarrow{\text{p}} 0. \tag{1.213}$$

Theorem 1.54 (SLLN for independent random variables)
Let $X_1, X_2, \ldots$ be a sequence of independent random variables such for some constant $p \in [1, 2]$,

$$\sum_{i=1}^\infty \frac{\mathrm{E}(|X_i|^p)}{i^p} < \infty.$$

Then

$$\frac{1}{n}\left(S_n - \sum_{i=1}^n \mathrm{E}(X_i)\right) \xrightarrow{\text{a.s.}} 0. \tag{1.214}$$

We notice that the normalizing term in all of the laws of large numbers has been n^{-1}. We recall that the normalizing term in the simple central limit theorem 1.38 (and in the central limit theorems we will consider in the next section) is $n^{-1/2}$. Since the central limit theorems give convergence to a non-degenerate distribution, when the normalizing factor is as small as $n^{-1/2}$, we cannot expect convergence in probability, and so certainly not almost sure convergence. We might ask if there is some sequence a_n with $n^{-1/2} < a_n < n^{-1}$, such that when a_n is used as a normalizing factor, we have convergence in probability and possibly almost sure convergence. The "Law of the Iterated Logarithm (LIL)" provides a very interesting answer for iid random variables with finite variance. The sequence involves the iterated logarithm, $\log(\log(n))$; specifically, $a_n = (n \log(\log(n)))^{-1/2}$.

Without loss of generality we will assume the random variables have mean 0.

Theorem 1.55 (Law of the Iterated Logarithm)
Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables with $\mathrm{E}(X_i) = 0$ and $\mathrm{V}(X_i) = \sigma^2 < \infty$. Then

$$\frac{1}{\sigma\sqrt{2}} \frac{1}{\sqrt{n \log(\log(n))}} S_n \xrightarrow{\mathrm{p}} 0 \tag{1.215}$$

and

$$\limsup \frac{1}{\sigma\sqrt{2}} \frac{1}{\sqrt{n \log(\log(n))}} S_n \overset{\mathrm{a.s.}}{=} 1. \tag{1.216}$$

Proof. *See Billingsley (1995).* ∎

Further generalizations of laws of large numbers apply to sequences of random variables in triangular arrays, as in the definition of infinite divisibility, Definition 1.32. We will use triangular arrays in an important central limit theorem in the next section.

1.4.2 Central Limit Theorems for Independent Sequences

Central limit theorems give conditions that imply that certain standardized sequences converge to a normal distribution. We will be interested in sequences of the form of equation (1.207):

$$a_n \left(\sum_{i=1}^{n} X_i - b_n \right).$$

where $\{a_n\}$ is a sequence of positive real constants and $\{b_n\}$ is a sequence of real constants.

The simplest central limit theorems apply to iid random variables. More complicated ones apply to independent random variables that are not necessarily identically distributed and/or that are not necessarily independent. In

this section we consider central limit theorems for independent sequences. On page 134 we will consider a central limit theorem in which the sequences are not necessarily independent.

The central limit theorems require finite second moments.

The de Moivre Laplace Central Limit Theorem

The first central limit theorem, called the de Moivre Laplace central limit theorem followed soon after Bernoulli's theorem, and like Bernoulli's theorem, it applies to S_n that has a binomial distribution with parameters n and π (because it is the sum of n iid Bernoullis with parameter π.

Theorem 1.56 (De Moivre Laplace Central Limit Theorem)
If S_n has a binomial distribution with parameters n and π, then

$$\frac{1}{\sqrt{\pi(1-\pi)}}\frac{1}{\sqrt{n}}(S_n - n\pi) \xrightarrow{\mathrm{d}} \mathrm{N}(0,1). \tag{1.217}$$

This central limit theorem is a special case of the classical central limit theorem for iid random variables with finite mean and variance.

Notice that Bernoulli's theorem and the de Moivre Laplace central limit theorem, which are stated in terms of binomial random variables, apply to normalized limits of sums of Bernoulli random variables. This is the usual form of these kinds of limit theorems; that is, they apply to normalized limits of sums of random variables. The first generalizations apply to sums of iid random variables, and then further generalizations apply to sums of just independent random variables.

The Central Limit Theorem for iid Scalar Random Variables with Finite Mean and Variance

Theorem 1.57
Let $X_1, X_2, \ldots$ be a sequence of independent random variables that are identically distributed with mean μ and variance $\sigma^2 > 0$. Then

$$\frac{1}{\sigma}\frac{1}{\sqrt{n}}\left(\sum_{i=1}^{n} X_i - n\mu\right) \xrightarrow{\mathrm{d}} \mathrm{N}(0,1). \tag{1.218}$$

A proof of this uses a limit of a characteristic function and the uniqueness of the characteristic function (see page 87).

Independent but Not Identical; Triangular Arrays

The more general central limit theorems apply to a triangular array; that is, to a sequence of finite subsequences. The variances of the sums of the subsequences is what is used to standardize the sequence so that it is convergent. We define the sequence and the subsequences as follows.

Let $\{X_{nj}, j = 1, 2, \ldots, k_n\}$ be independent random variables with $k_n \to \infty$ as $n \to \infty$. We let

$$R_n = \sum_{j=1}^{k_n} X_{nj}$$

represent "row sums", as we visualize the sequence in an array:

$$\begin{array}{llll} X_{11}, X_{12}, \ldots X_{1k_1} & & & R_1 \\ X_{21}, X_{22}, \ldots\ldots X_{2k_2} & & & R_2 \\ X_{31}, X_{32}, \ldots\ldots\ldots X_{3k_3} & & & R_3 \\ \vdots \quad \vdots & & \vdots & \vdots \end{array} \tag{1.219}$$

There is no requirement that $k_j > k_i$ when $j > i$, but since $k_n \to \infty$ as $n \to \infty$, that is certainly the trend. Note that within a row of this triangular array, the elements are independent, but between rows, there is no such requirement.

Let $\sigma_n^2 = \mathrm{V}(R_n)$, and assume $0 < \sigma_n^2$.

Notice that aside from the finite variance, the only assumption is that within a row, the elements are independent. There is no assumption regarding different rows; they may be independent and they may or may not have identical distributions.

In order to say anything about the asymptotic distribution of some function of $\{X_{nj}\}$ we need to impose conditions on the moments. There are three standard conditions that we consider for sequences satisfying the general conditions on k_n and σ_n^2.

- **Lyapunov's condition.** Lyapunov's condition applies uniformly to the sum of central moments of order $(2 + \delta)$. Lyapunov's condition is

$$\sum_{j=1}^{k_n} \mathrm{E}\left(|X_{nj} - \mathrm{E}(X_{nj})|^{2+\delta}\right) \in \mathrm{o}(\sigma_n^{2+\delta}) \quad \text{for some } \delta > 0. \tag{1.220}$$

- **Lindeberg's condition.** Lindeberg's condition is

$$\sum_{j=1}^{k_n} \mathrm{E}\left((X_{nj} - \mathrm{E}(X_{nj}))^2 \mathrm{I}_{\{|X_{nj} - \mathrm{E}X_{nj}| > \epsilon\sigma_n\}}(X_{nj})\right) \in \mathrm{o}(\sigma_n^2) \; \forall \, \epsilon > 0, \tag{1.221}$$

  Instead of a strong uniform condition on a power in terms of a positive addition δ to 2, as in Lyapunov's condition, Lindeberg's condition applies to a fixed power of 2 over an interval controlled by ϵ. Lindeberg's condition requires that the sum of the second central moments over the full support minus the squared central differences near the mean is ultimately dominated by the variance of the sum. (That is to say, the sum of the tail components of the variance is dominated by the variance of the sum. This means that the distributions cannot be too heavy-tailed.) The requirement is in terms of an ϵ that tells how much of the central region to remove before computing the individual central moments.

Clearly, Lyapunov's condition implies Lindeberg's condition.
Although Lyapunov's condition is more stringent than Lindeberg's condition, it is sometimes easier to establish Lyapunov's condition than Lindeberg's condition.

- **Feller's condition.** Lindeberg's condition (or Lyapunov's condition, of course) implies *Feller's condition*, which is:

$$\lim_{n\to\infty} \max_{j\le k_n} \frac{\sigma_{nj}^2}{\sigma_n^2} = 0. \tag{1.222}$$

This condition comes up in the proof of Lindeberg's central limit theorem.

A Central Limit Theorem for Independent Scalar Random Variables with Finite Mean and Variance

A more general central limit theorem is called *Lindeberg's central limit theorem.* It is stated in terms of a sequence of the finite subsequences of a triangular array, as we encountered in the definition of infinite divisibility on page 61.

Theorem 1.58 (Lindeberg's Central Limit Theorem)
For given n, let $\{X_{nj}, j = 1, 2, \ldots, k_n\}$ be independent random variables with $0 < \sigma_n^2$, where $\sigma_n^2 = \mathrm{V}(\sum_{j=1}^{k_n} X_{nj})$ and $k_n \to \infty$ as $n \to \infty$. If the Lindeberg condition (1.221) holds, then

$$\frac{1}{\sigma_n} \sum_{j=1}^{k_n} (X_{nj} - \mathrm{E}(X_{nj})) \xrightarrow{\mathrm{d}} \mathrm{N}(0, 1). \tag{1.223}$$

Proof. *** From inequality (1.95), we have for each X_{nj},

$$\left|\varphi_{X_{nj}}(t) - (-t^2\mathrm{V}(X_{nj})/2)\right| \le \mathrm{E}\left(\min\left(|tX_{nj}|^2, |tX_{nj}|^3/6\right)\right).$$

*** ∎

Multivariate Central Limit Theorems for Independent Random Variables with Finite Mean and Variance

The central limit theorems stated above have multivariate extensions that are relatively straightforward. The complications arise from the variance-covariance matrices, which must replace the simple scalars σ_n^2.

The simplest situation is the iid case where each member of the sequence $\{X_n\}$ of random k-vectors has the finite variance-covariance matrix Σ. In that case, similar to equation (1.218) for iid scalar random variables, we have

$$\frac{\sum_{i=1}^{n}(X_i - \mathrm{E}(X_i))}{\sqrt{n}} \xrightarrow{\mathrm{d}} \mathrm{N}_k(0, \Sigma). \tag{1.224}$$

Another type of multivariate central limit theorem can be formed by thinking of the subsequences in equation (1.223) as multivariate random variables. Let $\{k_n\}$ be a sequence of constants such that $k_n \to \infty$ as $n \to \infty$. Let $X_{ni} \in \mathbb{R}^{m_i}$, where $m_i \leq m$ for some fixed integer m and for $i = 1, \ldots, k_n$, be independent with

$$\inf_{i,n} \lambda_{ni} > 0,$$

where λ_{ni} is the smallest eigenvalue of $\mathrm{V}(X_{ni})$. (Note that this is saying that variance-covariance matrix is positive definite for every n and i; but it's saying a little more than that.) Also suppose that for some $\delta > 0$, we have

$$\sup_{i,n} \mathrm{V}(\|X_{ni}\|^{2+\delta}) < \infty.$$

Now, let c_{ni} be a sequence in $\mathbb{R}^{m_i}$ with the property that it is diffuse:

$$\lim_{n\to\infty} \left(\max_{1\leq i\leq k_n} \|c_{ni}\|^2 \Big/ \sum_{i=1}^{k_n} \|c_{ni}\|^2 \right) = 0.$$

Then we have something similar to equation (1.223):

$$\sum_{j=1}^{k_n} c_{ni}^{\mathrm{T}} \left(X_{nj} - \mathrm{E}(X_{nj})\right) \Big/ \left(\sum_{j=1}^{k_n} \mathrm{V}(c_{ni}^{\mathrm{T}} X_{nj}) \right)^{1/2} \xrightarrow{\mathrm{d}} \mathrm{N}(0,1). \tag{1.225}$$

1.4.3 Extreme Value Distributions

In Theorem 1.48 we saw that the asymptotic joint distribution of a set of central order statistics obeys the central limit theorem. The asymptotic distribution of extreme order statistics, however, is not normal. We have already considered asymptotic distributions of extreme order statistics in special cases in Examples 1.27 and 1.28 for random variables with bounded ranges and in Example 1.29 for a random variable with unbounded range. The latter is the more interesting case, of course.

Given a sequence of random variables $\{X_i \,|\, i = 1, \ldots, n\}$, we are interested in the limiting distribution of functions of the form in expression (1.208), that is,

$$a_n \left(X_{(n:n)} - b_n\right).$$

The first question, of course, is what conditions on a_n and b_n will yield a limiting distribution that is nondegenerate. These conditions clearly must depend on the distributions of the X_i, and must take into account any dependencies within the sequence. We will consider only the simple case; that is, we will assume that the X_i are iid. Let F be the CDF of each X_i. The problem now is to find a CDF G for a nondegenerate distribution, such that for each point of continuity x of G,

$$\lim_{n\to\infty} F^n(x/a_n + b_n) = G(x). \tag{1.226}$$

Fisher and Tippett (1928) The most general answer to this is given in the following theorem.

Theorem 1.59 (extreme value distribution)
A CDF satisfying equation (1.226) must have one of the following three forms in which $\alpha > 0$:

- $$G(x) = \begin{cases} 0, & x \le 0 \\ \exp(-x^{-\alpha}), & x > 0; \end{cases} \tag{1.227}$$
- $$G(x) = \exp(-\mathrm{e}^{-x}); \tag{1.228}$$
- $$G(x) = \begin{cases} \exp(-(x)^{\alpha}), & x < 0 \\ 1, & x \ge 0. \end{cases} \tag{1.229}$$

The proof of a version of this theorem was given by Fisher and Tippett (1928), and a more careful statement along with a proof was given by Gnedenko (1943). See de Haan and Ferreira (2006) for a proof, and see David and Nagaraja (2003) for further discussion.

*** Combine these in one express and introduce the *extreme value index*

*** Give names to the three classes.

domain of attraction

1.4.4 Other Limiting Distributions

Asymptotic distributions are very important in statistical applications because, while the exact distribution of a function of a finite set of random variables may be very complicated, often the asymptotic distribution is uncomplicated. Often, as we have seen, the limiting distribution is normal if the sequence is properly normalized. If a normal distribution can be used, even as an approximation, there is a wealth of statistical theory that can be applied to the problem.

The random variables of interest are often functions $g(X)$ of simpler random variables X. If we know the limiting distribution of $\{X_n\}$ we can often work out the limiting distribution of $\{g(X_n)\}$, depending on the nature of the function g. A simple example of this is equation (1.195) for the delta method. In this case we start with $\{X_n\}$ that has a limiting normal distribution and we get that the limiting distribution of $\{g(X_n)\}$ is also normal.

We also can often get useful limiting distributions from the central limit theorem and the distributions of functions of normal random variables such as chi-squared, t, or F, as discussed in Section 2.9.2.

1.5 Conditional Probability

The concept of conditional distributions provides the basis for the analysis of relationships among variables.

A simple way of developing the ideas begins by defining the conditional probability of event A, given event B. If $\Pr(B) \neq 0$, the conditional probability of event A given event B is

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}, \tag{1.230}$$

which leads to the useful multiplication rule

$$\Pr(A \cap B) = \Pr(B)\Pr(A|B). \tag{1.231}$$

We see from this that if A and B are independent

$$\Pr(A|B) = \Pr(A).$$

If we interpret all of this in the context of the probability space $(\Omega, \mathcal{F}, P)$, we can define a new "conditioned" probability space, $(\Omega, \mathcal{F}, P_B)$, where we define P_B by

$$P_B(A) = \Pr(A \cap B),$$

for any $A \in \mathcal{F}$. From this conditional probability space we could then proceed to develop "conditional" versions of the concepts discussed in the previous sections.

This approach, however, is not entirely satisfactory because of the requirement that $\Pr(B) \neq 0$. More importantly, this approach in terms of events does not provide a basis for the development of conditional probability density functions.

Another approach is to make use of a concept of conditional expectation, and that is what we will proceed to do. In this approach, we develop several basic ideas before we finally speak of distributions of conditional random variables in Section 1.5.4.

1.5.1 Conditional Expectation: Definition and Properties

The definition of conditional expectation of one random variable given another random variable is developed in two stages. First, we define conditional expectation over a sub-σ-field and consider some of its properties, and then we define conditional expectation with respect to another measurable function (a random variable, for example) in terms of the conditional expectation over the sub-σ-field generated by the inverse image of the function.

A major difference in conditional expectations and unconditional expectations is that conditional expectations may be nondegenerate random variables. When the expectation is conditioned on a random variable, relations involving the conditional expectations must be qualified as holding in probability, or holding with probability 1.

Conditional Expectation over a Sub-σ-Field

Definition 1.44 (conditional expectation over a sub-σ-field)
Let $(\Omega, \mathcal{F}, P)$ be a probability space, let $\mathcal{A}$ be a sub-σ-field of $\mathcal{F}$, and let X be an integrable random variable over Ω. The *conditional expectation* of X given $\mathcal{A}$, denoted by $\mathrm{E}(X|\mathcal{A})$, is an $\mathcal{A}$-measurable function from $(\Omega, \mathcal{F})$ to $(\mathbb{R}^d, \mathcal{B}^d)$ such that

$$\int_A \mathrm{E}(X|\mathcal{A})\,\mathrm{d}P = \int_A X\,\mathrm{d}P, \quad \forall\, A \in \mathcal{A}. \tag{1.232}$$

∎

Clearly, if $\mathcal{A} = \mathcal{F}$, then $\mathrm{E}(X|\mathcal{A}) = \mathrm{E}(X)$

Being a real $\mathcal{A}$-measurable function, the conditional expectation is a random variable from the space $(\Omega, \mathcal{A}, P)$. Such a random variable exists and is a.s. unique, as we will see below (Theorem 1.60).

Equation (1.232) in terms of an indicator function is

$$\int_A \mathrm{E}(X|\mathcal{A})\,\mathrm{d}P = \mathrm{E}(X\mathrm{I}_A), \ \forall\, A \in \mathcal{A}. \tag{1.233}$$

Another equivalent condition, in terms of bounded $\mathcal{A}$-measurable functions, is

$$\mathrm{E}(\mathrm{E}((X|\mathcal{A})Y)) = \mathrm{E}(XY) \tag{1.234}$$

for all bounded and $\mathcal{A}$-measurable Y for which XY is integrable.

Theorem 1.60
Let $(\Omega, \mathcal{F}, P)$ be a probability space, let $\mathcal{A}$ be a sub-σ-field of $\mathcal{F}$, and let X be an integrable random variable from Ω into $\mathbb{R}^d$. Then there is an a.s. unique d-variate random variable Y on $(\Omega, \mathcal{A}, P_{\mathcal{A}})$ such that

$$\int_A Y\,\mathrm{d}P_{\mathcal{A}} = \int_A X\,\mathrm{d}P, \quad \forall\, A \in \mathcal{A}.$$

Proof. Exercise. (*Use the Radon-Nikodym theorem 0.1.30, on page 739 to show the existence. For a.s. uniqueness, assume the $\mathcal{A}$-measurable functions Y_1 and Y_2 are such that $\int_A Y_1\,\mathrm{d}P = \int_A Y_2\,\mathrm{d}P\ \forall\, A \in \mathcal{A}$ and show that $Y_1 = Y_2$ a.s. $\mathcal{A}$ and P.*) ∎

Conditional Expectation with Respect to a Measurable Function

Definition 1.45 (with respect to another measurable function)
Let $(\Omega, \mathcal{F}, P)$ be a probability space, let $\mathcal{A}$ be a sub-σ-field of $\mathcal{F}$, let X be an integrable random variable over Ω, and let Y be a measurable function from $(\Omega, \mathcal{F}, P)$ to any measurable space $(\Lambda, \mathcal{G})$. Then the *conditional expectation* of X given Y, denoted by $\mathrm{E}(X|Y)$, is defined as the conditional expectation of X given the sub-σ-field generated by Y, that is,

$$\mathrm{E}(X|Y) = \mathrm{E}(X|\sigma(Y)). \tag{1.235}$$

∎

Definition 1.44 provides meaning for the expression $\mathrm{E}(X|\sigma(Y))$ in equation (1.235).

Sub-σ-fields generated by random variables, such as $\sigma(Y)$, play an important role in statistics. We can think of $\sigma(Y)$ as being the "information provided by Y". In an important type of time series, $Y_1, Y_2, \ldots$, we encounter a sequence $\sigma(Y_1) \subseteq \sigma(Y_2) \subseteq \cdots$ and we think of each random variable in the series as providing additional information.

Another view of conditional expectations in statistical applications is as approximations or predictions; see Section 1.5.3.

1.5.2 Some Properties of Conditional Expectations

Although the definition above may appear rather abstract, it is not too difficult to work with, and it yields the properties of conditional expectation that we have come to expect based on the limited definitions of elementary probability.

For example, we have the simple relationship with the unconditional expectation:

$$\mathrm{E}(\mathrm{E}(X|\mathcal{A})) = \mathrm{E}(X). \tag{1.236}$$

Also, if the individual conditional expectations exist, the conditional expectation is a linear operator:

$$\forall a \in \mathbb{R}, \mathrm{E}(aX + Y|\mathcal{A}) = a\mathrm{E}(X|\mathcal{A}) + \mathrm{E}(Y|\mathcal{A}) \text{ a.s.} \tag{1.237}$$

This fact follows immediately from the definition. For any $A \in \mathcal{A}$

$$\begin{aligned} \mathrm{E}(aX + Y|\mathcal{A}) &= \int_A aX + Y \, \mathrm{d}P \\ &= a\int_A X \, \mathrm{d}P + \int_A Y \, \mathrm{d}P \\ &= a\mathrm{E}(X|\mathcal{A}) + \mathrm{E}(Y|\mathcal{A}) \end{aligned}$$

As with unconditional expectations, we have immediately from the definition:

$$X \leq Y \text{ a.s.} \quad \Rightarrow \quad \mathrm{E}(X|\mathcal{A}) \leq \mathrm{E}(Y|\mathcal{A}) \text{ a.s.}. \tag{1.238}$$

We can establish conditional versions of the three theorems stated on page 89 that relate to the interchange of an integration operation and a limit operation (monotone convergence, Fatou's lemma, and dominated convergence). These extensions are fairly straightforward.

- monotone convergence:
  for $0 \leq X_1 \leq X_2 \cdots$ a.s.
  $$X_n \xrightarrow{\text{a.s.}} X \quad \Rightarrow \quad \mathrm{E}(X_n|\mathcal{A}) \xrightarrow{\text{a.s.}} \mathrm{E}(X|\mathcal{A}). \tag{1.239}$$
- Fatou's lemma:
  $$0 \leq X_n \,\forall\, n \quad \Rightarrow \quad \mathrm{E}(\liminf_n X_n|\mathcal{A}) \leq \liminf_n \mathrm{E}(X_n|\mathcal{A}) \text{ a.s.}. \tag{1.240}$$
- dominated convergence:
  given a fixed Y with $\mathrm{E}(Y|\mathcal{A}) < \infty$,
  $$|X_n| \leq Y \,\forall\, n \text{ and } X_n \xrightarrow{\text{a.s.}} X \quad \Rightarrow \quad \mathrm{E}(X_n|\mathcal{A}) \xrightarrow{\text{a.s.}} \mathrm{E}(X|\mathcal{A}). \tag{1.241}$$

Another useful fact is that if Y is $\mathcal{A}$-measurable and $|XY|$ and $|X|$ are integrable (notice this latter is stronger than what is required to define $\mathrm{E}(X|\mathcal{A})$), then

$$\mathrm{E}(XY|\mathcal{A}) = Y\mathrm{E}(X|\mathcal{A}) \text{ a.s.} \tag{1.242}$$

Some Useful Conditional Expectations

There are some conditional expectations that arise often, and which we should immediately recognize. The simplest one is

$$\mathrm{E}\big(\mathrm{E}(Y|X)\big) = \mathrm{E}(Y). \tag{1.243}$$

Note that the expectation operator is based on a probability distribution, and so anytime we see "E", we need to ask "with respect to what probability distribution?" In notation such as that above, the distributions are implicit and all relate to the same probability space. The inner expectation on the left is with respect to the conditional distribution of Y given X, and so is a function of X. The outer expectation is with respect to the marginal distribution of X.

Approaching this slightly differently, we consider a random variable Z that is a function of the random variables X and Y:

$$Z = f(X, Y).$$

We have

$$\mathrm{E}(f(X,Y)) = \mathrm{E}_Y\left(\mathrm{E}_{X|Y}(f(X,Y)|Y)\right) = \mathrm{E}_X\left(\mathrm{E}_{Y|X}(f(X,Y)|X)\right). \tag{1.244}$$

Another useful conditional expectation relates adjusted variances to "total" variances:

$$\mathrm{V}(Y) = \mathrm{V}\big(\mathrm{E}(Y|X)\big) + \mathrm{E}\big(\mathrm{V}(Y|X)\big). \tag{1.245}$$

This is intuitive, although you should be able to prove it formally. The intuitive explanation is: the total variation in Y is the sum of the variation of its mean given X and its average variation about X (or given X). (Think of SST = SSR + SSE in regression analysis.)

This equality implies the Rao-Blackwell inequality (drop the second term on the right).

Exchangeability, Conditioning, and Independence

De Finetti's representation theorem (Theorem 1.30 on page 75) requires an infinite sequence, and does not hold for finite sequences. For example, consider an urn containing one red ball and one blue ball from which we draw the balls without replacement. Let $R_i = 1$ if a red ball is drawn on the i^{th} draw and $R_i = 0$ otherwise. (This is the Polya's urn of Example 1.6 on page 24 with $r = b = 1$ and $c = -1$.) Clearly, the sequence R_1, R_2 is exchangeable. Because

$$\Pr(R_1 = 1, R_2 = 1) = 0,$$

if there were a measure μ as in de Finetti's representation theorem, then we would have

$$0 = \int_0^1 \pi^2 \mathrm{d}\mu(\pi),$$

which means that μ must put mass 1 at the point 0. But also

$$\Pr(R_1 = 0, R_2 = 0) = 0,$$

which would mean that

$$0 = \int_0^1 (1-\pi)^2 \mathrm{d}\mu(\pi).$$

That would not be possible if μ satisfies the previous requirement. There are, however, finite versions of de Finetti's theorem; see, for example, Diaconis (1977) or Schervish (1995).

An alternate statement of de Finetti's theorem identifies a random variable with the distribution P, and in that way provides a more direct connection to its use in statistical inference.

Theorem 1.61 (de Finetti's representation theorem (alternate))
The sequence $\{X_i\}_{i=1}^{\infty}$ of binary random variables is exchangeable iff there is a random variable Π such that, conditional on $\Pi = \pi$, the $\{X_i\}_{i=1}^{\infty}$ are iid Bernoulli random variables with parameter π. Furthermore, if $\{X_i\}_{i=1}^{\infty}$ is exchangeable, then the distribution of Π is unique and $\overline{X}_n = \sum_{i=1}^{n} X_i/n$ converges to Π almost surely.

Example 1.30 exchangeable Bernoulli random variables that are conditionally iid Bernoullis (Schervish, 1995)
Suppose $\{X_n\}_{n=1}^{\infty}$ are exchangeable Bernoulli random variables such that for each n and for $k = 0, 1, \ldots, n$,

$$\Pr\left(\sum_{i=1}^{n} = k\right) = \frac{1}{n+1}.$$

Now $\overline{X}_n \xrightarrow{\text{a.s.}} \Pi$, where Π is as in Theorem 1.61, and so $\overline{X}_n \xrightarrow{\text{d}} \Pi$. To determine the distribution of Π, we write the CDF of $\overline{X}_n$ as

$$F_n(t) = \frac{\lfloor nt \rfloor + 1}{n+1};$$

hence, $\lim F_n(t) = t$, which is the CDF of Π. Therefore, Π has a $\mathrm{U}(0,1)$ distribution. The X_i are conditionally iid Bernoulli(π) for $\Pi = \pi$.

The distributions in this example will be used in Examples 4.2 and 4.6 in Chapter 4 to illustrate methods in Bayesian data analysis. ∎

Conditional expectations also are important in approximations of one random variable by another random variable, and in "predicting" one random variable using another, as we see in the next section.

1.5.3 Projections

Use of one distribution or one random variable as an approximation of another distribution or random variable is a very useful technique in probability and statistics. It is a basic technique for establishing asymptotic results, and it underlies statistical applications involving regression analysis and prediction.

Given two scalar random variables X and Y, consider the question of what Borel function g is such that $g(X)$ is closest to Y in some sense. A common way to define closeness of random variables is by use of the expected squared distance, $\mathrm{E}((Y-g(X))^2)$. This leads to the least squares criterion for determining the optimal $g(X)$.

First, we must consider whether or under what conditions, the problem has a solution under this criterion.

Theorem 1.62
Let X and Y be scalar random variables over a common measurable space and assume $\mathrm{E}(Y^2) < \infty$. Then there exists a Borel measurable function g_0 with $\mathrm{E}((g_0(X))^2) < \infty$ such that

$$\mathrm{E}((Y - g_0(X))^2) = \inf\{\mathrm{E}((Y - g(X))^2) \mid g(X) \in \mathcal{G}_0\}, \tag{1.246}$$

where $\mathcal{G}_0 = \{g(X) \mid g : \mathbb{R} \mapsto \mathbb{R}$ is Borel measurable and $\mathrm{E}((g_0(X))^2) < \infty\}$.

Proof. ***fix ∎

Although Theorem 1.62 is stated in terms of scalar random variables, a similar result holds for vector-valued random variables. The next theorem identifies a g_0 that minimizes the L_2 norm for vector-valued random variables.

Theorem 1.63
Let Y be a d-variate random variable such that $\mathrm{E}(\|Y\|_2) < \infty$ and let G be the set of all Borel measurable functions from $\mathbb{R}^k$ into $\mathbb{R}^d$. Let X be a k-variate random variable such that $\mathrm{E}(\|\mathrm{E}(Y|X)\|_2) < \infty$. Let $g_0(X) = \mathrm{E}(Y|X)$. Then

$$g_0(X) = \arg\min_{g \in G} \mathrm{E}(\|Y - g(X)\|_2). \tag{1.247}$$

Proof. Exercise. Compare this with Theorem 1.13 on page 27, in which the corresponding solution is $g_0(X) = \mathrm{E}(Y|a) = \mathrm{E}(Y)$. ∎

By the general definition of projection (Definition 0.0.9 on page 637), we see that conditional expectation can be viewed as a projection in a linear space defined by the square-integrable random variables over a given probability space and the inner product $\langle Y, X\rangle = \mathrm{E}(YX)$ and its induced norm. (In fact, some people define conditional expectation this way instead of the way we have in Definitions 1.44 and 1.45.)

In regression applications in statistics using least squares, as we discuss on page 438, "$\widehat{Y}$", or the "predicted" Y given X, that is, $\mathrm{E}(Y|X)$ is the projection of Y onto X. For given fixed values of Y and X the predicted Y given X is the vector projection, in the sense of Definition 0.0.9.

We now formally define projection for random variables in a manner analogous to Definition 0.0.9. Note that the random variable space is the range of the functions in G in Theorem 1.63.

Definition 1.46 (projection of a random variable onto a space of random variables)

Let Y be a random variable and let $\mathcal{X}$ be a random variable space defined on the same probability space. A random variable $X_p \in \mathcal{X}$ such that

$$\mathrm{E}(\|Y - X_p\|_2) \leq \mathrm{E}(\|Y - X\|_2) \; \forall X \in \mathcal{X} \tag{1.248}$$

is called a *projection of* Y *onto* $\mathcal{X}$. ∎

The most interesting random variable spaces are linear spaces, and in the following we will assume that $\mathcal{X}$ is a linear space, and hence the norm arises from an inner product so that the terms in inequality (1.248) involve variances and covariances.

*** existence, closure of space in second norm (see page 35).

*** treat vector variables differently: $\mathrm{E}(\|Y - \mathrm{E}(Y)\|_2)$ is not the variance**** make this distinction earlier

When $\mathcal{X}$ is a linear space, we have the following result for projections.

Theorem 1.64
Let $\mathcal{X}$ be a linear space of random variables with finite second moments. Then X_p is a projection of Y onto $\mathcal{X}$ iff $X_p \in \mathcal{X}$ and

$$\mathrm{E}\left((Y - X_p)^{\mathrm{T}} X\right) = 0 \; \forall X \in \mathcal{X}. \tag{1.249}$$

Proof.
For any $X, X_p \in \mathcal{X}$ we have

$$\begin{aligned}\mathrm{E}((Y - X)^{\mathrm{T}}(Y - X)) = {} & \mathrm{E}((Y - X_p)^{\mathrm{T}}(Y - X_p)) \\ & + 2\mathrm{E}((Y - X_p)^{\mathrm{T}}(X_p - X)) \\ & + \mathrm{E}((X_p - X)^{\mathrm{T}}(X_p - X))\end{aligned}$$

If equation (1.249) holds then the middle term is zero and so $\mathrm{E}((Y-X)^{\mathrm{T}}(Y-X)) \geq \mathrm{E}((Y-X_p)^{\mathrm{T}}(Y-X_p))\ \forall X \in \mathcal{X}$; that is, X_p is a projection of Y onto $\mathcal{X}$.

Now, for any real scalar a and any $X, X_p \in \mathcal{X}$, we have

$$\mathrm{E}((Y-X_p-aX)^{\mathrm{T}}(Y-X_p-aX)) - \mathrm{E}((Y-X_p)^{\mathrm{T}}(Y-X_p)) = -2a\mathrm{E}((Y-X_p)^{\mathrm{T}}X) + a^2\mathrm{E}(X^{\mathrm{T}}X).$$

If X_p is a projection of Y onto $\mathcal{X}$, the term on the left side of the equation is nonnegative for every a. But the term on the right side of the equation can be nonnegative for every a only if the orthogonality condition of equation (1.249) holds; hence, we conclude that that is the case. ∎

Because a linear space contains the constants, we have the following corollary.

Corollary 1.64.1
Let $\mathcal{X}$ be a linear space of random variables with finite second moments. and let X_p be a projection of the random variable Y onto $\mathcal{X}$. Then,

$$\mathrm{E}(X_p) = \mathrm{E}(Y), \tag{1.250}$$

$$\mathrm{Cov}(Y-X_p, X) = 0\ \forall X \in \mathcal{X}, \tag{1.251}$$

and

$$\mathrm{Cov}(Y, X) = \mathrm{Cov}(X_p, X)\ \forall X \in \mathcal{X}. \tag{1.252}$$

***fix ** add uniqueness etc. $\mathrm{E}(Y) = \mathrm{E}(X_p)$ and $\mathrm{Cov}(Y-X_p, X) = 0\ \forall X \in \mathcal{X}$.

Definition 1.47 (projection of a function of random variables)
Let $Y_1, \ldots, Y_n$ be a set of random variables. The *projection of the statistic $T_n(Y_1, \ldots, Y_n)$ onto the k_n random variables $X_1, \ldots, X_{k_n}$* is

$$\widetilde{T}_n = \mathrm{E}(T_n) + \sum_{i=1}^{k_n} \left(\mathrm{E}(T_n|X_i) - \mathrm{E}(T_n)\right). \tag{1.253}$$

∎

An interesting projection is one in which the $Y_1, \ldots, Y_{k_n}$ in Definition 1.47 are the same as $X_1, \ldots, X_n$. In that case, if T_n is a symmetric function of the $X_1, \ldots, X_n$ (for example, the $X_1, \ldots, X_n$ are iid), then the $\mathrm{E}(T_n|X_i)$ are iid with mean $\mathrm{E}(T_n)$. The residual, $T_n - \widetilde{T}_n$, is often of interest. Writing

$$T_n - \widetilde{T}_n = T_n - \mathrm{E}(T_n) - \sum_{i=1}^{n}(\mathrm{E}(T_n|X_i) - \mathrm{E}(T_n)),$$

we see that $\mathrm{E}(T_n - \widetilde{T}_n) = 0$. Hence, we have

$$\mathrm{E}(\widetilde{T}_n) = \mathrm{E}(T_n), \tag{1.254}$$

and if $\mathrm{V}(T_n) < \infty$

$$\mathrm{V}(\widetilde{T}_n) = n\mathrm{V}(\mathrm{E}(T_n|X_i)) \tag{1.255}$$

(exercise).

If $\mathrm{V}(\mathrm{E}(T_n|X_i)) > 0$, by the central limit theorem, we have

$$\frac{1}{\sqrt{n\mathrm{V}(\mathrm{E}(T_n|X_i))}}(\widetilde{T}_n - \mathrm{E}(T_n)) \xrightarrow{\mathrm{d}} \mathrm{N}(0,1).$$

We also have an interesting relationship between the variances of T_n and $\widetilde{T}_n$, that is, $\mathrm{V}(\widetilde{T}_n) \leq \mathrm{V}(T_n)$, as the next theorem shows.

Theorem 1.65
If T_n is symmetric and $\mathrm{V}(T_n) < \infty$ for every n, and $\widetilde{T}_n$ is the projection of T_n onto $X_1, \ldots, X_n$, then

$$\mathrm{E}((T_n - \widetilde{T}_n)^2) = \mathrm{V}(T_n) - \mathrm{V}(\widetilde{T}_n).$$

Proof. Because $\mathrm{E}(T_n) = \mathrm{E}(\widetilde{T}_n)$, we have

$$\mathrm{E}((T_n - \widetilde{T}_n)^2) = \mathrm{V}(T_n) + \mathrm{V}(\widetilde{T}_n) - 2\mathrm{Cov}(T_n, \widetilde{T}_n). \tag{1.256}$$

But

$$\begin{aligned}
\mathrm{Cov}(T_n, \widetilde{T}_n) &= \mathrm{E}(T_n\widetilde{T}_n) - (\mathrm{E}(T_n))^2 \\
&= \mathrm{E}(T_n\mathrm{E}(T_n)) + \mathrm{E}\left(T_n \sum_{i=1}^{n} \mathrm{E}(T_n|X_i)\right) - n\mathrm{E}(T_n\mathrm{E}(T_n))) - (\mathrm{E}(T_n))^2 \\
&= n\mathrm{E}\left(T_n\mathrm{E}(T_n|X_i)\right) - n(\mathrm{E}(T_n))^2 \\
&= n\mathrm{V}(\mathrm{E}(T_n|X_i)) \\
&= \mathrm{V}(\widetilde{T}_n),
\end{aligned}$$

and the desired result follows from equation (1.256) above. ∎

The relevance of these facts, if we can show that $\widetilde{T}_n \to T_n$ in some appropriate way, then we can work out the asymptotic distribution of T_n. (The use of projections of U-statistics beginning on page 413 is an example.)

Partial Correlations

***fix

1.5.4 Conditional Probability and Probability Distributions

We now are in a position to define conditional probability. It is based on a conditional expectation.

Definition 1.48 (conditional probability given a sub-σ-field)
Let $(\Omega, \mathcal{F}, P)$ be a probability space, let $\mathcal{A}$ be a sub-σ-field of $\mathcal{F}$, and let $B \in \mathcal{F}$. The *conditional probability* of B given $\mathcal{A}$, denoted by $\Pr(B|\mathcal{A})$, is defined as $\mathrm{E}(\mathrm{I}_B|\mathcal{A})$. ∎

The concept of conditional probability given a sub-σ-field immediately yields the concepts of conditional probability given an event and conditional probability given a random variable. For a probability space $(\Omega, \mathcal{F}, P)$ with $A, B \in \mathcal{F}$, the *conditional probability* of B given A, denoted by $\Pr(B|A)$, is $\mathrm{E}(\mathrm{I}_B|\sigma(A))$.

Furthermore, if X is a random variable defined on $(\Omega, \mathcal{F}, P)$, the *conditional probability* of B given X, denoted by $\Pr(B|A)$, is $\mathrm{E}(\mathrm{I}_B|\sigma(X))$. This gives meaning to the concept of a conditional distribution of one random variable given another random variable.

Conditional Distributions

We start with a probability space $(\mathbb{R}^m, \mathcal{B}^m, P_1)$ and define a probability measure on the measurable space $(\mathbb{R}^n \times \mathbb{R}^m, \sigma(\mathcal{B}^n \times \mathcal{B}^m))$. We first need the existence of such a probability measure (proved in Billingsley (1995), page 439).

For a random variable Y in $\mathbb{R}^m$, its (marginal) distribution is determined by P_1, which we denote as $P_Y(y)$. For $B \in \mathcal{B}^n$ and $C \in \mathcal{B}^m$, the conditional distribution is defined by identifying a probability measure, denoted as $P_{X|Y}(\cdot|y)$, on $(\mathbb{R}^n, \sigma(\mathcal{B}^n))$ for any fixed $y \in \mathbb{R}^m$.

The joint probability measure of (X, Y) over $\mathbb{R}^n \times \mathbb{R}^m$ is defined as

$$P_{XY} = \int_C P_{X|Y}(\cdot|y)\mathrm{d}P_Y(y),$$

where $C \in \mathcal{B}^m$.

For distributions with PDFs the conditional, joint, and marginal PDFs have the simple relationship

$$f_{X|Y}(x|y) = \frac{f_{XY}(x, y)}{f_Y(y)},$$

so long as $f_Y(y) > 0$.

Independence

Theorem 1.66
The random variables X and Y are independent iff the conditional distribution equals the marginal distribution; that is, for the d-variate random variable X, iff

$$\forall A \subseteq \mathbb{R}^d, \Pr(X \in A|Y) = \Pr(X \in A).$$

Proof. Exercise. ∎

This theorem means that we can factor the joint PDF or CDF of independent random variables.

Conditional Independence

The ideas of conditional distributions also lead to the concept of conditional independence.

Definition 1.49 (conditional independence)
X and Y are conditionally independent given Z iff the joint conditional distribution equals the joint marginal distribution; that is, for the d-variate random variable X, iff

$$\forall A \subseteq \mathbb{R}^d, \Pr(X \in A|Y, Z) = \Pr(X \in A|Y).$$

∎

When two independent random variables are added to a third independent variable, the resulting sums are conditionally independent, given the third (common) random variable.

Theorem 1.67
Suppose the random variables X, Y, and Z are independent. Let $U = X + Z$ and $V = Y + Z$. Then $U|Z$ and $V|Z$ are independent; that is, U and V are conditionally independent given Z.

Proof. Exercise. ∎

Copulas

One of the most important uses of copulas is to combine two marginal distributions to form a joint distribution with known bivariate characteristics. We can build the joint distribution from a marginal and a conditional.

We begin with two $\mathrm{U}(0, 1)$ random variables U and V. For a given association between U and V specified by the copula $C(u, v)$, from Sklar's theorem, we can see that

$$P_{U|V}(u|v) = \frac{\partial}{\partial v} C(u, v)|_v. \tag{1.257}$$

We denote $\frac{\partial}{\partial v} C(u, v)|_v$ by $C_v(u)$.

Conditional Entropy

We define the conditional entropy of X given Y in two ways. The first meaning just follows the definition of entropy in equation (1.81) on page 42 with the conditional PDF $p_{X|Y}$ used in place of the marginal PDF p_X. This leads to an entropy that is a random variable or an entropy for a fixed value $Y = y$. In the more common usage, we define the conditional entropy of X given Y (which is also called the equivocation of X about Y) as the expected value of the term described above; that is,

$$H(X|Y) = -\sum_y p_Y(y) \sum_x p_{X|Y}(x|y) \log(p_{X|Y}(x|y)). \tag{1.258}$$

As before, the basic definition is made in terms of a PDF derived by a counting measure, but we extend it to any PDF.

From the definition we see that

$$H(X|Y) = H(X,Y) - H(Y) \tag{1.259}$$

or

$$H(X,Y) = H(X|Y) + H(Y).$$

Interpret $H(X,Y)$ as "total entropy", and compare the latter expression with equation (1.245).

1.6 Stochastic Processes

Many interesting statistical problems concern *stochastic processes*, which we can think of as a measurable function

$$X \ : \ \mathcal{I} \times \Omega \mapsto \mathbb{R}^d, \tag{1.260}$$

where $\mathcal{I}$ is some index set ($\mathcal{I}$ could be any ordered set).

In the expression above, X is a random variable, and for each $i \in \mathcal{I}$, X_i is a random variable. If the stochastic process is viewed as evolving in time, we usually denote the index by t and we may denote the process as $\{X_t\}$. In view of equation (1.260), it is also appropriate and common to use the notation $\{X(t,\omega)\}$.

The main interest in stochastic processes is the relationship among the distributions of $\{X_t\}$ for different values of t.

The sequences we discussed in Section 1.3 are of course stochastic processes. The sequences considered in that section did not have any particular structure, however. In some cases, we required that they have no structure; that is, that the elements in the sequence were independent. There are many special types of interesting stochastic processes with various structures, such as Markov chains, martingales, and other types of time series. In this section, we will just give some basic definitions, and then discuss briefly two important classes of stochastic process.

States, Times, Notation, and Basic Definitions

The smallest set of measure 1 is called the *state space* of a stochastic process; that is, the range of X is called the state space. Any point in the state space is called a *state*.

If the index set of a stochastic process is countable, we say the process is a *discrete time* stochastic process. We can index a discrete time process by $0, 1, 2, \ldots$, especially if there is a fixed starting point, although often $\ldots, -2, -1, 0, 1, 2, \ldots$ is more appropriate.

In many applications, however, the index of a stochastic process ranges over a continuous interval. In that case, we often use a slightly different notation for the index set. We often consider the index set to be the interval $[0, T]$, which of course could be transformed into any finite closed interval. If the index set is a real interval we say the process is a *continuous time* stochastic process. For continuous time stochastic processes, we sometimes use the notation $X(t)$, although we also use X_t. We will discuss continuous time processes in Section 1.6.2 below and consider a simple continuous time process in Example 1.32.

A property that seems to occur often in applications and, when it does, affords considerable simplifications for analyses is the conditional independence of the future on the past given the present. This property, called the *Markov property*, can be made precise.

Definition 1.50 (Markov property)
Suppose in the sequence $\{X_t\}$, for any set $t_0 < t_1 < \cdots < t_n < t$ and any x, we have

$$\Pr(X_t \leq x \mid X_{t_0}, X_{t_1}, \ldots, X_{t_n}) = \Pr(X_t \leq x \mid X_{t_n}). \tag{1.261}$$

Then $\{X_t\}$ is said to be a *Markov sequence* or the sequence is said to be *Markovian*. The condition expressed in equation (1.261) is called the *Markov property*. ∎

Definition 1.51 (homogeneous process)
If the marginal distribution of $X(t)$ is independent of t, the process is said to be *homogeneous*. ∎

***fix Many concepts are more easily defined for discrete time processes, although most have analogs for continuous time processes.

Definition 1.52 (stopping time)
Given a discrete time stochastic process ***fix change to continuous time

$$X : \{0, 1, 2, \ldots\} \times \Omega \mapsto \mathbb{R},$$

a random variable

$$T : \Omega \mapsto \{0, 1, 2, \ldots\} \tag{1.262}$$

is called a *stopping time* if the event $\{T = t\}$ depends only on $X_0, \ldots, X_t$ for $n = 0, 1, 2, \ldots$. ∎

Stopping times have several important characteristics, such as the fact that the Markov property holds at stopping times.

Definition 1.53 (first passage time)
A special stopping time is the *first passage time* defined (for discrete time processes) as

$$T_j = \min\{t \geq 1 : X_t = j\}, \tag{1.263}$$

if this set is nonempty; otherwise, $T_j = \infty$. ∎

There are other types of useful properties that simplify the analyses of processes with those properties. While a first passage time depends on the concept of a beginning point in the process, in the following we will usually allow the discrete time index to assume the values $\ldots, -2, -1, 0, 1, 2, \ldots$.

One of the most interesting properties of a stochastic process is the relationship of terms in the sequence to each other.

Definition 1.54 (autocovariance and autocorrelation)
For the process

$$\{X_t : t = \ldots, -2, -1, 0, 1, 2, \ldots\}$$

$\mathrm{E}(X_t) = \mu_t < \infty$, the function

$$\begin{aligned} \gamma(s,t) &= \mathrm{E}((X_s - \mu_s)(X_t - \mu_t)) \\ &= \mathrm{Cov}(X_s, X_t) \end{aligned} \tag{1.264}$$

if it is finite, is called the *autocovariance function*.

If the autocovariance exists, the function

$$\rho(s,t) = \frac{\gamma(s,t)}{\sqrt{\gamma(s,s)\gamma(t,t)}} \tag{1.265}$$

is called the *autocorrelation function*. The autocorrelation function, which is generally more useful than the autocovariance function, is also called the ACF. ∎

Definition 1.55 (white noise process)
A process

$$\{X_t : t = \ldots, -2, -1, 0, 1, 2, \ldots\}$$

with $\mathrm{E}(X_t) = 0 \quad \text{and} \quad \mathrm{V}(X_t) = \sigma^2 < \infty \; \forall \, t$ and such that $\rho(s,t) = 0 \; \forall \, s \neq t$ is called *white noise*. ∎

A zero-correlated process with constant finite mean and variance is also sometimes called white noise process even if the mean is nonzero. We denote a white noise process by $X_t \sim \mathrm{WN}(0, \sigma^2)$ or $X_t \sim \mathrm{WN}(\mu, \sigma^2)$.

Notice that the terms in a white noise process do not necessarily have identical distributions.

Definition 1.56 ((weakly) stationary process)
Suppose

$$\{X_t : t = \ldots, -2, -1, 0, 1, 2, \ldots\}$$

is such that $\mathrm{E}(X_t) = \mu$ and $\mathrm{V}(X_t) < \infty \; \forall \, t$ and $\gamma(s,t)$ is constant for any fixed value of $|s-t|$. Then the process $\{X_t\}$ is said to be *weakly stationary.* ∎

A white noise is clearly stationary.

In the case of a stationary process, the autocovariance function can be indexed by a single quantity, $h = |s-t|$, and we often write it as γ_h.

It is clear that in a stationary process, $\mathrm{V}(X_t) = \mathrm{V}(X_s)$; that is, the variance is also constant. The variance is γ_0 in the notation above.

Just because the means, variances, and autocovariances are constant, the distributions are not necessarily the same, so a stationary process is not necessarily homogeneous. Likewise, marginal distributions being equal does not insure that the autocovariances are constant, so a homogeneous process is not necessarily stationary.

The concept of stationarity can be made stricter.

Definition 1.57 (strictly stationary process)
Suppose

$$\{X_t : t = \ldots, -2, -1, 0, 1, 2, \ldots\}$$

is such that for any k, any set $t_1, \ldots, t_k$, and any h the joint distribution of

$$X_{t_1}, \ldots, X_{t_k}$$

is identical to the joint distribution of

$$X_{t_1+h}, \ldots, X_{t_k+h}.$$

Then the process $\{X_t\}$ is said to be *strictly stationary.* ∎

A strictly stationary process is stationary, but the converse statement does not necessarily hold. If the distribution of each X_t is normal, however, and if the process is stationary, then it is strictly stationary.

As noted above, a homogeneous process is not necessarily stationary. On the other hand, a strictly stationary process is homogeneous, as we see by choosing $k = 1$.

Example 1.31 a central limit theorem for a stationary process
Suppose $X_1, X_2, \ldots$ is a stationary process with $\mathrm{E}(X_t) = \mu$ and $\mathrm{V}(X_t) = \sigma^2$. We have

$$\begin{aligned} \mathrm{V}(\sqrt{n}(\overline{X} - \mu)) &= \sigma^2 + \frac{1}{n} \sum_{i \neq j = 1}^{n} \mathrm{Cov}(X_i, X_j) \\ &= \sigma^2 + \frac{2}{n} \sum_{h=1}^{n} (n-h)\gamma_h. \end{aligned} \tag{1.266}$$

(Exercise.) Now, if $\lim_{n\to\infty} \frac{2}{n}\sum_{h=1}^{n}(n-h)\gamma_h = \tau^2 < \infty$, then

$$\sqrt{n}(\overline{X} - \mu) \xrightarrow{\sim} \mathrm{N}(0, \sigma^2 + \tau^2).$$

∎

1.6.1 Probability Models for Stochastic Processes

A model for a stochastic process posits a sampling sequence over a sample space Ω. This yields a *path* or *trajectory*, $(\omega_1, \omega_2, \ldots)$. In continuous time we generally denote a path trajectory as $\omega(t)$. The sample space for the stochastic process becomes the set of paths. We denote this by $\Omega_{\mathcal{T}}$.

We think of a stochastic process in terms of a random variable, X_t, and an associated σ-field $\mathcal{F}_t$ in which X_t is measurable.

fix motivate and prove this ... a sequence of random variables $\{X_n\}$ for any n, the joint CDF of $X_1, \ldots, X_n$ is * uniqueness

Theorem 1.68 (Kolmogorov extension theorem)
For any positive integer k and any $t_1, \ldots, t_k \in \mathcal{T}$, let $P_{t_1,\ldots,t_k}$ be probability measures on $\mathbb{R}^{nt}$ such that for any Borel sets $B_1, \ldots, B_k \in \mathbb{R}^n$,

$$P_{\Pi(t_1),\ldots,\Pi(t_k)}(B_1 \times \cdots \times B_k) = P_{t_1,\ldots,t_k}(B_{\Pi^{-1}(1)} \times \cdots \times B_{\Pi^{-1}(k)}), \quad (1.267)$$

for all permutations Π on $\{1, \ldots, k\}$, and for all positive integers m,

$$P_{t_1,\ldots,t_k}(B_1 \times \cdots \times B_k) = P_{t_1,\ldots,t_k,t_{k+1},\ldots,t_{k+m}}(B_{\Pi^{-1}(1)} \times \cdots \times B_{\Pi^{-1}(k)} \times \mathbb{R}^n \cdots \times \mathbb{R}^n). \quad (1.268)$$

Then there exists a probability space $(\Omega, \mathcal{F}, P)$ and a stochastic process $\{X_t\}$ on Ω, $X_t : \Omega \mapsto \mathbb{R}^n$ such that

$$P_{t_1,\ldots,t_k}(B_1 \times \cdots \times B_k) = P(X_{t_1} \in B_1, \ldots, X_{t_k} \in B_k) \quad (1.269)$$

for all positive integers k, for all $t_i \in \mathcal{T}$, and for all Borel sets B_i.

Proof.
***fix ∎

Evolution of σ-Fields

In many applications, we assume an evolution of σ-fields, which, under the interpretation of a σ-field as containing all events of interest, is equivalent to an evolution of information. This leads to the concept of a filtration and a stochastic process adapted to the filtration.

Definition 1.58 (filtration; adaptation)
Let $\{(\Omega, \mathcal{F}_t, P)\}$ be a sequence of probability spaces such that if $s \le t$, then $\mathcal{F}_s \subseteq \mathcal{F}_t$. The sequence $\{\mathcal{F}_t\}$ is called a *filtration*.

For each t let X_t be a real function on Ω measurable wrt $\mathcal{F}_t$. The stochastic process $\{X_t\}$ is said to be *adapted to* the filtration $\{\mathcal{F}_t\}$. ∎

If $\{X_t\}$ is adapted to the filtration $\{\mathcal{F}_t\}$, we often write $X_t \in \mathcal{F}_t$. We also call the process $\{X_t\}$ *nonanticipating*, for obvious reasons.

Definition 1.59 (filtered probability space)
Given a probability space $(\Omega, \mathcal{F}, P)$ and a filtration $\{\mathcal{F}_t\}$ of sub-σ-fields of $\mathcal{F}$, we form the *filtered probability space* $(\Omega, \mathcal{F}, \{\mathcal{F}_t : t \in [0, \infty[\}, P)$. ∎

1.6.2 Continuous Time Processes

For a stochastic process over a continuous index set $\mathcal{I}$ we must be concerned about the continuity of the process in time. The problem arises because the countably-additive property of a measure (equation (0.1.8)) does not carry over to uncountable unions. For a process $X(t, \omega)$ where t is in uncountable index set, say, for example, an interval, we will be faced with the necessity to evaluate probabilities of sets of the form $\cup_{t \geq 0} A_t$. Such unions are not necessarily in the underlying σ-field.

** continuation motivation

We can define continuity of $X(t, \omega)$ on $\mathcal{I}$ in the usual way at a given point $\omega_0 \in \Omega$. Next, we consider continuity of a stochastic process over Ω.

Definition 1.60 (sample continuous)
Given a probability space $(\Omega, \mathcal{F}, P)$ and a function

$$X : \mathcal{I} \times \Omega \mapsto \mathbb{R},$$

we say X is *sample continuous* if $X(\omega) : \mathcal{I} \mapsto \mathbb{R}$ is continuous for almost all ω (with respect to P). ∎

The phrase *almost surely continuous*, or just *continuous*, is often used instead of sample continuous.

*** add more ... examples

The path of a stochastic process may be continuous, but many useful stochastic processes are mixtures of continuous distributions and discrete jumps. In such cases, in order to assign any reasonable value to the path at the point of discontinuity, we naturally assume that time is unidirectional and the discontinuity occurs at the time of the jump, and then the path evolves continuously from that point; that is, after the fact, the path is continuous from the right. The last value from the left is a limit of a continuous function. In French, we would describe this as *continu à droite, limité à gauche*; that is *cadlag*. Most models of stochastic processes are assumed to be cadlag.

1.6.3 Markov Chains

The simplest stochastic process is a sequence of exchangeable random variables; that is, a sequence with no structure. A simple structure can be imposed by substituting conditioning for independence. A sequence of random

variables with the Markov property is called a *Markov process*. A Markov process in which the state space is countable is called a *Markov chain*. (The term "Markov chain" is also sometimes used to refer to any Markov process, as in the phrase "Markov chain Monte Carlo", in applications of which the state space is often continuous.)

The theory of Markov chains is usually developed first for *discrete-time* chains, that is, those with a countable index set, and then extended to *continuous-time* chains.

If the state space is countable, it is equivalent to $\mathcal{X} = \{1, 2, \ldots\}$. If X is a random variable from some sample space to $\mathcal{X}$, and

$$\pi_i = \Pr(X = i), \tag{1.270}$$

then the vector $\pi = (\pi_1, \pi_2, \ldots)$ defines a distribution of X on $\mathcal{X}$. Formally, we define a Markov chain (of random variables) $X_0, X_1, \ldots$ in terms of an initial distribution π and a conditional distribution for X_{t+1} given X_t. Let X_0 have distribution π, and given $X_t = j$, let X_{t+1} have distribution $(p_{ij}; i \in \mathcal{X})$; that is, p_{ij} is the probability of a transition from state j at time t to state i at time $t + 1$, and $K = (p_{ij})$ is called the *transition matrix* of the chain. The initial distribution π and the transition matrix K characterize the chain, which we sometimes denote as $Markov(\pi, K)$. It is clear that K is a stochastic matrix, and hence $\rho(K) = \|K\|_\infty = 1$, and $(1, 1)$ is an eigenpair of K.

If K does not depend on the time (and our notation indicates that we are assuming this), the Markov chain is stationary.

A discrete-time Markov chain $\{X_t\}$ with discrete state space $\{x_1, x_2, \ldots\}$ can be characterized by the probabilities $p_{ij} = \Pr(X_{t+1} = x_i \mid X_t = x_j)$. Clearly, $\sum_{i \in \mathcal{I}} p_{ij} = 1$. A vector such as p_{*j} whose elements sum to 1 is called a *stochastic vector* or a distribution vector.

Because for each j, $\sum_{i \in \mathcal{I}} p_{ij} = 1$, K is a *right stochastic matrix*.

The properties of a Markov chain are determined by the properties of the transition matrix. Transition matrices have a number of special properties, which we discuss in Section 0.3.6, beginning on page 818.

(Note that many people who work with Markov chains define the transition matrix as the transpose of K above. This is not a good idea, because in applications with state vectors, the state vectors would naturally have to be row vectors. Until about the middle of the twentieth century, many mathematicians thought of vectors as row vectors; that is, a system of linear equations would be written as $xA = b$. Nowadays, almost all mathematicians think of vectors as column vectors in matrix algebra. Even in some of my previous writings, e.g., Gentle (2007), I have called the transpose of K the transition matrix, and I defined a stochastic matrix in terms of the transpose. The transpose of a right stochastic matrix is a left stochastic matrix, which is what is commonly meant by the unqualified phrase "stochastic matrix". I think that it is time to adopt a notation that is more consistent with current matrix/vector notation. This is merely a change in notation; no concepts require any change.)

If we assume that X_t is a random variable taking values in $\{x_1, x_2, \ldots\}$ and with a PDF (or probability mass function) given by

$$\Pr(X_t = x_i) = \pi_i^{(t)}, \tag{1.271}$$

and we write $\pi^{(t)} = (\pi_1^{(t)}, \pi_2^{(t)}, \ldots)$, then the PDF at time $t+1$ is

$$\pi^{(t+1)} = K\pi^{(t)}. \tag{1.272}$$

Many properties of a Markov chain depend on whether the transition matrix is reducible or not.

Because 1 is an eigenvalue and the vector 1 is the eigenvector associated with 1, from equation (0.3.70), we have

$$\lim_{t\to\infty} K^t = 1\pi_s, \tag{1.273}$$

where π_s is the Perron vector of K^{T}.

This also gives us the *limiting distribution* for an irreducible, primitive Markov chain,

$$\lim_{t\to\infty} \pi^{(t)} = \pi_s.$$

The Perron vector has the property $\pi_s = K^{\mathrm{T}}\pi_s$ of course, so this distribution is the *invariant distribution* of the chain.

The definition means that $(1, 1)$ is an eigenpair of any stochastic matrix. It is also clear that if K is a stochastic matrix, then $\|K\|_\infty = 1$, and because $\rho(K) \leq \|K\|$ for any norm and 1 is an eigenvalue of K, we have $\rho(K) = 1$.

A stochastic matrix may not be positive, and it may be reducible or irreducible. (Hence, $(1, 1)$ may not be the Perron root and Perron eigenvector.)

If the state space is countably infinite, the vectors and matrices have infinite order; that is, they have "infinite dimension". (Note that this use of "dimension" is different from our standard definition that is based on linear independence.)

We write the initial distribution as $\pi^{(0)}$. A distribution at time t can be expressed in terms of $\pi^{(0)}$ and K:

$$\pi^{(t)} = K^t\pi^{(0)}. \tag{1.274}$$

K^t is often called the *t-step transition matrix*.

The transition matrix determines various relationships among the states of a Markov chain. State i is said to be *accessible* from state j if it can be reached from state j in a finite number of steps. This is equivalent to $(K^t)_{ij} > 0$ for some t. If state i is *accessible* from state j and state j is *accessible* from state i, states i and j are said to *communicate*. Communication is clearly an equivalence relation. The set of all states that communicate with each other is an *equivalence class*. States belonging to different equivalence classes do not communicate, although a state in one class may be accessible from a state

in a different class. If all states in a Markov chain are in a single equivalence class, the chain is said to be *irreducible*.

The limiting behavior of the Markov chain is of interest. This of course can be analyzed in terms of $\lim_{t\to\infty} K^t$. Whether or not this limit exists depends on the properties of K.

Galton-Watson Process

An interesting class of Markov chains are branching processes, which model numbers of particles generated by existing particles. One of the simplest branching processes is the Galton-Watson process, in which at time t each particle is assumed to be replaced by $0, 1, 2, \ldots$ particles with probabilities $\pi_0, \pi_1, \pi_2, \ldots$, where $\pi_k \geq 0$, $\pi_0 + \pi_1 < 1$, and $\sum \pi_k = 1$. The replacements of all particles at any time t are independent of each other. The condition $\pi_0 + \pi_1 < 1$ prevents the process from being trivial.

*** add more

Continuous Time Markov Chains

In many cases it seems natural to allow the index of the Markov process to range over a continuous interval. The simplest type of continuous time Markov chain is a Poisson process.

Example 1.32 Poisson process
Consider a sequence of iid random variables, $Y_1, Y_2, \ldots$ distributed as exponential$(0, \theta)$, and build the random variables $T_k = \sum_{i=1}^{k} Y_i$. (The Y_is are the exponential spacings as in Example 1.18.)

*** prove Markov property

*** complete ∎

birth process

***add

$$K(t) = \mathrm{e}^{tR}$$

***fix R intensity rate. r_{ii} nonpositive, r_{ij} for $i \neq j$ nonnegative, $\sum_{i\in\mathcal{I}} r_{ij} = 0$ for all j.

1.6.4 Lévy Processes and Brownian Motion

Many applications of stochastic processes, such as models of stock prices, focus on the increments between two points in time. One of the most widely-used models makes three assumptions about these increments. These assumptions define a Lévy process.

Definition 1.61 (Lévy process)
Given a filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{0\leq t<\infty}, P)$. An adapted process $\{X(t) : t \in [0, \infty[\}$ with $X(0) \overset{\text{a.s.}}{=} 0$ is a *Lévy process* iff

(i) $X(t) - X(s)$ is independent of $\mathcal{F}_s$, for $0 \leq s < t < \infty$.
(ii) $X(t) - X(s) \overset{\mathrm{d}}{=} X(t-s)$ for $0 \leq s < t < \infty$.
(iii) $X(t) \xrightarrow{\mathrm{p}} X(s)$ as $t \to s$.

One of the most commonly used Lévy processes is *Brownian motion* also called a *Bachelier-Wiener process* (see Section 0.2.1 on page 766).

Definition 1.62 (Brownian motion)

(i) $X(t)$ is continuous in t almost surely.
(ii) $\mathrm{E}(X_t) \overset{\mathrm{a.s.}}{=} X(0)$.
(iii) $X(t) - X(s)$ for $0 \leq s < t$ has a normal distribution with variance $t - s$.

A Bachelier-Wiener process is also called a *Brownian motion*.
*** properties: covariance, etc.; existence, etc.

1.6.5 Brownian Bridges

*** definition, properties
Doob's transformation: If $\{Y(t)\}$ is a Brownian bridge and

$$X(t) = (1+t)Y(t/(1+t)) \quad \text{for } t \geq 0, \tag{1.275}$$

then $\{X(t)\}$ is a Brownian motion.

1.6.6 Martingales

Martingales are an important class of stochastic processes. The concept of conditional expectation is important in developing a theory of martingales. Martingales are special sequences of random variables that have applications in various processes that evolve over time.

Definition 1.63 (martingale, submartingale, supermartingale)
Let $\{\mathcal{F}_t\}$ be a filtration and let $\{X_t\}$ be adapted to the filtration $\{\mathcal{F}_t\}$. We say the sequence $\{(X_t, \mathcal{F}_t) : t \in \mathcal{T}\}$ is a *martingale* iff

$$\mathrm{E}(X_t | \mathcal{F}_{t-1}) \overset{\mathrm{a.s.}}{=} X_{t-1}. \tag{1.276}$$

We say the sequence $\{(X_t, \mathcal{F}_t) : t \in \mathcal{T}\}$ is a *submartingale* iff

$$\mathrm{E}(X_t | \mathcal{F}_{t-1}) \overset{\mathrm{a.s.}}{\geq} X_{t-1}. \tag{1.277}$$

We say the sequence $\{(X_t, \mathcal{F}_t) : t \in \mathcal{T}\}$ is a *supermartingale* iff

$$\mathrm{E}(X_t | \mathcal{F}_{t-1}) \overset{\mathrm{a.s.}}{\leq} X_{t-1}. \tag{1.278}$$

∎

We also refer to a sequence of random variables $\{X_t : t \in \mathcal{T}\}$ as a (sub, super)martingale if $\{(X_t, \sigma(\{X_s : s \leq t\})) : t \in \mathcal{T}\}$ is a (sub, super)martingale; that is, the martingale is the sequence $\{X_t\}$ instead of $\{(X_t, \mathcal{F}_t)\}$, and a corresponding sequence $\{\mathcal{F}_t\}$ is implicitly defined as $\{\sigma(\{X_s : s \leq t\})\}$.

This is consistent with the definition of $\{(X_t, \mathcal{F}_t) : t \in \mathcal{T}\}$ as a (sub, super)martingale because clearly

$$\sigma(\{X_s : s \leq r\})_r \subseteq \sigma(\{X_s : s \leq t\})_t \quad \text{if } r \leq t$$

(and so $\{\sigma(\{X_s : s \leq t\})_t\}$ is a filtration), and furthermore $\{X_t\}$ is adapted to the filtration $\{\sigma(\{X_s : s \leq t\})_t\}$.

We often refer to the type of (sub, super)martingale defined above as a *forward (sub, super)martingale*. We define a *reverse martingale* analogously with the conditions $\mathcal{F}_t \supset \mathcal{F}_{t+1} \supset \cdots$ and $\mathrm{E}(X_{t-1}|\mathcal{F}_t) \overset{\text{a.s.}}{=} X_t$.

The sequence of sub-σ-fields, which is a filtration, is integral to the definition of martingales. Given a sequence of random variables $\{X_t\}$, we may be interested in another sequence of random variables $\{Y_t\}$ that are related to the Xs. We say that $\{Y_t\}$ is a martingale with respect to $\{X_t\}$ if

$$\mathrm{E}(Y_t|\{X_\tau : \tau \leq s\}) \overset{\text{a.s.}}{=} Y_s, \ \forall s \leq t. \tag{1.279}$$

We also sometimes define martingales in terms of a more general sequence of σ-fields. We may say that $\{X_t : t \in \mathcal{T}\}$ is a martingale relative to the sequence of σ-fields $\{\mathcal{D}_t : t \in \mathcal{T}\}$ in some probability space $(\Omega, \mathcal{F}, P)$, if

$$X_s = \mathrm{E}(X_t|\mathcal{D}_t) \text{ for } s > t. \tag{1.280}$$

Submartingales and supermartingales relative to $\{\mathcal{D}_t : t \in \mathcal{T}\}$ may be defined analogously.

Example 1.33 Polya's urn process

Consider an urn that initially contains r red and b blue balls, and Polya's urn process (Example 1.6 on page 24). In this process, one ball is chosen randomly from the urn, and its color noted. The ball is then put back into the urn together with c balls of the same color. Let X_n be the number of red balls in the urn after n iterations of this procedure, and let $Y_n = X_n/(nc + r + b)$. Then the sequence $\{Y_n\}$ is a martingale (Exercise 1.82).

Interestingly, if $c > 0$, then $\{Y_n\}$ converges to the beta distribution with parameters r/c and b/c; see Freedman (1965). Freedman also discusses a variation on Polya's urn process called Friedman's urn process, which is the same as Polya's, except that at each draw in addition to the c balls of the same color being added to the urn, d balls of the opposite color are added to the urn. Remarkably, the behavior is radically different, and, in fact, if $c > 0$ and $d > 0$, then $Y_n \overset{\text{a.s.}}{\to} 1/2$. ∎

Example 1.34 likelihood ratios
Let f and g be probability densities. Let $X_1, X_2, \ldots$ be an iid sequence of random variables whose range is within the intersection of the domains of f and g. Let

$$Y_n = \prod_{i=1}^{n} g(X_i)/f(X_i). \tag{1.281}$$

(This is called a "likelihood ratio" and has applications in statistics. Note that $f(x)$ and $g(x)$ are likelihoods, as defined in equation (1.19) on page 20, although the "parameters" are the functions themselves.) Now suppose that f is the PDF of the X_i. Then $\{Y_n : n = 1, 2, 3, \ldots\}$ is a martingale with respect to $\{X_n : n = 1, 2, 3, \ldots\}$. ∎

The martingale in Example 1.34 has some remarkable properties. Robbins (1970) showed that for any $\epsilon > 1$,

$$\Pr(Y_n \geq \epsilon \text{ for some } n \geq 1) \leq 1/\epsilon. \tag{1.282}$$

Robbins's proof of (1.282) is straightforward. Let N be the first $n \geq 1$ such that $\prod_{i=1}^{n} g(X_i) \geq \epsilon \prod_{i=1}^{n} f(X_i)$, with $N = \infty$ if no such n occurs. Also, let $g_n(t) = \prod_{i=1}^{n} g(t_i)$ and $f_n(t) = \prod_{i=1}^{n} f(t_i)$.

$$\begin{aligned}
\Pr(Y_n \geq \epsilon \text{ for some } n \geq 1) &= \Pr(N < \infty) \\
&= \sum_{i=1}^{\infty} \int \mathrm{I}_{\{n\}}(N) f_n(t) \mathrm{d}t \\
&\leq \frac{1}{\epsilon} \sum_{i=1}^{\infty} \int \mathrm{I}_{\{n\}}(N) g_n(t) \mathrm{d}t \\
&\leq \frac{1}{\epsilon}.
\end{aligned}$$

Another important property of the martingale in Example 1.34 is

$$Y_n \xrightarrow{\text{a.s.}} 0. \tag{1.283}$$

You are asked to show this in Exercise 1.83.

Example 1.35 Bachelier-Wiener process
If $\{W(t) : t \in [0, \infty[\}$ is a Bachelier-Wiener process, then $W^2(t) - t$ is a martingale. (Exercise.) ∎

Example 1.36 A martingale that is not Markovian and a Markov process that is not a martingale
The Markov property is based on *conditional independence of distributions* and the martingale property is based on *equality of expectations*. Thus it is easy to construct a martingale that is not a Markov chain beginning with

X_0 has any given distribution with $\mathrm{V}(X_0) > 0$. The sequence $\{X_t : \mathrm{E}X_t = \mathrm{E}X_{t-1}, \mathrm{V}X_t = \sum_{k=0}^{t-1} \mathrm{V}X_k\}$ is not a Markov chain.

A Markov chain that is not a martingale, for example, is $\{X_t : X_t \stackrel{\mathrm{d}}{=} 2X_{t-1}\}$, where X_0 has any given distribution with $\mathrm{E}(X_0) \neq 0$. ∎

A common application of martingales is as a model for stock prices. As a concrete example, we can think of a random variable X_1 as an initial sum (say, of money), and a sequence of events in which $X_2, X_3, \ldots$ represents a sequence of sums with the property that each event is a "fair game"; that is, $\mathrm{E}(X_2|X_1) = X_1$ a.s., $\mathrm{E}(X_3|X_1, X_2) = X_2$ a.s., ... We can generalize this somewhat by letting $\mathcal{D}_n = \sigma(X_1, \ldots, X_n)$, and requiring that the sequence be such that $\mathrm{E}(X_n|\mathcal{D}_{n-1}) \stackrel{\text{a.s.}}{=} X_{n-1}$.

Doob's Martingale Inequality

A useful property of submartingales is Doob's martingale inequality. This inequality is a more general case of Kolmogorov's inequality (B.11), page 849, and the Hájek-Rènyi inequality (B.12), both of which involve partial sums that are martingales.

Theorem 1.69 (Doob's Martingale Inequality)
Let $\{X_t : t \in [0,T]\}$ be a submartingale relative to $\{\mathcal{D}_t : t \in [0,T]\}$ taking nonnegative real values; that is, $0 \leq X_s \leq \mathrm{E}(X_t|\mathcal{D}_t)$ for s, t. Then for any constant $\epsilon > 0$ and $p \geq 1$,

$$\Pr\left(\sup_{0\leq t\leq T} X_t \geq \epsilon\right) \leq \frac{1}{\epsilon^p}\mathrm{E}(|X_T|^p). \tag{1.284}$$

Proof. ***fix ∎

Notice that Doob's martingale inequality implies Robbins's likelihood ratio martingale inequality (1.282).

Azuma's Inequality

extension of Hoeffding's inequality (B.10), page 848

1.6.7 Empirical Processes and Limit Theorems

For a given random sample, the relationship of the ECDF F_n to the CDF F of the underlying distribution is of interest. At a given x, the normalized difference

$$G_n(x) = \sqrt{n}(F_n(x) - F(x)) \tag{1.285}$$

is called an *empirical process*. The convergence of this process will be studied below.

Martingale Central Limit Theorem

Most of the central limit theorems we discussed in Section 1.4.2 required identical distributions, and all required independence. Can we relax the independence assumption?

We focus on partial sums as in equation (1.223).

Theorem 1.70 (Martingale Central Limit Theorem)
Let

$$Y_n = \begin{cases} \sum_{j=1}^{k_n}(X_{nj} - \mathrm{E}(X_{nj})) & \text{if } n \leq k_n \\ \sum_{j=1}^{k_n}(X_{k_n j} - \mathrm{E}(X_{k_n j})) & \text{if } n > k_n. \end{cases} \tag{1.286}$$

Now, assume $\{Y_n\}$ is a martingale.

Next, starting with a fixed value for each subsequence, say $X_{n0} = 0$, assume the sum of the normalized conditional variances converge to 1:

$$\frac{1}{\sigma_n}\sum_{j=2}^{k_n} \mathrm{E}\left((X_{nj} - \mathrm{E}(X_{nj}))^2 | X_{n1}, \ldots, X_{n,j-1}\right) \xrightarrow{\mathrm{p}} 1,$$

where, as before, $\sigma_n^2 = \mathrm{V}(\sum_{j=1}^{k_n} X_{nj})$. Then we have

$$\frac{1}{\sigma_n}\sum_{j=1}^{k_n}(X_{nj} - \mathrm{E}(X_{nj})) \xrightarrow{\mathrm{d}} \mathrm{N}(0,1). \tag{1.287}$$

The addends in Y_n are called a triangular array as in the buildup to Lindeberg's Central Limit Theorem (see page 106), and the result (1.287) is the same as in Lindeberg's Central Limit Theorem on page 107.

Proof. ***fix ∎

Convergence of Empirical Processes

Although we may write the ECDF as F_n or $F_n(x)$, it is important to remember that it is a random variable. We may use the notation $F_n(x, \omega)$ to indicate that the ECDF is a random variable, yet to allow it to have an argument just as the CDF does. I will use this notation occasionally, but usually I will just write $F_n(x)$. The randomness comes in the definition of $F_n(x)$, which is based on the random sample.

The distribution of $nF_n(x)$ (at the fixed point x) is binomial, and so the pointwise properties of the ECDF are easy to see. From the SLLN, we see that it strongly converges pointwise to the CDF, and from the CLT, we have, at the point x,

$$\sqrt{n}(F_n(x) - F(x)) \xrightarrow{\mathrm{d}} \mathrm{N}\left(0, F(x)(1 - F(x))\right). \tag{1.288}$$

Although the pointwise properties of the ECDF are useful, its global relationship to the CDF is one of the most important properties of the ECDF. Our interest will be in the convergence of F_n, or more precisely, in the convergence of a metric on F_n and F. When we consider the convergence of metrics on functions, the arguments of the functions are sequences of random variables, yet the metric integrates out the argument.

An important property of empirical processes is a stochastic bound on its sup norm that is called the Dvoretzky/Kiefer/Wolfowitz (DKW) inequality, after the authors of the paper in which a form of it was given (Dvoretzky et al., 1956). This inequality provides a bound for the probability that the sup distance of the ECDF from the CDF exceeds a given value. Massart (1990) tightened the bound and gave a more useful form of the inequality. In one-dimension, for any positive z, the Dvoretzky/Kiefer/Wolfowitz/Massart inequality states

$$\Pr(\sup_x(\sqrt{n}|F_n(x,\omega) - F(x)|) > z) \leq 2\mathrm{e}^{-2z^2}. \tag{1.289}$$

This inequality is useful in proving various convergence results for the ECDF. For a proof of the inequality itself, see Massart (1990).

A particularly important fact regards the strong convergence of the sup distance of the ECDF from the CDF to zero; that is, the ECDF converges strongly and uniformly to the CDF. This is stated in the following theorem. The DKW inequality can be used to prove the theorem, but the proof below does not use it directly.

Theorem 1.71 (Glivenko-Cantelli) *If $X_1, \ldots, X_n$ are iid with CDF F and ECDF F_n, then* $\sup_x(|F_n(x,\omega) - F(x)|) \overset{\text{wp1}}{\to} 0$.

Proof. First, note by the SLLN and the binomial distribution of F_n, $\forall$ (fixed) x, $F_n(x,\omega) \overset{\text{wp1}}{\to} F(x)$; that is,

$$\lim_{n\to\infty} F_n(x,\omega) = F(x)$$

$\forall x$, except $x \in A_x$, where $\Pr(A_x) = 0$.

The problem here is that A_x depends on x and so there are uncountably many such sets. The probability of their union may possibly be positive. So we must be careful.

We will work on the CDF and ECDF from the other side of x (the discontinuous side). Again, by the SLLN, we have

$$\lim_{n\to\infty} F_n(x-,\omega) = F(x-)$$

$\forall x$, except $x \in B_x$, where $\Pr(B_x) = 0$.
Now, let

$$\phi(u) = \inf\{x\ ;\ u \leq F(x)\} \quad \text{for } 0 < u \leq 1.$$

(Notice $F(\phi(u)-) \leq u \leq F(\phi(u))$. Sketch the picture.)

Now consider $x_{m,k} = \phi(k/m)$ for positive integers m and k with $1 \leq k \leq m$. (There are countably many $x_{m,k}$, and so when we consider $F_n(x_{m,k}, \omega)$ and $F(x_{m,k})$, there are countably many null-probability sets, $A_{x_{m,k}}$ and $B_{x_{m,k}}$, where the functions differ in the limit.)

We immediately have the three relations:

$$F(x_{m,k}-) - F(x_{m,k-1}) \leq m^{-1}$$

$$F(x_{m,1}-) \leq m^{-1}$$

and

$$F(x_{m,m}) \geq 1 - m^{-1},$$

and, of course, F is nondecreasing.

Now let $D_{m,n}(\omega)$ be the maximum over all $k = 1, \ldots, m$ of

$$|F_n(x_{m,k}, \omega) - F(x_{m,k})|$$

and

$$|F_n(x_{m,k}-, \omega) - F(x_{m,k}-)|.$$

(Compare $D_n(\omega)$.)

We now consider three ranges for x:

$$\begin{array}{c}]-\infty\ ,\ x_{m,1}[\\ [x_{m,k-1}\ ,\ x_{m,k}[\text{ for } k = 1, \ldots, m \\ [x_{m,m}\ ,\ \infty[\end{array}$$

Consider $x \in [x_{m,k-1}, x_{m,k}[$. In this interval,

$$\begin{aligned} F_n(x, \omega) &\leq F_n(x_{m,k}-, \omega) \\ &\leq F(x_{m,k}-) + D_{m,n}(\omega) \\ &\leq F(x) + m^{-1} + D_{m,n}(\omega) \end{aligned}$$

and

$$\begin{aligned} F_n(x, \omega) &\geq F_n(x_{m,k-1}, \omega) \\ &\geq F(x_{m,k-1}) - D_{m,n}(\omega) \\ &\geq F(x) - m^{-1} - D_{m,n}(\omega) \end{aligned}$$

Hence, in these intervals, we have

$$\begin{aligned} D_{m,n}(\omega) + m^{-1} &\geq \sup_x |F_n(x, \omega) - F(x)| \\ &= D_n(\omega). \end{aligned}$$

We can get this same inequality in each of the other two intervals.

Now, $\forall m$, except on the unions over k of $A_{x_{m,k}}$ and $B_{x_{m,k}}$, $\lim_n D_{m,n}(\omega) = 0$, and so $\lim_n D_n(\omega) = 0$, except on a set of probability measure 0 (the countable unions of the $A_{x_{m,k}}$ and $B_{x_{m,k}}$.) Hence, we have the convergence wp1; i.e., a.s. convergence. ∎

The sup norm on the empirical process always exists because both functions are bounded. Other norms may not be finite.

Theorem 1.72 *If $X_1, \ldots, X_n$ are iid with CDF $F \in \mathcal{L}^1$ and ECDF F_n, then* $\|F_n(x,\omega) - F(x)\|_p \overset{\text{wp1}}{\to} 0$.

Proof. ******* use relationship between Lp norms ∎

********** add stuff Donsker's theorem

Notes and Further Reading

Probability theory is the most directly relevant mathematical background for mathematical statistics. Probability is a very large subfield of mathematics. The objective of this chapter is just to provide some of the most relevant material for statistics.

The CDF

I first want to emphasize how important the CDF is in probability and statistics.

This is also a good point to review the notation used in connection with functions relating to the CDF. The meaning of the notation in at least two cases (inverse and convolution) is slightly different from the usual meaning of that same notation in other contexts.

If we denote the CDF by F,

- $\overline{F}(x) = 1 - F(x)$;
- $F^{-1}(p) = \inf\{x, \text{ s.t. } F(x) \geq p\}$ for $p \in]0, 1[$;
- $F^{(2)}(x) = F \star F(x) = \int F(x-t)\mathrm{d}F(t)$;
- $F_n(x)$ is the ECDF of an iid sample of size n from distribution with CDF F.
- $f(x) = \mathrm{d}F(x)/\mathrm{d}x$.

Foundations

I began this chapter by expressing my opinion that probability theory is an area of pure mathematics: given a consistent axiomatic framework, "beliefs" are irrelevant. That attitude was maintained throughout the discussions in this chapter. Yet the literature on applications of probability theory is replete with interpretations of the meaning of "probability" by "frequentists", by "objectivists", and by "subjectivists", and discussions of the relative importance

of independence and exchangeability; see, for example, Hamaker (1977) and de Finetti (1979), just to cite two of the most eloquent (and opinionated) of the interlocutors. While the airing of some of these issues may just be furious sound, there are truly some foundational issues in application of probability theory to decisions made in everyday life. There are various to statistical inference that differ in fundamental ways (as alluded to by Hamaker (1977) and de Finetti (1979)) in whether or not prior "beliefs" or "subjective probabilities" are incorporated formally into the decision process.

While the phrase "subjective probability" is current, that concept does not fall within the scope ot this chapter, but it will be relevant in later chapters on statistical applications of probability theory.

There are, however, different ways of developing the concept of probability as a set measure that all lead to the same set of results discussed in this chapter. I will now briefly mention these alternatives.

Alternative Developments of a Probability Measure

Probability as a concept had been used by mathematicians and other scientists well before it was given a mathematical treatment. The first major attempt to provide a mathematical framework was Laplace's *Théorie Analytique des Probabilités* in 1812. More solid advances were made in the latter half of the 19th Century by Chebyshev, Markov, and Lyapunov at the University of St. Petersburg, but this work was not well known. (Lyapunov in 1892 gave a form of a central limit theorem. He developed this in the next few years into a central limit theorem similar to Lindeberg's, which appeared in 1920 in a form very similar to Theorem 1.58.) Despite these developments, von Mises (v. Mises) (1919a) said that "probability theory is not a mathematical science" (my translation), and set out to help to make it such. Indicating his ignorance of the work of both Lyapunov and Lindeberg, von Mises (v. Mises) (1919a) gives a more limited central limit theorem, but von Mises (v. Mises) (1919b) is a direct attempt to give a mathematical meaning to probability. In the "Grundlagen" he begins with a primitive concept of collective (or set), then defines probability as the limit of a frequency ratio, and formulates two postulates that essentially require invariance of the limit under any selections within the collective. This notion came to be called "statistical probability". Two years later, Keynes (1921) developed a concept of probability in terms of the relative support one statement leads to another statement. This idea was called "inductive probability". As Kolmogorov's axiomatic approach (see below) came to define probability theory and statistical inference for most mathematicians and statisticians, the disconnect between statistical probability and inductive probability continued to be of concern. Leblanc (1962) attempted to reconcile the two concepts, and his little book is recommended as a good, but somewhat overwrought, discussion of the issues.

Define probability as a special type of measure
We have developed the concept of probability by first defining a measurable space, then defining a measure, and finally defining a special measure as a probability measure.

Define probability by a set of axioms
Alternatively, the concept of probability over a given measurable space could be stated as axioms. In this approach, there would be four axioms: nonnegativity, additivity over disjoint sets, probability of 1 for the sample space, and equality of the limit of probabilities of a monotonic sequence of sets to the probability of the limit of the sets. The axiomatic development of probability theory is due to Kolmogorov in the 1920s and 1930s. In Kolmogorov (1956), he starts with a sample space and a collection of subsets and gives six axioms that characterize a probability space. (Four axioms are the same or similar to those above, and the other two characterize the collection of subsets as a σ-field.)

Define probability from a coherent ordering
Given a sample space Ω and a collection of subsets $\mathcal{A}$, we can define a total ordering on $\mathcal{A}$. (In some developments following this approach, $\mathcal{A}$ is required to be a σ-field; in other approaches, it is not.) The ordering consists of the relations "$\prec$", "$\preceq$", "$\sim$", "$\succ$", and "$\succeq$". The ordering is defined by five axioms it must satisfy. ("Five" depends on how you count, of course; in the five laid out below, which is the most common way the axioms are stated, some express multiple conditions.) For any sets, $A, A_i, B, B_i \in \mathcal{A}$ whose unions and intersections are in $\mathcal{A}$ (if $\mathcal{A}$ is a a σ-field this clause is unnecessary), the axioms are:

1. Exactly one of the following relations holds: $A \succ B$, $A \sim B$, or $A \prec B$.
2. Let A_1, A_2, B_1, B_2 be such that $A_1 \cap A_2 = \emptyset$, $B_1 \cap B_2 = \emptyset$, $A_1 \preceq B_1$, and $A_2 \preceq B_2$. Then $A_1 \cup A_2 \preceq B_1 \cup B_2$. Furthermore, if either $A_1 \prec B_1$ or $A_2 \prec B_2$, then $A_1 \cup A_2 \prec B_1 \cup B_2$.
3. $\emptyset \preceq A$ and $\emptyset \prec \Omega$.
4. If $A_1 \supseteq A_2 \supseteq \cdots$ and for each i, $A_i \succeq B$, then $\cap_i A_i \succeq B$.
5. Let $U \sim \mathrm{U}(0,1)$ and associate the ordering $(\prec, \preceq, \sim, \succ, \succeq)$ with Lebesgue measure on $[0,1]$. Then for any interval $I \subseteq [0,1]$, either $A \succ I$, $A \sim I$, or $A \prec I$.

These axioms define a linear, or total, ordering on $\mathcal{A}$. (Exercise 1.89).

Given these axioms for a "coherent" ordering on Ω, we can define a probability measure P on $\mathcal{A}$ by $P(A) \leq P(B)$ iff $A \preceq B$, and so on. It can be shown that such a measure exists, satisfies the Kolmogorov axioms, and is unique.

At first it was thought that the first 4 axioms were sufficient to define a probability measure that satisfied Kolmogorov's axioms, but Kraft et al. (1959) exhibited an example that showed that more was required.

A good exposition of this approach based on coherency that includes a proof of the existence and uniqueness of the probability measure is given by DeGroot (1970).

Define probability from expectations of random variables

Although the measurable spaces of Sections 1.1.1 and 0.1 (beginning on page 692) do not necessarily consist of real numbers, we defined real-valued functions (random variables) that are the basis of further development of probability theory. From the axioms characterizing probability (or equivalently from the definition of the concept of a probability measure), we developed expectation and various unifying objects such as distributions of random variables.

An alternate approach to developing a probability theory can begin with a sample space and random variables defined on it. (Recall our definition of random variables did not require a definition of probability.) From this beginning, we can base a development of probability theory on expectation, rather than on a probability measure as we have done in this chapter. (This would be somewhat similar to our development of conditional probability from conditional expectation in Section 1.5.)

In this approach we could define expectation in the usual way as an integral, or we can go even further and define it in terms of characterizing properties. We characterize an expectation operator E on a random variable X (and X_1 and X_2) by four axioms:

1. If $X \geq 0$, then $\mathrm{E}(X) \geq 0$.
2. If c is a constant in $\mathbb{R}$, then $\mathrm{E}(cX_1 + X_2) = c\mathrm{E}(X_1) + \mathrm{E}(X_2)$.
3. $\mathrm{E}(1) = 1$.
4. If a sequence of random variables $\{X_n\}$ increases monotonically to a limit $\{X\}$, then $\mathrm{E}(X) = \lim_{n\to\infty} \mathrm{E}(X_n)$.

(In these axioms, we have assumed a scalar-valued random variable, although with some modifications, we could have developed the axioms in terms of random variables in $\mathbb{R}^d$.) From these axioms, after defining the probability of a set as

$$\Pr(A) = \mathrm{E}(\mathrm{I}_A(\omega)),$$

we can develop the same probability theory as we did starting from a characterization of the probability measure. According to Hacking (1975), prior to about 1750 expectation was taken as a more basic concept than probability, and he suggests that it is more natural to develop probability theory from expectation. An interesting text that takes this approach is Whittle (2000).

We have already seen a similar approach. This was in our development of conditional probability. In order to develop an idea of conditional probability and conditional distributions, we began by defining conditional expectation with respect to a σ-field, and then defined conditional probability. While the most common kind of conditioning is with respect to a σ-field, a conditional

expectation with respect to a σ-lattice (see Definition 0.1.6 on page 695) can sometimes be useful; see Brunk (1963) and Brunk (1965).

Transformations of Random Variables

One of the most common steps in application of probability theory is to work out the distribution of some function of a random variable. The three methods mentioned in Section 1.1.10 are useful. Of those methods, the change-of-variables method in equation (1.124), which includes the convolution form, is probably the one that can be used most often. In that method, we use the Jacobian of the inverse transformation. Why the inverse transformation? Think of the density as a differential; that is, it has a factor $\mathrm{d}x$, so in the density for Y, we want a factor $\mathrm{d}y$. Under pressure you may forget exactly how this goes, or want a quick confirmation of the transformation. You should be able to construct a simple example quickly. An easy one is the right-triangular distribution; that is, the distribution with density $p_X(x) = 2x$, for $0 < x < 1$. Let $y = 2x$, so $x = \frac{1}{2}y$. Sketch the density of Y, and think of what transformations are necessary to get the expression $p_Y(y) = \frac{1}{2}y$, for $0 < y < 2$.

Structure of Random Variables

Most useful probability distributions involve scalar random variables. The extension to random vectors is generally straightforward, although the moments of vector random variables are quite different in structure from that of the random variable itself. Nevertheless, we find such random vectors as multivariate normal, Dirichlet, and multinomial very useful.

Copulas provide useful methods for relating the distribution of a multivariate random variable to the marginal distributions of its components and for understanding, or at least modeling, the relationships among the components of the random variable. Balakrishnan and Lai (2009) use copulas extensively in discussions of a large number of bivariate distributions, many of which are extensions of familiar univariate distributions. Nelson (2006) provides an extensive coverage of copulas.

The recent popularity of copulas in certain fields, such as finance, has probably led to some inappropriate use in probability models. See Mikosch (2006) and the discussion that follows his article.

Most approaches to multivariate statistical analysis are based on random vectors. There are some cases in which random matrices are useful. The most common family of random matrices is the Wishart, whose range is limited to symmetric nonnegative definite matrices. An obvious way to construct a random matrices is by an iid random sample of random variables. The random sample approach, of course, would not increase the structural complexity of the covariance. If instead of a random sample, however, the random matrix would be constructed from random vectors that are not iid, its covariance would have a complicated structure. Kollo and von Rosen (2005) use random matrices, rather than random vectors, as the basis of multivariate analysis.

Characteristic Functions

Characteristic functions play a major role in probability theory. Their use provides simple proofs for important facts, such as Geary's theorem. (The proof given for this theorem given on page 189 is based on one by Lukacs. The theorem, with the additional requirement that moments of all orders exist, was first proved by Geary (1936), using methods that are much more complicated.)

Gnedenko and Kolmogorov (1954) utilize characteristic functions throughout their development of limiting distributions. Lukacs (1970) provides a thorough exposition of characteristic functions and their various applications, including use of methods of differential equations in characteristic function theory, as in the proof of Geary's theorem.

The Problem of Moments

Thomas Stieltjes studied the problem of determining a nondecreasing function F, given a sequence of numbers $\nu_0, \nu_1, \nu_2, \ldots$ such that

$$\nu_k = \int_0^\infty x^k \mathrm{d}F(x), \quad k = 0, 1, 2, \ldots$$

Stieltjes called this the "moment problem". It is now often called the "Stieltjes problem", and the related problem with limits of integration 0 and 1 is called the "Hausdorff problem" and with limits $-\infty$ and ∞ is called the "Hamburger problem". These problems and the existence of a solution in each case are discussed by Shohat and Tamarkin (1943). (Although this is an older monograph, it is readily available in various reprinted editions.) In applications in probability, the existence of a solution is the question of whether there exists a probability distribution with a given set of moments.

After existence of a solution, the next question is whether the solution is unique, or in our formulation of the problem, whether the moments uniquely determine the probability distribution.

Many of the results concerning the moment problem involve probability distributions that are not widely used. Heyde (1963) was the first to show that a particular interesting distribution, namely the lognormal distribution, was not uniquely determined by its moments. (This is Exercise 1.28.)

Corollary 1.18.1 is due to Thomas Stieltjes who proved it without use of Theorem 1.18. Proofs and further discussion of the theorem and corollary can be found in Shohat and Tamarkin (1943).

Sequences and Limit Theorems

The various forms of the central limit theorem have a long history of both both the theory and the applications. Petrov (1995) provides an extensive

coverage. Dudley (1999) discusses many of the intricacies of the theorems and gives extensions of the theory.

The most important seminal result on the limiting distributions of extreme values was obtained by Fisher and Tippett (1928). von Mises (de Misès) (1939) and Gnedenko (1943) cleaned up some of the details, and Theorem 1.59 is essentially in the form stated in Gnedenko (1943). The limiting distributions of extreme values are discussed at some length by David and Nagaraja (2003), de Haan and Ferreira (2006), and Galambos (1978); and as mentioned in the text, a proof of Theorem 1.59, though not the same as given by Gnedenko is given by de Haan and Ferreira.

De Finetti's theorem allows the extension of certain results for independent sequences to similar results for exchangeable sequences. Taylor et al. (1985) prove a number of limit theorems for sums of exchangeable random variables.

A quote from the Preface of Gnedenko and Kolmogorov (1954) is appropriate:

> In the formal construction of a course in the theory of probability, limit theorems appear as a kind of superstructure over elementary chapters, in which all problems have finite purely arithmetical character. In reality, however, the epistemological value of the theory of probability is revealed only by limit theorems. Moreover, without limit theorems it is impossible to understand the real content of the primary concept of all our sciences — the concept of probability. In fact, all epistemologic value of the theory of probability is based on this: that large-scale random phenomena in their collective action create strict, nonrandom regularity. The very concept of mathematical *probability* would be fruitless if it did not find its realization in the *frequency* of occurrence of events under large-scale repetition of uniform conditions ....

Approximations and Expansions

The central limit theorems provide a basis for asymptotic approximations in terms of the normal distribution. Serfling (1980) and Bhattacharya and Ranga Rao (1976) discuss a number of approximations.

Other useful approximations are based on series representations of the PDF, CDF, or CF of the given distributions. Most of these series involve the normal PDF, CDF, or CF. Bhattacharya and Ranga Rao (1976) discuss a number of these series approximations. Hall (1992), especially Chapters 2 and 3, provides an extensive coverage of series expansions. Power series expansions are not now used as often in probability theory as they once were.

Cadlag Functions

We referred to the common assumption for models of stochastic processes that the functions are cadlag with respect to the time argument. The term cadlag

also applies to functions of arguments other than time that have this property. A CDF is a cadlag (or "càlàg") function. Analysis of continuous CDFs is relatively straightforward, but analysis of CDFs with discrete jumps is more challenging. Derivation and proofs of convergence properties of CDFs (such as may be encountered it central limits of stochastic processes) are sometimes difficult because of the discontinuities from the left. General results for a space of cadlag functions with a special metric have been developed. The space is called a Skorokhod space and the metric is called a Skorokhod metric. (Cadlag is synonymous with the English-derived acronyms RCCL, "right continuous [with] left limits", and corlol, "continuous on [the] right limit on [the] left", but there seems to be a preference for the French-derived acronym.)

Markov Chains

There are many other interesting properties of Markov chains that follow from various properties of nonnegative matrices (see Gentle (2007)). For more information on the properties of Markov chains, we refer the interested reader to the second edition of the classic text on Markov chains, Meyn and Tweedie (2009).

There are many special Markov chains that are motivated by applications. Branching process, for example, have applications in modeling such distinct areas biological populations and elementary particles. Harris (1989) developed many properties of such processes, and Athreya and Ney (1972) extended the theory. Some Markov chains are martingales, but of course not all are; conversely, not all martingales are Markov chains (see Example 1.36).

Martingales

Many of the basic ideas in martingale theory were developed by Joseph Doob, who gave this rather odd name to a class of stochastic processes after a type of betting system. Doob (1953) is still the classic text on stochastic processes generally and martingales in particular. Hall and Heyde (1980) cover many important limit theorems about martingales. Some of the most important applications of martingale theory are in financial modeling, in which a martingale model is equivalent to a no-arbitrage assumption. See Baxter and Rennie (1996) for applications of martingale theory in options pricing.

Empirical Processes

The standard texts on empirical processes are Shorack and Wellner (2009) and, especially for limit theorems relating to them, Dudley (1999).

Massart (1990) *** a tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality.

Additional References for Chapter 1

Among the references on probability in the bibliography beginning on page 873, some have not been mentioned specifically in the text. These include the general references on probability theory: Ash and Doleans-Dade (1999), Athreya and Lahiri (2006), Barndorff-Nielson and Cox (1994), Billingsley (1995), Breiman (1968), Chung (2000), Dudley (2002), Feller (1957) and Feller (1971), Gnedenko (1997), Gnedenko and Kolmogorov (1954), Gut (2005), and Pollard (2003).

The two books by Feller, which are quite different from each other but which together provide a rather comprehensive coverage of probability theory, the book by Breiman, and the book by Chung (the first edition of 1968) were among the main books I used in learning about probability years ago. They may be somewhat older than the others in the bibliography, but I'd probably still start out with them.

Exercises

1.1. For the measurable space $(\mathbb{R}, \mathcal{B})$, show that the collection of all open subsets of $\mathbb{R}$ is a determining class for probability measures on $(\mathbb{R}, \mathcal{B})$.

1.2. Prove Theorem 1.1.

1.3. For any theorem you should think carefully about the relevance of the hypotheses, and whenever appropriate consider the consequences of weakening the hypotheses. For the weakened hypotheses, you should construct a counterexample that shows the relevance of the omitted portion of the hypotheses. In Theorem 1.1, omit the condition that $\forall i \in \mathcal{I}, A, B \in \mathcal{C}_i \Rightarrow A \cap B \in \mathcal{C}_i$, and give a counterexample to show that without this, the hypothesis is not sufficient.

1.4. Prove Theorem 1.3.

1.5. Let $\Omega = \{1, 2, \ldots\}$ and let $\mathcal{F}$ be the collection of all subsets of Ω. Prove or disprove:

$$P(A) = \lim_{n\to\infty} \inf \frac{\#(A \cap \{1, \ldots, n\})}{n},$$

where $\#$ is the counting measure, is a probability measure on $(\Omega, \mathcal{F})$.

1.6. Let A, B, and C be independent events. Show that if D is any event in $\sigma(\{B, C\})$ then A and D are independent.

1.7. Prove Theorem 1.4.

1.8. Let X and Y be random variables. Prove that $\sigma(X) \subseteq \sigma(X, Y)$.

1.9. Given a random variable X defined on the probability space $(\Omega, \mathcal{F}, P)$, show that $P \circ X^{-1}$ is a probability measure.

1.10. Show that $X \overset{\text{a.s.}}{=} Y \Longrightarrow X \overset{\text{d}}{=} Y$.

1.11. Write out a proof of Theorem 1.6.

1.12. Let $F(x)$ be the Cantor function (0.1.30) (page 723) extended below the unit interval to be 0 and extended above the unit interval to be 1, as indicated in the text.

a) Show that $F(x)$ is a CDF.
b) Show that the distribution associated with this CDF does not have a PDF wrt Lebesgue measure. (What is the derivative?)
c) Does the distribution associated with this CDF have a PDF wrt counting measure?
d) Is the probability measure associated with this random variable dominated by Lebesgue measure? by the counting measure?

Let X be a random variable with this distribution.

e) What is $\Pr(X = 1/3)$?
f) What is $\Pr(X \leq 1/3)$?
g) What is $\Pr(1/3 \leq X \leq 2/3)$?

1.13. a) Show that F in equation (1.29) is a CDF.
b) Show that if each F_i in equation (1.29) is dominated by Lebesgue measure, then F is dominated by Lebesgue measure.
c) Show that if each F_i in equation (1.29) is dominated by the counting measure, then F is dominated by the counting measure.

1.14. Write out a proof of Theorem 1.8.

1.15. Write out a proof of Theorem 1.9.

1.16. Write out a proof of Theorem 1.10.

1.17. Write out a proof of Theorem 1.11.

1.18. Write out a proof of Theorem 1.12.

1.19. a) Show that the random variables R_1, R_2, R_3, R_4 in Example 1.6 are exchangeable.
b) Use induction to show that the sequence $R_1, R_2, \ldots$ in Example 1.6 is exchangeable.

1.20. Give an example in which the linearity of the expectation operator (equation (1.38)) breaks down.

1.21. Write out a proof of Theorem 1.13.

1.22. Show that if the scalar random variables X and Y are independent, then $\mathrm{Cov}(X, Y) = \mathrm{Cor}(X, Y) = 0$.

1.23. a) Let X be a random variable such that it is not the case that $X = \mathrm{E}(X)$ a.s. Prove $\mathrm{V}(X) > 0$.
b) Let $X = (X_1, \ldots, X_d)$ such that $\mathrm{V}(X_i) < \infty$, and assume that it is not the case that $X_i = \mathrm{E}(X_i)$ a.s. for any i nor that $\exists a_j, b_j$ for any element X_i of the vector X such that

$$X_i = \sum_{j \neq i} (a_j + b_j X_j) \quad \text{a.s.}$$

Prove that $\mathrm{V}(X)$ is full rank.

1.24. Show that the second raw moment $\mathrm{E}(X^2)$ for the Cauchy distribution (equation (1.37)) does not exist.

1.25. Expected values and quantile functions.
a) Write a formal proof of equation (1.44).
b) Extend equation (1.44) to $\mathrm{E}(X)$, where g is a bijective Borel function of X.

1.26. Expected values.
   a) Write a formal proof that equation (1.46) follows from the conditions stated.
   b) Write a formal proof that equation (1.48) follows from the conditions stated.
   c) Write a formal proof that equation (1.49) follows from the conditions stated.

1.27. Let X be a random variable. Show that the map

$$P_X : B \mapsto \Pr(X \in B),$$

where B is a Borel set, is a probability measure on the Borel σ-field.

1.28. Let X be a random variable with PDF

$$p_X(x) = \frac{1}{\sqrt{2\pi}} \frac{1}{x} \exp(-(\log(x))^2/2) \mathrm{I}_{\mathbb{R}_+}(x), \tag{1.290}$$

and let Y be a random variable with PDF

$$p_Y(y) = p_X(y)(1 + \alpha \sin(2\pi \log(y))) \mathrm{I}_{\mathbb{R}_+}(y),$$

where α is a constant and $0 < |\alpha| \leq 1$.
Notice the similarity of the PDF of Y to the PDF given in equation (1.53).
   a) Show that X and Y have different distributions.
   b) Show that for $r = 1, 2, \ldots$, $\mathrm{E}(X^r) = \mathrm{E}(Y^r)$.
   c) Notice that the PDF (1.290) is that of the lognormal distribution, which, of course, is an absolutely continuous distribution. Now consider the discrete random variable Y_a whose distribution, for given $a > 0$, is defined by

$$\Pr(Y_a = a\mathrm{e}^k) = c_a \mathrm{e}^{-k^2/2}/a^k, \quad \text{for } k = 0, 1, 2, \ldots,$$

   where c_a is an appropriate normalizing constant. Show that this discrete distribution has the same moments as the lognormal.
   Hint: First identify the support of this distribution. Then multiply the reciprocal of the r^{th} moment of the lognormal by the r^{th} of Y_a.

1.29. a) Prove equation (1.72):

$$\mathrm{V}(aX + Y) = a^2\mathrm{V}(X) + \mathrm{V}(Y) + 2a\mathrm{Cov}(X, Y).$$

   b) Prove equation (1.73):

$$\mathrm{Cor}(aX + Y, X + Z) = a\mathrm{V}(X) + a\mathrm{Cov}(X, Z) + \mathrm{Cov}(X, Y) + \mathrm{Cov}(Y, Z).$$

1.30. Prove the converse portion of Sklar's theorem (Theorem 1.19).

1.31. Let X and Y be random variables with (marginal) CDFs P_X and P_Y respectively, and suppose X and Y are connected by the copula C_{XY}. Prove:

$$\Pr(\max(X,Y) \le t) = C_{XY}(P_X(t), P_Y(t))$$

and

$$\Pr(\min(X,Y) \le t) = P_X(t) + P_Y(t) - C_{XY}(P_X(t), P_Y(t)).$$

1.32. a) Let X be a random variable that is normally distributed with mean μ and variance σ^2. Determine the entropy of X.
b) Let Y be a random variable that is distributed as beta(α, β). Determine the entropy of Y in terms of the digamma function (page 865). Make plots of the entropy for various values of α and β. An R function that evaluates the entropy is

```
entropy<-function(a,b){
   lbeta(a,b)-
     (a-1)*(digamma(a)-digamma(a+b))-
     (b-1)*(digamma(b)-digamma(a+b))
}
```

1.33. a) For the scalar random variable X prove equations (1.84) and (1.96).
b) For the random variable X prove equations (1.97) and (1.98).
c) Given the random variables X and Y with CF $\varphi_{X,Y}(t_1, t_2)$ write out $\mathrm{Cov}(X,Y)$ in terms of derivatives of the CF.

1.34. a) Show that the moment-generating function does not exist for a Cauchy distribution.
b) Determine the characteristic function for a Cauchy distribution, and show that it is not differentiable at 0.

1.35. In Example 1.10 we showed that the moment-generating function does not exist for a lognormal distribution. Determine the characteristic function for a lognormal distribution. Simplify the expression.

1.36. Consider the the distribution with PDF

$$p(x) = \frac{c}{2x^2 \log(|x|)} \mathrm{I}_{\{\pm 2, \pm 3, \ldots\}}(x).$$

Show that the characteristic function has a finite first derivative at 0, yet that the first moment does not exist (Zygmund, 1947).

1.37. Write an expression similar to equation (1.96) for the cumulants, if they exist, in terms of the cumulant-generating function.

1.38. Show that equations (1.99), (1.100), and (1.101) are correct.

1.39. Show that equation (1.107) is correct.

1.40. a) Let X and Y be iid N$(0,1)$. Work out the PDF of $(X-Y)^2/Y^2$.
b) Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ be iid N$(0,1)$. Work out the PDF of $\sum_i (X_i - Y_i)^2 / \sum_i Y_i^2$.

1.41. Show that the distributions of the random variables X and Y in Example 1.7 are the same as, respectively, the ratio of two standard exponential random variables and the ratio of two standard normal random variables.

1.42. Show that equation (1.132) is correct.

1.43. Stable distributions.

a) Show that an infinitely divisible family of distributions is stable.
b) Show that the converse of the previous statement is not true. (*Hint:* Show that the Poisson family is a family of distributions that is infinitely divisible, but not stable.)
c) Show that the definition of stability based on equation (1.136) is equivalent to Definition 1.33.
d) Let X, X_1, X_2 be as in Definition 1.33. Show that $Y = X_1 - X_2$ has a stable distribution, and show that the distribution of Y is symmetric about 0. (Y has a symmetric stable distribution).
e) Show that the normal family of distributions is stable with characteristic exponent of 2.
f) Show that the standard Cauchy distribution is stable with characteristic exponent of 1.

1.44. Prove Theorem 1.25.

1.45. Provide a heuristic justification for equation (1.138).

1.46. Show that the PDF of the joint distribution of all order statistic in equation (1.140) is equal to the PDF of the joint distribution of all of the (unordered) random variables, $\prod f(x_i)$.

1.47. Show that the Y_i in Example 1.19 on page 64 are independent of both $X_{(1)}$ and $X_{(n)}$.

1.48. a) Let $X_{(1)}, \ldots, X_{(n)}$ be the order statistics in a sample of size n, let $\mu_{(k:n)} = \mathrm{E}(X_{(k:n)})$, and let X be a random variable with the distribution of the sample. Show that $\mu_{(k:n)}$ exists and is finite if $\mathrm{E}(X)$ exists and is finite.
b) Let n be an odd integer, $n = 2k+1$, and consider a sample of size n from a Cauchy distribution with PDF $f_X = 1/(\pi(1+(x-\theta)^2))$. Show that the PDF of $X_{(k+1)}$, the sample median, is

$$f_{X_{(k+1)}} = \frac{n!}{(k!)^2\pi}\left(\frac{1}{4} - \frac{1}{\pi^2}(\arctan(x-\theta))^2\right)^k \frac{1}{1+(x-\theta)^2}.$$

What is $\mu_{(k:n)}$ in this case?

1.49. a) Prove equation (1.142).
b) Prove the following generalization of equation (1.142):

$$(n-k)\mathrm{E}\left(X^p_{(k:n)}\right) + k\mathrm{E}\left(X^p_{(k+1:n)}\right) = n\mathrm{E}\left(X^p_{(k:n-1)}\right).$$

See David and Nagaraja (2003).

1.50. Given the sequence of events $A_1, A_2, \ldots$, show that a tail event of $\{A_n\}$ occurs infinitely often.

1.51. Given the random variables $X_1, X_2, \ldots$ and X on a common probability space. For $m = 1, 2, \ldots$, and for any $\epsilon > 0$, let $A_{m,\epsilon}$ be the event that $\|X_m - X\| > \epsilon$. Show that almost sure convergence of $\{X_n\}$ to X is equivalent to

$$\lim_{n\to\infty} \Pr\left(\cup_{m=n}^{\infty} A_{m,\epsilon}\right) = 0,$$

for every $\epsilon > 0$.
Hint: For $j = 1, 2, \ldots$, consider the events

$$B_j = \cup_{n=1}^{\infty} \cap_{m=n}^{\infty} A^c_{m,1/j}.$$

1.52. Show that a convergence-determining class is a determining class.

1.53. a) Show that the collection of all finite open intervals in $\mathbb{R}$ that do not include 0 (as in Example 1.20) is a determining class for probability measures on $(\mathbb{R}, \mathcal{B})$. (Compare Exercise 1.1.)
b) Show that the collection of all finite open intervals in $\mathbb{R}$ is a convergence-determining class for probability measures on $(\mathbb{R}, \mathcal{B})$.

1.54. a) Give an example of a sequence of random variables that converges in probability to X, but does not converge to X a.s.
b) Give an example of a sequence of random variables that converges in probability to X, but does not converge to X in second moment.

1.55. Prove Theorem 1.31.

1.56. a) Weaken the hypothesis in Theorem 1.31 to $X_n \xrightarrow{\mathrm{d}} X$, and give a counterexample.
b) Under what condition does convergence in distribution imply convergence in probability?

1.57. Let $X_1, \ldots, X_n \overset{\mathrm{iid}}{\sim} \text{Bernoulli}(\pi)$. Let $Y_n = \sum_{i=1}^n X_i$. Show that as $n \to \infty$ and $\pi \to 0$ in such a way that $n\pi \to \theta > 0$, $Y_n \xrightarrow{\mathrm{d}} Z$ where Z has a Poisson distribution with parameter θ.

1.58. Given a sequence of scalar random variables $\{X_n\}$, prove that if

$$\mathrm{E}((X_n - c)^2) \to 0$$

then X_n converges in probability to c.

1.59. A sufficient condition for a sequence X_n of random variables to converge to 0 a.s. is that, for every $\epsilon > 0$, $\sum_{n=1}^{\infty} \Pr(|X_n| > \epsilon) < \infty$. Let U be uniformly distributed over $(0, 1)$ and define

$$X_n = \begin{cases} 1 \text{ if } U < \frac{1}{n} \\ 0 \text{ otherwise.} \end{cases}$$

Use this sequence to show that the condition $\sum_{n=1}^{\infty} \Pr(|X_n| > \epsilon) < \infty$ is not a necessary condition for the sequence X_n to converge to 0 a.s.

1.60. a) Show that if $X_n \xrightarrow{\mathrm{d}} X$ for any random variable X, then $X_n \in \mathrm{O_P}(1)$.
b) Show that if $X_n \in \mathrm{o_P}(1)$, then also $X_n \in \mathrm{O_P}(1)$.

1.61. Show the statements (1.165) through (1.174) (page 84) are correct.

1.62. Show the statements (1.175) and (1.176) are correct.

1.63. Show that the relationship in statement (1.177) is correct.

1.64. Show that equation (1.196) for the second-order delta method follows from Theorem 1.47 with $m = 2$.

1.65. a) Prove equation (1.206).

b) Show that the expression in equation (1.206) is a CDF.

1.66. Prove Theorem 1.49.

1.67. Show that Lindeberg's condition, equation (1.221), implies Feller's condition, equation (1.222).

1.68. The inequalities in Appendix B are stated in terms of unconditional probabilities and expectations. They all hold as well for conditional probabilities and expectations. In the following, assume the basic probability space $(\Omega, \mathcal{F}, P)$. Assume that $\mathcal{A}$ is a sub-σ-field of $\mathcal{F}$ and Z is a random variable in the given probability space. Prove:

a) An extension of Theorem B.3.1:
For $\epsilon > 0$, $k > 0$, and r.v. $X \ni \mathrm{E}(|X|^k)$ exists,

$$\Pr(|X| \geq \epsilon \,|\, \mathcal{A}) \leq \frac{1}{\epsilon^k} \mathrm{E}\left(|X|^k \,|\, \mathcal{A}\right).$$

b) An extension of Theorem B.4.1:
For f a convex function over the support of the r.v. X (and all expectations shown exist),

$$f(\mathrm{E}(X \,|\, Z)) \leq \mathrm{E}(f(X) \,|\, Z).$$

c) An extension of Corollary B.5.1.4:
If the second moments of X and Y are finite, then

$$\left(\mathrm{Cov}(X, Y \,|\, Z)\right)^2 \leq \mathrm{V}(X \,|\, Z)\, \mathrm{V}(Y \,|\, Z).$$

1.69. a) Show that the alternative conditions given in equations (1.232), (1.233), and (1.234) for defining conditional expectation are equivalent.

b) Show that equation (1.242) follows from equation (1.236).

1.70. Show that if $\mathrm{E}(X)$ exists, then so does $\mathrm{E}(X|B)$ for any event B such that $\Pr(B) \neq 0$.

1.71. Let X and Y be random variables over the same probability space. Show that $\sigma(X|Y) \subseteq \sigma(X)$. (Compare equation (0.1.7).)

1.72. Prove Theorem 1.60.

1.73. Modify your proof of Theorem 1.13 (Exercise 1.21) to prove Theorem 1.63.

1.74. Let T_n be a function of the iid random variables $X_1, \ldots, X_n$, with $\mathrm{V}(X_i) < \infty$, $\mathrm{V}(T_n) < \infty$, and $\mathrm{V}(\mathrm{E}(T_n|X_i)) > 0$. Now let $\widetilde{T}_n$ be the projection of T_n onto $X_1, \ldots, X_n$. Derive equations (1.254) and (1.255):

$$\mathrm{E}(\widetilde{T}_n) = \mathrm{E}(T_n),$$

and

$$\mathrm{V}(\widetilde{T}_n) = n\mathrm{V}(\mathrm{E}(T_n|X_i)).$$

1.75. Prove: The random variables X and Y are independent iff the conditional distribution of X given Y (or of Y given X) equals the marginal distribution of X (or of Y) (Theorem 1.66).

1.76. Prove Theorem 1.67.

1.77. Let X be a nonnegative integrable random variable on $(\Omega, \mathcal{F}, P)$ and let $\mathcal{A} \subseteq \mathcal{F}$ be a σ-field. Prove that $\mathrm{E}(X|\mathcal{A}) = \int_0^\infty \Pr(X > t|\mathcal{A})\,\mathrm{d}t$ a.s.

1.78. Show that equation (1.266) is correct.

1.79. Let $\{X_n : n = 1, 2, \ldots\}$ be a sequence of iid random variables with mean 0 and finite variance σ^2. Let $S_n \overset{\text{a.s.}}{=} X_1 + \cdots + X_n$, and let

$$Y_n \overset{\text{a.s.}}{=} S_n^2 - n\sigma^2.$$

Prove that $\{Y_n\}$ is a martingale with respect to $\{S_n : n = 1, 2, \ldots\}$.

1.80. Let $\{Z_i\}$ be an iid sequence of Bernoulli(π) random variables. Let $X_0 = 0$, and for $n = 1, 2, \ldots$, let

$$X_n \overset{\text{a.s.}}{=} X_{n-1} + 2Z_n - 1,$$

and let

$$Y_n \overset{\text{a.s.}}{=} ((1-\pi)/\pi)^{X_n}.$$

Show that $\{Y_n : n = 1, 2, 3, \ldots\}$ is a martingale with respect to $\{X_n : n = 1, 2, 3, \ldots\}$. (This is sometimes called de Moivre's martingale.)

1.81. Show that $\{X_n : n = 1, 2, 3, \ldots\}$ of equation (1.184) is a martingale.

1.82. Show that $\{Y_n : n = 1, 2, 3, \ldots\}$ of Polya's urn process (Example 1.33, page 131) is a martingale with respect to $\{X_n\}$.

1.83. Show that the likelihood-ratio martingale, equation (1.281), converges almost surely to 0.
Hint: Take logarithms and use Jensen's inequality and equation (1.63).

1.84. Show that if $\{W(t) : t \in [0, \infty[\}$ is a Bachelier-Wiener process, then $W^2(t) - t$ is a martingale.

1.85. Show that Doob's martingale inequality (1.284) implies Robbins's likelihood ratio martingale inequality (1.282).

1.86. Let $\{M_n\}$ be a martingale, and let $\{C_n\}$ be adapted to $\{\sigma(M_t : t \leq n)\}$. Let $\widetilde{M}_0 = 0$, and for $n \geq 1$, let

$$\widetilde{M}_n = \sum_{j=1}^{n} C_j(M_j - M_{j-1}).$$

Show that $\widetilde{M}_n$ is a martingale.
The sequence $\widetilde{M}_n$ is called the *martingale transform* of $\{M_n\}$ by $\{C_n\}$.

1.87. Let $X_1, X_2, \ldots$ be a sequence of independent random variables over a common probability space such that for each $\mathrm{E}(X_i^2) < \infty$. Show that the sequence of partial sums

$$Y_n = \sum_{i=1}^{n} (X_i - \mathrm{E}(X_i))$$

is a martingale.

1.88. Let X be a random variable that is normally distributed with mean μ and variance σ^2. Show that the entropy of X is at least as great as the entropy of any random variable with finite mean μ and finite variance σ^2 and having a PDF that is dominated by Lebesgue measure. (See Exercise 1.32a.)

1.89. Show that the axioms for coherency given on page 139 define a linear ordering, that is, a total ordering, on $\mathcal{A}$. (See page 620.)

2

Distribution Theory and Statistical Models

Given a measurable space, $(\Omega, \mathcal{F})$, different choices of a probability measure lead to different probability triples, $(\Omega, \mathcal{F}, P)$. A set of measures $\mathcal{P} = \{P\}$ associated with a fixed $(\Omega, \mathcal{F})$ is called a family of distributions. Families can be defined in various ways. For example, for Ω a real interval and $\mathcal{F} = \mathcal{B}_\Omega$, a very broad family is $\mathcal{P}_c = \{P : P \ll \nu\}$, where ν is the Lebesgue measure. An example of a very specific family for $\Omega = \{0, 1\}$ and $\mathcal{F} = 2^\Omega$ is the probability measure $P_\pi(\{1\}) = \pi$ and $P_\pi(\{0\}) = 1 - \pi$. The probability measures in this family, the Bernoulli distributions, are dominated by the counting measure.

Certain families of distributions have proven to be very useful as models of observable random processes. Familiar families include the normal or Gaussian family of distributions, the Poisson family of distributions, the binomial family of distributions, and so on. A list of some of the important families of distributions is given in Appendix A, beginning on page 835. Occasionally, as part of a parametric approach, transformations on the observations are used so that a standard distribution, such as the normal, models the phenomena better.

A semi-parametric approach uses broader families whose distribution functions can take on a much wider range of forms. In this approach, a differential equation may be developed to model a limiting case of some discrete frequency model. The Pearson system is an example of this approach (in which the basic differential equation arises as a limiting case of a hypergeometric distribution). Other broad families of distributional forms have been developed by Johnson, by Burr, and by Tukey. The objective is to be able to represent a wide range of distributional properties (mean, variance, skewness, shape, etc.) with a small number of parameters, and then to fit a specific case by proper choice of these parameters.

Statistical inference, which is the main topic of this book, can be thought of as a process whose purpose is to use observational data within the context of an *assumed* family of probability distributions $\mathcal{P}$ to *infer* that the observations are associated with a subfamily $\mathcal{P}_H \subseteq \mathcal{P}$, or else to decide that the assumed family is not an adequate model for the observed data. For example, we may

assume that the data-generating process giving rise to a particular set of data is in the Poisson family of distributions, and based on our methods of inference decide it is the Poisson distribution with $\theta = 5$. (See Appendix A for how θ parameterizes the Poisson family.)

A very basic distinction is the nature of the values the random variable assumes. If the set of values is countable, we call the distribution "discrete"; otherwise, we call it "continuous".

With a family of probability distributions is associated a random variable space whose properties depend on those of the family. For example, the random variable space associated with a location-scale family (defined below) is a linear space.

Discrete Distributions

The probability measures of discrete distributions are dominated by the counting measure.

One of the simplest types of discrete distribution is the discrete uniform. In this distribution, the random variable assumes one of m distinct values with probability $1/m$.

Another basic discrete distribution is the Bernoulli, in which random variable takes the value 1 with probability π and the value 0 with probability $1 - \pi$. There are two common distributions that arise from the Bernoulli: the binomial, which is the sum of n iid Bernoullis, and the negative binomial, which is the number of Bernoulli trials before r 1's are obtained. A special version of the negative binomial with $r = 1$ is called the geometric distribution. A generalization of the binomial to sums of multiple independent Bernoullis with different values of π is called the multinomial distribution.

The random variable in the Poisson distribution takes the number of events within a finite time interval that occur independently and with constant probability in any infinitesimal period of time.

A hypergeometric distribution models the number of selections of a certain type out of a given number of selections.

A logarithmic distribution (also called a logarithmic series distribution) models phenomena with probabilities that fall off logarithmically, such as first digits in decimal values representing physical measures.

Continuous Distributions

The probability measures of continuous distributions are dominated by the Lebesgue measure.

Continuous distributions may be categorized first of all by the nature of their support. The most common and a very general distribution with a finite interval as support is the beta distribution. Although we usually think of the support as $[0, 1]$, it can easily be scaled into any finite interval $[a, b]$. Two parameters determine the shape of the PDF. It can have a U shape, a J shape,

a backwards J shape, or a unimodal shape with the mode anywhere in the interval and with or without an inflection point on either side of the mode. A special case of the beta has a constant PDF over the support.

Another distribution with a finite interval as support is the von Mises distribution, which provides useful models of a random variable whose value is to be interpreted as an angle.

One of the most commonly-used distributions of all is the normal or Gaussian distribution. Its support is $]-\infty, \infty[$. There are a number of distributions, called stable distributions, that are similar to the normal. The normal has a multivariate extension that is one of the simplest multivariate distributions, in the sense that the second-order moments have intuitive interpretations. It is the prototypic member of an important class of multivariate distributions, the elliptically symmetric family.

From the normal distribution, several useful sampling distributions can be derived. These include the chi-squared, the t, the F, and the Wishart, which come from sample statistics from a normal distribution with zero mean. There are noncentral analogues of these that come from statistics that have not been centered on zero. Two other common distributions that are related to the normal are the lognormal and the inverse Gaussian distribution. These are related by applications. The lognormal is an exponentiated normal. (Recall two interesting properties of the lognormal: the moments do not determine the distribution (Exercise 1.28) and although the moments of all orders exist, the moment-generating function does not exist (Example 1.10).) The inverse Gaussian models the length of time required by Brownian motion to achieve a certain distance, while the normal distribution models the distance achieved in a certain time.

Two other distributions that relate to the normal are the inverted chi-squared and the inverted Wishart. They are useful because they are conjugate priors and they are also related to the reciprocal or inverse of statistics formed from samples from a normal. They are also called "inverse" distributions, but their origins are not at all related to that of the standard inverse Gaussian distribution.

Continuous Distributions with Point Masses

We can form a mixture of a continuous distribution and a discrete distribution. Such a distribution is said to be continuous with point masses. The probability measure is not dominated by either a counting measure or Lebesgue measure. A common example of such a distribution is the ϵ-mixture distribution, whose CDF is given in equation (2.45) on page 194.

Entropy

It is often of interest to know the entropy of a given probability distribution. In some applications we seek distributions with maximum or "large" entropy to

use as probability models. It is interesting to note that of all distributions with given first and second moments and having a PDF dominated by Lebesgue measure, the one with maximum entropy is the normal (Exercise 1.88).

Characterizing a Family of Distributions

A probability family or family of distributions, $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$, is a set of probability distributions of a random variable that is defined over a given sample space Ω. The index of the distributions may be just that, an arbitrary index in some given set Θ which may be uncountable, or it may be some specific point in a given set Θ in which the value of θ carries some descriptive information about the distribution; for example, θ may be a 2-vector in which one element is the mean of the distribution and the other element is the variance of the distribution.

The distribution functions corresponding to the members of most interesting families of distributions that we will discuss below do not constitute a distribution function space as defined on page 754. This is because mixtures of distributions in most interesting families of distributions are not members of the same family. That is, distributions defined by convex linear combinations of CDFs generally are not members of the same family of distributions. On the other hand, often linear combinations of random variables do have distributions in the same family of distributions as that of the individual random variables. (The sum of two normals is normal; but a mixture of two normals is not normal.) Table 1.1 on page 58 lists a number of families of distributions that are closed under addition of independent random variables.

Likelihood Functions

The problem of fundamental interest in statistics is to identify a particular distribution within some family of distributions, given observed values of the random variable. Hence, in statistics, we may think of θ or P_θ as a *variable*. A likelihood function is a function of that variable.

Definition 2.1 (likelihood function)
Given a PDF f_θ, which is a function whose argument is a value of a random variable x, we define a *likelihood function* as a function of θ for the fixed x:

$$L(\theta \,|\, x) = f_\theta(x).$$

∎

The PDF $f_\theta(x)$ is a function whose argument is a value of a random variable x for a fixed θ; the likelihood function $L(\theta \,|\, x)$ is a function of θ for a fixed x; see Figure 1.2 on page 20.

In statistical applications we may be faced with the problem of choosing between two distributions P_{θ_1} and P_{θ_2}. For a given value of x, we may base

our choice on the two likelihoods, $L(\theta_1 \mid x)$ and $L(\theta_2 \mid x)$, perhaps using the *likelihood ratio*

$$\lambda(\theta_1, \theta_2|x) = \frac{L(\theta_2 \mid x)}{L(\theta_1 \mid x)}.$$

We have seen in equation (1.65) that the expectation of the likelihood ratio, taken wrt to distribution in the denominator, is 1.

Parametric Families, Parameters, and Parameter Spaces

In Definition 1.13, we say that a family of distributions on a measurable space $(\Omega, \mathcal{F})$ with probability measures P_θ for $\theta \in \Theta$ is called a *parametric family* if $\Theta \subseteq \mathbb{R}^d$ for some fixed positive integer d and θ fully determines the measure. In that case, we call θ the *parameter* and Θ the parameter space.

A family that cannot be indexed in this way is called a nonparametric family. In nonparametric methods, our analysis usually results in some general description of the distribution, such as that the CDF is continuous or that the distribution has finite moments or is continuous, rather than in a specification of the distribution.

The type of a family of distributions depends on the parameters that characterize the distribution. A "parameter" is a real number that can take on more than one value within a parameter space. If the parameter space contains only one point, the corresponding quantity characterizing the distribution is not a parameter.

In most cases of interest the parameter space Θ is an open convex subset of $\mathbb{R}^d$. In the $\mathrm{N}(\mu, \sigma^2)$ family, for example, $\Theta = \mathbb{R} \times \mathbb{R}_+$. In the binomial$(n, \pi)$ family $\Theta =]0, 1[$ and n is usually not considered a "parameter" because in most applications it is assumed to be fixed and known. In many cases, for a family with PDF $p_\theta(x)$, the function $\partial p_\theta(x)/\partial\theta$ exists and is an important characteristic of the family (see page 168).

An example of a family of distributions whose parameter space is neither open nor convex is the hypergeometric(N, M, n) family (see page 838). In this family, as in the binomial, n is usually not considered a parameter because in most applications it is assumed known. Also, in most applications, either N or M is assumed known, but if they are both taken to be parameters, then $\Theta = \{(i, j) \ : \ i = 2, 3, \ldots, j = 1, \ldots, i\}$. Obviously, in the case of the hypergeometric family, the function $\partial p_\theta(x)/\partial\theta$ does not exist.

Many common families are multi-parameter, and specialized subfamilies are defined by special values of one or more parameters. As we have mentioned and illustrated, a certain parameter may be referred to as a "location parameter" because it identifies a point in the support that generally locates the support within the set of reals. A location parameter may be a boundary point of the support of it may be the mean or a median of the distribution. Another parameter may be referred to as a "scale parameter" because it is associated with scale transformations of the random variable. The standard deviation of a normal random variable, for example, is the scale parameter of

that distribution. Other parameters may also have some common interpretation, such as "shape". For example, in the "three-parameter gamma" family of distributions there are three parameters, γ, called the "location"; β, called the "scale"; and α, called the "shape". Its PDF is

$$(\Gamma(\alpha))^{-1}\beta^{-\alpha}(x-\gamma)^{\alpha-1}\mathrm{e}^{-(x-\gamma)/\beta}\mathrm{I}_{[\gamma,\infty[}(x).$$

This family of distributions is sometimes called the three-parameter gamma, because often γ is taken to be a fixed value, usually 0.

Specific values of the parameters determine special subfamilies of distributions. For example, in the three-parameter gamma, if α is fixed at 1, the resulting distribution is the two-parameter exponential, and if, additionally, γ is fixed at 0, the resulting distribution is what most people call an exponential distribution.

(Oddly, teachers of mathematical statistics many years ago chose the two-parameter exponential, with location and scale, to be "the exponential", and chose the two-parameter gamma, with shape and scale, to be "the gamma". The convenient result was that the exponential could be used as an example of a distribution that is not a member of the exponential class but is a member of the location-scale class, and the gamma could be used as an example of a distribution that is a member of the exponential class but is not a member of the location-scale class. This terminology is not nonstandard, and it seems somewhat odd to choose to include the location parameter in the definition of the exponential family of distributions and not in the definition of the more general gamma family of distributions. As noted, of course, it is just so we can have convenient examples of specific types of families of distributions.)

Notation in Parametric Families

I use notation of the form "$\mathrm{N}(\mu, \sigma^2)$" or "gamma$(\alpha, \beta, \gamma)$" to represent a parametric family. The notation for the parameters is positional, and follows the notations of the tables in Appendix A beginning on page 838. I generally use a Greek letter to represent a parameter. Sometimes a distribution depends on an additional quantity that is not a parameter in the usual sense of that term, and I use a Latin letter to represent such a quantity, as in "binomial(n, π)" for example. (The notation for the hypergeometric, with parameters N and M, one of which is usually assigned a fixed value, is an exception.)

As noted above, if a parameter is assigned a fixed value, then it ceases to be a parameter. If a parameter is fixed at some known value, I use a subscript to indicate that fact, for example "$\mathrm{N}(\mu, \sigma_0^2)$" may represent a normal distribution with variance known to be σ_0^2. In that case, σ_0^2 is not a parameter. (A word of caution, however: I may use subscripts to distinguish between two distributional families, for example, $\mathrm{N}(\mu_1, \sigma_1^2)$ and $\mathrm{N}(\mu_2, \sigma_2^2)$.)

Whether or not a particular characteristic of a distribution is a parameter is important in determining the class of a particular family. For example,

the three-parameter gamma is not a member of the exponential class (see Section 2.4); it is a member of the parametric-support class (see Section 2.5). The standard two-parameter gamma, however, with γ fixed at 0, is a member of the exponential class. If γ is fixed at any value γ_0, the gamma family is a member of the exponential class. Another example is the normal distribution, $\mathrm{N}(\mu, \sigma^2)$, which is a complete family (see Section 2.1); however, $\mathrm{N}(\mu_0, \sigma^2)$ is not a complete family.

Types of Families

A reason for identifying a family of distributions is so that we can state interesting properties that hold for all distributions within the family. The statements that specify the family are the hypotheses for important theorems. These statements may be very specific: "if $X_1, X_2, \ldots$ is a random sample from a *normal distribution*...", or they may be more general: "if $X_1, X_2, \ldots$ is a random sample from a *distribution with finite second moment*..."

Some simple characterization such as "having finite second moment" is easy to state each time its need arises, so there is little to be gained by defining such a class of distributions. On the other hand, if the characteristics are more complicated to state in each theorem that refers to that family of distributions, it is worthwhile giving a name to the set of characteristics.

Because in statistical applications we are faced with the problem of choosing the particular distributions P_{θ_0} from a family of distributions, $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, the behavior of the CDFs or PDFs as functions of θ are of interest. It may be important, for example, that the PDFs in this family be continuous with respect to θ or that derivatives of a specified order with respect to θ exist.

We identify certain collections of families of distributions for which we can derive general results. Although I would prefer to call such a collection a "class", most people call it a "family", and so I will too, at least sometimes. Calling these collections of families "families" leads to some confusion, because we can have a situation such as "exponential family" with two different meanings.

The most important class is the exponential class, or "exponential family". This family has a number of useful properties that identify optimal procedures for statistical inference, as we will see in later chapters.

Another important type of family of distributions is a group family, of which there are three important instances: a scale family, a location family, and a location-scale family.

There are various other types of families characterized by their shape or by other aspects useful in specific applications or that lead to optimal standard statistical procedures.

Parametric Modeling Considerations

In statistical applications we work with families of probability distributions that seem to correspond to observed frequency distributions of a data-generating process. The first considerations have to do with the nature of the observed measurements. The *structure* of the observation, just as the structure of a random variable, as discussed on page 37, is one of the most relevant properties. As a practical matter, however, we will emphasize the basic scalar structure, and attempt to model more complicated structures by imposition of relationships among the individual components. Another important property is whether or not the measurement is within the set of integers. Often the measurement is a count, such as the number of accidents in a given period of time or such as the number of defective products in a batch of a given size. A PDF dominated by a counting measure would be appropriate in such cases. On the other hand, if the measurement could in principle be arbitrarily close to any of an uncountable set of irrational numbers, then a PDF dominated by Lebesgue would be more appropriate.

The next consideration is the range of the measurements. This determines the support of a probability distribution used as a model. It is convenient to focus on three ranges, $]-\infty, \infty[$, $[a, \infty[$, and $[a, b]$ where $-\infty < a < b < \infty$. For integer-valued measurements within these three types of ranges, the Poisson family or the binomial family provide flexible models. Both the Poisson and the binomial are unimodal, and so we may need to consider other distributions. Often, however, mixtures of members of one of these families can model more complicated situations.

For continuous measurements over these three types of ranges, the normal family, the gamma family, and the beta family, respectively, provide flexible models. Mixtures of members of one of these families provide even more flexibility. The generality of the shapes of these distributions make them very useful for approximation of functions, and the most common series of orthogonal polynomials are based on them. (See Table 0.2 on page 752.)

2.1 Complete Families

A family of distributions $\mathcal{P}$ is said to be *complete* iff for any Borel function h that does not involve $P \in \mathcal{P}$

$$\mathrm{E}(h(X)) = 0 \; \forall P \in \mathcal{P} \quad \Longrightarrow \quad h(t) = 0 \text{ a.e. } \mathcal{P}.$$

A slightly weaker condition, "bounded completeness", is defined as above, but only for bounded Borel functions h.

Full rank exponential families are complete (exercise). The following example shows that a nonfull rank exponential family may not be complete.

Example 2.1 complete and incomplete family
Let

$$\mathcal{P}_1 = \{\text{distributions with densities of the form } (\sqrt{2\pi}\sigma)^{-1} \exp(x^2/(2\sigma^2))\}.$$

(This is the $\mathrm{N}(0, \sigma^2)$ family.) It is clear that $\mathrm{E}(h(X)) = 0$ for $h(x) = x$, yet clearly it is not the case that $h(t) = 0$ a.e. λ, where λ is Lebesgue measure. Hence, this family, the family of normals with known mean, is not complete. This example of course would apply to any symmetric distribution with known mean.

With some work, we can see that the family

$$\mathcal{P}_2 = \{\text{distributions with densities of the form } (\sqrt{2\pi}\sigma)^{-1} \exp((x-\mu)^2/(2\sigma^2))\}$$

is complete. ∎

Notice in the example that $\mathcal{P}_1 \subseteq \mathcal{P}_2$; and $\mathcal{P}_2$ is complete, but $\mathcal{P}_1$ is not. This is a common situation.

Going in the opposite direction, we have the following theorem.

Theorem 2.1
Let $\mathcal{P}_2$ be the family of distributions wrt which the expectation operator is defined and assume that $\mathcal{P}_2$ is complete. Now let $\mathcal{P}_2 \subseteq \mathcal{P}_1$, where all distributions in $\mathcal{P}_1$ have common support. Then the family $\mathcal{P}_1$ is complete.

Proof. Exercise. ∎

2.2 Shapes of the Probability Density

The general shape of a probability density may determine properties of statistical inference procedures. We can easily identify various aspects of a probability distribution that has a continuous density function. For discrete distributions, some of the concepts carry over in an intuitive fashion, and some do not apply.

In the following, we will assume that X is a random variable (or vector) with distribution in the family

$$\mathcal{P} = \{P_\theta : \theta \in \Theta \subseteq \mathbb{R}^k\}$$

that is dominated by a σ-finite measure ν, and we let

$$f_\theta(x) = \mathrm{d}P_\theta/\mathrm{d}\nu.$$

First of all, we consider the shape only as a function of the value of the random variable, that is, for a fixed member of the family the shape of the PDF.

In some cases, the shape characteristic that we consider has (simple) meaning only for random variables in $\mathbb{R}$.

Empirical Distributions and Kernels of Power Laws

Many important probability distributions that we identify and give names to arise out of observations on the behavior of some data-generating process. For example, in the process of forming coherent sentences on some specific topics there is an interesting data-generating process that yields the number of the most often used word, the number of the second most often used word and so on; that is, the observed data are $x_1, x_2, \ldots$, where x_i is the count of the word that occurs as the i^{th} most frequent. The linguist George Kingsley Zipf studied this data-generating process and observed a remarkable empirical relationship. In a given corpus of written documents, the second most commonly-used word occurs approximately one-half as often as the most common-used word, the third most commonly-used word occurs approximately one-third as often as the second most common-used word. (This general kind of relationship had been known before Zipf, but he studied it more extensively.) A probability-generating function that expresses this empirical relationship has the kernel

$$k(x) = x^{-\alpha}, \quad x = 1, 2, \ldots$$

where $\alpha > 1$.

The salient characteristic, which determines the shape of the PDF, is that the relative frequency is a function of the value raised to some power. This kind of situation is observed often, both in naturally occurring phenomena such as magnitudes of earthquakes or of solar flares, and in measures of human artifacts such as sizes of cities or of corporations. This is called a "power law". Use of the kernel above leads to a Zipf distribution, also called a zeta distribution because the partition function is the (real) zeta function, $\zeta(s) = \sum_{i=1}^{\infty} z^s$. (The Riemann zeta function is the analytic continuation of series, and obviously it is much more interesting than the real series.) The PDF, for $\alpha > 1$, is

$$f(x) = \frac{1}{\zeta(\alpha)} x^{-\alpha}, \quad x = 1, 2, \ldots$$

power law distribution Pareto distribution Benford distribution power function distribution (not power series distribution)

$$f(x) = c(\alpha, \theta) x^{-\alpha} \theta^{x-1}, \quad x = 1, 2, \ldots$$

Symmetric Family

A symmetric family is one for which for any given θ there is a constant τ that may depend on θ, such that

$$f_\theta(\tau + x) = f_\theta(\tau - x), \quad \forall x.$$

In this case, we say the distribution is symmetric about τ.

In a symmetric family, the third standardized moment, η_3, if it exists is 0; however, *skewness coefficient*. If $\eta_3 = 0$, the distribution is not necessarily symmetric.

The characteristic function of distribution that is symmetric about 0 is real, and any distribution whose characteristic function is real must have symmetries about 0 within the periods of the sine function (see equation (1.91) on page 46).

Unimodal Family

A family of distributions is said to be *unimodal* if for any given θ the mode of the distribution exists and is unique. This condition is sometimes referred to as *strictly unimodal*, and the term unimodal is used even with the mode of the distribution is not unique.

A family of distributions with Lebesgue PDF p is unimodal if for any given θ, $f_\theta(x)$ is strictly concave in x (exercise). This fact can be generalized to families with superharmonic Lebesgue PDFs (see Definition 0.0.14 on page 659).

Theorem 2.2
A probability distribution with a Lebesgue PDF that is superharmonic is unimodal.

Proof. Exercise. ∎

If the PDF is twice differentiable, by Theorem 0.0.15 unimodality can be characterized by the Laplacian. For densities that are not twice differentiable, negative curvature along the principal axes is sometimes called orthounimodality.

Logconcave Family

If $\log f_\theta(x)$ is strictly concave in x for any θ, the family is called a logconcave family. It is also called a *strongly unimodal family*. A strongly unimodal family is unimodal; that is, if $\log f_\theta(x)$ is concave in x, then $f_\theta(x)$ is unimodal (exercise). Strong unimodality is a special case of total positivity (see below).

The relevance of strong unimodality for location families, that is, for families in which $f_\theta(x) = g(x - \theta)$, is that the likelihood ratio is monotone in x (see below) iff the distribution is strongly unimodal for a fixed value of θ (exercise).

Heavy-tailed Family

A heavy-tailed family of probability distributions is one in which there is a relatively large probability in a region that includes $]-\infty, b[$ or $]b, \infty[$ for some finite b. This general characterization has various explicit instantiations, and

one finds in the literature various definitions of "heavy-tailed". A standard definition of that term is not important, but various specific cases are worth study. A heavy-tailed distribution is also called an *outlier-generating distribution*, and it is because of "outliers" that such distributions find interesting applications.

The concept of a heavy tail is equally applicable to the "left" or the "right" tail, or even a mixture in the case of a random variable over $\mathbb{R}^d$ when $d > 1$. We will, however, consider only the right tail; that is, a region $]b, \infty[$.

Most characterizations of heavy-tailed distributions can be stated in terms of the behavior of the tail CDF. It is informative to recall the relationship of the first moment of a positive-valued random variable in terms of the tail CDF (equation (1.46)):

$$\mathrm{E}(X) = \int_0^\infty \overline{F}(t)\mathrm{d}t.$$

If for some constant b, $x > b$ implies

$$f(x) > c\exp(-x^{\mathrm{T}}Ax), \tag{2.1}$$

where c is some positive constant and A is some positive definite matrix, the distribution with PDF f is said to be *heavy-tailed*.

Equivalent to the condition (2.1) in terms of the tail CDF is

$$\lim_{x\to\infty} \mathrm{e}^{a^{\mathrm{T}}x}\overline{F}(x) = \infty \quad \forall a > 0. \tag{2.2}$$

Another interesting condition in terms of the tail CDF that implies a heavy-tailed distribution is

$$\lim_{x\to\infty} \overline{F}(x+t) = \overline{F}(x). \tag{2.3}$$

Distributions with this condition are sometimes called "long-tailed" distributions because of the "flatness" of the tail in the left-hand support of the distribution. This condition states that $\overline{F}(\log(x))$ is a slowly varing function of x at ∞. (A function g is said to be *slowly varying* at ∞ if for any $a > 0$, $\lim_{x\to\infty} g(ax)/g(x) = 1$.)

Condition (2.3) implies condition (2.2), but the converse is not true (Exercise 2.3).

Most heavy-tailed distributions of interest are univariate or else product distributions. A common family of distributions that are heavy-tailed is the Cauchy family. Another common example is the Pareto family with $\gamma = 0$.

Subexponential Family

Another condition that makes a family of distributions heavy-tailed is

$$\lim_{x\to\infty} \frac{1 - F^{(2)}(x)}{1 - F(x)} = 2. \tag{2.4}$$

A family of distributions satisfying this condition is called a subexponential family (because the condition can be expressed as $\lim_{x\to\infty} \mathrm{e}^{-a^{\mathrm{T}}x}/\overline{F}(x) = 0$).

Condition (2.4) implies condition (2.3), but the converse is not true (Exercise 2.4).

Monotone Likelihood Ratio Family

The shape of parametric probability densities as a function of both the values of the random variable and the parameter may be important in statistical applications. Here and in the next section, we define some families based on the shape of the density over the cross product of the support and the parameter space. These characteristics are most easily expressed for the case of scalar parameters ($k = 1$), and they are also most useful in that case.

Let $y(x)$ be a scalar-valued function. The family $\mathcal{P}$ is said to have a *monotone likelihood ratio* iff for any $\theta_1 \neq \theta_2$, the likelihood ratio,

$$\lambda(\theta_1, \theta_2|x) = f_{\theta_2}(x)/f_{\theta_1}(x)$$

is a monotone function of x for all values of x for which $f_{\theta_1}(x)$ is positive.

We also say that the family has a *monotone likelihood ratio in* $y(x)$ iff the likelihood ratio is a monotone function of $y(x)$ for all values of x for which $f_{\theta_1}(x)$ is positive.

Some common distributions that have monotone likelihood ratios are shown in Table 2.1. See also Exercise 2.5.

Table 2.1. Some Common One-Parameter Families of Distributions with Monotone Likelihood Ratios

normal(μ, σ_0^2)
uniform(θ_0, θ), uniform$(\theta, \theta + \theta_0)$
exponential(θ) or exponential(α_0, θ)
double exponential(θ) or double exponential(μ_0, θ)
binomial(n, π) (n is assumed known)
Poisson(θ)

A subscript on a symbol for a parameter indicates that the symbol represents a known fixed quantity. See Appendix A for meanings of symbols.

Families with monotone likelihood ratios are of particular interest because they are easy to work with in testing composite hypotheses (see the discussion in Chapter 7 beginning on page 520).

The concept of a monotone likelihood ratio family can be extended to families of distributions with multivariate parameter spaces, but the applications in hypothesis testing are not as useful because we are usually interested in each element of the parameter separately.

Totally Positive Family

A totally positive family of distributions is defined in terms of the total positivity of the PDF, treating it as a function of two variables, θ and x. In this sense, a family is totally positive of order r iff for all $x_1 < \cdots < x_n$ and $\theta_1 < \cdots < \theta_n$,

$$\begin{vmatrix} f_{\theta_1}(x_1) & \cdots & f_{\theta_1}(x_n) \\ \vdots & \vdots & \vdots \\ f_{\theta_n}(x_1) & \cdots & f_{\theta_n}(x_n) \end{vmatrix} \geq 0 \quad \forall\, n = 1, \ldots, r. \tag{2.5}$$

A totally positive family with $r = 2$ is a monotone likelihood ratio family.

2.3 "Regular" Families

Conditions that characterize a set of objects for which a theorem applies are called "regularity conditions". I do not know the origin of this term, but it occurs in many areas of mathematics. In statistics there are a few sets of regularity conditions that define classes of interesting probability distributions.

We will often use the term "regularity conditions" to refer to continuity and differentiability of the PDF wrt the parameter.

2.3.1 The Fisher Information Regularity Conditions

The most important set of regularity conditions in statistics are some that allow us to put a lower bound on the variance of an unbiased estimator (see inequality (B.25) and Sections 3.1.3 and 5.1). Consider the family of distributions $\mathcal{P} = \{P_\theta; \theta \in \Theta\}$ that have densities f_θ.

There are generally three conditions that together are called the *Fisher information regularity conditions*:

- The parameter space $\Theta \subseteq \mathbb{R}^k$ is convex and contains an open set.
- For any x in the support and $\theta \in \Theta^\circ$, $\partial f_\theta(x)/\partial\theta$ and $\partial^2 f_\theta(x)/\partial\theta^2$ exist and are finite, and $\partial^2 f_\theta(x)/\partial\theta^2$ is continuous in θ.
- The support is independent of θ; that is, all P_θ have a common support.

The latter two conditions ensure that the operations of integration and differentiation can be interchanged twice.

Because the Fisher information regularity conditions are so important, the phrase "regularity conditions" is often taken to mean "Fisher information regularity conditions". The phrase "Fisher regularity conditions" is also used synonymously, as is "FI regularity conditions".

2.3.2 The Le Cam Regularity Conditions

The Le Cam regularity conditions are the first two of the usual FI regularity conditions plus the following.

- The Fisher information matrix (see equation (1.82)) is positive definite for any fixed $\theta \in \Theta$.
- There exists a positive number c_θ and a positive function h_θ such that $\mathrm{E}(h_\theta(X)) < \infty$ and

$$\sup_{\gamma:\|\gamma-\theta\|<c_\theta} \left\| \frac{\partial^2 \log f_\gamma(x)}{\partial\gamma(\partial\gamma)^{\mathrm{T}}} \right\|_{\mathrm{F}} \leq h_\theta(x) \text{ a.e.} \tag{2.6}$$

  where $f_\theta(x)$ is a PDF wrt a σ-finite measure, and "a.e." is taken wrt the same measure.

2.3.3 Quadratic Mean Differentiability

The Fisher information regularity conditions are often stronger than is needed to ensure certain useful properties. The double exponential distribution with Lebesgue PDF $\frac{1}{2\theta}\mathrm{e}^{-|y-\mu|/\theta}$, for example, has many properties that make it a useful model, yet it is not differentiable wrt μ at the point $x = \mu$, and so the FI regularity conditions do not hold. A slightly weaker regularity condition may be more useful.

Quadratic mean differentiability is expressed in terms of the square root of the density. As with differentiability generally, we first consider the property at one point, and then we apply the term to the function, or in this case, family, if the differentiability holds at all points in the domain.

Consider again a family of distributions $\mathcal{P} = \{P_\theta; \theta \in \Theta \subseteq \mathbb{R}^k\}$ that have densities f_θ. This family is said to be quadratic mean differentiable at θ_0 iif there exists a real k-vector function $\eta(x, \theta_0) = (\eta_1(x, \theta_0), \ldots, \eta_k(x, \theta_0))$ such that

$$***fix \int (*****)^2 \, \mathrm{d}x \in \mathrm{o}(|h|^2) \quad \text{as } |h| \to 0.$$

Compare quadratic mean differentiability with Fréchet differentiability (Definition 0.1.57, on page 760).

If each member of a family of distributions (specified by θ) is quadratic mean differentiable at θ, then the family is said to be quadratic mean differentiable, or QMD.

2.4 The Exponential Class of Families

The exponential class is a set of families of distributions that have some particularly useful properties for statistical inference. The important characteristic

of a family of distributions in the exponential class is the way in which the parameter and the value of the random variable can be separated in the density function. Another important characteristic of the exponential family is that the support of a distribution in this family does not depend on any "unknown" parameter.

Definition 2.2 (exponential class of families)
A member of a family of distributions in the exponential class is one with densities that can be written in the form

$$p_\theta(x) = \exp\left((\eta(\theta))^{\mathrm{T}} T(x) - \xi(\theta)\right) h(x), \tag{2.7}$$

where $\theta \in \Theta$, and where $T(x)$ is not constant in x. ∎

Notice that all members of a given family of distributions in the exponential class have the same support. Any restrictions on the range may depend on x through $h(x)$, but they cannot depend on the parameter.

Many of the common families of distributions used as probability models of data-generating processes are in the exponential class. In Table 2.2, I list some families of distributions in the exponential class.

Table 2.2. Some Common Families of Distributions in the Exponential Class

Discrete Distributions
binomial(n, π) (n is assumed known)
multinomial(n, π) (n is assumed known)
negative binomial(n, π) (n is assumed known)
Poisson(θ)
power series$(\theta, \{h_y\})$ ($\{h_y\}$ is assumed known)

Continuous Distributions
normal(μ, σ^2), normal(μ_0, σ^2), or normal(μ, σ_0^2)
log-normal(μ, σ^2), log-normal(μ_0, σ^2), or log-normal(μ, σ_0^2)
inverse Gaussian(μ, λ)
beta(α, β)
Dirichlet(α)
exponential(θ) or exponential(α_0, θ)
double exponential(θ) or double exponential(μ_0, θ)
gamma(α, β) or gamma$(\alpha, \beta, \gamma_0)$
gamma(α_0, β) (which includes the exponential)
gamma(α, β_0) (which includes the chi-squared)
inverted chi-squared(ν_0)
Weibull(α, β_0)
Pareto(α, γ_0)
logistic(μ, β)

A subscript on a symbol for a parameter indicates that the symbol represents a known fixed quantity. See Appendix A for meanings of symbols.

Note that the binomial, negative binomial, and Poisson families in Table 2.2 are all special cases of the general power series distributions whose PDF may be formed directly from equation (2.7); see page 175.

In Table 2.3, I list some common families of distributions that are not in the exponential class. Notice that some families listed in Table 2.2, such as Pareto(α, γ_0), for which some measure of the distribution is considered to be a fixed constant are no longer in the exponential class if we consider that fixed constant to be a parameter, as in the case of the two-parameter Pareto Pareto(α, γ).

Table 2.3. Some Common Families of Distributions Not in the Exponential Class

exponential(α, θ)
gamma(α, β, γ)
Weibull(α, β)
uniform(θ_1, θ_2)
double exponential(μ, θ)
Cauchy(γ, β), Cauchy(γ_0, β), or Cauchy(β)
Pareto(α, γ)
t(ν_0)

A family of distributions in the exponential class is called an exponential family, but do not confuse an "exponential family" in this sense with *the* "exponential family", that is, the parametric family with density of the form $\frac{1}{b}\mathrm{e}^{-x/b}\,\mathrm{I}_{[0,\infty[}(x)$. (This is the usual form of the exponential family, and it is a member of the exponential class. In courses in mathematical statistics, it is common to define the exponential family to be the two-parameter family with density $\frac{1}{b}\mathrm{e}^{-(x-a)/b}\,\mathrm{I}_{[a,\infty[}(x)$. This two-parameter form is not used very often, but it is popular in courses in mathematical statistics because this exponential family is not an exponential family(!) because of the range dependency.)

The form of the expression for the PDF depends on the σ-finite dominating measure that defines it. If the expression above results from

$$p_\theta = \frac{\mathrm{d}P_\theta}{\mathrm{d}\nu}$$

and we define a measure λ by $\lambda(A) = \int_A h\mathrm{d}\nu \;\forall A \in \mathcal{F}$, then we could write the PDF as

$$\frac{\mathrm{d}P_\theta}{\mathrm{d}\lambda} = \exp\left((\eta(\theta))^{\mathrm{T}}T(x) - \xi(\theta)\right). \tag{2.8}$$

Whatever the particular form of the PDF, an essential characteristic of it is the form of the decomposition as in equation (1.17). Formed from equation (2.7), this is

$$p_\theta(x) = \exp(-\xi(\theta))\exp\left((\eta(\theta))^{\mathrm{T}}T(x)\right)h(x); \tag{2.9}$$

that is, the kernel has the form $k(x) = \exp\left((\eta(\theta))^{\mathrm{T}} T(x)\right) h(x)$. The important thing to note is that the elements of the parameter vector $\eta(\theta) = (\eta(\theta)_1, \ldots, \eta(\theta)_k)$ appear in the kernel only in an exponential and as a linear combination of functions of x.

In the notation of equation (2.9), we see that the partition function is

$$\exp(\xi(\theta)) = \int \exp\left((\eta(\theta))^{\mathrm{T}} T(x)\right) h(x) \mathrm{d}x.$$

The form of the expression also depends on the *parametrization*; that is, the particular choice of the form of the parameters. First, notice that the only identifiable parameters must be in the elements of $\eta(\theta)$. The other function of the parameters, $\xi(\theta)$, which forms the partition, cannot introduce any more identifiable parameters; in fact, it can be written simply as

$$\xi(\theta) = \log\left(\int_{\mathcal{X}} \exp\left((\eta(\theta))^{\mathrm{T}} T(x)\right) h(x) \mathrm{d}x\right).$$

The expression

$$\mu = \mathrm{E}_\theta(T(x)) \tag{2.10}$$

is called the *mean-value parameter*, and use of μ for $\eta(\theta)$ is called the *mean-value parametrization*. We can develop an explicit expression for $\mathrm{E}_\theta(T(x))$ as

$$\mathrm{E}(T(X)) = \xi'(\theta)/\eta'(\theta).$$

(See Section 2.4.7.)

If a family of distributions has parameters α and β, we could equivalently say the family has parameters α and γ, where $\gamma = \alpha + \beta$; that is,

$$\begin{pmatrix} \alpha \\ \gamma \end{pmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}.$$

In this case of course we would have to replace $T(x) = (T_1(x), T_2(x))$

$$\widetilde{T}(x) = (T_1(x) - T_2(x), T_2(x)).$$

In fact, if $\eta(\theta) \in \mathbb{R}^d$, and D is any nonsingular $d \times d$ matrix, then with $\tilde{\eta} = D\eta(\theta)$, we can write an equivalent form of $(\eta(\theta))^{\mathrm{T}} T(x)$. To do so of course, we must transform $T(x)$ also. So $(\eta(\theta))^{\mathrm{T}} T(x) = \tilde{\eta}^{\mathrm{T}} \widetilde{T}(x)$, where $\widetilde{T}(x) = (D^{\mathrm{T}})^{-1} T(x)$.

A PDF of the form $f(x;\theta)\mathrm{I}(x;\theta)$ with respect to a σ-finite measure λ (where $\mathrm{I}(x;\theta)$ is an indicator function such that for some given x_0, $\exists \theta_1, \theta_2 \in \Theta \ni \mathrm{I}(x;\theta_1) = 0, \mathrm{I}(x;\theta_2) = 1$) cannot be put in the form $c\exp(g(x;\theta))h(x)$ because $c\exp(g(x;\theta)) > 0$ λ-a.e. (because the PDF must be bounded λ-a.e.).

2.4.1 The Natural Parameter Space of Exponential Families

In the expression for the density, it might be more natural to think of the parameter as η rather than θ; that way we would have an expression of form $\eta^{\mathrm{T}}T(x)$ rather than $(\eta(\theta))^{\mathrm{T}}T(x)$. We call the form

$$p_\theta(x) = \exp\left((\eta^{\mathrm{T}}T(x) - \zeta(\eta)\right)h(x) \tag{2.11}$$

the *canonical exponential form*, and we call

$$\mathrm{H} = \{\eta : \int \mathrm{e}^{\eta^{\mathrm{T}}T(x)}h(x)\mathrm{d}x < \infty\} \tag{2.12}$$

the *natural parameter space*. (Notice that H is the upper-case form of η.) The conditions in equation (2.12) are necessary to ensure that a $\zeta(\eta)$ exists such that $p_\theta(x)$ is a PDF. Another characterization of H is

$$\mathrm{H} = \{\eta : \eta = \eta(\theta), \theta \in \Theta\}$$

(under the assumption that Θ is properly defined, of course).

2.4.2 The Natural Exponential Families

An interesting subclass of exponential families is the class of exponential families in which $T(x)$ in the defining expression (2.7) is linear. This subclass is variously called the "natural exponential families", the "linear exponential families", or the "canonical exponential families".

Given a random variable X whose distribution is in any exponential family, the random variable $Y = T(X)$ has a distribution in the natural exponential family.

The cumulant-generating and probability-generating functions of natural exponential families have several simple properties (see Brown (1986) or Morris and Lock (2009)).

2.4.3 One-Parameter Exponential Families

An important subfamily of exponential families are those in which $\eta(\theta) \in \mathbb{R}$, that is, those whose parameter is a scalar (or effectively a scalar). This family is called a one-parameter exponential.

Theorem 2.3
Suppose a PDF $p(x|\theta)$ can be written as $\exp(g(x;\theta))h(x)$. where

$$g(x;\theta) = \eta(\theta)T(x) - \xi(\theta),$$

with $\eta(\theta) \in \mathbb{R}$, and Let x_1, x_2, x_3, x_4 be any values of x for which $p(x|\theta) > 0$. Then a necessary and sufficient condition that the distribution with the given PDF is in a one-parameter exponential family is that

$$\frac{g(x_1;\theta) - g(x_2;\theta)}{g(x_3;\theta) - g(x_4;\theta)} \tag{2.13}$$

is constant with respect to θ.

Proof.
It is clear from the definition that this condition is sufficient.

To show that it is necessary, first observe

$$g(x_i;\theta) - g(x_j;\theta) = \eta(\theta)(T(x_i) - T(x_j))$$

Now, for x_3 and x_4 such that $g(x_3;\theta) \neq g(x_4;\theta)$, we see that the ratio (2.13) must be constant in θ, because $\eta(\theta) \in \mathbb{R}$. ∎

An example of an application of Theorem 2.3 is to show that the one-parameter Cauchy family of distributions is not in the exponential class. (In these distributions the scale parameter $\beta = 1$.)

Example 2.2 the Cauchy family is not an exponential family
The PDF of the Cauchy is

$$\begin{aligned} p(x|\gamma) &= \frac{1}{\pi\left(1 + (x-\gamma)^2\right)} \\ &= \exp\left(-\log(\pi) - \log\left(1 + (x\gamma)^2\right)\right). \end{aligned}$$

Thus,

$$g(x;\theta) = -\log(\pi) - \log\left(1 + (x-\gamma)^2\right)$$

and for the four distinct points x_1, x_2, x_3, x_4,

$$\frac{g(x_1;\theta) - g(x_2;\theta)}{g(x_3;\theta) - g(x_4;\theta)} = \frac{-\log\left(1 + (x_1-\gamma)^2\right) + \log\left(1 + (x_2-\gamma)^2\right)}{-\log\left(1 + (x_3-\gamma)^2\right) + \log\left(1 + (x_4-\gamma)^2\right)}$$

is not constant in γ; hence the one-parameter Cauchy family of distributions is not in the exponential class. ∎

We often express the PDF for a member of a one-parameter exponential family as

$$p_\eta(x) = \beta(\eta)\mathrm{e}^{\eta T(x)} h(x). \tag{2.14}$$

In some cases if the support is $\mathbb{R}$, we can write the PDF as

$$p_\eta(x) = \beta(\eta)\mathrm{e}^{\eta T(x)}. \tag{2.15}$$

One-parameter exponential families are monotone likelihood ratio families (exercise), and have useful applications in statistical hypothesis testing.

2.4.4 Discrete Power Series Exponential Families

Various common discrete distributions can be formed directly from the general form of the PDF of one-parameter exponential families. In the notation of equation (2.14), let $\theta = \mathrm{e}^{\eta}$, and suppose $T(x) = x$. If $h(x)$ (or h_x) is such that

$$\sum_{x=0}^{\infty} h_x \theta^x = c(\theta) < \infty, \quad \text{for } \theta \in \Theta \subseteq \mathbb{R}_+,$$

we have the probability mass function

$$p_\theta(x) = \frac{h_x}{c(\theta)} \theta^x \mathrm{I}_{\{0,1,\ldots\}}(x), \tag{2.16}$$

where $\theta \in \Theta$. A family of distributions with PDFs of the form (2.16) is called a discrete power series family. Many of the common discrete families of distributions, such as the Poisson, the binomial, and the negative binomial, are of the power series class (Exercise 2.13).

2.4.5 Quadratic Variance Functions

An interesting class of exponential families are those whose variance is at most a quadratic function of its mean. For example, in the binomial distribution with parameters n and π, the mean is $\mu = n\pi$ and the variance is $n\pi(1-\pi)$. The variance as a function of μ is

$$n\pi(1-\pi) = -\mu^2/n + \mu = v(\mu)$$

As another example, in the normal distribution $\mathrm{N}(\mu, \sigma^2)$, the variance is at most a quadratic function of the mean because, in fact, it is constant with respect to the mean. In the Poisson distribution, the variance is a linear function of the mean; in the gamma with parameters α and β, we have $v(\mu) = \mu^2/\alpha$; and in the negative binomial distribution with parameters r and π, we have $v(\mu) = \mu^2/n + \mu$.

The normal, Poisson, gamma, binomial, negative binomial distributions, and one other family are in fact the only univariate natural exponential families with quadratic variance functions. (The other family is formed from hyperbolic secant distributions, and is not often used.) The quadratic variance property can be used to identify several other interesting properties, including infinite divisibility, cumulants, orthogonal polynomials, large deviations, and limits in distribution.

2.4.6 Full Rank and Curved Exponential Families

We say the exponential family is of *full rank* if the natural parameter space contains an open set.

An exponential family that is not of full rank may also be *degenerate*, meaning that there exists a vector a and a constant r such that

$$\int_{a^{\mathrm{T}}x=r} p_\theta(x)\mathrm{d}x = 1.$$

(The term "degenerate" in this sense is also applied to any distribution, whether in an exponential family or not.) The support of a degenerate distribution within $\mathbb{R}^d$ is effectively within $\mathbb{R}^k$ for $k < d$. An example of a nonfull rank exponential family that is also a degenerate family is the family of multinomial distributions (page 838). A continuous degenerate distribution is also called a singular distribution.

An example of a family of distributions that is a nonfull rank exponential family is the normal family $\mathrm{N}(\mu, \mu^2)$.

A nonfull rank exponential family is also called a *curved exponential family*.

2.4.7 Properties of Exponential Families

Exponential families have a number of useful properties. First of all, we note that an exponential family satisfies the Fisher information regularity conditions. This means that we can interchange the operations of differentiation and integration, a fact that we will exploit below. Other implications of the Fisher information regularity conditions allow us to derive optimal statistical inference procedures, a fact that we will exploit in later chapters.

In the following, we will use the usual form of the PDF,

$$f_\theta(x) = \exp(\eta(\theta)^{\mathrm{T}}T(x) - \xi(\theta))h(x),$$

and we will assume that it is of full rank.

We first of all differentiate both sides of the identity, wrt θ,

$$\int f_\theta(x)\,\mathrm{d}x = 1 \tag{2.17}$$

Carrying the differentiation on the left side under the integral, we have

$$\int \Big(\mathrm{J}_\eta(\theta)T(x) - \nabla\xi(\theta)\Big)\exp\big(\eta(\theta)T(x) - \xi(\theta)\big)h(x)\,\mathrm{d}x = 0.$$

Hence, because by assumption $\mathrm{J}_\eta(\theta)$ is of full rank, by rearranging terms under the integral and integrating out terms not involving x, we get the useful fact

$$\mathrm{E}(T(X)) = (\mathrm{J}_\eta(\theta))^{-1}\nabla\xi(\theta). \tag{2.18}$$

We now consider $\mathrm{E}(T(X))$. As it turns out, this is a much more difficult situation. Differentiation yields more complicated objects. (See Gentle (2007), page 152, for derivatives of a matrix wrt a vector.) Let us first consider the scalar case; that is, $\eta(\theta)$ and $T(x)$ are scalars, so $\eta(\theta)^{\mathrm{T}}T(x)$ is just $\eta(\theta)T(x)$.

In this case, differentiating a second time with respect to θ, we get

$$\int T(x)(\eta'(\theta)T(x)-\xi'(\theta))\exp(\eta(\theta)T(x)-\xi(\theta))h(x)\,\mathrm{d}x = \xi''(\theta)\eta'(\theta)/(\eta'(\theta))^2-\xi'(\theta)\eta''(\theta)/(\eta'(\theta))^2,$$

or

$$\eta'(\theta)\mathrm{E}((T(X))^2) - \xi'(\theta)\mathrm{E}(T(X)) = \xi''(\theta)/\eta'(\theta) - \xi'(\theta)\eta''(\theta)/(\eta'(\theta))^2$$

or

$$\mathrm{E}((T(X))^2) = (\xi'(\theta))^2/(\eta'(\theta))^2 - \xi''(\theta)/(\eta'(\theta))^2 - \xi'(\theta)\eta''(\theta)/(\eta'(\theta))^3.$$

Finally, collecting terms, we have

$$\begin{aligned}\mathrm{V}(T(X)) &= \mathrm{E}((T(X))^2) - (\mathrm{E}((T(X)))^2 \\ &= \frac{\xi''(\theta)}{(\eta'(\theta))^2} - \frac{\xi'(\theta)\eta''(\theta)}{(\eta'(\theta))^3}. \end{aligned} \tag{2.19}$$

********* proposition 3.2 in shao, page 171. look at two parameterizations; natural and mean.

$$\mathrm{V}(T(X)) = \mathrm{H}_\zeta(\eta) * 8,$$

where $\mathrm{H}_\zeta(\eta)$ is the matrix of second derivatives of ζ with respect to η.

It is often a simple matter to determine if a member of the exponential class of distributions is a monotone likelihood ratio family. If $\eta(\theta)$ and $T(x)$ in equation (2.7) for the PDF of a distribution in the exponential class are scalars, and if $\eta(\theta)$ is monotone in θ, then the family has a monotone likelihood ratio in $T(x)$.

2.5 Parametric-Support Families

Parametric-support families have simple range dependencies, that is, these are distributions whose supports depend on parameters. A distribution in any of these families has a PDF in the general form

$$p_\theta(x) = f(x,\theta)\mathrm{I}_{[f_1(\theta),f_2(\theta)]}(x). \tag{2.20}$$

These families are also called "truncation families", but most people use the term "truncated family" to refer to a family that is artificially truncated (for example, due to censoring; see Section 2.10.1). For example, to refer to the three-parameter gamma as a truncated distribution would be to confuse it with the more standard terminology in which a truncated gamma is the distribution formed from a two-parameter distribution with PDF

$$\frac{c}{\Gamma(\alpha)\beta^\alpha}x^{\alpha-1}\mathrm{e}^{-x/\beta}\mathrm{I}_{[\tau_1,\tau_2]}(x),$$

where c is just the normalizing constant, which is a function of α, β, τ_1, and τ_2. In applications, the truncation points, τ_1 and τ_2, are often known fixed values. If they are treated as parameters, of course, then the truncated distribution is a parametric-support family.

Parametric-support families, such as the family of two-parameter exponentials, are not exponential families; likewise, exponential families, such as the family of one-parameter exponentials, are not parametric-support families.

In some cases the parameters can be separated so some apply to the support and others are independent of the support. If the parameters are functionally independent, various properties of the distribution may be identified with respect to some parameters only. Also different statistical methods may be used for different parameters. For example, in the three-parameter gamma with PDF

$$\frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} \mathrm{e}^{-x/\beta} \mathrm{I}_{[\gamma,\infty[}(x),$$

some aspects of the distribution are those of a family in the exponential class, while other aspects can be related to a simple uniform distribution, $\mathrm{U}(0,\theta)$.

2.6 Transformation Group Families

"Group" families are distributions that have a certain invariance with respect to a group of transformations on the random variable. If g is a transformation within a group $\mathcal{G}$ of transformations (see Example 0.0.4 on page 630), and X is a random variable whose distribution is in the family $\mathcal{P}_\mathcal{G}$ and if the random variable $g(X)$ also has a distribution in the family $\mathcal{P}_\mathcal{G}$, the family $\mathcal{P}_\mathcal{G}$ is said to be invariant with respect to $\mathcal{G}$.

Transformations on the Sample Space and the Parameter Space

Let $\mathcal{G}$ be a group of transformations that map the probability space onto itself. For $g \in \mathcal{G}$ X and $g(X)$ are random variables that are based on the same underlying measure, so the probability spaces are the same; the transformation is a member of a transformation group, so the domain and the range are equal and the transformations are one-to-one.

$$g : \mathcal{X} \mapsto \mathcal{X}, \quad 1:1 \text{ and onto}$$

For given $g \in \mathcal{G}$ above, let $\tilde{g}$ be a 1:1 function that maps the parameter space onto itself, $\tilde{g} : \Theta \mapsto \Theta$, in such a way that for any set A,

$$\Pr_\theta(g(X) \in A) = \Pr_{\tilde{g}(\theta)}(X \in A). \tag{2.21}$$

If this is the case we say $\tilde{g}$ *preserves* Θ. Any two functions that preserve the parameter space form a group of functions that preserve the parameter space.

The set of all such $\tilde{g}$ together with the induced structure is a group, $\widetilde{\mathcal{G}}$. We write

$$\tilde{g}(\theta) = \tilde{\theta}. \tag{2.22}$$

$$\tilde{g} : \Theta \mapsto \Theta, \quad 1:1 \text{ and onto}$$

We may refer to $\widetilde{\mathcal{G}}$ as the *induced* group under $\mathcal{G}$. The group $\widetilde{\mathcal{G}}$ is *transitive* in the sense defined on page 755.

Example 2.3 Transformations in a binomial distribution
Suppose X is a random variable with a distribution in the binomial(n, π) family. In applications, the random variable is often taken as the sum of binary variables in which the 0 value is interpreted as one value of a binary state ("off-on", "good-bad", etc.). If the meaning of the binary state were changed, the binomial model for the application would remain unchanged. Instead of the original random variable, however, we would have $g(X) = n - X$. A further transformation $\tilde{g}(\pi) = 1 - \pi$ establishes the effect on the parameter space occasioned by the transformation on the probability space.

In the notation above G in $\mathcal{G} = (G, \circ)$ is given by

$$G = \{g(x) = x, g(x) = n - x \, : \, x \in \{0, \ldots, n\}\},$$

and $\widetilde{G}$ in $\widetilde{\mathcal{G}}$ is given by

$$\widetilde{G} = \{\tilde{g}(x) = x, \tilde{g}(x) = 1 - x \, : \, x \in]0, 1[\}.$$

It is easy to see that both $\mathcal{G}$ and $\widetilde{\mathcal{G}}$ are groups (exercise). ∎

Formation of Transformation Group Families

A group family can be formed from any family of distributions. (Notice that the preceding statement does not mean that any family is a group family; that depends on what variable parameters define the family.) The usual one-parameter exponential family of distributions, which is of the exponential class, is a transformation group family where the transformation is a scale (multiplicative) transformation, but is not a transformation group family where the transformation is a location and scale transformation. This family can be made into a location-scale group family by adding a location parameter. The resulting two-parameter exponential family is, of course, not of the exponential class.

The random variable space associated with a transformation group of probability distributions is closed with respect to that class of transformations.

2.6.1 Location-Scale Families

The most common group is the group of linear transformations, and this yields a location-scale group family, or just *location-scale family*, the general form of which is defined below.

Given a d-variate random variable X, a $d \times d$ positive-definite matrix Σ and a d-vector μ, it is clear that if $F(x)$ is the CDF associated with the random variable X, then $|\Sigma^{-1/2}|F(\Sigma^{-1/2}(x-\mu))$ is the CDF associated with Y. The class of all distributions characterized by CDFs that can be formed in this way is of interest.

Definition 2.3 (location-scale families)
Let X be a random variable on $\mathbb{R}^k$, let $\mathcal{V} \subseteq \mathbb{R}^k$, and let $\mathcal{M}_k$ be the collection of $k \times k$ symmetric positive definite matrices. The family of distributions of the random variables of the form

$$Y = \Sigma^{1/2}X + \mu, \text{ for } \mu \in \mathcal{V}, \Sigma \in \mathcal{M}_k \tag{2.23}$$

is called a *location-scale* family. The group of linear transformations $y = g(x)$ in equation (2.23) is also called the location-scale group. ∎

The random variable space associated with a location-scale family is a linear space.

If the PDF of a distribution in a location-scale family is $f(x)$, the PDF of any other distribution in that family is $|\Sigma^{-1/2}|f(\Sigma^{-1/2}(x-\mu))$. In the case of a scalar x, this simplifies to $f((x-\mu)/\sigma)/\sigma$. Thus, in a location-scale family the kernel of the PDF is invariant under linear transformations (see Definition 0.1.103 on page 755). The probability measure itself is invariant to the location transformation and equivariant to the scale transformation.

We often use

$$f((x-\mu)/\sigma)/\sigma \tag{2.24}$$

generically to represent the PDF of a distribution in a location-scale family.

While we can always form a location-scale family beginning with any distribution, our interest is in which of the usual families of distributions are location-scale families. Clearly, a location-scale family must have enough parameters and parameters of the right form in order for the location-scale transformation to result in a distribution in the same family. For example, a three-parameter gamma distribution is a location-scale family, but a two-parameter gamma (without the range dependency) is not.

In Table 2.4, I list some common distribution families in which we can identify a location parameter. While the usual form of the family has more than one parameter, if all but one of the parameters are considered to be fixed (that is, effectively, they are not parameters), the remaining parameter is a location parameter.

An interesting property of a location family is that the likelihood function is the same as the PDF. Figure 1.2 on page 20 illustrates the difference in a likelihood function and a corresponding PDF. In that case, the distribution family was exponential$(0, \theta)$, which of course is not a location family. A similar pair of plots for exponential(α, θ_0), which is a location family, would be identical to each other (for appropriate choices of α on the one hand and x on the other, of course).

Table 2.4. Some Common One-Parameter Location Group Families of Distributions

normal(μ, σ_0^2)
exponential(α, θ_0)
double exponential(μ, θ_0)
gamma$(\alpha_0, \beta_0, \gamma)$
Cauchy(γ, β_0)
logistic(μ, β_0)
uniform$(\theta - \theta_0, \theta + \theta_0)$

A subscript on a symbol for a parameter indicates that the symbol represents a known fixed quantity. See Appendix A for meanings of symbols.

Table 2.5. Some Common One-Parameter Scale Group Families of Distributions

normal(μ_0, σ^2)
inverse Gaussian(μ, λ)
exponential(α_0, θ)
double exponential(μ_0, θ)
gamma$(\alpha_0, \beta, \gamma_0)$
Cauchy(γ_0, β)
logistic(μ_0, β)
uniform$(\theta_0 - \theta, \theta_0 + \theta)$

A subscript on a symbol for a parameter indicates that the symbol represents a known fixed quantity. See Appendix A for meanings of symbols.

Often, a particular parameter in a parametric family can be identified as a "location parameter" or as a "scale parameter", and the location-scale transformation affects these two parameters in the obvious way. In some cases, however, a location transformation or a scale transformation alone affects more than one parameter. For example, a scale transformation σX on a random variable with distribution inverse Gaussian(μ, λ) results in a random variable with distribution inverse Gaussian$(\sigma\mu, \sigma\lambda)$. (There is an alternative parametrization of the inverse Gaussian with $\tilde{\lambda} = \lambda$ and $\tilde{\mu} = \sqrt{\lambda/\mu}$. In that notation, the scaling affects only the $\tilde{\lambda}$.)

Many of the common parametric families are both location and scale families; that is, they are location-scale group families. The families in both Tables 2.4 and 2.5 can be combined into two-parameter families that are location-scale group families. The normal(μ, σ^2), for example, is a location-scale group family.

Some standard parametric families that are not location-scale group families are the usual one-parameter exponential family, the binomial family, and the Poisson family.

Note that the only families of distributions that are in both the exponential class and the transformation group class are the normal, the inverse Gaussian, and the gamma.

We have used the term "scale" to refer to a positive (or positive definite) quantity. There are other group families formed by a larger transformation group than those of equation (2.23). The transformations

$$h(x) = a + Bx, \tag{2.25}$$

where B is a nonsingular matrix (but not necessarily positive definite), forms a transformation group, and the multivariate normal family $\mathrm{N}_d(\mu, \Sigma)$ is a group family with respect to this group of transformation. Notice that the transformations in the group $\mathcal{G} = \{h : h(x) = a + Bx,\ B \text{ nonsingular}\}$ can be formed by a smaller group, such as the same transformations in which B is a nonsingular lower triangular matrix (that is, one with nonnegative diagonal elements).

2.6.2 Invariant Parametric Families

Our interest in group families is often motivated by a certain symmetry in the sample space and the parameter space. That symmetry is expressed in the relationships between $\mathcal{G}$ the group of transformations that map the probability space onto itself and $\widetilde{\mathcal{G}}$ the induced group under $\mathcal{G}$.

Definition 2.4 (invariant class of families)
Let $\mathcal{P} = \{p_\theta : \theta \in \Theta\}$ be a parametric family of distributions on $(\mathcal{X}, \mathcal{B})$. Let X be a random variable with CDF $P_{\theta_0} \in \mathcal{P}$. Let $\mathcal{G}$ be a group of transformations on $\mathcal{X}$. If for each $g \in \mathcal{G}$ there is a 1:1 function $\tilde{g}$ on Θ such that the random variable $g(X)$ has CDF $P_{\tilde{g}(\theta_0)} \in \mathcal{P}$, then $\mathcal{P}$ is said to be an *invariant parametric family under* $\mathcal{G}$. ∎

Some common families of distributions that are in the invariant parametric class under the location-scale group with their regular parametrization include the normal, the double exponential, the exponential, the uniform (even with parametric ranges), and the Cauchy.

As suggested above, an invariant class under some transformation group can be generated by any distribution. This is not always possible for a specified group of transformations, however. For example, the (usual single-parameter) exponential family is not a member of a location invariant class.

Other Invariant Distributions

Given independent random variables X and Y, the distributions of X and Y may be such that there is a nontrivial transformation involving X and Y that yields a random variable that has the same distribution as X. that is, the distribution of X and $g(X, Y)$ is the same. For this to be the case,

clearly g must be a many-to-one map, so it is not an arbitrary member of a transformation group.

An important property of the $\mathrm{U}(0,1)$ distribution is the following. If a random variable is distributed as $\mathrm{U}(0,1)$ and is scaled or convolved with a random variable that has any uniform distribution whose range includes $[0,1]$, and then reduced modulo 1, the resulting random variable has a $\mathrm{U}(0,1)$ distribution. To state this more clearly, let $X \sim \mathrm{U}(0,1)$ and Y be distributed independently as $\mathrm{U}(a,b)$ where $a < 0$ and $b > 1$ and let $c > 1$. Now let

$$Z = (cX + Y) \mod 1.$$

Then $Z \overset{\mathrm{d}}{=} X$.

Another example of a distribution and transformation that is invariant (or nearly so) is the distribution of first digits and a positive scaling transformation. The digital representation of a random variable X is

$$X = D_1 b^{K-1} + D_2 b^{K-2} + D_3 b^{K-3} + \cdots$$

where $b \geq 3$ is a fixed integer, K is an integer-valued random variable, D_i is a nonnegative integer-valued random variable less than b, and $D_1 > 0$. If X has a uniform distribution over (b^{k-1}, b^k), the distribution of the first digit D_1 is not uniform over $\{1, \ldots, b-1\}$ as one might guess at first glance. With a larger range of X, remarkably, the distribution of D_1 is invariant to scale transformations of X (so long as the range of the scaled random variable includes a range of powers of b. (Note that the first digit in the digital representation of aX is not necessarily aD_1.)

A wellknown example of this type of distribution and transformation is known as "Benford's law". See Exercise 2.10.

2.7 Infinitely Divisible and Stable Families

The concept of divisibility was put forth in Definitions 1.31 and 1.32 on page 60. Distributions that are infinitely divisible are of most interest because they yield tractable models with a wide range of applications, especially in stochastic processes.

If $\{X_t : t \in [0, \infty[\}$ is a Lévy process, then any random variable $X(t)$ is infinitely divisible.

Stable Families

Stability of random variables was defined in Definition 1.33 on page 61.

Stable families are closely related to infinitely divisible families. All stable families are infinitely divisible, but an infinitely divisible family is not necessarily stable. The Poisson family is an example (exercise).

Most members of the stable family do not have PDFs or CDFs that can be represented in a closed form. The family is defined by a cumulant-generating function of the form

$$K(t) = \mathrm{i}\mu t - |\sigma t|^{\alpha}\left(1 - \mathrm{i}\beta\,\mathrm{sign}(t)\omega(\alpha, t)\right), \tag{2.26}$$

where $\omega(\alpha, t) = (2/\pi)\log(|t|)$ for $\alpha = 1$ and $t \neq 0$, and $\omega(\alpha, t) = \tan(\alpha\pi/2)$ for $\alpha \neq 1$. The parameters are analogous to parameters in other distributions that represent the mean, standard deviation, skewness, and kurtosis, although except in some special cases by the usual definitions of such measures of a distribution (based on expectations) these quantities do not exist or else are infinite in the stable family of distributions. The parameters are

- $\alpha \in]0, 2]$: stability coefficient of equation (1.137)
- $\beta \in [-1, 1]$: skewness parameter (not necessarily related to a third moment)
- $\sigma \in \mathbb{R}_+$: scale parameter (not necessarily related to a second moment)
- $\mu \in \mathbb{R}$: location parameter (not necessarily related to a first moment)

If $\beta = 0$ the distribution is symmetric (and note if also $\mu = 0$, the cumulant-generating function and hence, the characteristic function, is real).

The symmetric α-stable families, as α ranges from 1, which is the Cauchy distribution, to 2, which is the normal, has progressively lighter tails.

2.8 Families of Distributions with Heavy Tails

Exponential power family of distributions, also called the generalized error family of distributions.

Kernel

$$k(x) = \mathrm{e}^{-|x/\beta|^{\alpha}}$$

The Pareto distribution has relatively heavy tails; for some values of the parameter, the mean exists but the variance does not. A "Pareto-type" distribution is one whose distribution function satisfies the relationship

$$P(x) = 1 - x^{-\gamma}g(x),$$

where $g(x)$ is a *slowly varying function*; that is, for fixed $t > 0$,

$$\lim_{x\to\infty} \frac{g(tx)}{g(x)} = 1.$$

The Burr distribution with the CDF given in (2.54) is of the Pareto type, with $\gamma = \alpha\mathrm{B}$.

The stable family of distributions is a flexible family of generally heavy-tailed distributions. This family includes the normal distribution at one extreme value of one of the parameters and the Cauchy distribution at the other extreme value. There are various parameterizations of the stable distributions.

Depending on one of the parameters, α, the index of stability, the characteristic function (equation (**??**), page **??**) for random variables of this family of distributions has one of two forms:

$$\phi(t \mid \alpha, \sigma, \mathrm{B}, \mu) = \exp\Big(-\sigma^\alpha |t|^\alpha \big(1 - \mathrm{iBsign}(t)\tan(\pi\alpha/2)\big) + \mathrm{i}\mu t\Big) \quad \text{if } \alpha \neq 1,$$

or

$$\phi(t \mid 1, \sigma, \mathrm{B}, \mu) = \exp\Big(-\sigma|t|\big(1 + 2\mathrm{iBsign}(t)\log(t)/\pi\big) + \mathrm{i}\mu t\Big) \quad \text{if } \alpha = 1$$

for $0 < \alpha \leq 2$, $0 \leq \sigma$, and $-1 \leq \mathrm{B} \leq 1$. For $\alpha = 2$, this is the normal distribution (in which case B is irrelevant), and for $\alpha = 1$ and $\mathrm{B} = 0$, this is the Cauchy distribution.

The member of the stable family with $\alpha = 1$ and $\mathrm{B} = 1$ is called the Landau distribution, which has applications in modeling fluctuation of energy loss in a system of charged particles.

2.9 The Family of Normal Distributions

The normal distribution is probably the most important probability distribution. The normal family is in the exponential class. It is a complete family, a regular family, a group family, an infinitely divisible family, a stable family, and an elliptical family. One reason that the family of normal distributions is so important in statistical applications is the central limit theorem that gives the normal as the limiting distribution of properly normalized sequences of random variables with other distributions.

The family of normal distributions has a number of interesting and useful properties. One involves independence and covariance. It is easy to see that if the scalar random variables X and Y are independent, then $\mathrm{Cov}(X, Y) = \mathrm{Cor}(X, Y) = 0$, no matter how X and Y are distributed. An important property of the normal distribution is that if X and Y have a bivariate normal distribution and $\mathrm{Cov}(X, Y) = \mathrm{Cor}(X, Y) = 0$, then X and Y are independent. This is also easy to see by merely factoring the joint PDF. In addition to the bivariate normal, there are various other bivariate distributions for which zero correlation implies independence. Lancaster (1959) gave necessary and sufficient conditions for this implication.

Interestingly, X and Y having normal marginal distributions and zero correlation is not sufficient for X and Y to be independent. This, of course, must mean that although the marginals are normal, the joint distribution is not bivariate normal. A simple example of this is the case in which $X \sim \mathrm{N}(0, 1)$, Z is a random variable such that $\Pr(Z = -1) = \Pr(Z = 1) = \frac{1}{2}$, and $Y = ZX$. Clearly, $Y \sim \mathrm{N}(0, 1)$ and $\mathrm{Cor}(X, Y) = 0$, yet X and Y are not independent. We also conclude that X and Y cannot be jointly normal.

There are a number of interesting and useful properties that only the normal distribution has; that is, these properties *characterize* the normal distribution. We will consider some of these properties in the next section.

Some properties of the normal family of distributions form the basis for many statistical methods, such as the use of the Student's t for testing hypotheses or setting confidence limits, or the use of the F distribution in the analysis of variance. Many statistical methods depend on an assumption of a normal distribution.

2.9.1 Multivariate and Matrix Normal Distribution

The d-variate normal distribution, which we denote as $\mathrm{N}_d(\mu, \Sigma)$, has the PDF

$$\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \mathrm{e}^{-(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)/2},$$

where $\mu \in \mathbb{R}^d$ and $\Sigma \succ 0 \in \mathbb{R}^{d\times d}$.

Notice that the exponent in this expression could alternatively be written as

$$-\mathrm{tr}(\Sigma^{-1}(x-\mu)(x-\mu)^{\mathrm{T}})/2.$$

This form is often useful.

As we noted above, each element of a random d-vector X may have a marginal normal distribution, yet X itself may not have a d-variate normal distribution.

Generally, a "multivariate distribution" refers to the distribution of a random vector. If the random object has some other structure, however, a distribution that recognizes the relationships within the structure may be more useful. One structure of interest is a matrix. Some random objects, such as a Wishart matrix (see page 841), arise naturally from other distributions. Another useful random matrix is one in which all elements have a joint normal distribution and the columns of the matrix have one correlational structure and the rows have another correlational structure. This is called a multivariate matrix distribution, which we denote as $\mathrm{MN}_{n\times m}(M, \Psi, \Sigma)$. The PDF for the random $n \times m$ random matrix X is

$$\frac{1}{(2\pi)^{nm/2}|\Psi|^{n/2}|\Sigma|^{m/2}} \mathrm{e}^{-\mathrm{tr}(\Psi^{-1}(X-M)^{\mathrm{T}}\Sigma^{-1}(X-M))/2},$$

where $M \in \mathbb{R}^{n\times m}$, $\Psi \succ 0 \in \mathbb{R}^{m\times m}$, and $\Sigma \succ 0 \in \mathbb{R}^{n\times n}$.

The variance-covariance matrix of X is $\mathrm{V}(X) = \mathrm{V}(\mathrm{vec}(X)) = \Psi \otimes \Sigma$. The variance-covariance matrix of each row of X is Ψ, and the variance-covariance matrix of each column of X is Σ.

The multivariate matrix normal distribution of the matrix X with PDF as given above is related to the ordinary multivariate normal for the vector $\mathrm{vec}(X)$ with PDF

$$\frac{1}{(2\pi)^{nm/2}|\Psi \otimes \Sigma|^{nm/2}} \mathrm{e}^{-\mathrm{vec}(X-M)^{\mathrm{T}}(\Psi\otimes\Sigma)^{-1}\mathrm{vec}(X-M)/2}.$$

Complex Multivariate Normal Distribution

Consider the random d-vector Z, where

$$Z = X + \mathrm{i}Y.$$

The vector Z has a complex d-variate normal distribution if (X, Y) has a real $2d$-variate normal distribution. The PDF of Z has the form

$$\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \mathrm{e}^{-(x-\mu)^{\mathrm{H}}\Sigma^{-1}(x-\mu)/2},$$

where $\mu \in \mathbb{C}^d$ and $\Sigma \succ 0 \in \mathbb{C}^{d\times d}$.

2.9.2 Functions of Normal Random Variables

One reason that the normal distribution is useful is that the distributions of certain functions of normal random variables are easy to derive and they have nice properties. These distributions can often be worked out from the CF of the normal distribution $\mathrm{N}(\mu, \sigma^2)$, which has a particularly simple form:

$$\varphi(t) = \mathrm{e}^{\mathrm{i}\mu t - \sigma^2 t^2/2}.$$

Given n iid $\mathrm{N}(\mu, \sigma^2)$ random variables, $X_1, X_2, \ldots, X_n$, the sample mean and sample variance

$$\overline{X} = \sum_{i=1}^{n} X_i/n \tag{2.27}$$

and

$$S^2 = \sum_{i=1}^{n} \left(X_i - \sum_{i=1}^{n} X_i/n \right)^2 /(n-1) \tag{2.28}$$

are important functions.

Using the CF and equations (1.99) and (1.100), it is easy to see that

$$\overline{X} \sim \mathrm{N}(\mu, \sigma^2/n). \tag{2.29}$$

In Example 1.16, we saw that the sum of squares of n iid standard normal random variables is chi-squared with n degrees of freedom. Using properties of sums of independent normal random variables and of chi-squared random variables we see that $\overline{X}$ and S^2 are independent and furthermore that

$$(n-1)S^2/\sigma^2 \overset{\mathrm{d}}{=} \chi^2_{n-1}. \tag{2.30}$$

Another way to establish the independence of $\overline{X}$ and S^2 and to get the distribution of S^2 is by use of the elegant Helmert transformation. We first define

$$Y_k = X_k - \overline{X}, \quad k = 1, \ldots, n-1 \tag{2.31}$$

and

$$Y_n = -Y_1 - \cdots - Y_{n-1}. \tag{2.32}$$

The joint density of $\overline{X}, Y_1, \ldots, Y_n$ is proportional to

$$\exp\left(-n(\bar{x}-\mu)^2)/2\sigma^2\right) \exp\left(-(y_1^2 + \cdots + y_n^2)/2\sigma^2\right); \tag{2.33}$$

hence, we see that $\overline{X}$ is independent of $Y_1, \ldots, Y_n$, or any function of them, including S^2.

The Helmert transformations are

$$\begin{aligned} W_1 &= \sqrt{2}\left(Y_1 + \tfrac{1}{2}Y_2 + \cdots + \tfrac{1}{2}Y_{n-1}\right) \\ W_2 &= \sqrt{\tfrac{3}{2}}\left(Y_2 + \tfrac{1}{3}Y_3 + \cdots + \tfrac{1}{3}Y_{n-1}\right) \\ W_3 &= \sqrt{\tfrac{4}{3}}\left(Y_3 + \tfrac{1}{4}Y_4 + \cdots + \tfrac{1}{4}Y_{n-1}\right) \\ &\cdots \\ W_{n-1} &= \sqrt{\tfrac{n}{n-1}}Y_{n-1} \end{aligned} \tag{2.34}$$

We have

$$\sum_{k=1}^{n-1} W_k^2 = (n-1)S^2. \tag{2.35}$$

Because the joint density of $W_1, \ldots, W_{n-1}$ is the same as $n-1$ iid $\mathrm{N}(0, \sigma^2)$ random variables (exercise), we have that $(n-1)S^2/\sigma^2$ is distributed as χ^2_{n-1}.

If X is distributed as $\mathrm{N}_d(\mu, I_d)$, and for $i = 1, \ldots k$, A_i is a $d \times d$ symmetric matrix with rank r_i such that $\sum_i A_i = I_d$, then we can write

$$X^{\mathrm{T}}X = X^{\mathrm{T}}A_1X + \cdots + X^{\mathrm{T}}A_kX,$$

and the $X^{\mathrm{T}}A_iX$ have independent noncentral chi-squared distributions $\chi^2_{r_i}(\delta_i)$ with $\delta_i = \mu^{\mathrm{T}}A_i\mu$ if and only if $\sum_i r_i = d$. This result is known as Cochran's theorem. This form of the theorem and various preliminary forms leading up to it are proved beginning on page 430.

From the family of central chi-squared distributions together with an independent normal family, we get the family of t distributions (central or noncentral, depending on the mean of the normal). From the family of chi-squared distributions (central or noncentral) we get the family of F distributions (central, or singly or doubly noncentral; see Example 1.15 on page 59 for the central distributions).

The expectations of reciprocals of normal random variables have interesting properties. First of all, we see that for $X \sim \mathrm{N}(0,1)$, $\mathrm{E}(1/X)$ does not exist. Now, for $X \sim \mathrm{N}_d(0, I)$ consider

$$\mathrm{E}\left(\frac{1}{\|X\|_2^2}\right). \tag{2.36}$$

For $d \leq 2$, this expectation is infinite (exercise). For $d \geq 3$, however, this expectation is finite (exercise).

2.9.3 Characterizations of the Normal Family of Distributions

A simple characterization of a normal distribution was proven by Cramér in 1936:

Theorem 2.4
Let X_1 and X_2 be independent random variables. Then X_1 and X_2 have normal distributions if and only if their sum $X_1 + X_2$ has a normal distribution.

Proof. ***fix ∎

The independence of certain functions of random variables imply that those random variables have a normal distribution; that is, the independence of certain functions of random variables characterize the normal distribution.

Theorem 2.5 (Bernstein's theorem)
Let X_1 and X_2 be iid random variables with nondegenerate distributions, and let $Y_1 = X_1 + X_2$ and $Y_2 = X_1 - X_2$. If Y_1 and Y_2 are also independent then X_1 and X_2 have normal distributions.

Proof. ***fix ∎

An extension of Bernstein's theorem is the Darmois theorem, also called the Darmois-Skitovich theorem.

Theorem 2.6 (Darmois theorem)
Let $X_1, \ldots, X_n$ be iid random variables with nondegenerate distributions, and let

$$Y_1 = \sum_{i=1}^{n} b_i X_i$$

and

$$Y_2 = \sum_{i=1}^{n} c_i X_i,$$

where the b_i and c_i are nonzero real constants. If Y_1 and Y_2 are also independent then $X_1, \ldots, X_n$ have normal distributions.

The proof of the Darmois theorem proceeds along similar lines as that of Bernstein's theorem.

The following theorem is a remarkable fact that provides a characterization of the normal distribution in terms of the sample mean $\overline{X}$ and the sample variance S^2.

Theorem 2.7 (Geary's theorem)
Let $X_1, X_2, \ldots, X_n$ be iid with PDF f with finite variance, σ^2. A necessary and sufficient condition that the parent distribution be a normal distribution is that the sample mean and the sample variance be independent.

Proof. First, we note that the sufficiency can be established directly quite easily.

Now for the necessity. Assume that

$$\overline{X} = \sum_{i=1}^{n} X_i/n$$

and

$$S^2 = \sum_{i=1}^{n} (X_i - \overline{X})^2/(n-1)$$

are independent.

We will work with various characteristic functions, all of which are determined by E_X (see page 49). We will adopt a simple notation for these CFs. Our main interest will be the CF of the joint distribution of $\overline{X}$ and S^2, so we denote it simply as $\varphi(t_1, t_2)$; that is,

$$\begin{aligned}\varphi(t_1, t_2) &= \varphi_{\overline{X},S^2}(t_1, t_2) \\ &= \int \mathrm{e}^{\mathrm{i}t_1\bar{x}+\mathrm{i}t_2 s^2} \prod f(x_i)\mathrm{d}x_i.\end{aligned}$$

We denote the separate CFs as $\varphi_1(t_1)$ and $\varphi_2(t_2)$:

$$\varphi_1(t_1) = \varphi_{\overline{X}}(t_1) = \varphi_{\overline{X},S^2}(t_1, 0)$$

and

$$\varphi_2(t_2) = \varphi_{S^2}(t_2) = \varphi_{\overline{X},S^2}(0, t_2).$$

Finally, we let

$$\varphi_X(t)$$

be the CF of each X_i.

From equation (1.131) (after dividing Y by n), we have

$$\varphi_1(t_1) = (\varphi_X(t/n))^n.$$

From equation (1.104), the independence of $\overline{X}$ and S^2 implies that

$$\varphi(t_1, t_2) = \varphi_1(t_1)\varphi_2(t_2),$$

and we have

$$\begin{aligned}\left.\frac{\partial\varphi(t_1, t_2)}{\partial t_2}\right|_{t_2=0} &= \varphi_1(t_1) \left.\frac{\partial\varphi_2(t_2)}{\partial t_2}\right|_{t_2=0} \\ &= (\varphi_X(t/n))^n \left.\frac{\partial\varphi_2(t_2)}{\partial t_2}\right|_{t_2=0} \qquad (2.37)\end{aligned}$$

Directly from the definition of $\varphi(t_1, t_2)$, we have

$$\left.\frac{\partial\varphi(t_1,t_2)}{\partial t_2}\right|_{t_2=0} = \mathrm{i}\int s^2\mathrm{e}^{\mathrm{i}t_1\bar{x}}\prod f(x_i)\mathrm{d}x_i. \tag{2.38}$$

Now, substituting $s^2 = g(x_1,\ldots,x_n)$ and $\bar{x} = h(x_1,\ldots,x_n)$ into this latter equation, we get

$$\begin{aligned}\left.\frac{\partial\varphi(t_1,t_2)}{\partial t_2}\right|_{t_2=0} &= \mathrm{i}(\varphi_X(t_1/n))^{n-1}\int x^2\mathrm{e}^{\mathrm{i}t_1x/n}f(x)\mathrm{d}x \\ &\quad -\mathrm{i}(\varphi_X(t_1/n))^{n-2}\left(\int x\mathrm{e}^{\mathrm{i}t_1x/n}f(x)\mathrm{d}x\right)^2. \end{aligned}\tag{2.39}$$

Furthermore, because $\mathrm{E}(S^2) = \sigma^2$, we have

$$\left.\frac{\partial\varphi_2(t_2)}{\partial t_2}\right|_{t_2=0} = \mathrm{i}\sigma^2. \tag{2.40}$$

Now, substituting (2.39) and (2.40) into (2.37) and writing $t = t_1/n$, we have

$$\varphi_X(t)\int x^2\mathrm{e}^{\mathrm{i}tx}f(x)\mathrm{d}x - \left(\int x\mathrm{e}^{\mathrm{i}tx}f(x)\mathrm{d}x\right)^2 = (\varphi_X(t))^2\sigma^2. \tag{2.41}$$

Note that i^k times the integrals in this latter equation are of the form $\mathrm{d}^k\varphi_X(t)/\mathrm{d}t^k$, so we may re express equation (2.41) as the differential equation

$$-\varphi_X(t)\varphi''_X(t) + (\varphi'_X(t))^2 = (\varphi_X(t))^2\sigma^2. \tag{2.42}$$

Now, solving this differential equation with the initial conditions

$$\varphi_X(0) = 1, \quad \text{and} \quad \varphi'_X(0) = \mathrm{i}\mu,$$

where $\mu = \mathrm{E}(X)$, we have

$$\varphi_X(t) = \mathrm{e}^{\mathrm{i}\mu t}\mathrm{e}^{-\mathrm{i}\sigma^2t^2/2}. \tag{2.43}$$

(The ordinary differential equation (2.42) is second order and second degree, so the solution is difficult. We can confirm that equation (2.43) is the solution by differentiation and substitution.)

Equation (2.43) is the characteristic function of the normal distribution $\mathrm{N}(\mu,\sigma^2)$, and so the theorem is proved. ∎

This theorem can easily be extended to the multivariate case where the X_i are d-vectors. Because only first and second moments are involved, the details of the proof are similar (see Exercise 2.31).

2.10 Generalized Distributions and Mixture Distributions

Probability distributions used in statistical applications are chosen both for their general properties that determine the properties of the statistical methods used and for their similarities to the frequency of observed data. Statistical methods based on a distributional family in the exponential, for example, can yield optimal unbiased procedures. On the other hand, statistical inference based on a family of distributions whose moments correspond to the sample moments of observed data would have greater credence than inference based on an assumption of a probability distribution with properties that differ from the frequencies observed.

Various families of probability distributions have been identified that are useful models of observed frequency distributions. The only specific family that we consider in this chapter is the normal family studied in Section 2.9. We list other important families in Appendix A. In this section, we consider general modifications of distributions that may yield more realistic models of observed data. We may find, for example, that a normal distribution fits observed data well over some ranges but not over others. The data may be censored (Section 2.10.1). On the other hand, it may be that the data fall into different groups, some of which have frequencies corresponding to one normal probability distribution, and others have frequencies corresponding to a different normal distribution or even to a distribution in a different family (Section 2.10.2). Another possibility is that frequencies of the observational data are quite similar to a normal distribution, but that they are skewed one way or another (Section 2.10.3).

Given a well-studied family of distributions such as the gamma or some other family in Appendix A, we may seek to generalize the family by incorporating another parameter. (A simple example for the two-parameter gamma family is just to add a "starting" parameter, as in Table A.6.)

More generally, we make seek to define a distribution with given properties, such as skewness or kurtosis. We may define a PDF or CDF that matches given quantiles, for example. We discuss some of these approaches in Section 2.10.4.

The need to develop a probability distribution that models the frequency distribution of observed data has led to many useful distributions. Another motivation to developing useful and tractable probability distributions is to have meaningful prior distributions in Bayesian analysis (see Chapter 4).

2.10.1 Truncated and Censored Distributions

Often the support of a standard family is truncated to yield a family whose support is a proper subset of the standard family's. The infinite support of a normal distribution may be truncated to a finite interval, for example, because in a given application, no data will be, or can be, observed outside of some finite interval. Another common example is a truncated Poisson (also called

"positive Poisson") in which the point 0 has been removed from the support. This may be because a realization of 0 does not make sense in the particular application.

The kernel of the PDF of a truncated distribution over its support is the same as kernel of the original distribution over its support. The partition function is adjusted as appropriate. The PDF of the positive Poisson distribution, for example, is

$$f(x) = \frac{1}{\mathrm{e}^\theta - 1}\frac{\theta^x}{x!}\mathrm{I}_{\{1,2,\ldots\}}(x), \tag{2.44}$$

The support of a distribution may also be changed by censoring. "Censoring" refers what is done to a random variable, so strictly speaking, we do not study a "censored distribution", but rather the distribution of a random variable that has been censored.

There are various types of censoring. One type is similar to the truncation of a distribution, except that if a realization of the random variable occurs outside of the truncated support, that fact is observed, but the actual value of the realized is not known. This type of censoring is called "type I" fixed censoring, and in the case that the support is an interval, the censoring is called "right" or "left" if the truncated region of the support is on the right (that is, large values) or on the left (small values). A common situation in which type I fixed censoring occurs is when the random variable is a survival time, and several observational units are available to generate data. Various probability distributions such as exponential, gamma, Weibull, or lognormal may be used to model the survival time. If an observational unit survives beyond some fixed time, say t_c, only that fact is recorded and observation of the unit ceases.

In another kind of fixed censoring, also illustrated by observation of failure times of a given set of say n units, the realized failure times are recorded until say r units have failed. This type of censoring is called "type II" fixed censoring.

If an observational unit is removed prior to its realized value being observed for no particular reason relating to that unobserved value, the censoring is called "random censoring".

Censoring in general refers to a failure to observe the realized value of a random variable but rather to observe only some characteristic of that value. As another example, again one that may occur in studies of survival times, suppose we have independent random variables T_1 and T_2 with some assumed distributions. Instead of observing T_1 and T_2, however, we observe $X = \min(T_1, T_2)$ and G, an indicator of whether $X = T_1$ or $X = T_2$. In this case, T_1 and T_2 are censored, but the joint distribution of X and G may be relevant, and it may be determined based on the distributions of T_1 and T_2.

2.10.2 Mixture Families

In applications it is often the case that a single distribution models the observed data adequately. Sometimes two or more distributions from a single family of distributions provide a good fit of the observations, but in other cases, more than one distributional family is required to provide an adequate fit. In some cases most of the data seem to come from one population but a small number seem to be extreme outliers. Some distributions, such as a Cauchy, are said to be "outlier-generating", but often such distributions are difficult to work with (because they have infinite moments, for example). Mixtures of distributions, such as the ϵ-mixture distribution (see page 601), are often useful for modeling data with anomalous observations.

A mixture family can be defined in terms of a set of CDFs $\mathcal{P}_0$. The CDF of a mixture is $\sum w_i P_i$, where $P_i \in \mathcal{P}_0$, $0 \leq w_i \leq 1$, and $\sum w_i = 1$. The set $\mathcal{P}$ of all such mixture CDFs is called a distribution function space (see page 754). If each the probability measure associated with each P_i is dominated by the measure ν, then the probability measure associated with $\sum w_i P_i$ is dominated by the ν.

One family that is useful in robustness studies is the ϵ-mixture distribution family, which is characterized by a given family with CDF P that we refer to as the reference distribution, together with a point x_c and a weight ϵ. The CDF of a ϵ-mixture distribution family is

$$P_{x_c,\epsilon}(x) = (1-\epsilon)P(x) + \epsilon \mathrm{I}_{[x_c,\infty[}(x), \tag{2.45}$$

where $0 \leq \epsilon \leq 1$. The point x_c may be thought of as a "contaminant" in the distribution with CDF P. In a common example of this kind of mixture, the probability measure associated with P is dominated by the Lebesgue measure μ, and in that case the probability measure associated with $P_{x_c,\epsilon}$ is dominated by $\mu + \delta_{x_c}$, where δ_{x_c} is the dirac measure concentrated at x_c.

Another type of mixture family is composed of two distributions dominated by Lebesgue measure that have CDFs P_1 and P_2 such that at the point x_c, $P_1(x_c) < P_2(x_c)$ and whose CDF is given by

$$P(x) = \begin{cases} P_1(x) & -\infty < x < x_c \\ P_2(x) & x_c \leq x < \infty \end{cases} \tag{2.46}$$

The probability measure associated with P is dominated by $\mu + \delta_{x_c}$.

Finally, we consider a mixture family formed by censoring. Let $Y_1, \ldots, Y_n$ be iid with CDF P, and let

$$X_i = \begin{cases} Y_i & \text{if } Y_i \geq c \\ c & \text{if } Y_i < c \end{cases} \quad i = 1, \ldots, n. \tag{2.47}$$

If the distribution of the Y_i is dominated by Lebesgue measure, then the distribution of the X_i is dominated by $\mu + \delta_c$.

2.10.3 Skewed Distributions

Most of the common skewed distributions, such as the gamma, the log normal, and the Weibull, have semi-infinite range. The common distributions that have range $(-\infty, \infty)$, such as the normal and the t, are symmetric.

There are several ways to form a skewed distribution from a symmetric one. Two simple ways are

- *CDF-skewing:* take a random variable as the maximum (or minimum) of two independent and identically symmetrically distributed random variables, or
- *differential scaling:* for some constant $\xi \neq 0$, scale a symmetric random variable by ξ if it is less than its mean and by $1/\xi$ if it is greater than its mean.

In each case, it may be desirable to shift and scale the skewed random variable so that it has a mean of 0 and a variance of 1. (We can then easily shift and scale the random variable so as to have any desired mean and variance.)

CDF-Skewing

If a random variable has PDF $f(x)$ and CDF $F(x)$, from equation (**??**), the PDF of the maximum of two independent random variables with that distribution has the PDF

$$2F(x)f(x). \tag{2.48}$$

Intuitively, we see that the maximum of two symmetrically distributed random variables has a skewed distribution.

We can generalize this form by scaling the argument in the CDF,

$$2F(\alpha x)f(x). \tag{2.49}$$

A negative scaling, that is, $\alpha < 0$ yields a negative skewness. (In the case of the normal distribution, the value $\alpha = -1$ is equivalent to the minimum of two independent normal random variables.) Values of α larger in absolute value yield greater degrees of skewness. The scaling also changes the kurtosis, with larger absolute values of α yielding a larger kurtosis.

Specifically, for the standard normal distribution, we form the PDF of a skewed normal as

$$f_{\mathrm{SN}_1}(x;\alpha) = 2\Phi(\alpha x)\phi(x), \tag{2.50}$$

$\phi(x)$ denotes the PDF of the standard normal distribution and $\Phi(x)$ denotes the CDF.

Obviously, CDF-skewing can be applied to other distributions, such as the t or the generalized error distribution, and of course including normals as in equation (2.50) with other means and variances.

Differential Scaling

Another way of forming a skewed distribution is by scaling the random variable differently on different sides of the mean or some other central point. Specifically, for the normal distribution, we can form the PDF of a skewed normal as

$$f_{\mathrm{SN}_2}(x;\xi) = \frac{2}{\xi + \frac{1}{\xi}} \begin{cases} \phi(x/\xi) \text{ for } x < 0 \\ \\ \phi(\xi x) \quad \text{for } x \geq 0, \end{cases} \tag{2.51}$$

where, as before, $\phi(x)$ denotes the PDF of the standard normal distribution. Values of ξ less than one produce a positive skew, and values greater than one produce a negative skew. In either case, the excess kurtosis is positive.

The dividing point for differential scaling obviously could be chosen arbitrarily. To form a skewed distribution from a unimodal symmetric distribution, an obvious dividing point would be the mode of the distribution.

Distributions formed by differential scaling are sometimes called "two piece distributions".

2.10.4 Flexible Families of Distributions Useful in Modeling

Some of the useful families of probability distributions arise from simple processes of "random" events. This is one way we naturally define the Bernoulli or binomial family, the hypergeometric family. the Poisson family, or the exponential family, as examples. These discrete families can be generalized by developing a differential equation that models a limiting case of the discrete frequency model. The Pearson system is an example of this approach (in which the basic differential equation arises as a limiting case of a hypergeometric distribution). Other broad families of distributional forms have been developed by Johnson, by Burr, and by Tukey. The objective is to be able to represent a wide range of distributional properties (mean, variance, skewness, shape, etc.) with a small number of parameters, and then to fit a specific case by proper choice of these parameters.

A special type of mixture distribution is a *probability-skewed distribution*, in which the mixing weights are the values of a CDF. The skew-normal distribution is a good example.

The (standard) skew-normal distribution has density

$$g(x) = \frac{2}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2} \Phi(\lambda x) \quad \text{for } -\infty \leq x \leq \infty, \tag{2.52}$$

where $\Phi(\cdot)$ is the standard normal CDF, and λ is a constant such that $-\infty < \lambda < \infty$. For $\lambda = 0$, the skew-normal distribution is the normal distribution, and in general, if $|\lambda|$ is relatively small, the distribution is close to the normal. For larger $|\lambda|$, the distribution is more skewed, either positively or negatively.

Other distributions symmetric about 0 can also be skewed by a CDF in this manner. The kernel of the probability density is

$$k(x) = p(x)P(\lambda x),$$

where $p(\cdot)$ is the density of the underlying symmetric distribution, and $P(\cdot)$ is a CDF. (It is not necessary that the CDF be for the same distribution.) The idea also extends to multivariate distributions.

In most cases, if $|\lambda|$ is relatively small, generation of random variables from a probability-skewed symmetric distribution using an acceptance/rejection method with the underlying symmetric distribution as the majorizing density is entirely adequate. For larger values of $|\lambda|$, it is necessary to divide the support into two or more intervals. It is still generally possible to use the same majorizing density, but the multiplicative constant can be different in different intervals.

Some commonly used ones are the Pearson family, the Johnson family, the generalized lambda family, and the Burr family. The Pearson family is probably the best known of these distributions. A specific member of the family is determined by the first four moments, so a common way of fitting a distribution to an observed set of data is by matching the moments of the distribution to those of the sample.

Another widely used general family of distributions is the Johnson family. A specific member of this family is also determined by the first four or five moments, depending on the parametrization.

A generalized lambda family of distributions was described by Ramberg and Schmeiser (1974). This system, which is a generalization of a system introduced by John Tukey, has four parameters that can be chosen to fit a variety of distributional shapes. They specify the distribution in terms of the inverse of its distribution function,

$$P^{-1}(u) = \lambda_1 + \frac{u^{\lambda_3} - (1-u)^{\lambda_4}}{\lambda_2}. \tag{2.53}$$

The distribution function itself cannot be written in closed form, but the inverse allows deviates from this distribution to be generated easily by the inverse CDF method; just generate u and apply equation (2.53).

Albert, Delampady, and Polasek (1991) defined a family of distributions that is very similar to the lambda distributions and is particularly useful in Bayesian analysis with location-scale models.

Another family of distributions that is very flexible and that can have a wide range of shapes is the Burr family of distributions (Burr, 1942). One of the common forms (Burr and Cislak, 1968) has the CDF

$$P(x) = 1 - \frac{1}{(1+x^{\alpha})^{\mathrm{B}}} \quad \text{for } 0 \le x \le \infty;\ \alpha, \mathrm{B} > 0, \tag{2.54}$$

which is easily inverted. Other forms of the Burr family have more parameters, allowing modeling of a wider range of empirical distributions.

Fleishman (1978) suggested representing the random variable of interest as a polynomial in a standard normal random variable, in which the coefficients

are determined so that the moments match specific values. If Z has a $\mathrm{N}(0,1)$ distribution, then the random variable of interest, X, is expressed as

$$X = c_0 + c_1 Z + \cdots + c_k Z^k. \tag{2.55}$$

If m moments are to be matched to prespecified values, then k can be chosen as $m-1$, and the cs can be determined from m equations in m unknowns that involve expectations of powers of a $\mathrm{N}(0,1)$ random variable. Fleishman used this representation to match four moments; hence, he used a third-degree polynomial in a standard normal random variable.

The motivation for some of the early work with general families of distributions was to use them as approximations to some standard distribution, such as a gamma, for which it is more difficult to generate deviates. As methods for the standard distributions have improved, it is more common just to generate directly from the distribution of interest. The general families, however, often provide more flexibility in choosing a distribution that better matches sample data. The distribution is fit to the sample data using either percentiles or moments.

2.11 Multivariate Distributions

While our previous discussions have generally applied to multivariate distributions, the dimensionality of the range of a random variable may limit our studies or in other ways may have major effects on the properties of the distribution that affect statistical analysis.

2.11.1 Marginal Distributions

Characterizations of families of multivariate probability distributions are often more difficult or less intuitive. The covariance matrix is the common measure that relates the individual components of a random variable to each other in a pairwise manner. This is a very useful distribution measure for multivariate normal families, but it is much less useful for other multivariate families; consider, for example, a multivariate gamma family characterized by vectors α and β (generalizing the univariate parameters) and some variance-covariance matrix. It is not clear what that matrix would be and how it would be incorporated in a simple manner into the PDF. In applications, copulas are often used in an ad hoc sense to express the relationship of the individual components of a random variable to each other.

2.11.2 Elliptical Families

Spherical and elliptical families are important in multivariate statistics. A d-variate random variable X is said to have a *spherical* distribution iff $Q^{\mathrm{T}}X \overset{\mathrm{d}}{=} X$ for every $d \times d$ orthogonal matrix Q.

Because $Q^{\mathrm{T}}Q = I$, the density function must depend on $X = x$ only through xx^{T}. Spherical distributions include multivariate normals with diagonal variance-covariance matrices and multivariate t distributions formed from iid scalar t variates.

If the k-variate random variable Y has a spherical distribution, and for $d \leq k$, μ is a fixed d-vector and A is a fixed $d \times k$ matrix of full rank, then

$$X = \mu + AY$$

is said to have an *elliptical* distribution.

Any elliptical family is a location-scale family.

General multivariate normal distributions, multivariate t distributions, and multivariate Cauchy distributions are members of elliptical families.

2.11.3 Higher Dimensions

Another problem with multivariate distributions arises from the properties of $\mathbb{R}^d$ as d increases. A simple instance of this can be seen in the multivariate normal distribution as d increases from 2 to 3 (see page 273). We can state a general result nontechnically: *in a nondegenerate multivariate family, as the dimension increases, every observation becomes an outlier*. See Exercise 2.27 for a specific example of this result.

Notes and Further Reading

Distribution Theory

In many applications of probability theory in statistics, the first step is to associate a phenomenon of interest with a random variable. The statistical analysis then becomes a study of the distribution of the random variable. The "study" involves collecting or assimilating data, exploration of the data, transformations of the data, comparisons of the observed data with regions and quantiles of probability distributions or families of distributions, and finally inference about some specific distribution or family of distributions. "Statistics" is the science of the methods of the study.

Study of the characteristics of the probability distributions used in statistical analysis is part of the subject of probability theory.

Although we do not want to draw a hard line between probability and statistics — drawing hard lines between any disciplines impedes the advancement of science — distribution theory *per sé* is within the domain of probability theory, rather than of statistical theory. That is why, for example, discussion of the exponential class of distributions is included in this chapter on probability, rather than placed in a later chapter on statistical theory.

Regular Families

When some interesting properties apply to certain cases but not others, those cases for which the properties hold may be referred to as "regular" cases. The regularity conditions are essentially the hypotheses of a theorem that states that the particular properties hold. A theorem can seem more important if its conclusions hold for all except "irregular" cases.

In statistics, the most common usage of the phrase "regularity conditions" or "regular families of distributions" is in connection with Fisher information.)

Quadratic mean differentiable families play important roles in much of the asymptotic theory developed by Lucien Le Cam. Properties of these families are considered at some length by Le Cam and Yang (2000) and in TSH3, Chapter 12.

The Exponential Class

Extensive discussions of exponential families are provided by Barndorff-Nielson (1978) and Brown (1986). Morris (1982) defined the natural exponential family with quadratic variance function (NEF-QVF) class of distributions and showed that much theory could be unified by appeal to the quadratic variance property. (See also Morris and Lock (2009).)

Heavy-Tailed Families

Various types of heavy-tailed distributions have been extensively studied, often because of their applications in financial analysis.

Some of the basic results of subexponential families were developed by Teugels (1975), who also considered their applications in renewal theory. Multivariate subexponential families with somewhat similar properties can be identified, but their definition is not as simple as the convergence of the ratio in expression (2.4) (see Cline and Resnick (1992)).

Infinitely Divisible and Stable Families

Steutel and van Harn (2004) provide a general coverage of infinitely divisible distributions in $\mathbb{R}$. Infinitely divisible distributions arise often in applications of stochastic processes. Janicki and Weron (1994) discuss such distributions in this context and in other areas of application such as density estimation. Steutel (1970) considers mixtures of infinitely divisible distributions.

The discussion of stable distributions in this chapter generally follows the development by Feller (1971), but is also heavily influenced by Breiman (1968). Stable distributions also provide useful models for heavy-tailed processes. Samorodnitsky and Taqqu (1994) provide an extensive discussion of stable distributions in this context. Stable distributions are useful in studies of the robustness of statistical procedures.

Exercises

2.1. Prove Theorem 2.2.

2.2. State the conditions on the parameters of a beta(α, β) distribution for its PDF to be
  a) subharmonic
  b) superharmonic
  c) harmonic

2.3. a) Show that condition (2.3) on page 166 implies condition (2.2).
  b) Find a counterexample to show that the converse is not true.

2.4. a) Show that condition (2.4) on page 166 implies condition (2.3); that is, a subexponential family is a long-tailed family.
  Hint: Note that condition (2.3) is $\lim_{x\to\infty} \overline{F}(x+t)/\overline{F}(x) = 1$. (See Athreya and Ney (1972) page 148, and Pitman (1980).)
  b) Find a counterexample to show that the converse is not true.

2.5. a) Show that the following families of distributions are monotone likelihood ratio families.
    i. The one-parameter exponential class, with PDF

$$\exp\left(\eta(\theta)T(x) - \xi(\theta)\right) h(x),$$

    with $\theta \in]a, b[\subseteq \mathbb{R}$, where a and b are known and may be infinite, and $\eta(\theta)$ is a monotone scalar function of θ.
    ii. U$(0, \theta)$, with PDF

$$\frac{1}{\theta} \mathrm{I}_{[0,\theta]}(x).$$

    iii. U$(\theta, \theta+1)$, with PDF

$$\mathrm{I}_{[\theta,\theta+1]}(x).$$

  b) Show that the one-parameter Cauchy family is not a monotone likelihood ratio family. The Lebesgue PDF is

$$\frac{1}{\pi\beta(1 + x/\beta)^2}.$$

2.6. Show that a totally positive family with $r = 2$ (see equation (2.5)) is a monotone likelihood ratio family.

2.7. Assume that $\log p_\theta(x)$, where $p_\theta(x)$ is a PDF, is strictly concave in x.
  a) Show that $p_\theta(x)$ is unimodal. Now, generalize this result (beyond the log function) and prove the generalization.
  b) Give an example of a family of distributions that is (strictly) unimodal but not strongly unimodal.
  c) Now for $\theta \in \mathbb{R}$, show that $p_\theta(x)$ is a monotone likelihood ratio family in θ.

2.8. Write the likelihood rato for each of the one-parameter families of distributions in Table 2.1, and show that it is monotone in the relevant variable.

2.9. Show that the Cauchy(γ, β_0) family (with β_0 known and fixed) does not have a monotone likelihood rato.

2.10. Benford's law is used as a model of the probability distribution of the digits, particularly the first digit, in the decimal representation of "naturally occurring" numbers, such as the lengths of rivers of the world, the areas of countries, and so on. By Benford's law, the probability that the first digit is $d = 1, 2, \ldots, 9$ is

$$p(d) = \log_{10}(d+1) - \log_{10}(d), \quad d = 1, 2, \ldots, 9.$$

This law has been used to detect artificially constructed data because data generated by a person tends to have first digits that are uniformly distributed over $\{1, 2, \ldots, 9\}$.

a) There are many instances in which this law does not apply, of course. If all data in some specific area are between 300 and 500, say, then obviously the law would not be applicable. What is needed to know the distribution of the number being represented.
Derive the probability function $p(d)$ for the case that d is the first digit in the decimal representation of the realization of a random variable with distribution U$(0, 1)$.

b) Of course, if this law is to correspond to naturally occurring numbers such as lengths of rivers, it must be invariant to the unit or measurement. To show exact invariance would require an assumption about the distribution of the number being represented (the lengths of rivers, for example). This in itself is not a straightforward task.
Rather than making specific statements about the underlying distributions, develop a heuristic argument that the Benford's law probability function is approximately invariant to unit of measurement, and that it is the unique probability function with this approximate invariance property. (These approximate properties are often stated as facts (theorems), and a heuristic argument, probably similar to yours is given as their "proofs".)

2.11. a) Write the PDF of each family of distributions listed in Table 2.2 in the form of equation (2.7).

b) Write the PDF of each family of distributions listed in Table 2.2 in the form of equation (2.11), and identify the natural parameter space H.

c) Which of the families of distributions listed in Table 2.2 are natural (linear) exponential families?

2.12. Show that each of the distributions listed in Table 2.3 is not a member of the exponential class.

2.13. Represent the probability mass functions of the Poisson, the binomial, and the negative binomial distributions as members of the discrete power series class; that is, for each of these distributions, identify θ, h_x, and $c(\theta)$ in equation (2.16).

2.14. Show that the positive Poisson family, with PDF as given in equation (2.44), is of the discrete power series class, and hence of the exponential class.

2.15. Suppose that $X_1, \ldots, X_n$ are iid as a discrete power series(θ) distribution, with given series $\{h_x\}$ and normalizing constant $c(\theta)$. Show that $T = \sum_{i=1}^{n} X_i$ has a discrete power series distribution, and identify the corresponding terms in its probability mass function.

2.16. Consider the family of 2-variate distributions with PDF

$$\frac{1}{(2\pi)|\Sigma|^{1/2}} \exp\left(-(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)/2\right),$$

where μ is a 2-vector of constants and Σ is a 2×2 positive definite matrix of constants.

Show that this family is of the exponential class, and express the density in the canonical (or natural) form of the exponential class.

2.17. Show that both $\mathcal{G}$ and $\widetilde{\mathcal{G}}$ in Exmaple 2.3 on page 179 are groups.

2.18. a) Show that each of the distributions listed in Table 2.4 is a location family.

b) Show that each of the distributions listed in Table 2.5 is a scale family.

2.19. Show that the likelihood function and the PDF for a location familiy are the same function. Produce graphs similar to those in Figure 1.2 for the exponential(α, θ_0) family.

2.20. Show that these distributions are not location-scale families:
the usual one-parameter exponential family, the binomial family, and the Poisson family.

2.21. Show that a full rank exponential family is complete.

2.22. Prove Theorem 2.1.

2.23. Show that the normal, the Cauchy, and the Poisson families of distributions are all infinitely divisible.

2.24. Show that the Poisson family of distributions is not stable.

2.25. Show that the normal and the Cauchy families of distributions are stable, and show that their indexes of stability are 2 and 1 respectively.

2.26. Express the PDF of the curved normal family $\mathrm{N}(\mu, \mu^2)$ in the canonical form of an exponential family.

2.27. Higher dimensions.

a) Let the random variable X have a uniform distribution within the ball $\|x\|_2 \leq 1$ in $\mathbb{R}^d$. This is a spherical distribution. Now, for given $0 < \delta < 1$, show that

$$\Pr(\|X\| > 1 - \delta) \to 1$$

as d increases without bound.

b) Let the random variable X have a d-variate standard normal distribution distribution, $\mathrm{N}_d(0, I_d)$. Determine

$$\lim_{d \to \infty} \Pr(\|X\| > 1).$$

2.28. a) Show that the joint density of $\overline{X}, Y_1, \ldots, Y_n$ given in equations (2.27), (2.31), and (2.32) is proportional to

$$\exp\left(-n(\bar{x}-\mu)^2\right)/2\sigma^2\right) \exp\left(-(y_1^2+\cdots+y_n^2)/2\sigma^2\right).$$

b) Show that the joint density of $W_1, \ldots, W_{n-1}$ given in equation (2.35) is the same as $n-1$ iid $\mathrm{N}(0,\sigma^2)$ random variables.

2.29. Higher dimensions. Let $X \sim \mathrm{N}_d(0, I)$ and consider

$$\mathrm{E}\left(\frac{1}{\|X\|_2^2}\right).$$

a) Show that for $d \leq 2$, this expectation is infinite.

b) Show that for $d \geq 3$, this expectation is finite. What is the value of this expectation as $d \to \infty$?

2.30. Prove Theorem 2.4.

2.31. Work through the details of the proof of Geary's theorem (Theorem 2.7) for the case that $X_1, \ldots, X_n$ are iid d-vectors.

3

Basic Statistical Theory

The field of statistics includes various areas, such as *descriptive statistics* including statistical graphics, *official statistics*, *exploratory data analysis* including data mining and statistical learning, and *statistical inference*, including forecasting or predictive inference. Statistical learning generally involves predictive inference, so in that sense it is also part of the broad area of statistical inference. Mathematical statistics is generally concerned with the theory underlying statistical inference. Most of the important advances in mathematical statistics have been driven by real applications.

In probability theory, we develop models of probability distributions and consider the characteristics of random variables generated by such models. The field of statistics is concerned with an inverse problem; we have realizations of random variables from which we want to infer characteristics of the models of probability distributions.

We develop methods of statistical inference using probability theory. Statistical inference, as I describe it, requires data. Some people describe any probabilistic reasoning as "statistical inference", and they actually use the term "no-data problem" to describe such a process of computing expected values under various scenarios.

Data are the *observable* output from some *data-generating process*. The data-generating process can often be described in terms of a physical model. For statistical analysis, however, the data-generating process is described by an abstract probability distribution P, and this model may involve *unobservable* quantities such as *parameters* or *latent variables*. The objective in statistical inference is to make decisions about unknown aspects of either the data-generating process itself or the probability distribution P. Whether we emphasize the data-generating process or the assumed probability distribution may affect our methods of inference. An issue that will arise from time to time is whether or not all aspects of the data-generating process should affect the inference about P, or whether the inference should be based solely on the data and the assumed probability distribution P. If we have only the data and no knowledge of the process by which it was generated (that is, we

lack "metadata"), then we may be limited in the methods of inference that we can use.

The Canonical Problems in Statistical Inference

The basic problem in statistical inference is to develop more precise models of the data-generating process that gave rise to a set of observed data. The problem we generally address, however, is that of refining the probability distribution P by making decisions about unknown aspects of this assumed distribution. The models themselves are probability distributions or classes of probability distributions. Formal statistical inference (which I will just call "statistical inference") focuses on the probability distribution.

Statistical inference is the process of using observational data from a population that is in an assumed family of distributions $\mathcal{P}$ to identify another family $\mathcal{P}_H$ that "more likely" contains the population from which the data arose. In a restricted form of this problem, we have two families $\mathcal{P}_{H_0}$ and $\mathcal{P}_{H_1}$, and based on available data, we wish to decide which of these gave rise to the data. (This setup follows the Neyman-Pearson paradigm of hypothesis testing; there are various approaches, however.) In another restricted form of this problem, we have a single hypothesized family $\mathcal{P}_{H_0}$, and based on available data, we wish to determine the plausibility that this family gave rise to the data. (This setup follows the Fisherian paradigm of significance testing; there are various approaches, however.)

The assumed family $\mathcal{P}$ must be broad enough to allow reasonable inferences from observational data. The choice of $\mathcal{P}$ may be somewhat subjective. It is based on whatever may be known or assumed about the data-generating process being studied. Generally, $\mathcal{P}_H$ is a subfamily, $\mathcal{P}_H \subseteq \mathcal{P}$, but it may be the case that the data indicates that the original family $\mathcal{P}$ is not rich enough to contain a family of distributions that matches the observational data.

Another way of describing the problem is to assume that we have a family of probability distributions $\mathcal{P} = \{P_\Theta\}$, where Θ may be some parameter in a real-valued parameter space Θ ("parametric inference"), or Θ may just be some index in an index set $\mathcal{I}$ to distinguish one distribution, P_{Θ_1}, from another, P_{Θ_2} ("nonparametric inference"). The parameter or the index is not observable; however, we assume $P_{\Theta_1} \neq P_{\Theta_2}$ if $\Theta_1 \neq \Theta_2$ a.s. (This assumption guarantees "identifiability". The almost sure condition is given so as not to exclude the possibility that Θ is a function.)

Often the observable variable of interest has associated covariates. The inference problems described above involve development of probability models that include the covariates. When there are covariates, there are two related problems of statistical inference. One is to use use a set of observed data to develop a model to *predict* some aspect of *future* observations for a given value of the covariates. Another is to develop a rule to "estimate" an unobservable random variable, given observed covariates. (A common instance of this latter problem is called "classification".)

How Does $\mathcal{P}_H$ Differ from $\mathcal{P}$?

What we know about Θ, or more to the point, what we assume about Θ determine some of the details of the inference procedures. We may, for example, assume that $\Theta = \theta$, some fixed but unknown real quantity. In that case, whether we view Θ as a parameter space or some more general index set, we may state our objective in statistical inference as to move from

$$\mathcal{P} = \{P_\theta \mid \theta \in \Theta\} \tag{3.1}$$

to

$$\mathcal{P}_H = \{P_\theta \mid \theta \in \Theta_H\} \tag{3.2}$$

by using observed data. On the other hand, we may assume that Θ is some Borel-measurable function. If Θ is a random variable, our interest may be in its distribution. In that case, the canonical problem in statistical inference, begins with a class of populations

$$\mathcal{P} = \{P_\Theta \mid \Theta \sim Q_0 \in \mathcal{Q}\}, \tag{3.3}$$

where Θ is a random variable and Q_0 is some "prior distribution", and, using observed data, arrives at the class of populations

$$\mathcal{P}_H = \{P_\Theta \mid \Theta \sim Q_H \in \mathcal{Q}\}, \tag{3.4}$$

where Q_H is some "posterior distribution" conditional on the observations.

As we mentioned above, the choice of $\mathcal{P}$, whether in the form of (3.1) or (3.3), is rather subjective. The form of equation (3.3) is "more subjective", in the sense that Q_0 allows direct incorporation of prior beliefs or subjective evidence. Statistical inference in the paradigm of equations (3.3) and (3.4) is sometimes referred to as "subjective inference", and is said to based on "subjective probability". We will consider this approach in more detail in Chapter 4.

Confidence, Significance, and Posterior Conditional Distributions

Statistical inference is a process of making decisions in the face of uncertainty. If there is no uncertainty, statistical inference is not a relevant activity. Given the uncertainty, the decision that is made may be "wrong".

Statistical inference must be accompanied by some quantification of how "likely" the decision is to be "correct". Exactly how this should be done is a very deep question whose answers may involve careful consideration of the philosophical foundations of statistical inference. In this book we will not get involved in these foundations. Rather, we will consider some specific ways of trying to come to grips with the question of quantifying the uncertainty in our decisions. Two ways this is done involve the paradigm of repetitive sampling from a stationary data-generating process, which leads to the concepts of

"confidence" and "significance". The formulation of the problem of statistical inference as in equations (3.3) and (3.4) avoids a direct confrontation of the question of how likely the decision is to be correct or incorrect. The conclusion in that kind of setup is the simple statement that the conditional distribution of Θ is Q_H, given that its marginal distribution is Q_0 and the conditional distribution of the data is P_Θ.

Examples

As an example of the approach indicated in equations (3.1) and (3.2), assume that a given sample $y_1, \ldots, y_n$ is taken in some prescribed manner from some member of a family of distributions

$$\mathcal{P} = \{\mathrm{N}(\mu, \sigma^2) \mid \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+\}.$$

Statistical inference in this situation may lead us to place the population giving rise to the observed sample in the family of distributions

$$\mathcal{P}_H = \{\mathrm{N}(\mu, \sigma^2) \mid \mu \in [\mu_1, \mu_2], \sigma^2 \in \mathbb{R}_+\}$$

(think confidence intervals!). The process of identifying the subfamily may be associated with various auxiliary statements (such as level of "confidence").

As another example, we assume that a given sample $y_1, \ldots, y_n$ is taken independently from some member of a family of distributions

$$\mathcal{P} = \{P \mid P \ll \lambda\},$$

where λ is the Lebesgue measure, and our inferential process may lead us to decide that the sample arose from the family

$$\mathcal{P}_H = \left\{P \mid P \ll \lambda \text{ and } \int_{-\infty}^{t} \mathrm{d}P = .5 \Rightarrow t \geq 0\right\}$$

(think hypothesis tests concerning the median!).

Notice that "P" in the example above is used to denote both a population and the associated probability measure; this is a notational convention that we adopted in Chapter 1 and which we use throughout this book.

Statistical inference following the setup of equations (3.3) and (3.4) can be illustrated by observable data that follows a Bernoulli distribution with a parameter Π which, in turn, has a beta marginal distribution with parameters α and β. (That is, in equation (3.3), Θ is Π, P_Π is a Bernoulli distribution with parameter Π, and Q_0 is a beta distribution with parameters α and β.) Given the single observation $X = x$, we can work out the conditional distribution of Π to be a beta with parameters $x + \alpha$ and $1 - x + \beta$.

It is interesting to note the evolutionary nature of this latter example. Suppose that we began with Q_0 (of equation (3.3)) being a beta with parameters $x_1 + \alpha$ and $1 - x_1 + \beta$, where x_1 is the observed value as described

above. (That is, the marginal "prior" is in the same parametric family of distributions as the original conditional "posterior".) Now, given the single observation $X = x_2$, we can work out the conditional distribution of Π to be a beta with parameters $x_1 + x_2 + \alpha$ and $2 - x_1 - x_2 + \beta$, which of course would be the same conclusion we would have reached if we had begun as originally described, but had observed a sample of size 2, $\{x_1, x_2\}$.

Covariates

Often a problem in statistical inference may involve other observable quantities. In that case, we may represent the family of distributions for an observable random variable Y (which, of course, may be a vector) in the form

$$\mathcal{P} = \{P_{\theta,x}\},$$

where $\theta \subseteq \mathbb{R}^d$ is an unobservable parameter or index and x is an observable concomitant variable.

There are two common types of interesting problems when we have observable covariates. In one case, we have a model of the form

$$Y = g(x;\theta) + E,$$

where g is a function of known form, and our objective is to make inferences concerning θ.

The other common problem is to predict or "estimate" Y for a given x. If we assume an additive model such as that above, predictive inference on Y is based on the inferences concerning θ. In other situations, such as in the problem of classification, the random variable Y represents some class of data, and given x, the objective is to "estimate" Y.

Asymptotic Considerations

Statistical inference is based on an observed sample of the random variable of interest along with observed values of any concomitant variable. Often the precision with which we can state conclusions depends on the dimensions of the parameter and any concomitant variables (that is on the structure of $\mathcal{P}$), as well as on the size of the sample. For any statistical procedure, it is of interest to know how the precision improves with increasing sample size.

Although we cannot have an infinite sample size, we often carry the mathematical analysis to the limit. This may give us an indication of how well the statistical methods perform with increasing sample sizes. Another reason that we often consider limiting cases is that asymptotic properties are often more mathematically tractable than properties for finite samples, and we use the asymptotic properties as approximations for the actual finite case.

Statistical Reasoning

We have indicated above that there may be a difference in inference regarding a data-generating process of interest and inference regarding an assumed probability distribution governing the data-generating process, or at least some aspects of that process.

Statistical inference leads us to select a subfamily of distributions. The particular subfamily may depend on how we formulate the inference problem and how we use a standard method of statistical inference. We must ensure that the formulation of the problem will lead to a reasonable conclusion.

An example in which the usual conclusions of a statistical inference procedure may not be justified is in common types of tests of statistical hypotheses. We must be careful in the specification of the alternative hypotheses. The two sample Wilcoxon statistic or the Mann-Whitney statistic is often used to test whether one population has a larger median than another population. While this test does have a connection with the medians, the procedure actually tests that the distributions of two populations are the same versus the alternative that a realization from one distribution is typically smaller (or larger) than a realization from the other distribution. If the distributions have quite different shapes, a typical value from the first population may tend to be smaller than a typical value from the second population, and the relationships between the medians may not affect the results of the test. See Example 5.22 for further discussion of these issues.

****

Data-generating process versus a given probability distribution

****

waiting time paradox

Renewal process paradox

****

An example in which clear statistical reasoning may not follow the lines of "common sense" is the Monte Hall problem. The name of this problem is derived from the host of a television game show that was first aired in 1963. While the problem itself does not involve statistical inference, it illustrates use of a probability model to make a decision.

****

Subjective Statistical Inference

In Chapter 1, I referred to the role of "beliefs" in applications of probability theory and the differing attitudes of "objectivists" and "subjectivists". The two different starting points of statistical inference expressed in equations (3.1) and (3.3) establish the different attitudes. These differences lead to differences in the fundamental ways in which statistical inference is approached. The differences involve how prior beliefs are incorporated into the inference process. The differences also lead to differences in how the concept

of probability (however *it* is interpreted) is used in the interpretation of the results of statistical analyses. In an objective approach, we may speak of the probability, given a data-generating process, of observing a given set of values. In a subjective approach, given a set of values, we may speak of the probability that the data-generating process has certain properties. This latter setting may lead to the use of the phrase "inverse probability", although this phrase is often used in different ways.

The objective and subjective approaches also differ in how the overall data-generating process affects the inferences on the underlying probability distribution P. (This difference is associated with the "likelihood principle", which I shall mention from time to time.) In this chapter I give a brief overview of statistical inference without much emphasis on whether the approach is "objective" or "subjective".

Ranks of Mathematical Objects

In statistical inference we deal with various types of mathematical objects. We would like to develop concepts and methods that are independent of the type of the underlying objects, but that is not always possible. Occasionally we will find it necessary to discuss scalar objects, rank one objects (vectors), and rank two objects (matrices) separately. In general, most degree-one properties, such as expectations of linear functions, can be considered uniformly across the different types of mathematical objects. Degree-two properties, such as variances, however, must usually be considered separately for scalars, vectors, and matrices.

Matrices often require special consideration because of the richness of that kind of structure. Sometimes we must consider the special cases of symmetric matrices, full-rank matrices, and positive-definite matrices.

3.1 Inferential Information in Statistics

In statistics, we generally assume that we have a *random sample* of observations $X_1, \ldots, X_n$ on a random variable X. We usually assume either that they are independent or that they are exchangeable, although we may assume other relations among the variables that depend on the *sampling design*.

We will often use X to denote a random sample on the random variable X. (This may sound confusing, but it is always clear from the context.) The common distribution of the variables is called the *parent distribution* of the random sample, and a common objective in statistics is to use the sample to make inferences about properties of the parent distribution.

In many statistical applications we also have covariates associated with the observations on the random variable of interest. In this case, in order to conform to common usage in statistics, I will use Y to represent the random variable of interest and X or x to represent the covariates or concomitant

variables. Both the random variable of interest and any covariates are assumed to be observable. There may also be unobservable variables, called "latent variables", associated with the observed data. Such variables are artifacts of the statistical model and may or may not correspond to phenomena of interest.

When covariates are present our interest usually is in the conditional distribution of Y, given X. For making statistical inferences, we generally assume that the conditional distributions of $Y_1|X_1, \ldots, Y_n|X_n$ are either conditionally independent or at least conditionally exchangeable.

A *statistic* is any function T of the observables that does not involve any unobservable values. We often use a subscript T_n to indicate the number of observations, but usually a statistic is defined as some formula that applies to a general number of observations (such as the sample mean). While we most often work with statistics based on a random sample, that is, an iid set of variables, or at least based on an exchangeable sample, we may have a statistic that is a function of a general set of random variables, $X_1, \ldots, X_n$. We see that if the random variables are exchangeable, then the statistic is *symmetric*, in the sense that $T(X_{k_1}, \ldots, X_{k_n}) = T(X_1, \ldots, X_n)$ for any indices $k_1, \ldots, k_n$ such that $\{k_1, \ldots, k_n\} = \{1, \ldots, n\}$.

Statistical Models

We assume that the sample arose from some distribution P_θ, which is a member of some family of probability distributions $\mathcal{P}$. The family of probability distributions $\mathcal{P}$ is a *statistical model*. We fully specify the family $\mathcal{P}$ (it can be a very large family), but we assume some aspects of P_θ are unknown. (If the distribution P_θ that yielded the sample is fully known, while there may be some interesting questions about probability, there are no interesting statistical questions.) Our objective in statistical inference is to determine a specific $P_\theta \in \mathcal{P}$, or some subfamily $\mathcal{P}_\theta \subseteq \mathcal{P}$, that could likely have generated the sample.

The distribution may also depend on other observable variables. In general, we assume we have observations $Y_1, \ldots, Y_n$ on Y, together with associated observations on any related variable X or x. We refer to the associated variables as "covariates". In this context, a *statistic*, which in our common use of the term is a function that does not involve any unobserved values, may also involve the observed covariates.

A general statistical model that includes covariates is

$$Y = f(x\,;\,\theta) + E, \tag{3.5}$$

where Y and x are observable variables, f is some unknown function, θ is an unknown parameter, and E is an unobservable random variable with unknown distribution P_τ independent of other quantities in the model. In the usual setup, Y is a scalar random random variable, and x is a p-vector. Given independent observations $(Y_1, x_1), \ldots, (Y_n, x_n)$, we often use the notation Y

to represent an n-vector, X to represent an $n \times p$ matrix whose rows are the x_i^{T}, and E to represent an n-vector of iid random variables. A model such as (3.5) is often called a regression model.

A common statistical model that expresses a relationship of an observable random variable and other observable variables is the linear model

$$Y = \beta^{\mathrm{T}} x + E, \tag{3.6}$$

where Y is the observable random variable, x is an observable p-vector of covariates, β is an unknown and unobservable p-vector of parameters, and E is an unobservable random variable with $\mathrm{E}(E) = 0$ and $\mathrm{V}(E) = \sigma^2 I$. The parameter space for β is $B \subseteq \mathbb{R}^p$.

A random sample may be written in the vector-matrix form

$$Y = X\beta + E, \tag{3.7}$$

where Y and E are n-vectors, X is an $n \times p$ matrix whose rows are the x_i^{T}, and β is the p-vector above. (The notation "$\beta^{\mathrm{T}} x$" in equation (3.6) and "$X\beta$" in equation (3.7) is more natural in the separate contexts. All vectors are considered to be column vectors. We could of course write "$x^{\mathrm{T}}\beta$" in equation (3.6).)

Because the linear model is familiar from applications of statistics, we will refer to it from time to time, but we will not study it systematically until Section 5.5.1.

In statistical inference, we distinguish observable random variables and "parameters", but we are not always careful in referring to parameters. We think of two kinds of parameters; "known" and "unknown". A statistic is a function of observable random variables that does not involve any unknown parameters.

Algorithmic Statistical Models

There are various types of models that may have different purposes. A common form of a model is a mathematical equation or a system of equations. If the purpose of the model is to enhance the understanding of some phenomenon, there would be a large premium on simplicity of the model. If the model is very complicated, it may correspond very well to the reality being studied, but it is unlikely to be understandable. If its primary purpose is to aid understanding, an *equation model* should be relatively simple..

A model may be embedded in a computer program. In this case, the model itself is not ordinarily scrutinized; only its input and output are studied. The complexity of the model is not of essential consequence. Especially if the objective is prediction of a response given values of the associated variables, and if there is a large premium on making accurate predictions or classifications in a very short time, an *algorithmic model* may be appropriate. An algorithmic model prioritizes prediction accuracy. The details of the model may be very different from the details of the data-generating process being modeled. That

is not relevant; the important thing is how well the output of the algorithmic model compares to the output of the data-generating process being modeled when they are given the same input.

The asymmetric relationship between a random variable Y and a variable x may be represented as a black box that accepts x as input and outputs Y:

$$Y \leftarrow \boxed{\text{unknown process}} \leftarrow x. \tag{3.8}$$

The relationship might also be described by a statement of the form

$$Y \leftarrow f(x)$$

or

$$Y \approx f(x). \tag{3.9}$$

Asymmetry of Statistical Models; Systematic and Random Components

If f has an inverse, the model (3.9) appears symmetric. Even in that case, however, there is an asymmetry that results from the role of random variables in the model. We model the response as a random variable and our methods of analysis would not apply to the model

$$x \approx f^{-1}(Y).$$

We may think of $f(x)$ as a *systematic* effect and write the model with an additive adjustment, or error, as

$$Y = f(x) + E \tag{3.10}$$

or with a multiplicative error as

$$Y = f(x)\Delta, \tag{3.11}$$

where E and Δ are assumed to be random variables. (The "E" is the Greek uppercase epsilon.) We refer to these as "errors", although this word does not indicate a mistake. In additive models, E is also called the "residual". The model therefore is composed of a *systematic component* related to the values of x and a *random component* that accounts for the indeterminacy between Y and $f(x)$.

An objective in statistical analysis often is to understand the systematic and random components better. The relative contribution to the variability in the observed Y due to the systematic component and due to the random component is called the *signal to noise ratio*. (Notice that this is a nontechnical term here; we could quantify it more precisely in certain classes of models. We can view an F ratio in analysis of variance as a type of quantification of a signal to noise ratio.)

An additive model has the advantage of separability of first order expectations of the two components no matter what assumptions are made about joint probability distributions of elements of the one component and those of the other. Note a questionable requirement for this separability: the variance of the residual component must be constant no matter what the magnitude of the expectation of the systematic component. Despite these issues, in the following, we will concentrate on models with additive random effects.

In the case of the black-box model (3.8), both the systematic and random components are embedded in the box. The objectives of statistical analysis may be to identify the individual components or, more often, to determine "average" or "most likely" output Y for given input x.

Because the functional form f of the relationship between Y and x may contain a *parameter*, we may write the equation in the model as

$$Y = f(x; \theta) + E, \tag{3.12}$$

where θ is a parameter whose value determines a specific relationship within the family specified by f. In most cases, θ is a vector. In the usual linear regression model, for example, the parameter is a vector with two more elements than the number of elements in x,

$$Y = \beta_0 + x^{\mathrm{T}}\beta + E, \tag{3.13}$$

where $\theta = (\beta_0, \beta, \sigma^2)$.

3.1.1 Statistical Inference: Point Estimation

Statistical inference is a process of identifying a family of distributions that generated a given set of observations. The process begins with an assumed family of distributions $\mathcal{P}$. This family may be very large; for example, it may be the family of distributions with probability measures dominated by the counting measure. Often the assumed family is narrowly defined; for example, it may be the family of univariate normal distributions. In any event, the objective of statistical inference is to identify a subfamily, $\mathcal{P}_H \subseteq \mathcal{P}$, that contains the population from which the data arose.

Types of Statistical Inference

There are various types of inference related to the problem of determining the specific $P_\theta \in \mathcal{P}$ or else some specific element of θ. Some of these are *point estimation*, *hypothesis tests*, *confidence sets*, and *function estimation* (estimation of the PDF, for example). Hypothesis tests and confidence sets are associated with probability statements that depend on P_θ.

Beginning on page 12 and later in Chapter 2, we have distinguished families of probability distributions as either parametric or nonparametric. In

statistical inference, we also speak of methods as being either parametric or nonparametric. The meanings in this case are somewhat nebulous, however. We often refer to a method as "nonparametric" if it does not make use of specific properties of the assumed family of distributions. We also use the term "nonparametric" if the assumed family of distributions is very broad, or if it includes more than one of the standard parametric families.

In parametric settings, each of the types of inference listed above concerns a parameter, θ, in a parameter space, $\Theta \subseteq \mathbb{R}^k$. In this case, function estimation is essentially estimation of a parameter that determines the function.

In some cases whether the inference is "parametric" or "nonparametric" is not determined by the nature of the assumed family of probability distributions. Parametric statistical inference often may involve only some element of the parameter (for example, estimation only of the mean and not the variance), and so we may, effectively, perform parametric inference in a nonparametric family by an ad hoc definition of a "parameter", say the mean or the median of a distribution in the family.

In parametric inference about a parameter θ, we generally identify a Borel function, say g or h, of that parameter and then consider inference specifically on $g(\theta)$ or $h(\theta)$.

Whether the object of the statistical inference is a scalar or some other element of a vector space, often makes a difference in the simplicity of the inference procedures. We will often emphasize the "one-parameter" case because it is simpler. There are a large number of useful families of distributions that are indeed characterized by a single parameter, such as the binomial or the (one-parameter) exponential.

Two related problems in inference are *prediction* and *causal inference*. For either of these problems, in addition to the random variable Y with the probability triple $(\Omega, \mathcal{F}, P)$, we have a measurable function X that maps $(\Omega, \mathcal{F})$ to $(\Lambda, \mathcal{G})$. Our interest is in the probability measure on the $(\Lambda, \mathcal{G})$ space. This is the conditional distribution of $Y|X$. In either case, the strength of the inference depends on the extent of the difference in the conditional distribution of $Y|X$ and the marginal distribution of Y.

For prediction, given X, we wish to predict Y; that is, we seek a Borel function g such that $\mathrm{E}(g(X))$ is "close to" $\mathrm{E}(Y)$. We will discuss issues in prediction in Section 3.1.6.

In causal inference, as in prediction, we have an associated measurable function XY, but the values of this function may be said to "cause" the conditional distribution of $Y|X$ to be different from the marginal distribution of Y. The methods of causal inference invoke an intermediary variable that depends on X and on which Y depends. We will not consider causal inference in this text. It is a popular topic in research in the social "sciences"; see, for example, Morgan and Winship (2007).

The Basic Paradigm of Point Estimation

We will focus our attention in the remainder of this section on point estimation.

The setting for point estimation is a real-valued observable random variable X that has a distribution that depends in some way on a real-valued statistical function (in the sense of a distribution measure, as in Section 1.1.9). Often the statistical function can be viewed as a parameter θ that takes a value in the set Θ. We assume we have observations $X_1, \ldots, X_n$ on X.

The statistical function to be estimated is called the *estimand*. Although it may be an underlying natural parameter, it is often a Borel function of that parameter. We will use $g(\theta)$ to represent the estimand. (Some authors use the symbol ϑ for the function of θ that is the estimand.) We want to estimate $g(\theta)$ using observations on X. We denote the estimator as $T(X)$. We think of T as a rule or formula. We also use T to denote a decision in hypothesis testing. We also denote the rule as $\delta(X)$. In many cases, there is an observable covariate, and so the notation we have just adopted would be modified to indicate the presence of the covariate.

Approaches to Estimation

There are several approaches to estimation of $g(\theta)$. We generally assume that a specific value of $g(\theta)$ results in a specific distribution for X. If $\widehat{g(\theta)}$ is an estimate of $g(\theta)$ and we make the substitution $g(\theta) = \widehat{g(\theta)}$ we have a specific family of distributions with CDF $P_{\widehat{g(\theta)}}$, say.

In most cases we assume that the set Θ is open, and hence the range of $g(\theta)$ is open. We generally allow a point estimator to be in the closure of those spaces. For example, in the case of a binomial(n, π) distribution, the parameter space for π is usually taken to be $]0, 1[$; however, a "good" estimator of π may take the value 0 or 1. A common approach to estimation, called maximum likelihood estimation, explicitly requires that the estimator be in the closure of the parameter space instead of the parameter space itself. (See page 242.) Other "good" estimators may not even be in the closure of the parameter space; see Example 5.31 on page 436.

A good estimation scheme is one that specifies a distribution of X that corresponds in some sense to the observed values of X. We start on the problem by defining some computable, heuristic estimation procedure, and then analytically study the properties of that procedure under various scenarios, that is, under different assumed distributions.

Optimal Point Estimators

We seek an estimator with "good" properties. We will briefly discuss some desirable properties for estimators: small bias, small mean squared error, Pitman

closeness, and equivariance under transformations. In this section we consider these properties only in the context of estimation, but the same properties have meaning in other types of statistical inference, although the specific definition may be different. "Good" can be defined either in terms of various measures associated with an estimator (bias, mean squared error, Pitman closeness), or as a property that the estimator either possesses or does not (equivariance).

Bias

The *bias* of $T(X)$ for estimating $g(\theta)$ is

$$\mathrm{E}(T(X)) - g(\theta). \tag{3.14}$$

One of the most commonly required desirable properties of point estimators is *unbiasedness*.

Definition 3.1 (unbiased point estimator)
The *estimator* $T(X)$ *is unbiased for* $g(\theta)$ if

$$\mathrm{E}(T(X)) = g(\theta) \quad \forall \theta \in \Theta.$$

∎

Unbiasedness is a *uniform property* of the expected value.

We can also define other types of unbiasedness in terms of other aspects of a probability distribution. For example, an estimator whose median is the estimand is said to be *median-unbiased*.

Unbiasedness has different definitions in other settings (estimating functions, for example, see page 255) and for other types of statistical inference (for example, testing, see page 296, and determining confidence sets, see page 300), but the meanings are similar.

If two estimators are unbiased, we would reasonably prefer one with smaller variance.

Mean Squared Error

Another measure of the goodness of a scalar estimator is the mean squared error or MSE,

$$\mathrm{MSE}(T(x)) = \mathrm{E}((T(X) - g(\theta))^2), \tag{3.15}$$

which is the square of the bias plus the variance:

$$\mathrm{MSE}(T(x)) = (\mathrm{E}(T(X)) - g(\theta))^2 + \mathrm{V}(T(X)). \tag{3.16}$$

An example due to C. R. Rao (Rao, 1981) causes us to realize that we often face a dilemma in finding a good estimate. A "bad" estimator may have a smaller MSE than a "good" estimator.

Example 3.1 an unreasonable estimator with a smaller MSE
Suppose we have n observations $X_1, \ldots, X_n$ from a distribution with mean μ_1 and finite standard deviation σ. We wish to estimate μ_1. An obvious estimator is the sample mean $\overline{X}$. (We will see that this is generally a good estimator under most criteria.) The MSE of $\overline{X}$ is σ^2/n. Now, suppose we have m observations $Y_1, \ldots, Y_m$ from a different distribution with mean $\mu_2 = \mu_1 + \delta\sigma$ and the same standard deviation σ. Let

$$T = (n\overline{X} + m\overline{Y})/(n+m),$$

so we have

$$\mathrm{E}((T-\mu_1)^2) = \frac{\sigma^2}{n+m}\left(1 + \frac{m^2\delta^2}{n+m}\right).$$

Now if $\delta^2 < m^{-1} + n^{-1}$, then

$$\mathrm{MSE}(T) < \mathrm{MSE}(\overline{X});$$

that is, in this case, the MSE is improved by using spurious observations. If $\delta < 1$, just using a single spurious observation improves the MSE. ∎

Pitman Closeness

While the MSE gives us some sense of how "close" the estimator is to the estimand, another way of thinking about closeness is terms of the probability that $|T(X) - g(\theta)|$ is less than some small value ϵ:

$$\Pr\big(|T(X) - g(\theta)| < \epsilon \big| \theta\big) \quad \epsilon > 0. \tag{3.17}$$

This type of measure is called *Pitman closeness*.

Definition 3.2 (Pitman-closer; Pitman-closest)
Given two estimators $T_1(X)$ and $T_2(X)$ of $g(\theta)$, we say that $T_1(X)$ is *Pitman-closer* than $T_2(X)$, if

$$\Pr\big(|T_1(X) - g(\theta)| \le |T_2(X) - g(\theta)| \ \big|\theta\big) \ \ge \ \frac{1}{2} \tag{3.18}$$

for all $\theta \in \Theta$ and for some $\theta_0 \in \Theta$

$$\Pr\big(|T_1(X) - g(\theta)| < |T_2(X) - g(\theta)| \ \big|\theta_0\big) \ > \ \frac{1}{2}. \tag{3.19}$$

We say that $T_1(X)$ is the *Pitman-closest* estimator, if $T_1(X)$ is *Pitman-closer* than $T(X)$ for any other statistic $T(X)$. ∎

Pitman closeness is affected more by the properties of the distribution in the region of interest, rather than by the behavior of statistics in the tail regions. Measures such as MSE, for example, may be unduly affected by the properties of the distribution in the tail regions.

Example 3.2 Lack of transitivity of Pitman closeness
Although Pitman closeness is a useful measure in evaluating an estimator, the measure lacks the desirable property of transitivity; that is, $T_1(X)$ may be *Pitman-closer* than $T_2(X)$ and $T_2(X)$ *Pitman-closer* than $T_3(X)$, but yet $T_3(X)$ may be *Pitman-closer* than $T_1(X)$. It is easy to construct an example to illustrate that this is possible. Rather than trying to construct a realistic distribution and statistics, let us just consider three independent random variables T_1, T_2, and T_3 and assign probability distributions to them (following Blyth (1972)):

$\Pr(T_1 = 3) = 1.0$
$\Pr(T_2 = 1) = 0.4, \ \Pr(T_2 = 4) = 0.6$
$\Pr(T_3 = 2) = 0.6, \ \Pr(T_3 = 5) = 0.4$

We see that

$$\Pr(T_1 < T_2) = 0.6, \quad \Pr(T_2 < T_3) = 0.64, \quad \Pr(T_1 < T_3) = 0.4.$$

The point is that probability itself is not transitive over inequalities of random variables.

This example illustrates the fact that there is often no Pitman-closest estimator. ∎

Example 3.3 Pitman closeness of shrunken estimators
Efron (1975) gives an example of an otherwise "good" estimator that is not as close in the Pitman sense as a biased estimator.

Consider the problem of estimating the mean μ in a normal distribution $\mathrm{N}(\mu, 1)$, given a random sample $X_1, \ldots, X_n$. The usual estimator, the sample mean $\overline{X}$, is unbiased and has minimum variance among all unbiased estimators, so clearly it is a "good" estimator. Consider, however, the biased estimator

$$T(X) = \overline{X} - \Delta_n(\overline{X}), \tag{3.20}$$

where

$$\Delta_n(u) = \frac{\min(u\sqrt{n}, \ \Phi(-u\sqrt{n}))}{2\sqrt{n}}, \quad \text{for } u \geq 0, \tag{3.21}$$

in which $\Phi(\cdot)$ is the standard normal CDF. This "shrinkage" of $\overline{X}$ toward 0 yields an estimator that is Pitman-closer to the population mean μ than the sample mean $\overline{X}$. (Exercise.) ∎

On page 273, we will encounter a more dramatic example of the effect of shrinking the sample mean in a multivariate normal distributional model.

Equivariance

In Section 2.6, beginning on page 178, we discussed families of distributions that are characterized as equivalence classes under groups of transformations

of the random variables. In a parametric setting the group $\mathcal{G}$ of transformations of the random variable can be associated with a group $\overline{\mathcal{G}}$ of transformations of the parameter. Likewise, we consider a group of transformations on the estimator, $\mathcal{G}^*$. For $g \in \mathcal{G}$ and $g^* \in \mathcal{G}^*$ an estimator $T(X)$ is equivariant if

$$T(g(X)) = g^*(T(X)). \tag{3.22}$$

Some people use the terms "invariant" and "invariance" for equivariant and equivariance, but I prefer the latter terms unless, indeed there is *no change* in the statistical procedure.

For equivariant or invariant statistical procedures, there are issues that relate to other properties of the estimator that must be considered (see, for example, the discussion of L-invariance on page 266). We will discuss the equivariance property of statistical procedures in more detail in Section 3.4.

Uniform Properties

If the goodness of an estimator does not depend on the parameter, we say the estimator is *uniformly* good (and, of course, in this statement we would be more precise in what we mean by "good"). All discussions of statistical inference are in the context of some family of distributions, and when we speak of a "uniform" property, we mean a property that holds for all members of the family.

Unbiasedness, by definition, is a uniform property. We will see, however, that many other desirable properties cannot be uniform.

Statements of Probability Associated with Statistics

Although much of the development of inferential methods emphasizes the expected value of statistics, often it is useful to consider the probabilities of statistics being in certain regions. Pitman closeness is an example of the use of probabilities associated with estimators. Two other approaches involve the probabilities of various sets of values that the statistics may take on. These approaches lead to statistical *tests of hypotheses* and determination of *confidence sets*. These topics will be discussed in Section 3.5, and more thoroughly in later chapters.

3.1.2 Sufficiency, Ancillarity, Minimality, and Completeness

There are important properties of statistics, such as sufficiency and complete sufficiency, that determine the usefulness of those statistics in statistical inference. These general properties often can be used as guides in seeking optimal statistical procedures.

Sufficiency

The most fundamental concept in statistical inference is *sufficiency*, because it relates functions of observations to the object of the inference.

Definition 3.3 (sufficient statistic)
Let X be a sample from a population $P \in \mathcal{P}$. A statistic $T(X)$ is *sufficient* for $P \in \mathcal{P}$ if and only if the conditional distribution of X given T does not depend on P. ∎

In general terms, sufficiency involves the conditional independence from the parameter of the distribution of any other function of the random variable, given the sufficient statistic.

Sufficiency depends on the family of distributions, $\mathcal{P}$, wrt which the conditional expectation, which ultimately defines the conditional distribution, is defined. If a statistic is sufficient for $\mathcal{P}$, it may not be sufficient for a larger family, $\mathcal{P}_1$, where $\mathcal{P} \subseteq \mathcal{P}_1$.

Sufficiency determines the nature and extent of any reduction of data that can be made without sacrifice of information. Thus, a function of a sufficient statistic may not be sufficient, but if a sufficient statistic can be defined as a measurable function of another statistic, then that other statistic is necessarily sufficient (exercise).

We can establish sufficiency by the factorization criterion.

Theorem 3.1 (factorization criterion)
Let $\mathcal{P}$ be a family of distributions dominated by a σ-finite measure ν. A necessary and sufficient condition for a statistic T to be sufficient for $P \in \mathcal{P}$ is that there exist nonnegative Borel functions g_P and h, where h does not depend on P, such that

$$\mathrm{d}P/\mathrm{d}\nu(x) = g_P(T(x))h(x) \quad \text{a.e. } \nu. \tag{3.23}$$

Proof.
****
∎

If $X_1, \ldots, X_n$ are iid whose distribution is dominated by a σ-finite measure, the joint PDF of the order statistics given in equation (1.140) is the same as the joint distribution of all of the (unordered) random variables (see Exercise 1.46), we have the immediate and useful corollary to Theorem 3.1.

Corollary 3.1.1 (the order statistics are sufficient) *The order statistics of a random sample from a distribution dominated by a σ-finite measure are sufficient.*

Actually, the requirement that the distribution be dominated by a σ-finite measure is not necessary. The proof is a simple exercise in permutations.

If the family of distributions is characterized by a parameter θ, then instead of referring to a statistic as begin sufficient for the distribution, we may say that the statistic is sufficient for θ.

When the density can be written in the separable form $c(\theta)f(x)$, unless $c(\theta)$ is a constant, the support must be a function of θ, and a sufficient statistic for θ must be an extreme order statistic. When the support depends on the parameter, and that parameter does not in any other way characterize the distribution, then the extreme order statistic(s) at the boundary of the support determined by the parameter carry the full information about the parameter.

An important consequence of sufficiency in an estimation problem with convex loss is the Rao-Blackwell theorem (see Section 3.3.2).

The Sufficiency Principle

Sufficiency is such an important concept in statistical inference that it leads to a principle that often guides statistical decisions. The *sufficiency principle* states that if $T(X)$ is a sufficient statistic for P, and x_1 and x_2 are results of two independent experiments in P with

$$T(x_1) = T(x_2) \quad \text{a.e. } P, \tag{3.24}$$

then any decision about P based on one experiment should be in agreement with any decision about P based on the other experiment. Principles should, of course, be consistent with objectives. In the introductory section for this chapter, we stated that the objective in statistical inference is to make decisions about either the data-generating process or about P. The sufficiency principle cannot be consistent with an objective of understanding the data-generating process.

Ancillarity

Often a probability model contains a parameter of no interest for inference. Such a parameter is called a *nuisance parameter*. A statistic to be used for inferences about the parameters of interest should not depend on any nuisance parameter. This lack of dependence on a parameter is called ancillarity. Ancillarity is, in a way, the opposite of sufficiency.

Definition 3.4 (ancillary statistic; first-order ancillary statistic)
A statistic $U(X)$ is called *ancillary* for P (or θ) if the distribution of $U(X)$ does not depend on P (or θ). If $\mathrm{E}(U(X))$ does not depend on P (or θ), then $U(X)$ is said to be *first-order ancillary* for P (or θ). ∎

Restating the intuitive remark before the definition, we can say a statistic to be used for inferences about the parameters of interest should be *ancillary* for a nuisance parameter.

In a probability space $(\Omega, \mathcal{F}, P_\theta)$ and random variable X, if $U(X)$ is ancillary for P_θ, then the definition implies that

$$\sigma(U(X)) \subset \sigma(X), \tag{3.25}$$

and, further, for any set B,

$$P_\theta((U(X))^{-1}(B)) \text{ is constant.} \tag{3.26}$$

These facts merely state that the ancillary statistic provides no information about P_θ.

Minimal Sufficiency

While the whole sample is a sufficient statistic, sufficiency in this case is not very meaningful. We might more reasonably ask what, if any, statistics of lower dimension are also sufficient.

Definition 3.5 (minimially sufficient)
Let T be a given sufficient statistic for $P \in \mathcal{P}$. The statistic T is *minimal sufficient* if for any sufficient statistic for $P \in \mathcal{P}$, S, there is a measurable function h such that $T = h(S)$ a.s. $\mathcal{P}$. ∎

Minimal sufficiency has a heuristic appeal: it relates to the greatest amount of data reduction that is possible without losing information, in the sense of losing sufficiency.

When the range does not depend on the parameter, we can often establish minimality by use of one of the following two theorems.

Theorem 3.2 (minimal sufficiency I)
Let $\mathcal{P}$ be a family with densities $p_0, p_1, \ldots, p_k$, all with the same support. The statistic

$$T(X) = \left(\frac{p_1(X)}{p_0(X)}, \ldots, \frac{p_k(X)}{p_0(X)}\right) \tag{3.27}$$

is minimal sufficient.

Proof.
This follows from the following corollary of the factorization theorem. ∎

Corollary 3.1.1 (factorization theorem (page 222))
A necessary and sufficient condition for a statistic $T(X)$ to be sufficient for a family $\mathcal{P}$ of distributions of a sample X dominated by a σ-finite measure ν is that for any two densities p_1 and p_2 in $\mathcal{P}$, the ratio $p_1(x)/p_2(x)$ is a function only of $T(x)$.

Theorem 3.3 (minimal sufficiency II)
Let $\mathcal{P}$ be a family of distributions with the common support, and let $\mathcal{P}_0 \subseteq \mathcal{P}$. If T is minimal sufficient for $\mathcal{P}_0$ and is sufficient for $\mathcal{P}$, then T is minimal sufficient for $\mathcal{P}$.

Proof.
Consider any statistic U that is sufficient for $\mathcal{P}$. Then U must also be sufficient for $\mathcal{P}_0$, and since T is minimal sufficient for $\mathcal{P}_0$, T is a function of U. ∎

We can also establish minimal sufficiency by use of completeness, as we see below.

Completeness

A sufficient statistic T is particularly useful in a complete family or a boundedly complete family of distributions (see Section 2.1 beginning on page 162). In this case, for every Borel (bounded) function h that does not involve P,

$$\mathrm{E}_P(h(T)) = 0 \; \forall P \in \mathcal{P} \Rightarrow h(t) = 0 \text{ a.e. } \mathcal{P}.$$

Complete families are defined in terms of properties of any Borel function of a random variable that does not involve the particular distribution (that is, of a "statistic").

We are now interested in the completeness of a statistic, rather than the completeness of the family. A statistic can be complete even though the underlying family of distributions of the observables is not complete. (This is because of the definition of a complete family; see Example 2.1 on page 163.) The family of marginal distributions of the statistic itself must be complete.

We now give a definition of completeness for a statistic.

Definition 3.6 (complete statistic)
Given a family of distributions $\{P_\theta\}$ a statistic $T(X)$, where T is a nonconstant function, is said to be *complete* for $P \in \{P_\theta\}$ iff for any Borel function h that does not involve P

$$\mathrm{E}(h(T(X))) = 0 \; \forall P \in \mathcal{P} \Rightarrow h(T(x)) = 0 \text{ a.e. } \mathcal{P}.$$

∎

The weaker condition, "bounded completeness of a statistic", is defined in a similar manner, but only for bounded Borel functions h. A complete statistic is boundedly complete.

Completeness and sufficiency are different properties; either one can exist without the other. Sufficiency relates to a statistic and a sample. There is always a sufficient statistic: the sample itself. There may or may not be a complete statistic within a given family.

We are generally interested in statistics that are complete and sufficient.

Complete sufficiency depends on $\mathcal{P}$, the family of distributions wrt which E is defined. If a statistic is complete and sufficient with respect to $\mathcal{P}$, and if it is sufficient for $\mathcal{P}_\infty$, where $\mathcal{P} \subseteq \mathcal{P}_1$ and all distributions in $\mathcal{P}_\infty$ have common support, then it is complete and sufficient for $\mathcal{P}_1$, because in this case, the condition a.s. $\mathcal{P}$ implies the condition a.s. $\mathcal{P}_1$.

We can establish complete sufficiency by the exponential criterion.

Theorem 3.4 (exponential criterion)
Let $\mathcal{P}$ be a family of distributions dominated by a σ-finite measure ν. Given a statistic T suppose there exist Borel functions c and q and a nonnegative Borel function h, where h does not depend on P, such that

$$\mathrm{d}P/\mathrm{d}\nu(x) = \exp((q(P))^{\mathrm{T}}T(x) - c(P)h(x) \quad \text{a.e. } \nu. \tag{3.28}$$

A sufficient condition for the statistic T to be complete and sufficient for $P \in \mathcal{P}$ is that $q(P)$ contain an interior point.

Complete sufficiency is useful for establishing independence using Basu's theorem (see below), and in estimation problems in which we seek an unbiased estimator that has minimum variance uniformly (UMVUE, discussed more fully in Section 5.1).

It is often relatively straightforward to identify complete sufficient statistics in certain families of distributions, such as those in the exponential class; see Example 3.6. In a parametric-support family, there may be a complete statistic. If so, it is usually an extreme order statistic; see Example 3.7.

Theorem 3.5 (minimal sufficiency III)
If T is a complete statistic in $\mathcal{P}$ and T is sufficient, then T is minimal sufficient.

Proof. Exercise (follows from definitions). ∎

Complete sufficiency implies minimal sufficiency, but minimal sufficiency does not imply completeness, as we see in the following example.

Example 3.4 minimal sufficient but not complete sufficient
Consider a sample X of size 1 from $\mathrm{U}(\theta, \theta+1)$. Clearly, X is minimal sufficient. Any bounded periodic function $h(x)$ with period 1 that is not a.e. 0 serves to show that X is not complete. Let $h(x) = \sin(2\pi x)$. Then

$$\mathrm{E}(h(X)) = \int_{\theta}^{\theta+1} \mathrm{d}x = 0.$$

Clearly, however $h(X)$ is not 0 a.e., so X is not complete. We can see from this that there can be no complete statistic in this case. ∎

We will later define completeness of a class of statistics called decision rules, and in that context, define minimal completeness of the class.

Basu's Theorem

Complete sufficiency, ancillarity, and independence are related.

Theorem 3.6 (Basu's theorem)
Let $T(X)$ and $U(X)$ be statistics from the population P_θ in the family $\mathcal{P}$ If $T(X)$ is a boundedly complete sufficient statistic for $P_\theta \in \mathcal{P}$, and if $U(X)$ is ancillary for $P_\theta \in \mathcal{P}$, then T and U are independent.

Proof.
If U is ancillary for P_θ and A is any set, then $\Pr(U \in A)$ is independent of P_θ. Now, consider $p_A(t) = \Pr(U \in A | T = t)$. We have

$$\mathrm{E}_{P_\theta}(p_A(T)) = \Pr(U \in A);$$

and so by completeness,

$$p_A(T) = \Pr(U \in A) \ a.e.\mathcal{P}.$$

Hence U and T are independent. ∎

An interesting example shows the importance of completeness in Basu's theorem. This example also shows that minimality does not imply completeness.

Example 3.5 minimal sufficient but ancillary is not independent
Let $X_1, \ldots, X_n$, with $n \geq 2$, be iid as $\mathrm{U}(\theta - 1/2,\ \theta + 1/2)$. It is clear that $T = \{X_{(1)}, X_{(n)}\}$ is sufficient; in fact, it is minimal sufficient. Now consider $U = X_{(n)} - X_{(1)}$, which we easily see is ancillary. It is clear that T and U are not independent (U is a function of T).

If T were complete, then Basu's theorem would say that T and U are independent, but writing $U = h(T)$, where h is a measurable function, we can see that T is not complete (although it is minimal). ∎

Sufficiency, Minimality, and Completeness in Various Families

We can use general properties of specific families of distributions to establish properties of statistics quickly and easily.

Complete sufficiency is often easy to show in exponential family.

Example 3.6 complete sufficient statistics in a normal distribution
Consider the normal family of distributions with parameter $\theta = (\mu, \sigma^2)$. Suppose we have observations $X_1, X_2, \ldots, X_n$. Let $T_1 = (\sum X_i, \sum X_i^2)$. Then $T_1(X)$ is sufficient and complete for θ. (Exercise)

Now, let $T_2 = (\overline{X}, S^2)$, where $\overline{X}$ and S^2 are respectively the sample mean and sample variance (equations (1.32) and (1.33)). Then $T_2(X)$ is also sufficient and complete for θ. (Exercise)

We have seen in Section 2.9.3 that $\overline{X}$ and S^2 are independent and we worked out their distributions there. ∎

Complete sufficiency is also often easy to show in distributions whose range depends on θ in a simple way. (We can relate any such range-dependent distribution to $\mathrm{U}(\theta_1, \theta_2)$.)

In general, proof of sufficiency is often easy, but proof of minimality or completeness is often difficult. We often must rely on the awkward use of the definitions of minimality and completeness. Completeness of course implies minimality.

Example 3.7 complete sufficient statistics in a uniform distribution
Consider the uniform family of distributions with parameter θ that is the upper bound of the support, $\mathrm{U}(0, \theta)$. Suppose we have observations $X_1, \ldots, X_n$. Then $T(X) = X_{(n)}$ is complete sufficient for θ. (Exercise) ∎

Parametric-support families (or "truncation families") have simple range dependencies. A distribution in any of these families has a PDF in the general form

$$f_\theta(x) = c(\theta)g(x)\mathrm{I}_{S(\theta)}(x).$$

The most useful example of distributions whose support depends on the parameter is the uniform $\mathrm{U}(0,\theta)$, as in Example 3.7. Many other distributions can be transformed into this one. For example, consider $X_1,\ldots,X_n$ iid as a shifted version of the standard exponential family of distributions with Lebesgue PDF

$$\mathrm{e}^{-(x-\alpha)}\mathrm{I}_{]\alpha,\infty[}(x),$$

and $Y_i = \mathrm{e}^{-X_i}$ and $\theta = \mathrm{e}^{-\alpha}$, then $Y_1,\ldots,Y_n$ are iid $\mathrm{U}(0,\theta)$; hence if we can handle one problem, we can handle the other. We can also handle distributions like $\mathrm{U}(\theta_1,\theta_2)$ a general shifted exponential, as well as some other related distributions, such as a shifted gamma.

We can show *completeness* using the fact that

$$\int_A |g|\,\mathrm{d}\mu = 0 \iff g = 0 \text{ a.e. on } A. \tag{3.29}$$

Another result we often need in going to a multiparameter problem is Fubini's theorem.

The sufficient statistic in the simple univariate case where $S(\theta) = (\theta_1,\theta_2)$ is $T(X) = (X_{(1)}, X_{(n)})$, as we can see from the the factorization theorem by writing the joint density of a sample as

$$c(\theta)g(x)\mathrm{I}_{]x_{(1)},x_{(n)}[}(x).$$

For example, for a distribution such as $\mathrm{U}(0,\theta)$ we see that $X_{(n)}$ is sufficient by writing the joint density of a sample as

$$\frac{1}{\theta}\mathrm{I}_{]0,x_{(n)}[}.$$

Example 3.8 complete sufficient statistics in a two-parameter exponential distribution

In Examples 1.11 and 1.18, we considered a shifted version of the exponential family of distributions, called the two-parameter exponential with parameter (α,θ). The Lebesgue PDF is

$$\theta^{-1}\mathrm{e}^{-(x-\alpha)/\theta}\mathrm{I}_{]\alpha,\infty[}(x)$$

Suppose we have observations $X_1, X_2, \ldots, X_n$.

In Examples 1.11 and 1.18, we worked out the distributions of $X_{(1)}$ and $\sum X_i - nX_{(1)}$. Now, we want to show that $T = (X_{(1)}, \sum X_i - nX_{(1)})$ is sufficient and complete for (α,θ).

****

The properties of a specific family of distributions are useful in identifying optimal methods of statistical inference. Exponential families are particularly useful for finding UMVU estimators. We will discuss UMVU estimators more fully in Section 5.1. A group family is useful in identifying equivariant and invariant statistical procedures. We will discuss procedures of these types in Section 3.4.

3.1.3 Information and the Information Inequality

The term "information" is used in various ways in statistical inference. In general, information relates to the variability in the probability distribution or to the variability in a random sample.

A common type of information is Shannon information, which for an event is the negative of the log of the probability of the event; see page 41. In this view, an observed event that is less probable than another event provides more information than that more probable event.

In parametric families of probability distributions, we also use the term "information" in another sense that relates to the extent of the difference between two PDFs in the same family, but with difference values of the parameters. This kind of information, called Fisher information, is measured by taking derivatives with respect to the parameters.

A fundamental question is how much information does a realization of the random variable X contain about the scalar parameter θ.

If a random variable X has a PDF $f(x;\theta)$ wrt a σ-finite measure that is differentiable in θ, the rate of change of the PDF at a given x with respect to different values of θ intuitively is an indication of the amount of information x provides. If the support of the random variable, however, depends on θ, that derivative may not be so useful. Let us restrict our attention to distributions that satisfy the first two Fisher information regularity conditions we defined on on page 168 for a family of distributions $\mathcal{P} = \{P_\theta; \theta \in \Theta\}$ that have densities f_θ:

- The parameter space Θ is real and connected and contains an open set (in one dimension, it is an interval with positive measure).
- For any x in the support and $\theta \in \Theta^\circ$, $\partial f_\theta(x)/\partial\theta$ exists and is finite.

For such distributions, we define the "information" (or "Fisher information") that X contains about θ as

$$I(\theta) = \mathrm{E}_\theta\left(\left(\frac{\partial \log f(X;\theta)}{\partial\theta}\right)\left(\frac{\partial \log f(X;\theta)}{\partial\theta}\right)^{\mathrm{T}}\right). \tag{3.30}$$

Information is larger when there is larger relative variation in the density as the parameter changes, but the information available from an estimator is less when the estimator exhibits large variation (i.e., has large variance), so we want to use statistics with smaller variance.

The third Fisher information regularity condition guarantees that integration and differentiation can be interchanged.

- The support is independent of θ; that is, all P_θ have a common support.

In Fisher information regular families, we have

$$\begin{aligned}\mathrm{E}\left(\frac{\partial \log(f(X,\theta))}{\partial\theta}\right) &= \int \frac{1}{f(x,\theta)}\frac{\partial f(x,\theta)}{\partial\theta} f(x,\theta)\mathrm{d}x \\ &= \frac{\partial}{\partial\theta}\int f(x,\theta)\mathrm{d}x \\ &= 0; \end{aligned} \tag{3.31}$$

therefore, the expectation in the information definition (3.30) is also the variance of $\partial \log(f(X,\theta))/\partial\theta$:

$$I(\theta) = \mathrm{V}\left(\frac{\partial \log(f(X,\theta))}{\partial\theta}\right). \tag{3.32}$$

If the second derivative with respect to θ also exists for all x and θ, and if it can be obtained by differentiation twice under the integral sign in (3.31), then we also have a relationship with the second derivative:

$$\begin{aligned} I(\theta) &= \mathrm{E}\left(\left(\frac{\partial \log f(X;\theta)}{\partial\theta}\right)\left(\frac{\partial \log f(X;\theta)}{\partial\theta}\right)^{\mathrm{T}}\right) \\ &= -\mathrm{E}\left(\frac{\partial^2 \log(f(X,\theta))}{\partial\theta^2}\right). \end{aligned} \tag{3.33}$$

We see this by writing

$$\begin{aligned}\frac{\partial^2 \log f(X;\theta)}{\partial\theta^2} &= \frac{\frac{\partial^2 f(X;\theta)}{\partial\theta^2}}{f(X;\theta)} - \frac{\left(\frac{\partial f(X;\theta)}{\partial\theta}\right)\left(\frac{\partial f(X;\theta)}{\partial\theta}\right)^{\mathrm{T}}}{(f(X;\theta))^2} \\ &= \frac{\frac{\partial^2 f(X;\theta)}{\partial\theta^2}}{f(X;\theta)} - \left(\frac{\partial \log f(X;\theta)}{\partial\theta}\right)\left(\frac{\partial \log f(X;\theta)}{\partial\theta}\right)^{\mathrm{T}}\end{aligned}$$

and taking the expectation of both sides, noting that the first term on the right is zero as before (after again interchanging differentiation and integration).

Example 3.9 Fisher information in a normal distribution
Consider the $\mathrm{N}(\mu,\sigma^2)$ distribution with $\theta = (\mu,\sigma)$ (which is simpler than for $\theta = (\mu,\sigma^2)$):

$$\log f_{(\mu,\sigma)}(x) = c - \log(\sigma) - (x-\mu)^2/(2\sigma^2).$$

We have

$$\frac{\partial}{\partial \mu} \log f_{(\mu,\sigma)}(x) = \frac{x-\mu}{\sigma^2}$$

and

$$\frac{\partial}{\partial \sigma} \log f_{(\mu,\sigma)}(x) = -\frac{1}{\sigma} + \frac{(x-\mu)^2}{\sigma^3},$$

so

$$\begin{aligned} I(\mu,\sigma) &= \mathrm{E}_\theta\left(\left(\frac{\partial \log f(X;\theta)}{\partial \theta}\right)\left(\frac{\partial \log f(X;\theta)}{\partial \theta}\right)^{\mathrm{T}}\right) \\ &= \mathrm{E}_{(\mu,\sigma)}\left(\begin{bmatrix} \frac{(X-\mu)^2}{(\sigma^2)^2} & \frac{X-\mu}{\sigma^2}\left(-\frac{1}{\sigma}+\frac{(x-\mu)^2}{\sigma^3}\right) \\ \frac{x-\mu}{\sigma^2}\left(-\frac{1}{\sigma}+\frac{(X-\mu)^2}{\sigma^3}\right) & +\left(-\frac{1}{\sigma}+\frac{(X-\mu)^2}{\sigma^3}\right)^2 \end{bmatrix}\right) \\ &= \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{bmatrix}. \end{aligned}$$

The normal is rather unusual among common multiparameter distributions in that the information matrix is diagonal.

Notice that the Fisher information matrix is dependent on the parametrization. The parametrization of the normal distribution in either the canonical exponential form or even $\theta = (\mu, \sigma^2)$ would result in a different Fisher information matrix (see Example 5.11 on page 400). ∎

Example 3.10 Fisher information in a gamma distribution
Consider the gamma(α, β) distribution. We have for $x > 0$

$$\log f_{(\alpha,\beta)}(x) = -\log(\Gamma(\alpha)) - \alpha \log(\beta) + (\alpha - 1)\log(x) - x/\beta.$$

This yields the Fisher information matrix

$$I(\theta) = \begin{bmatrix} \psi'(\alpha) & \frac{1}{\beta} \\ \frac{1}{\beta} & \frac{\alpha^2}{\beta^2} \end{bmatrix},$$

where $\psi(\alpha)$ is the digamma function, $\mathrm{d}\log(\Gamma(\alpha))/\mathrm{d}\alpha$, and $\psi'(\alpha)$ is the trigamma function, $\mathrm{d}\psi(\alpha)/\mathrm{d}\alpha$.

In the natural parameters, $\alpha - 1$ and $1/\beta$, obviously the Fisher information would be different. (Remember, derivatives are involved, so we cannot just substitute the transformed parameters in the information matrix.) ∎

Fisher Information in Families in the Exponential Class

Consider the general canonical exponential form for a distribution in the exponential class:

$$f_\theta(x) = \exp\left((\eta^{\mathrm{T}} T(x) - \zeta(\eta)\right) h(x)$$

(see page 173). If η is the mean-value parameter (see equation (2.10)), then

$$I(\theta) = V^{-1},$$

where

$$V = \mathrm{V}(T(X)).$$

********** prove this

Example 3.11 Fisher information in a beta distribution
Consider the beta(α, β) distribution. We have for $x \in]0, 1[$

$$\log f_{(\alpha,\beta)}(x) = \log(\Gamma(\alpha+\beta)) - \log(\Gamma(\alpha)) - \log(\Gamma(\beta)) + (\alpha-1)\log(x) + (\beta-1)\log(1-x).$$

This yields the Fisher information matrix

$$I(\theta) = \begin{bmatrix} \psi'(\alpha) - \psi'(\alpha+\beta) & -\psi'(\alpha+\beta) \\ -\psi'(\alpha+\beta) & \psi'(\beta) - \psi'(\alpha+\beta) \end{bmatrix}.$$

***** Show that this follows from above. ∎

Fisher Information in Location-Scale Families

The Fisher information for the two parameters $\theta = (\mu, \sigma)$ in a location-scale family with Lebesgue PDF

$$\frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$$

has a particularly simple form:

$$I(\theta) = \frac{n}{\sigma^2} \begin{bmatrix} \int \left(\frac{f'(x)}{f(x)}\right)^2 f(x)\mathrm{d}x & \int x \left(\frac{f'(x)}{f(x)}\right)^2 f(x)\mathrm{d}x \\ \int x \left(\frac{f'(x)}{f(x)}\right)^2 f(x)\mathrm{d}x & \int \left(\frac{xf'(x)}{f(x)} + 1\right)^2 f(x)\mathrm{d}x \end{bmatrix}. \tag{3.34}$$

The prime on $f'(x)$ indicates differentiation with respect to x of course. (The information matrix is defined in terms of differentiation with respect to the parameters followed by an expectation.)

Another expression for the information matrix for a location-scale family is

$$I(\theta) = \frac{n}{\sigma^2} \begin{bmatrix} \int \frac{(f'(x))^2}{f(x)}\mathrm{d}x & \int \frac{f'(x)\,(xf'(x)+f(x))}{f(x)}\mathrm{d}x \\ \int \frac{f'(x)\,(xf'(x)+f(x))}{f(x)}\mathrm{d}x & \int \frac{(xf'(x)+f(x))^2}{f(x)}\mathrm{d}x \end{bmatrix}. \tag{3.35}$$

(This is given in a slightly different form in Example 3.9 of MS2, which is Exercise 3.34, which is solved in his *Solutions*, using the form above, which is

a more straightforward expression from the derivation that begins by defining the function $g(\mu, \sigma, x) = \log(f((x-\mu)/\sigma)/\sigma)$, and then proceeding with the definition of the information matrix.)

Also we can see that in the location-scale family, if it is symmetric about the origin (that is about μ), the covariance term is 0.

The Information Inequality

For distributions satisfying the Fisher information regularity conditions we have the information inequality that relates the variance of a statistic to the Fisher information about the parameters. In the following we will consider a scalar statistic, $T(X)$, and a scalar function of θ, $g(\theta)$.

Following TPE2, we first give two lemmas for an unbiased estimator of $g(\theta)$.

Lemma 3.7.1
Given the Borel function $g(\theta)$ and statistics $T(X)$ and $S(X)$ such that $\mathrm{E}(T(X)) = g(\theta)$. A necessary and sufficient condition for $\mathrm{Cov}(T(X), S(X))$ to depend on θ only through $g(\theta)$ is that for all θ

$$\mathrm{Cov}_\theta(U, S(X)) = 0 \qquad \forall U \ni \mathrm{E}_\theta(U) = 0, \mathrm{E}_\theta(U^2) < \infty. \tag{3.36}$$

Proof.
We only need to show that for any $T_1(X)$ and $T_2(X)$ with $\mathrm{E}(T_1(X)) = \mathrm{E}(T_2(X)) = g(\theta)$, $\mathrm{Cov}_\theta(T_1(X), S(X)) = \mathrm{Cov}_\theta(T_2(X), S(X))$. We have

$$\begin{aligned}\mathrm{Cov}_\theta(T_1(X), S(X)) - \mathrm{Cov}_\theta(T_2(X), S(X)) &= \mathrm{Cov}_\theta(T_1(X) - T_2(X), S(X)) \\ &= \mathrm{Cov}_\theta(U, S(X)).\end{aligned}$$

We have therefore $\mathrm{Cov}_\theta(T_1(X), S(X)) = \mathrm{Cov}_\theta(T_2(X), S(X))$ for any $T_1(X)$ and $T_2(X)$ if and only if $\mathrm{Cov}_\theta(U, S(X)) = 0$ for all U as in equation (3.36). ∎

A comment about notation may be in order here. First, we have been writing $T(X)$ and $S(X)$ to emphasize the common random variable in the statistics. A simpler notation may be T and S. A more complicated notation would be $T(X, \theta)$ and $S(X, \theta)$ to emphasize the dependence of the distribution of T and S on θ, just as we have written $f(X, \theta)$ for the PDF above. In the next lemma, we will consider a vector of functions $S(X, \theta) = (S_i(X, \theta))$. As usual, this is a column vector, and so is $(\mathrm{Cov}(T(X), S_i(X))$, for example.

Lemma 3.7.2
Let $T(X)$ be an unbiased estimator of $g(\theta)$ and let $S(X, \theta) = (S_i(X, \theta))$, where the $S_i(X, \theta)$ are any stochastically independent functions with finite second moments. Then

$$\mathrm{V}(T(X)) \geq (\mathrm{Cov}(T(X), S_i(X)))^{\mathrm{T}} \ (\mathrm{V}(S))^{-1} \ \mathrm{Cov}(T(X), S_i(X)). \tag{3.37}$$

Proof.
Let $a_1, \ldots, a_k$ be any constants. From the covariance inequality for scalar random variables Y and Z,

$$\mathrm{V}(Y) \geq \frac{(\mathrm{Cov}(Y,Z))^2}{\mathrm{V}(Z)},$$

we have

$$\mathrm{V}(T(X)) \geq \frac{(\mathrm{Cov}(T(X), \sum_i a_i S_i(X)))^2}{\mathrm{V}(\sum_i a_i S_i(X))}. \tag{3.38}$$

Rewriting, we have,

$$\mathrm{Cov}(T(X), \sum_i a_i S_i(X)) = a^{\mathrm{T}} \mathrm{Cov}(T(X), S_i(X))$$

and

$$\mathrm{V}\left(\sum_i a_i S_i(X)\right) = a^{\mathrm{T}} \mathrm{V}(S) a.$$

Because (3.38) is true for any a, this yields

$$\mathrm{V}(T(X)) \geq \max_a \frac{\left(a^{\mathrm{T}} \mathrm{Cov}(T(X), S_i(X))\right)^2}{a^{\mathrm{T}} \mathrm{V}(S) a}.$$

Noting that

$$\left(a^{\mathrm{T}} \mathrm{Cov}(T(X), S_i(X))\right)^2 = a^{\mathrm{T}} \mathrm{Cov}(T(X), S_i(X)) (\mathrm{Cov}(T(X), S_i(X))^{\mathrm{T}} a,$$

from equation (0.3.17) on page 788, we have

$$\mathrm{V}(T(X)) \geq (\mathrm{Cov}(T(X), S_i(X)))^{\mathrm{T}} \ (\mathrm{V}(S))^{-1} \ \mathrm{Cov}(T(X), S_i(X)).$$

∎

Theorem 3.7 (information inequality)
Assume that the Fisher information regularity conditions hold for the distribution with PDF $f(x;\theta)$ and that $I(\theta)$ is positive definite, where θ is a k-vector. Let $T(X)$ be any scalar statistic with finite second moment, and assume for $i = 1, \ldots, k$, that $(\partial/\partial\theta_i)\mathrm{E}_\theta(T(X))$ exists and can be obtained by differentiating under the integral sign. Then $\mathrm{E}_\theta((\partial/\partial\theta_i)\log(f(x;\theta))) = 0$ and

$$\mathrm{V}(T(X)) \geq \left(\frac{\partial}{\partial\theta}\mathrm{E}(T(\theta))\right)^{\mathrm{T}} (I(\theta))^{-1} \frac{\partial}{\partial\theta}\mathrm{E}(T(\theta)). \tag{3.39}$$

Proof. Take the functions S_i in Lemma 3.7.2 to be $(\partial/\partial\theta_i)\log(f(x;\theta))$. ∎

An alternate direct proof of Theorem 3.7 can be constructed using equations (3.31) through (3.33) and the covariance inequality.

The right side of inequality (3.39) is called the information or the Cramér-Rao lower bound (CRLB). The CRLB results from the covariance inequality. The proof of the CRLB is an "easy piece" that every student should be able to provide quickly. (Notice that for the one-parameter case, this is Corollary B.5.1.7 of Hölder's inequality on page 852.)

This inequality plays a very important role in unbiased point estimation, as we will see in Section 5.1.5 on page 399.

3.1.4 "Approximate" Inference

When the exact distribution of a statistic is known (based, of course, on an assumption of a given underlying distribution of a random sample), use of the statistic for inferences about the underlying distribution is called exact inference.

Often the exact distribution of a statistic is not known, or is too complicated for practical use. In that case, we may resort to approximate inference. There are basically three types of approximate inference.

One type occurs when a simple distribution is very similar to another distribution. For example, the Kumaraswamy distribution with PDF

$$p(x) = \alpha\beta x^{\alpha-1}(1-x^{\alpha})^{\beta-1}\mathrm{I}_{[0,1]}(x) \tag{3.40}$$

may be used as an approximation to the beta distribution.

Another type of approximate inference, called computational inference, is used when an unknown distribution can be simulated by resampling of the given observations.

Asymptotic inference is probably the most commonly used type of approximate inference. In asymptotic approximate inference we are interested in the properties of T_n as the sample size increases. We focus our attention on the sequence $\{T_n\}$ for $n = 1, 2, \ldots$, and, in particular, consider the properties of $\{T_n\}$ as $n \to \infty$. We discuss asymptotic inference in more detail in Section 3.8.

3.1.5 Statistical Inference in Parametric Families

A real-valued observable random variable X has a distribution that may depend in some way on a real-valued parameter θ that takes a value in the set Θ, called the *parameter space*. This random variable is used to model some observable phenomenon.

As the parameter ranges over Θ it determines a family of distributions, $\mathcal{P}$. We denote a specific member of that family as P_θ for some fixed value of θ.

We often want to make inferences about the value of θ or about some function or transformation of an underlying parameter θ. To generalize our object of interest, we often denote it as ϑ, or $g(\theta)$ or $g(\theta; z)$, where g is some Borel function, and z may represent auxiliary data.

Summary of Sufficient Statistics and Their Distributions for Some Common Parametric Families

3.1.6 Prediction

In addition to the three different types of inference discussed in the preceding sections, which were related to the problem of determining the specific $P_\theta \in \mathcal{P}$, we may also want to *predict* the value that a random variable will realize.

In the prediction problem, we have a random variable Y with the probability triple $(\Omega, \mathcal{F}, P)$ and a measurable function X that maps $(\Omega, \mathcal{F}, P)$ to $(\Lambda, \mathcal{G})$. Given an observed value of X we wish to predict Y; that is, to find a Borel function g such that $\mathrm{E}(\|g(X)\|_2^2) < \infty$ and $\mathrm{E}(g(X))$ is "close to" $\mathrm{E}(Y)$.

A useful measure of closeness in the prediction problem is the *mean squared prediction error* or MSPE:

$$\mathrm{MSPE}(g) = \mathrm{E}(\|Y - g(X)\|_2^2). \tag{3.41}$$

Conditional expectation plays a major role in prediction. If $\mathrm{E}(Y^2) < \infty$, it may be of interest to determine the best predictor in the sense of minimizing the mean squared prediction error. Letting $\mathcal{T}$ be the class of all functions $g(X)$ such that $\mathrm{E}((g(X))^2) < \infty$ and assuming $\mathrm{E}(Y^2) < \infty$, we expand the mean-squared prediction error in a manner similar to the operations in inequality (B.1) on page 845. For $g(X) \in \mathcal{T}$, we have

$$\begin{aligned}
\mathrm{E}(\|Y - g(X)\|_2^2) &= \mathrm{E}(\|Y - \mathrm{E}(Y|X) + \mathrm{E}(Y|X) - g(X)\|_2^2) \\
&= \mathrm{E}(\|Y - \mathrm{E}(Y|X)\|_2^2) + \mathrm{E}(\|\mathrm{E}(Y|X) - g(X)\|_2^2) + \\
&\quad 2\mathrm{E}((Y - \mathrm{E}(Y|X))^{\mathrm{T}})(\mathrm{E}(Y|X) - g(X))) \\
&= \mathrm{E}(\|Y - \mathrm{E}(Y|X)\|_2^2) + \mathrm{E}(\|\mathrm{E}(Y|X) - g(X)\|_2^2) + \\
&\quad 2\mathrm{E}\left(\mathrm{E}((Y - \mathrm{E}(Y|X))^{\mathrm{T}})(\mathrm{E}(Y|X) - g(X)))|X\right) \\
&= \mathrm{E}(\|Y - \mathrm{E}(Y|X)\|_2^2) + \mathrm{E}(\|\mathrm{E}(Y|X) - g(X)\|_2^2) \\
&\geq \mathrm{E}(\|Y - \mathrm{E}(Y|X)\|_2^2).
\end{aligned} \tag{3.42}$$

(This proves Theorem 1.13 on page 27.)

3.1.7 Other Issues in Statistical Inference

In addition to properties of statistical methods that derive from tight assumptions about the underlying probability distribution, there are two additional considerations. One issue concerns the role of the assumed probability distribution, that is, how critical is that assumption to the performance of the statistical procedure. The other issue has to do with how the data are collected.

Robustness

We seek statistical methods that have optimal properties. We study these methods in the context of a family of probability distributions $\mathcal{P}$. An optimal method over one family of probability distributions may be far from optimal in other families. Hence, although focusing on one family we may also consider the performance over a larger class of probability families, $\{\mathcal{P}, \mathcal{Q}, \mathcal{R}, \ldots\}$. This is called "robustness".

Data and Sampling: Probability Distributions and Data-Generating Processes

For many statistical methods we begin by assuming that we have observations $X_1, \ldots, X_n$ on a random variable X from some distribution P_θ, which is a member of some family of probability distributions $\mathcal{P}$. We usually assume either that the observations (or the "data"), constitute a random sample and that they are independent or they are at least exchangeable.

The data-generating process that yields the sample depends on the probability distribution P_θ, but it may not be fully characterized by that underlying distribution. An issue is whether or not all aspects of the data-generating process should affect the inference about P_θ, or whether the inference should be based solely on the data and the assumed probability distribution P_θ.

In the case of sampling within a finite population, the observations are often taken according to some efficient scheme, such as stratification or clustering, in which the observations taken as a whole do not constitute a random sample of realizations of iid random variables. The method of analysis must take the nature of the data into consideration.

The problem of collecting data for making inferences concerning a Bernoulli parameter π provides a simple example of different data-generating processes.

Example 3.12 Sampling in a Bernoulli distribution
The family of Bernoulli distributions is that formed from the class of the probability measures $P_\pi(\{1\}) = \pi$ and $P_\pi(\{0\}) = 1 - \pi$ on the measurable space $(\Omega = \{0, 1\}, \mathcal{F} = 2^\Omega)$. A simple problem is statistical inference about π.

One approach is to take a random sample of size n, $X_1, \ldots, X_n$ from the Bernoulli(π), and then use some function of that sample as an estimator. An obvious statistic to use is the number of 1's in the sample, that is, $T = \sum X_i$. This is a sufficient statistic. The distribution of T is very simple; it is binomial with parameters n and π, and its PDF is

$$p_T(t\,;\, n, \pi) = \binom{n}{t} \pi^t (1-\pi)^{n-t}, \quad t = 0, 1, \ldots, n. \tag{3.43}$$

We could use this distribution, which depends only on π and the pre-chosen n, to form unbiased estimators, set confidence sets, or perform tests of hypotheses regarding π.

A very different approach is to take a sequential sample, $X_1, X_2, \ldots$, until a fixed number t of 1's have occurred. This is a different data-generating process. In this case, the size of the sample N is random, and its distribution is the negative binomial with parameters t and π, and its PDF is

$$p_N(n\,;\,t,\pi) = \binom{n-1}{t-1}\pi^t(1-\pi)^{n-t}, \quad n = t, t+1, \ldots. \tag{3.44}$$

(The negative binomial distribution is often defined slightly differently so that it is the distribution of the random variable $N - t$ above; either definition of course completely determines the other.) In this process, the sample size N is a sufficient statistic. We could use the distribution of N, which depends only on π and the pre-chosen t, to form unbiased estimators, set confidence sets, or perform tests of hypotheses regarding π.

In the description of the two experiments we have n (pre-chosen), N, T, and t (pre-chosen). If the realized value of N is n should the conclusions in the second experiment be the same as those in the first experiment if the realized value of T is t?

While one or the other approach may be better from either a practical or a theoretical standpoint, we may adopt the principle that similar results from the two experiments should lead to similar conclusions (where "similar results" is defined in the *likelihood principle*; see page 245). In Examples 4.3, 4.6 and 4.7, we see that a Bayesian analysis would lead to similar conclusions. In Example 6.1 on page 447 and in Example 6.4, we see that the maximum likelihood estimators of π are the same. ∎

Further comments on Example 3.12

We also consider this example in Example 4.1 on page 333, but there consider only the case in which the X_i are independent (or at least exchangeable). In that case, of course, the appropriate model is the binomial, and we can ignore the overall data-generating process. On the other hand, however, because one experiment was based on a stopping rule that was conditional on the data, perhaps different conclusions should be drawn. The sample is quite different; in the latter case, the X_i are not exchangeable, only the first $n-1$ are. Because the distributions are different, we may expect to reach different conclusions for many inferences that depend on expected values of the random observables. The simplest concern about expected values is just the bias, and indeed, the unbiased estimators of π are different (see Example 5.1 on page 390). We may also expect to reach different conclusions for many inferences that depend on quantiles of the random observables, and, indeed, hypothesis tests concerning π may be different (see Example 7.12 on page 539). You are asked to explore these issues further in Exercise 7.5. ∎

3.2 Statistical Inference: Approaches and Methods

If we assume that we have a random sample of observations $X_1, \ldots, X_n$ on a random variable X from some distribution P_θ, which is a member of some family of probability distributions $\mathcal{P}$, our objective in statistical inference is to determine a specific $P_\theta \in \mathcal{P}$, or some subfamily $\mathcal{P}_\theta \subseteq \mathcal{P}$, that could likely have generated the sample. In the previous section we have discussed various ways of assessing the information content of the sample, but we have yet to address the question of how to use the data in statistical inference.

How should we approach this problem?

Five Approaches to Statistical Inference

Five approaches to statistical inference are

- use of a likelihood function
  for example, maximum likelihood
  an estimator is an MLE
- use of the empirical cumulative distribution function (ECDF)
  for example, method of moments
  an estimator is an MME
- fitting expected values
  for example, least squares
  an estimator may be an LSE or a BLUE
- fitting a probability distribution
  for example, maximum entropy
- definition and use of a loss function; a "decision-theoretic" approach
  for example, uniform minimum variance unbiased estimation, and most Bayesian methods.
  an estimator may be a UMVUE, a UMRE (rarely!), or a UMREE (or UMRIE)

These approaches are associated with various philosophical/scientific principles, sometimes explicitly stated and sometimes not. The sufficiency principle (see page 223) guides most aspects of statistical inference, and is generally consistent with the more specific principles associated with various approaches to inference. Some of these principles, such as the substitution principle (see page 247) and the likelihood principle (see page 245), inform a major class of statistical methods, while other principles, such as the bootstrap principle (see page 249), are more local in application. Although some statisticians feel that an axiomatic approach to statistical inference should be based on universal principles, a substantial proportion of statisticians feel that abstract principles may be too general to guide inference in a specific case. Most general statistical principles focus on the *observed data* rather than on the *data-generating process*. Statisticians who emphasize general principles often characterize consideration of the data-generating process as "adhockery".

Certain elements that may be central to a particular one of these approaches may be found in other approaches; for example the concept of likelihood can be found in most approaches. The principles of data reduction and the inferential information in a sample that we discussed in the previous section obviously must be recognized in any approach to statistical inference. Finally, there are certain methods that are common to more than one of these approaches. Abstracting and studying the specific method itself may illuminate salient properties of the overall approach. An example of a method that can be studied in a unified manner is the use of estimating equations, which we discuss in Section 3.2.5.

In the following four subsections, 3.2.1 through 3.2.4, we will briefly discuss the first four of the approaches list above. The "decision theory" approach to statistical inference is based on a loss function, and we will discuss this important approach in Section 3.3.

Some Methods in Statistical Inference

Within the broad framework of a particular approach to statistical inference, there are various specific methods that may be applied. I do not attempt a comprehensive listing of these methods, but in order to emphasize the hierarchy of general approaches and specific methods, I will list a few.

- transformations
- transforms (functional transformations)
- asymptotic inference
  this includes a wide range of methods such as
  - the delta method (first order and second order)
  - various Taylor series expansions (of which the delta method is an example)
  - orthogonal series representations
- computational inference
  (this includes a wide range of methods, including MCMC)
- decomposition of variance into variance components
- Rao-Blackwellization
- scoring
- EM methods
- bootstrap
- jackknife
- empirical likelihood
- tilting
- use of saddlepoint approximations
- PDF decomposition

It is worthwhile to be familiar with a catalog of common operations in mathematical statistics. A list such as that above can be useful when working

in statistical theory (or applications, of course). In Section 0.0.9 beginning on page 676 we list several general methods to think of when doing mathematics.

We will illustrate these methods in various examples throughout this book.

3.2.1 Likelihood

Given a sample $X_1, \ldots, X_n$ from distributions with probability densities $p_i(x)$, where all PDFs are defined with respect to a common σ-finite measure, the *likelihood function* is

$$L_n(p_i\,;\,X) = c\prod_{i=1}^{n} p_i(X_i), \tag{3.45}$$

where $c \in \mathbb{R}_+$ is any constant independent of the p_i. A likelihood function, therefore, may be considered to be an equivalence class of functions. It is common to speak of $L_n(p_i\,;\,X)$ with $c = 1$ as "the" likelihood function.

We can view the sample either as a set of random variables or as a set of constants, the realized values of the random variables, in which case we usually use lower-case letters.

The likelihood function arises from a probability density, but it is not a probability density function. It does not in any way relate to a "probability" associated with the parameters or the model.

Although non-statisticians will often refer to the "likelihood of an observation", in statistics, we use the term "likelihood" to refer to a model or a distribution *given observations*.

The *log-likelihood function* is the log of the likelihood:

$$l_{L_n}(p_i\,;\,x) = \log L_n(p_i\,;\,x_i), \tag{3.46}$$

It is a sum rather than a product.

The n subscript serves to remind us of the sample size, and this is often very important in use of the likelihood or log-likelihood function particularly because of their asymptotic properties. We often drop the n subscript, however.

In many cases of interest, the sample is from a single parametric family. If the PDF is $p(x\,;\,\theta)$ then the likelihood and log-likelihood functions are written as

$$L(\theta\,;\,x) = \prod_{i=1}^{n} p(x_i\,;\,\theta), \tag{3.47}$$

and

$$l(\theta\,;\,x) = \log L(\theta\,;\,x). \tag{3.48}$$

The Parameter Is the Variable

Note that the likelihood is a function of θ for a given x, while the PDF is a function of x for a given θ. We sometimes write the expression for the

likelihood without the observations: $L(\theta)$. I like to think of the likelihood as a function of some dummy variable t that ranges over the parameter space Θ, and I write $L(t\,;\,x)$ or $l(t\,;\,x)$. While if we think of θ as a fixed, but unknown, value, it does not make sense to think of a function of that particular value, and if we have an expression in terms of that value, it does not make sense to perform operations such as differentiation with respect to that quantity.

For certain properties of statistics derived from a likelihood approach, it is necessary to consider the parameter space Θ to be closed (see, for example, Wald (1949)). Except for cases when those properties are important, we will not assume Θ to be closed, but may, however, consider the closure $\overline{\Theta}$.

In a multiparameter case, we may be interested in only some of the parameters. There are two ways of approaching this, use of a profile likelihood or of a conditional likelihood.

If $\theta = (\theta_1, \theta_2)$ and if θ_2 is fixed, the likelihood $L(\theta_1\,;\,\theta_2, x)$ is called a *profile likelihood* or *concentrated likelihood* of θ_1 for given θ_2 and x.

If the PDFs can be factored so that one factor includes θ_2 and some function of the sample, $S(x)$, and the other factor, given $S(x)$, is free of θ_2, then this factorization can be carried into the likelihood. Such a likelihood is called a *conditional likelihood* of θ_1 given $S(x)$.

Maximum Likelihood Estimation

The *maximum likelihood estimate* (MLE) of θ is defined as

$$\hat{\theta} = \arg\max_{\theta \in \overline{\Theta}} L(\theta\,;\,x), \tag{3.49}$$

if it exists (that is, if $\sup_{\theta \in \overline{\Theta}} L(\theta\,;\,x) \in \mathbb{R}$).

Because the logarithm is a strictly increasing function, the MLE is also the argmax of the log-likelihood. Also, of course, the maximum of $L(\theta)$ occurs at the same value of the argument as the maximum of $cL(\theta)$.

The MLE in general is not unbiased for its estimand. A simple example is the MLE of the variance σ^2 in a normal distribution with unknown mean.

Example 3.13 MLE of the mean and the variance in a normal distribution
Consider the normal family of distributions with parameters μ and σ^2. Suppose we have observations $x_1, x_2, \ldots, x_n$. The log-likelihood is

$$l_L(\mu, \sigma^2\,;\,x) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2}. \tag{3.50}$$

We seek values that could be substituted for μ and σ^2 so as to maximize this quantity. If μ and σ^2 are treated as variables, we see that the function in equation (3.50) is differentiable with respect to them. We have

$$\frac{\partial}{\partial \mu} l_L(\mu, \sigma^2 ; x) = \frac{\sum_{i=1}^n (x_i - \mu)}{\sigma^2} \tag{3.51}$$

$$\frac{\partial}{\partial \sigma^2} l_L(\mu, \sigma^2 ; x) = \frac{\sum_{i=1}^n (x_i - \mu)^2}{2(\sigma^2)^2} - \frac{n}{2\sigma^2} \tag{3.52}$$

We set the derivatives equal to zero to obtain the "likelihood equations", and solving that system, we obtain the stationary points

$$\begin{aligned} \hat{\mu} &= \bar{x} \qquad (3.53) \\ &\quad \text{and if } \exists\, i, j \in \{1, \ldots, n\} \ni x_i \neq x_j, \\ \hat{\sigma}^2 &= \sum_{i=1}^n (x_i - \bar{x})^2 / n. \qquad (3.54) \end{aligned}$$

Next, we compute the Hessian of $l_L(\mu, \sigma^2 ; x)$, and observe that it is negative definite at the stationary point; hence $\hat{\mu}$ and $\hat{\sigma}^2$ maximize the log-likelihood (exercise).

We know that $S^2 = \sum_{i=1}^n (X_i - \bar{X})^2/(n-1)$ where $X_1, X_2, \ldots, X_n$ is a random sample from a normal distribution with variance σ^2 is unbiased for σ^2; hence, the MLE is biased.

Note that if we had only one observation or if all observations had the same value, the log-likelihood would be unbounded when $\mu = x_1$ and σ^2 approaches zero. ∎

The MLE for σ^2 in a normal distribution with unknown mean is the same as the plug-in estimator or MME (3.64). Note that the plug-in estimator (or MME) is not based on an assumed underlying distribution, but the MLE is.

The MLE may have smaller MSE than an unbiased estimator. This is the case in the example above. The estimator S^2 in equation (3.65) is unbiased, and the MLE is $(n-1)S^2/n$. Consider any estimator of the form $c(n-1)S^2$. Note that $(n-1)S^2/\sigma^2$ has a χ^2_{n-1} distribution. In the case of $\mathrm{N}(\mu, \sigma^2)$ we have the MSE

$$\mathrm{E}((c(n-1)S^2 - \sigma^2)^2) = \sigma^4((n^2-1)c^2 - 2(n-1)c + 1). \tag{3.55}$$

From this we see that the MLE of σ^2, that is, where $c = 1/n$, has uniformly smaller MSE than the unbiased estimator S^2.

Likelihood Equation

In statistical estimation, the point at which the likelihood attains its maximum (which is, of course, the same point at which the log-likelihood attains its maximum) is of interest. We will consider this approach to estimation more thoroughly in Chapter 6.

If the likelihood is differentiable with respect to the parameter, we may be able to obtain the maximum by setting the derivative equal to zero and solving the resulting equation:

$$\frac{\partial l(\theta\,;\,x)}{\partial \theta} = 0. \tag{3.56}$$

This is called the *likelihood equation*. The derivative of the likelihood equated to zero, $\partial L(\theta\,;\,x)/\partial\theta = 0$, is also called the likelihood equation.

The roots of the likelihood equation are often of interest, even if these roots do not provide the maximum of the likelihood function.

Equation (3.56) is an *estimating equation*; that is, its solution, if it exists, is an estimator. *Note that this estimator is not necessarily MLE; it is a root of the likelihood equation, or RLE.* We will see in Chapter 6 that RLEs have desirable asymptotic properties.

It is often useful to define an estimator as the solution of some estimating equation. We will see other examples of estimating equations in subsequent sections.

Score Function

The derivative of the log-likelihood on the left side of equation (3.56) plays an important role in statistical inference. It is called the *score function*, and often denoted as $s_n(\theta\,;\,x)$:

$$s_n(\theta\,;\,x) = \frac{\partial l(\theta\,;\,x)}{\partial \theta}. \tag{3.57}$$

(I should point out that I use the notation "$\nabla l(\theta\,;\,x)$" and the slightly more precise "$\partial l(\theta\,;\,x)/\partial\theta$" more-or-less synonymously.)

Finding an RLE is called *scoring*.

In statistical inference, knowing how the likelihood or log-likelihood would vary if θ were to change is important. For a likelihood function (and hence, a log-likelihood function) that is differentiable with respect to the parameter, the score function represents this change.

Likelihood Ratio

When we consider two different distributions for a sample x, we have two different likelihoods, say L_0 and L_1. (Note the potential problems in interpreting the subscripts; here the subscripts refer to the two different distributions. For example L_0 may refer to $L(\theta_0 \,|\, x)$ in a notation consistent with that used above.) In this case, it may be of interest to compare the two likelihoods in order to make an inference about the two possible distributions. A simple comparison, of course, is the ratio. The ratio

$$\frac{L(\theta_0\,;\,x)}{L(\theta_1\,;\,x)}, \tag{3.58}$$

or L_0/L_1 in the simpler notation, is called the *likelihood ratio* with respect to the two possible distributions.

Although in most contexts we consider the likelihood to be a function of the parameter for given, fixed values of the observations, it may also be useful to consider the likelihood ratio to be a function of x. On page 167, we defined a family of distributions based on their having a "monotone" likelihood ratio. Monotonicity in this case is with respect to a function of x. In a family with a monotone likelihood ratio, for some scalar-valued function $y(x)$ and for any $\theta_1 < \theta_0$, the likelihood ratio is a nondecreasing function of $y(x)$ for all values of x for which $f_{\theta_1}(x)$ is positive.

The most important use of the likelihood ratio is as the basis for a statistical test.

Under certain conditions that we will detail later, with L_0 and L_1, with corresponding log-likelihoods l_0 and l_1, based on a random variable (that is, $L_i = L(p_i\,;\,X)$, instead of being based on a fixed x), the random variable

$$\begin{aligned}\lambda &= -2\log\left(\frac{L_0}{L_1}\right) \\ &= -2(l_0 - l_1)\end{aligned} \tag{3.59}$$

has an approximate chi-squared distribution with degrees of freedom whose number depends on the numbers of parameters. (We will discuss this more fully in Chapter 7.)

This quantity in a different setting is also called the *deviance*. We encounter the deviance in the analysis of generalized linear models, as well as in other contexts.

The likelihood ratio, or the log of the likelihood ratio, plays an important role in statistical inference. Given the data x, the log of the likelihood ratio is called the *support* of the hypothesis that the data came from the population that would yield the likelihood L_0 versus the hypothesis that the data came from the population that would yield the likelihood L_1. The support clearly is relative and ranges over $\mathbb{R}$. The support is also called the *experimental support*.

Likelihood Principle

The *likelihood principle* in statistical inference asserts that all of the information that the data provide concerning the relative merits of two hypotheses (two possible distributions that give rise to the data) is contained in the likelihood ratio of those hypotheses and the data. An alternative statement of the likelihood principle is that if for x and y,

$$\frac{L(\theta\,;\,x)}{L(\theta\,;\,y)} = c(x,y) \quad \forall\theta, \tag{3.60}$$

where $c(x,y)$ is constant for given x and y, then any inference about θ based on x should be in agreement with any inference about θ based on y.

3.2.2 The Empirical Cumulative Distribution Function

Given a sample $X_1 = x_1, \ldots, X_n = x_n$ as independent observations on a random variable X, we can form another random variable X^* that has a discrete uniform distribution with probability mass $1/n$ at each of the sample points. (In the case that $x_i = x_j$ for some i and j, X^* would not have a uniform distribution of course. Whether or not this is the case, however, for convenience, we will continue to refer to the distribution as uniform.) This discrete uniform distribution can be used for making inferences on the distribution of the random variable of interest

From observations on a random variable, $X_1, \ldots, X_n$, we can form an empirical cumulative distribution function, or ECDF, that corresponds in a natural way with the CDF of the random variable.

For the sample, $X_1, \ldots, X_n$, the ECDF is defined as the CDF of X^*; that is,

$$P_n(x) = \frac{\#\{X_i \leq x\}}{n}. \tag{3.61}$$

The ECDF is a random simple function, and often it is appropriate to treat the ECDF as a random variable. It is clear that the ECDF conditional on a given sample is itself a CDF. (Conditionally it is not a "random" variable; that is, it is a degenerate random variable.) It has the three properties that define a CDF:

- $\lim_{x\to-\infty} P_n(x) = 0$ and $\lim_{x\to\infty} P_n(x) = 1$.
- $P_n(x)$ is monotone increasing.
- $P_n(x)$ is continuous from the right.

The ECDF defines a discrete population with mass points at each value in the sample.

The ECDF is particularly useful in nonparametric inference. We will see below how it can allow us to "bootstrap" an unknown population and also how it can allow us to use the principles of likelihood even though we may be unable to write out the density of the population.

Plug-In Estimators; The Substitution Principle

As discussed in Section 1.1.9, many distribution parameters and other measures can be represented as a statistical function, that is, as a functional of the CDF. The functional of the CDF that defines a parameter defines a plug-in estimator of that parameter when the functional is applied to the ECDF. A functional of a population distribution function, $\Theta(P)$, that defines a parameter θ can usually be expressed as

$$\begin{aligned} \theta &= \Theta(P) \\ &= \int g(y)\,\mathrm{d}P(y). \end{aligned} \tag{3.62}$$

The plug-in estimator T is the same functional of the ECDF:

$$\begin{aligned} T &= \Theta(P_n) \\ &= \int g(y)\,\mathrm{d}P_n(y). \end{aligned} \tag{3.63}$$

(In both of these expressions, we are using the integral in a general sense. In the second expression, the integral is a finite sum. It is also a countable sum in the first expression if the random variable is discrete. Note also that we use the same symbol to denote the functional and the random variable.)

The use of the functional that defines a statistical function on the ECDF to make inferences on the statistical function is called the *substitution principle*. It is one of the most useful and most pervasive principles in statistical inference.

We may base inferences on properties of the distribution with CDF P by identifying the corresponding properties of the ECDF P_n. In some cases, it may not be clear what we mean by "corresponding". If a property of a distribution can be defined by a functional on the CDF, the corresponding property is the same functional applied to the ECDF. This is the underlying idea of the method of moments, for example.

The asymptotic properties of plug-in estimators can be developed by Taylor-series-type expansions of the statistical functions, as discussed on page 95 in Section 1.3.7. We consider this further on page 316.

Method of Moments Estimators

In the method of moments, sample moments, which are moments of the discrete population represented by the sample, are used for making inferences about population moments. The MME of the population mean, $\mathrm{E}(X)$, is the sample mean, $\overline{X}$. The thing to be estimated is the functional M in equation (1.109), and the estimator is M applied to the ECDF:

$$M(P_n) = \sum X_i P_n(X_i).$$

We call a method-of-moments estimator an MME.

The plug-in estimator $\Theta(P_n)$ in general is not unbiased for the associated statistical function $\Theta(P)$. A simple example is the variance,

$$\Sigma(P) = \sigma^2 = \int \left(x - \int x\,\mathrm{d}P\right)^2 \mathrm{d}P.$$

The plug-in estimator, which in this case is also a MME, is

$$\Sigma(P_n) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2. \tag{3.64}$$

We see that if $n \geq 2$, the MME $\Sigma(P_n) = (n-1)S^2/n$, where

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{3.65}$$

is the usual sample variance.

On the other hand, the plug-in estimator may have smaller MSE than an unbiased estimator, and, in fact, that is the case for the plug-in estimator of σ^2 (see equation (3.55)). Also, plug-in estimators often have good limiting and asymptotic properties, as we might expect based on convergence properties of the ECDF.

Convergence of the ECDF

The ECDF is one of the most useful statistics, especially in nonparametric and robust inference. It is essentially the same as the set of order statistics, so like them, it is a sufficient statistic. Although we may write the ECDF as F_n or $F_n(x)$, it is important to remember that it is a random variable.

The distribution of the ECDF at a point is binomial, and so the pointwise properties of the ECDF are easy to see. From the SLLN, we see that it strongly converges pointwise to the CDF. At the point x, by the CLT we have

$$\sqrt{n}(F_n(x) - F(x)) \xrightarrow{\mathrm{d}} \mathrm{N}\left(0, F(x)(1-F(x))\right).$$

Although the pointwise properties of the ECDF are useful, its global relationship to the CDF is one of the most important properties of the ECDF. Dvoretzky/Kiefer/Wolfowitz/Massart inequality (1.289)

$$\Pr(\rho_\infty(F_n, F) > z) \leq 2\mathrm{e}^{-2nz^2}$$

provides a tight bound on the difference in the ECDF and the CDF.

The Glivenko-Cantelli theorem (page 135) tells us that the sup distance of the ECDF from the CDF, $\rho_\infty(F_n, F)$, converges almost surely to zero; that is, the ECDF converges strongly and uniformly to the CDF.

The Bootstrap

The ECDF plays a major role in a bootstrap method, in which the population of interest is studied by sampling from the population defined by a given sample from the population of interest. This is a method of *resampling*.

Resampling methods involve the use of many samples, each taken from a single sample that was taken from the population of interest. Inference based on resampling makes use of the conditional sampling distribution of a new sample (the "resample") drawn from a given sample. Statistical functions on the given sample, a finite set, can easily be evaluated. Resampling methods

therefore can be useful even when very little is known about the underlying distribution.

A basic idea in bootstrap resampling is that, because the observed sample contains all the available information about the underlying population, the observed sample can be considered *to be* the population; hence, the distribution of any relevant test statistic can be simulated by using random samples from the "population" consisting of the original sample.

Suppose that a sample $y_1, \ldots, y_n$ is to be used to estimate a population parameter, θ. For a statistic T that estimates θ, as usual, we wish to know the sampling distribution so as to correct for any bias in our estimator or to set confidence intervals for our estimate of θ. The sampling distribution of T is often intractable in applications of interest.

A basic bootstrapping method formulated by Efron (1979) uses the discrete distribution represented by the sample to study the unknown distribution from which the sample came. The basic tool is the empirical cumulative distribution function. The ECDF is the CDF of the finite population that is used as a model of the underlying population of interest.

For a parameter θ of a distribution with CDF P defined as $\theta = \Theta(P)$, we can form a plug-in estimator T as $T = \Theta(P_n)$. Various properties of the distribution of T can be estimated by use of "bootstrap samples", each of the form $\{y_1^*, \ldots, y_n^*\}$, where the y_i^*'s are chosen from the original y_i's with replacement.

We define a *resampling vector*, p^*, corresponding to each bootstrap sample as the vector of proportions of the elements of the original sample in the given bootstrap sample. The resampling vector is a realization of a random vector P^* for which nP^* has an n-variate multinomial distribution with parameters n and $(1/n, \ldots, 1/n)$. The resampling vector has random components that sum to 1. For example, if the bootstrap sample $(y_1^*, y_2^*, y_3^*, y_4^*)$ happens to be the sample (y_2, y_2, y_4, y_3), the resampling vector p^* is

$$(0,\ 1/2,\ 1/4,\ 1/4).$$

The bootstrap replication of the estimator T is a function of p^*, $T(p^*)$. The resampling vector can be used to estimate the variance of the bootstrap estimator. By imposing constraints on the resampling vector, the variance of the bootstrap estimator can be reduced.

The Bootstrap Principle

The *bootstrap principle* involves repeating the process that leads from a population CDF to an ECDF. Taking the ECDF P_n to be the CDF of a population, and resampling, we have an ECDF for the new sample, $P_n^{(1)}$. (In this notation, we could write the ECDF of the original sample as $P_n^{(0)}$.) The difference is that we know more about $P_n^{(1)}$ than we know about P_n. Our knowledge about

$P_n^{(1)}$ comes from the simple discrete uniform distribution, whereas our knowledge about P_n depends on knowledge (or assumptions) about the underlying population.

The bootstrap resampling approach can be used to derive properties of statistics, regardless of whether any resampling is done. Most common uses of the bootstrap involve computer simulation of the resampling; hence, bootstrap methods are usually instances of computational inference.

Empirical Likelihood

The ECDF can also be used to form a probability density for a method based on a likelihood function.

3.2.3 Fitting Expected Values

Given a random sample $X_1, \ldots, X_n$ from distributions with probability densities $p_i(x_i; \theta)$, where all PDFs are defined with respect to a common σ-finite measure, if we have that $\mathrm{E}(X_i) = g_i(\theta)$, a reasonable approach to estimation of θ may be to choose a value $\hat{\theta}$ that makes the differences $\mathrm{E}(X_i) - g_i(\theta)$ close to zero. If the X_i are iid, then all $g_i(\theta)$ are the same, say $g(\theta)$.

We must define the sense in which the differences are close to zero. A simple way to do this is to define a nonnegative scalar-valued Borel function of scalars, $\rho(u, v)$, that is increasing in the absolute difference of its arguments. One simple choice is $\rho(u, v) = (u - v)^2$. We then define

$$S_n(\theta, x) = \sum_{i=1}^{n} \rho(x_i, g(\theta)). \tag{3.66}$$

For a random sample $X = X_1, \ldots, X_n$, an estimator fitted to the expected values is $g(T)$ where

$$T = \arg\min_{\theta \in \overline{\Theta}} S_n(\theta, X). \tag{3.67}$$

Compare this with the maximum likelihood estimate of θ, defined in equation (3.49).

As with solving the maximization of the likelihood, if the function to be optimized is differentiable, the solution to the minimization problem (3.67) may be obtained by solving

$$s_n(\theta\,;\, x) = \frac{\partial S_n(\theta\,;\, x)}{\partial \theta} = 0. \tag{3.68}$$

Equation (3.68) is an *estimating equation*; that is, its solution, if it exists, is taken as an estimator. *Note that this estimator is not necessarily a solution to the optimization problem (3.67).*

In common applications, we have *covariates*, $Z_1, \ldots, Z_n$, and the $\mathrm{E}(X_i)$ have a constant form that depends on the covariate: $\mathrm{E}(X_i) = g(Z_i, \theta)$.

Example 3.14 least squares in a linear model
Consider the linear model (3.7)

$$Y = X\beta + E,$$

where Y is the random variable that is observable, in the least squares setup of equations (3.66) and (3.67) we have

$$S_n(\beta\,;\, y, X) = \|y - X\beta\|_2. \tag{3.69}$$

In the case of the linear model, we have the estimating equation

$$s_n(\beta\,;\, y, X) = X^{\mathrm{T}}y - X^{\mathrm{T}}X\beta = 0. \tag{3.70}$$

This system of equations is called the "normal equations".

For the estimand $g(\beta) = l^{\mathrm{T}}\beta$ for some fixed l, the least squares estimator is $l^{\mathrm{T}}(X^{\mathrm{T}}X)^{-}X^{\mathrm{T}}Y$, where M^{-} denotes a generalized inverse of a matrix M. See page 424 and the following pages for a more thorough discussion of the linear model in this example. ∎

Example 3.14 illustrates a very simple and common application of estimation by fitting expected values; the expected values are those of the observable random variable. The next example is a somewhat less common situation of defining which expected values to focus on.

Example 3.15 estimation in a stable family; the empirical CF
Most members of the stable family of distributions are quite complicated. In general, there is no closed form for the CDF of the PDF, and none of the moments exist or else are infinite. The family of distributions is usually specified by means of the characteristic function (see equation (2.26)),

$$\varphi(t) = \exp\left(\mathrm{i}\mu t - |\sigma t|^{\alpha}\left(1 - \mathrm{i}\beta\,\mathrm{sign}(t)\omega(\alpha, t)\right)\right).$$

Because the characteristic function is an expected value, its sample analogue, that is, the empirical characteristic function can be formed easily from a sample, $x_1, \ldots, x_n$:

$$\varphi_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{e}^{\mathrm{i}tx_i}. \tag{3.71}$$

The empirical characteristic function can be computed at any point t.

The expected values would be fit by minimizing

$$S_n(x, r, \mu, \sigma, \alpha, \beta) = \|\varphi_n(t) - \varphi(t)\|_r \tag{3.72}$$

for given r at various values of t.

While this is a well-defined problem (for some given values of t) and the resulting estimators are strongly consistent (for the same reason that estimators based on the ECDF are strongly consistent), there are many practical

issues in the implementation of the method. Press (1972) proposed fitting moments as an approximation to the values that would be obtained by minimizing s_n in equation (3.72). For the case of $r = 2$, Koutrouvelis (1980) proposed a regression method that seems to perform fairly well in simulation studies. Kogan and Williams (1998) summarize these and other methods for estimating the parameters in a stable family. ∎

Quantile Based Estimators

The expected values of order statistics can often yield good estimators when sample quantiles are fit to them. In most cases, the statistical properties of these estimators are not as good as alternative estimators, but there are some cases where they are useful. One such situation is where the distribution is very complicated, such as the family of stable distributions in Example 3.15. Fama and Roll (1971) describe methods for estimation of the parameters in a symmetric stable family (that is, one in which $\beta = 0$).

Estimators based on quantiles are especially useful in heavy-tailed distributions. (The stable family is heavy-tailed.) Beginning on page 608, I will discuss a type of robust estimators, called L-estimators, that are linear combinations of order sample quantiles.

Regularization of Fits

The objective function for fitting expected values may not be well-behaved. Small variations in the sample may yield large differences in the estimates. The ill-conditioned objective function yield estimators with large variances. An approach to this problem is to "regularize" the objective function by modifying it to be better conditioned. In a minimization problem a simple way of making the objective function better conditioned is to add a penalty term for variation in the solution. This means that the solution is pulled toward some fixed value. Often there is no obvious fixed value to bias an estimator toward. A common procedure is merely to shrink the estimator toward 0. Given an objective function of the form (3.66), a modified objective function that shrinks the estimator toward 0 is

$$\widetilde{S}_n(\theta, x) = \sum_{i=1}^{n} \rho_1(x_i, g(\theta)) + \rho_2(\theta, \theta). \tag{3.73}$$

Use of this type of objective function in regression analysis leads to "ridge regression"; see Example 5.27.

The idea of regularization is also used in other estimation methods that involve a minimization or maximization, such as maximum likelihood estimation of a PDF as on page 581, for example.

3.2.4 Fitting Probability Distributions

Another approach to statistical inference is to use the observed data to fit the probability distribution over the full support. The fit is chosen to minimize some measure of the difference between it and the values of the PDF (or probability function). To pursue this idea, we need some measure of the difference.

The quantity

$$d(P,Q) = \int_{\mathbb{R}} \phi\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) \mathrm{d}Q, \tag{3.74}$$

if it exists, is called the *ϕ-divergence* from Q to P. The ϕ-divergence is also called the *f-divergence*.

The expression often has a more familiar form if both P and Q are dominated by Lebesgue measure and we write $p = \mathrm{d}P$ and $q = \mathrm{d}Q$.

A specific instance of ϕ-divergence is the *Kullback-Leibler measure*,

$$\int_{\mathbb{R}} p(x) \log\left(\frac{p(x)}{q(x)}\right) \mathrm{d}x. \tag{3.75}$$

(Recall from page 850 that this quantity is nonnegative.)

The ϕ-divergence is in general not a metric because it is not symmetric. One function is taken as the base from which the other function is measured. In Section 0.1.9 beginning on page 747, we discuss metrics and also ϕ-divergence in the more general context of comparing two functions.

While this idea can be used for any type of distribution, it is most useful in the case of a discrete distribution with a finite number of mass points. In the case of a distribution with d mass points with probabilities $\pi_1, \ldots, \pi_d$, the full information content of a sample $X_1, \ldots, X_n$ is the information in a sample $Y_1, \ldots, Y_d$ from a multinomial distribution with parameters n and $\pi_1, \ldots, \pi_d$.

Various measures of divergence in a multinomial distribution are well-known. The most commonly used measure is the chi-squared measure, which is given in a general form in equation (0.1.86) on page 748. . This measure has a simpler form in the multinomial case. It is also a member of the family of *power divergence measures*, which for $\lambda \in \mathbb{R}$, is

$$I_\lambda = \frac{1}{\lambda(\lambda+1)} \sum_{i=1}^{d} Y_i \left(\left(\frac{Y_i}{n\pi_i} \right)^\lambda - 1 \right). \tag{3.76}$$

There are equivalent forms of this measure that are scaled by d or by some other constant, and for given λ the constant factor plays no role in minimizing the divergence. For $\lambda = 1$, this is the same as the chi-squared discrepancy measure, For $\lambda = 0$ in the limit, this is the same as the log-likelihood ratio statistic, and for $\lambda = -1/2$, it is the Freeman-Tukey statistic. Cressie and Read (1984) studied this general family of power divergence measures, and suggested $\lambda = 2/3$ as a value that has some of the desirable properties of both the chi-squared and log-likelihood ratio statistics.

The parameters of the multinomial are just $\pi_1, \ldots, \pi_d$ and estimators of them based on a power divergence measure are straightforward. When a multinomial distribution is formed from another distribution, however, the estimation problem is more interesting.

Example 3.16 minimum distance estimation in a Poisson model
Suppose we have observations $x_1, x_2, \ldots, x_n$ from a Poisson distribution with unknown parameter θ. The sample values are all nonnegative integers and if θ is relatively small, there may be very few observations that exceed some small number. Suppose we form a multinomial model, as indicated above, with $d = 3$; that is, y_1 is the number of 0s observed, y_2 is the number of 1s observed, and y_3 is the number of observed values greater than or equal to 2. We have $\pi_1 = \mathrm{e}^{-\theta}$, $\pi_2 = \theta\mathrm{e}^{-\theta}$, and $\pi_3 = 1 - (1+\theta)\mathrm{e}^{-\theta}$.

The minimum power divergence estimator of θ is obtained by substituting the appropriate values of π in expression (3.76) and then minimizing it with respect to θ. Thus, given the observations y_1, y_2, y_3, the Cressie-Read estimate is

$$\begin{aligned}\arg\min_{\theta\in\mathbb{R}} \Bigg(& y_1 \left(\left(\frac{y_1/n}{\mathrm{e}^{-\theta}} \right)^{2/3} - 1 \right) \\ & + y_2 \left(\left(\frac{y_2/n}{\theta\mathrm{e}^{-\theta}} \right)^{2/3} - 1 \right) \\ & + y_3 \left(\left(\frac{y_3/n}{1-(1+\theta)\mathrm{e}^{-\theta}} \right)^{2/3} - 1 \right) \Bigg).\end{aligned}$$

∎

In an approach to statistical inference based on information theory, the true but unknown distribution is compared with information in the sample. The focus is on "information" or "entropy", in the sense discussed on page 42. The basic quantity is of the form $\mathrm{E}(-\log(\mathrm{d}P))$. The principle underlying methods of statistical inference using these concepts and quantities is called *maximum entropy*.

3.2.5 Estimating Equations

Equations (3.56) and (3.68) are *estimating equations*; that is, their solutions, if they exist, are taken as estimates. Note that the solutions to the estimating equations are not necessarily the solutions to the optimization problems that gave rise to them. They are both merely roots of estimating equations.

Estimating Functions and Generalized Estimating Equations

Estimating equations arise often in statistical inference. There are also several modifications of the basic equations; for example, sometimes we cannot form a tractable likelihood, so we form some kind of "quasi-likelihood". We therefore consider a generalized class of estimating equations.

We consider an independent sample $X_1, \ldots, X_n$ of random vectors with orders $d_1, \ldots, d_n$, with $\sup d_i < \infty$. We assume the distributions of the X_i are defined with respect to a common parameter $\theta \in \Theta \subseteq \mathbb{R}^k$. We now define Borel functions $\psi_i(X_i, t)$ and let

$$s_n(t\,;\,X) = \sum_{i=1}^{n} \psi_i(X_i, t) \quad t \in \Theta. \tag{3.77}$$

If $\mathrm{E}_\theta((s_n(\theta\,;\,X))^2) < \infty$, we call

$$s_n(t\,;\,X) \tag{3.78}$$

an *estimating function*. We often write the estimating function simply as $s_n(t)$. (Also, note that I am writing "t" instead of "θ" to emphasize that it is a variable in place of the unknown parameter.)

Two prototypic estimating functions are the score function, equation (3.57) and the function on the left side of the normal equations (3.70).

We call

$$s_n(t) = 0 \tag{3.79}$$

a *generalized estimating equation* (GEE) and we call a root of the generalized estimating equation a GEE estimator. If we take

$$\psi_i(X_i, t) = \partial\rho(X_i, t)/\partial t,$$

we note the similarity of the GEE to equation (3.68).

Unbiased Estimating Functions

The estimating function is usually chosen so that

$$\mathrm{E}_\theta(s_n(\theta\,;\,X)) = 0, \tag{3.80}$$

or else so that the asymptotic expectation of $\{s_n\}$ is zero.

If $s_n(\theta\,;\,X) = T(X) - g(\theta)$, the condition (3.80) is equivalent to the estimator $T(X)$ being unbiased for the estimand $g(\theta)$. This leads to a more general definition of unbiasedness for a function.

Definition 3.7 (unbiased estimating function)
The *estimating function* $s_n(\theta\,;\,X)$ *is unbiased* if

$$\mathrm{E}_\theta(s_n(\theta\,;\,X)) = 0 \quad \forall\theta \in \Theta. \tag{3.81}$$

∎

An unbiased estimating function does not necessarily lead to an unbiased estimator of $g(\theta)$, unless, of course, $s_n(\theta\,;\,X) = T(X) - g(\theta)$.

We also note that unbiased estimating functions are essentially members of equivalence classes formed by multiples that are independent of the random variable. That is, if $s_n(\theta\,;\,X)$ is unbiased, and $g(\theta)$ is a function that does not depend on X, then $g(\theta)s_n(\theta\,;\,X)$ is also unbiased.

Notice that equation (3.80) holds for the normal equations (3.70); therefore, the estimating function in the normal equations, $X^{\mathrm{T}}Y - X^{\mathrm{T}}X\beta$ is unbiased. On page 463, we will also see that the score function is unbiased.

Efficiency of Estimating Functions

The efficiency of a statistical procedure generally refers to the mean squared error of the procedure. For certain families of distributions, we can establish lower bounds on the variance of a statistic.

An approach to estimation that we have mentioned a few times already and will study more fully in later sections and chapters is to restrict attention to unbiased statistics and to determine one of those that minimizes the variance at all points in the parameter space. If $s_n(\theta\,;\,X)$ is unbiased, then

$$\mathrm{V}_\theta(s_n(\theta\,;\,X)) = \mathrm{E}_\theta((s_n(\theta\,;\,X))^2). \tag{3.82}$$

For the case of unbiased estimators in certain families of distributions, the lower bound on the variance takes on special importance.

For an unbiased estimator T of $g(\theta)$ in a family of densities satisfying the regularity conditions and such that T has a finite second moment, from inequality (3.39) on page 234, we have the matrix relationship

$$\mathrm{V}(T(X)) \geq \left(\frac{\partial}{\partial\theta}g(\theta)\right)^{\mathrm{T}} (I(\theta))^{-1} \frac{\partial}{\partial\theta}g(\theta), \tag{3.83}$$

where we assume the existence of all quantities in the expression.

Definition 3.8 (efficient estimator; Fisher efficient)
Given a family of distributions $\{P_\theta\}$ satisfying the FI regularity conditions, an unbiased estimator $T(X)$ of $g(\theta)$ is said to be *efficient* or *Fisher efficient* if $\mathrm{V}(T(X))$ attains the lower bound in inequality (3.83). ∎

Notice the slight difference in "efficiency" and "efficient"; while one meaning of "efficiency" is a relative term that is not restricted to unbiased estimators (or other unbiased procedures, as we will see later), "efficient" is absolute. "Efficient" only applies to unbiased estimators, and an estimator either is or is not efficient. The state of being efficient, of course is called "efficiency". This is another meaning of the term. The phrase "Fisher efficiency" helps to emphasis this difference.

To minimize the variance among all unbiased estimating functions leads to the trivial solution $(s_n(\theta\,;\,X)) \equiv 0$, because, as we noted above, any multiple of $s_n(\theta\,;\,X)$ that does not involve X is unbiased. We therefore seek other ways of defining optimality among a class of unbiased estimating functions.

We consider a generalization of the Fisher information (3.30) with $s_n(\theta\,;\,X) = \partial \log p(X;\theta)/\partial\theta$:

$$\mathrm{E}_\theta\left((s_n(\theta\,;\,X))\,(s_n(\theta\,;\,X))^{\mathrm{T}}\right).$$

Now we define efficiency of unbiased estimating functions in terms of this quantity.

Definition 3.9 (efficiency of unbiased estimating functions)
Let $s_n(\theta\,;\,X)$ be an unbiased estimating function that is differentiable in θ. The *efficiency* of s_n is

$$(\mathrm{E}_\theta(\partial s_n(\theta\,;\,X)/\partial\theta))^{\mathrm{T}}\left(\mathrm{E}_\theta\left((s_n(\theta\,;\,X))\,(s_n(\theta\,;\,X))^{\mathrm{T}}\right)\right)^{-1}(\mathrm{E}_\theta(\partial s_n(\theta\,;\,X)/\partial\theta)).$$

∎

The efficiency of unbiased estimating functions is sometimes called *Godambe efficiency*, after V. P. Godambe. Compare this expression for the efficiency of an unbiased estimating function with the CRLB, which is expressed in terms of a score function.

Notice that for estimating functions, we define efficiency only for unbiased functions. Just as in the case of point estimators, with estimating functions, we use the word "efficient" in the sense of "most efficient".

Definition 3.10 (efficient unbiased estimating functions)
Let $s_n^*(\theta\,;\,X)$ be an unbiased estimating function that is differentiable in θ. If the efficiency of s_n^* is at least as great as the efficiency of any other unbiased estimating function that is differentiable in θ, then we say s_n^* is *efficient*, or (synonymously) *Godambe efficient*. ∎

That is, while "efficiency" is a relative term, "efficient" is absolute. An efficient estimating function is not necessarily unique, however.

Definition 3.11 (martingale estimating function)
Let $\{(X_t, \mathcal{F}_t) : t \in \mathcal{T}\}$ be a forward martingale, and let $\{s_t(\theta\,;\,X_t) : t \in \mathcal{T}\}$ be adapted to the filtration $\{\mathcal{F}_t)\}$. Then $\{s_t(\theta\,;\,X_t) : t \in \mathcal{T}\}$ is called a *martingale estimating function* iff

$$s_0(\theta\,;\,X_0) \overset{\text{a.s.}}{=} 0$$

and

$$\mathrm{E}(s_t(\theta\,;\,X_t)|\mathcal{F}_{t-1}) \overset{\text{a.s.}}{=} s_{t-1}(\theta\,;\,X_{t-1}).$$

∎

Martingale estimating functions arise in applications of stochastic process models, for example, in the analysis of financial data.

Our interest in estimating functions is due to their use in forming estimating equations and subsequently in yielding estimators. We will consider some asymptotic properties of solutions to estimating equations in Section 3.8.1 (consistency) and in Section 6.3.4 (asymptotic normality).

3.2.6 Summary and Preview

We have discussed four general approaches to statistical inference, and have identified a fifth one that we implied would warrant more careful study later. At this point, let us review and summarize the procedures that we have discussed and briefly introduce the other approach, which we will discuss in Section 3.3.

- estimation based on the ECDF
  - estimate $g(\theta)$ so that the quantiles of $P_{\widehat{g(\theta)}}$ are close to the quantiles of the data
    How many and which quantiles to match?
    Use of a plug-in estimator from the empirical cumulative distribution function follows this approach, and in that case all quantiles from the data are used.
    This approach may involve questions of how to define sample quantiles. We will continue to use the term "sample quantile" of order π to refer to the order statistic $X_{(\lceil n\pi \rceil + 1:n)}$.
    An example of this approach is the requirement of median-unbiasedness (one specific quantile).
  - estimate $g(\theta)$ so that the moments of $P_{\widehat{g(\theta)}}$ are close to the sample moments
    How many and which moments to match?
    Do the population moments exist?
    Method-of-moments estimators may have large variances; hence, while this method may be simple (and widely-used), it is probably not a good method generally.
    An example of this approach is the requirement of unbiasedness (one specific moment).
- use the likelihood
  - estimate $g(\theta)$ as $g(\hat{\theta})$, where $\hat{\theta}$ maximizes the likelihood function, $L(\theta, x, z)$.
    Maximum likelihood estimation is closely related to minimum-residual-norm estimation. For the normal distribution, for example, MLE is the same as LS, and for the double exponential distribution, MLE is the same as LAV.
    If there is a sufficient statistic, a MLE is a function of it. (This does not say that every MLE is a function of the sufficient statistic.)
    MLEs often have very good statistical properties. The are particularly easy to work with in exponential families.
- estimation by fitting expected values
  - estimate $g(\theta)$ so that residuals $\|x_i - \mathrm{E}_{\widehat{g(\theta)}}(X_i, z_i)\|$ are small.
    An example of this approach is least squares (LS) estimation (the Euclidean norm of the vector of residuals, or square root of an inner prod-

uct of the vector with itself). If the expectation exists, least squares yields unbiasedness.
Another example of this approach is least absolute values (LAV) estimation, in which the L_1 norm of the vector of residuals is minimized. This yields median-unbiasedness.

- fit an empirical probability distribution
  This approach is somewhat similar to fitting an ECDF, but in the case of PDFs, the criterion of closeness of the fit must be based on regions of nonzero probability; that is, it can be based on divergence measures.
- define and use a loss function
  (This is an approach based on "decision theory", which we introduce formally in Section 3.3. The specific types of estimators that result from this approach are the subjects of several later chapters.)
  The loss function increases the more the estimator differs from the estimand, and then estimate $g(\theta)$ so as to minimize the expected value of the loss function (that is, the "risk") at points of interest in the parameter space.
  - require unbiasedness and minimize the variance at all points in the parameter space (this is UMVU estimation, which we discuss more fully in Section 5.1)
  - require equivariance and minimize the risk at all points in the parameter space (this is MRE or MRI estimation, which we discuss more fully in Section 3.4)
  - minimize the maximum risk over the full parameter space
  - define an a priori averaging function for the parameter, use the observed data to update the averaging function and minimize the risk defined by the updated averaging function.

3.3 The Decision Theory Approach to Statistical Inference

3.3.1 Decisions, Losses, Risks, and Optimal Actions

In the decision-theoretic approach to statistical inference, we call the inference a *decision* or an *action*, and we identify a *cost* or *loss* that depends on the decision and the true (but unknown) state of nature modeled by $P \in \mathcal{P}$. (Instead of loss, we could use its opposite, which is called *utility*.)

Our objective is to choose an action that minimizes the expected loss, or conversely maximizes the expected utility.

We call the set of allowable actions or decisions the *action space* or decision space, and we denote it as $\mathcal{A}$. We base the inference on the random variable X; hence, the decision is a mapping from $\mathcal{X}$, the range of X, to $\mathcal{A}$.

In estimation problems, the action space may be a set of real numbers corresponding to a parameter space. In tests of statistical hypotheses, we may

define the action space as $\mathcal{A} = \{0, 1\}$, in which 0 represents not rejecting and 1 represents rejecting.

If we observe X, we take the action $T(X) = a \in \mathcal{A}$. An action or a decision may be the assignment of a specific value to an estimator, that is, an *estimate*, or it may be to decide whether or not to reject a statistical hypothesis.

Decision Rules

Given a random variable X with associated measurable space $(\mathcal{X}, \mathcal{F}_{\mathcal{X}})$ and an action space $\mathcal{A}$ with a σ-field $\mathcal{F}_{\mathcal{A}}$, a decision rule is a function, T, from $\mathcal{X}$ to $\mathcal{A}$ that is measurable $\mathcal{F}_{\mathcal{X}}/\mathcal{F}_{\mathcal{A}}$.

A decision rule is also often denoted by δ or $\delta(X)$.

Randomized Decision Rules

Sometimes the available data, that is, the realization of X, does not provide sufficient evidence to make a decision. In such cases, of course, it would be best to obtain more data before making a decision. If a decision must be made, however, it may be desirable to choose an action randomly, perhaps under a probability model that reflects the available evidence. A *randomized decision rule* is a function δ over $\mathcal{X} \times \mathcal{F}_{\mathcal{A}}$ such that for every $A \in \mathcal{F}_{\mathcal{A}}$, $\delta(\cdot, A)$ is a Borel function, and for every $x \in \mathcal{X}$, $\delta(x, \cdot)$ is a probability measure on $(\mathcal{A}, \mathcal{F}_{\mathcal{A}})$.

To evaluate a randomized decision rule requires the realization of an additional random variable. As suggested above, this random variable may not be independent of the data. Randomized decision rules are rarely appropriate in actual applications, but an important use of randomized decision rules is to evaluate properties of statistical procedures. In the development of statistical theory, we often use randomized decision rules to show that certain deterministic rules do or do not have certain properties.

Loss Function

A *loss function*, L, is a mapping from $\mathcal{P} \times \mathcal{A}$ to $[0, \infty[$. The value of the function at a given distribution P for the action a is $L(P, a)$. More commonly, we refer to the loss function associated with a given nonrandomized decision rule $T(X)$ as composition $L(P, T(X))$. For a given rule T, we may also denote the loss function as $L_T(P)$. If the class of distributions is indexed by a parameter θ, we may use the equivalent notation $L(\theta, T)$ or $L_T(\theta)$.

Given a loss function $L(P, a)$, the loss function associated with a given randomized decision rule $\delta(X, A)$ is

$$L(P, \delta(X, A)) = \mathrm{E}_{\delta(X,\cdot)}(L(P, Y))$$

where Y is a random variable corresponding to the probability measure $\delta(X, \cdot)$.

If $\mathcal{P}$ indexed by θ, we can write the value of the function at a given value θ for the action a as $L(\theta, a)$.

The loss function is defined with respect to the objectives of the statistical inference in such a way that a small loss is desired.

Depending on Θ, $\mathcal{A}$, and our objectives, the loss function often is a function only of $a - g(\theta)$ or of $a/g(\theta)$, where if a and $g(\theta)$ are vectors, $a/g(\theta)$ may represent element-wise division or some other appropriate operation. We may have, for example,

$$L(\theta, a) = L_1(g(\theta) - a) = \|g(\theta) - a\|.$$

In this case, which might be appropriate for estimating $g(\theta)$,

$$\begin{aligned} L(\theta, a) &\geq 0 \quad \forall \theta, a \\ L(\theta, a) &= 0 \quad \text{if } a = g(\theta). \end{aligned}$$

Notice that the loss function is just a mathematical function associated with a function g of distribution measures. There are no assumed underlying random variables. It does not matter what θ and a are; they are just mathematical variables, or placeholders, taking values in Θ and $\mathcal{A}$. In this case, the loss function generally should be nondecreasing in $\|g(\theta) - a\|$. A loss function that is convex has nice mathematical properties. (There is some heuristic appeal to convexity, but we like it because of its mathematical tractability. There are lots of other properties of statistical procedures that are deemed interesting for this nonreason.)

While the loss function may take on various forms, in a situation where we assume the underlying class of probability distributions is $\mathcal{P}$, for any function L that is a loss function we will assume that for any action a, there is a subclass $\mathcal{P}_L \subseteq \mathcal{P}$ of positive measure (of the appropriate type, usually Lebesgue) such that

$$L(P, a) > 0, \quad \text{for } P \in \mathcal{P}_L. \tag{3.84}$$

Without this condition, the loss function would have no meaning for evaluating statistical procedures.

Common Forms of Loss Functions

In the following, for simplicity, we will assume that $g(\theta)$ and a are scalars. Each of these forms of loss functions can easily be extended to the vector case by use of an appropriate norm.

A particularly nice loss function, which is strictly convex, is the "squared-error loss":

$$L_2(\theta, a) = \|g(\theta) - a\|_2. \tag{3.85}$$

If $g(\theta)$ and a are scalars, the squared-error loss becomes

$$L_2(\theta, a) = (g(\theta) - a)^2.$$

Another loss function that is often appropriate is the "absolute-error loss":

$$L_1(\theta, a) = \|g(\theta) - a\|_1, \tag{3.86}$$

which is just the absolute value of the difference if $g(\theta)$ and a are scalars. The absolute-error loss, which is convex but not strictly convex, is not as mathematically tractable as the squared-error loss.

Sometimes, especially if $g(\theta)$ and a are scalars, it is appropriate that the loss function be asymmetric; that is, the cost if $g(\theta) > a$ increases more (or less) rapidly than if $g(\theta) < a$. A simple generalization of the absolute-error loss that provides this asymmetry is

$$L(\theta, a) = \begin{cases} c(a - g(\theta)) & \text{for } a \geq g(\theta) \\ (1 - c)(g(\theta) - a) & \text{for } a < g(\theta) \end{cases} \tag{3.87}$$

for $0 < c < 1$.

Another common loss function that is asymmetric is the so-called "linex" loss function,

$$L(\theta, a) = \mathrm{e}^{c(g(\theta)-a)} - c(g(\theta) - a) - 1, \tag{3.88}$$

for scalars $g(\theta)$ and a. If c is negative than the linex loss function increases linearly if $g(\theta) > a$ and exponentially if $g(\theta) < a$ (hence, the name "linex"), and just the opposite if c is positive.

When the action space is binary, that is, $\mathcal{A} = \{0, 1\}$, a reasonable loss function may be the 0-1 loss function

$$\begin{array}{ll} L_{0-1}(\theta, a) = 0 & \text{if } g(\theta) = a \\ L_{0-1}(\theta, a) = 1 & \text{otherwise.} \end{array} \tag{3.89}$$

Any strictly convex loss function over an unbounded interval is unbounded. Even when the action space is dense, it is not always realistic to use an unbounded loss function. In such a case we may use the 0-1 loss function defined as

$$\begin{array}{ll} L_{0-1}(\theta, a) = 0 & \text{if } |g(\theta) - a| \leq \alpha(n) \\ L_{0-1}(\theta, a) = 1 & \text{otherwise.} \end{array} \tag{3.90}$$

In a binary decision problem in which the true state is also 0 or 1, we often formulate a loss function of the form

$$L(\theta, a) = \begin{cases} c_a & \text{for 0 true} \\ b_a & \text{for 1 true} \end{cases} \tag{3.91}$$

where $c_1 > c_0$ and $b_0 > b_1$.

It is common to choose $c_0 = b_1 = 0$, $b_0 = 1$, and $c_1 = \gamma > 0$. In this case, this is called a *0-1-γ loss function*.

Another approach to account for all possibilities in the binary case, and to penalize errors differently depending on the true state and the decision is to define a weighted 0-1 loss function:

$$L(\theta, a) = \begin{cases} 0 & \text{if } a = 0 \text{ and } 0 \text{ true} \\ 0 & \text{if } a = 1 \text{ and } 1 \text{ true} \\ \alpha_0 & \text{if } a = 1 \text{ and } 0 \text{ true} \\ \alpha_1 & \text{if } a = 0 \text{ and } 1 \text{ true.} \end{cases} \tag{3.92}$$

This is sometimes called a α_0-α_1 loss or a weighted 0-1 loss.

Risk Function

To choose an action rule T so as to minimize the loss function is not a well-defined problem. The action itself depends on the random observations, so the action is $T(X)$, which is a random variable.

We can make the problem somewhat more precise by considering the expected loss based on the action $T(X)$, which we define to be the *risk*:

$$R(P, T) = \mathrm{E}\big(L(P, T(X))\big). \tag{3.93}$$

We also often write the risk $R(P, T)$ as $R_T(P)$.

The expectation that defines the risk is taken with respect to the distribution P, the "true", but unknown distribution; thus, the risk is a function of the distribution, both because the loss is a function of the distribution and because the expectation is taken with respect to the distribution.

If the family of distributions are indexed by a parameter θ, then the risk is a function of that parameter, and we may write $R(\theta, T)$.

Optimal Decision Rules

We compare decision rules based on their risk with respect to a given loss function and a given family of distributions. If a decision rule T^* has the property

$$R(P, T^*) \leq R(P, T) \quad \forall\, P \in \mathcal{P}, \tag{3.94}$$

for all T, then T^* is called an *optimal* decision rule.

Often we limit the set of possible rules. If

$$R(P, T^*) \leq R(P, T) \quad \forall\, P \in \mathcal{P} \text{ and } \forall\, T \in \mathcal{T}, \tag{3.95}$$

then T^* is called a $\mathcal{T}$*-optimal* decision rule.

For the case of a convex loss function, when a sufficient statistic exists an optimal decision rule depends on that sufficient statistic. This fact derives from Jensen's inequality (B.13) on page 849, and is codified in the Rao-Blackwell theorem:

Theorem 3.8 (Rao-Blackwell theorem)
Suppose the loss function is convex. Let T be a sufficient statistic for $P \in \mathcal{P}$, and let T_0 be a statistic with finite risk. Let

$$T_{\mathrm{RB}} = \mathrm{E}(T_0|T).$$

Then

$$R(P, T_{\mathrm{RB}}) \leq R(P, T_0) \quad \forall\, P \in \mathcal{P}.$$

The statistic T_{RB} is called a "Rao-Blackwellized" version of T_0.

Admissibility

Before considering specific definitions of minimum-risk procedures, we define another general desirable property for a decision rule, namely, admissibility. We define admissibility negatively in terms of dominating rules.

Definition 3.12 (dominating rules)
Given decision rules T and T^* for a family of distributions $\mathcal{P}$, with a risk function R. The rule T is said to *dominate* the rule T^* iff

$$R(P, T) \leq R(P, T^*) \quad \forall\, P \in \mathcal{P}, \tag{3.96}$$

and

$$R(P, T) < R(P, T^*) \quad \text{for some } P \in \mathcal{P}. \tag{3.97}$$

■

Definition 3.13 (admissible rules) A decision rule T^* is *admissible* if there does not exist a decision rule T that dominates T^*. ■

Note that admissibility depends on

- the loss function L
- $\mathcal{P}$, the family of distributions wrt which E is defined

For a given problem there may be no admissible decision rule.

The fact that a decision rule T^* is admissible does not mean that the risk curve of some other decision rule cannot dip below the risk curve of T^* at some points.

Often we limit the set of possible rules to a set $\mathcal{T}$, and we have $\mathcal{T}$-admissibility:
A decision rule T^* is *$\mathcal{T}$-admissible* if there does not exist a decision rule within the class of decision rules $\mathcal{T}$ that dominates T^*.

A slightly more general form of admissibility is λ-admissibility:
A decision rule T^* is *λ-admissible* if T^* is admissible almost everywhere with respect to the measure λ defined over the sample space.

Optimality of a decision rule under whatever criterion implies admissibility of the rule under that criterion.

Completeness of a Class of Decision Rules

We have defined completeness of distributions (on page 162) and of statistics. We now define completeness of a class of decision rules. A class of decision rules $\mathcal{T}$ is said to be *complete* if for any decision rule $T \notin \mathcal{T}$, there exists a rule in $\mathcal{T}$ that dominates T. A class is said to be *minimal complete* if it does not contain a complete proper subclass.

If two decision rules have identical risk functions, we would like to think of them as equivalent, but we do not want necessarily to include all such equivalent rules in a class of interest. We therefore define a class of rules $\mathcal{T}$ to be *essentially complete* if for any rule T there is a rule $T_0 \in \mathcal{T}$ such that $R(P, T_0) \leq R(P, T)\ \forall\, P$.

Let $\mathcal{T}$ be a class of decision rules and let $\mathcal{T}_0 \subseteq \mathcal{T}$. The class $\mathcal{T}_0$ is said to be $\mathcal{T}$*-complete* if $\forall T \in \mathcal{T} - \mathcal{T}_0, \exists T_0 \in \mathcal{T}_0$ that dominates T.

The class $\mathcal{T}_0$ is said to be $\mathcal{T}$*-minimal complete* if $\mathcal{T}_0$ is $\mathcal{T}$*-complete* and no proper subset of $\mathcal{T}_0$ is $\mathcal{T}$*-complete*.

It is easy to see (using the method of proving one set is equal to another by showing each is a subset of the other) that if a $\mathcal{T}$-minimal complete class exists, it is identical to the class of $\mathcal{T}$-admissible decision rule.

One of the most fundamental approaches to statistical inference is to identify a complete class of decision rules and then to seek rules within that class that have various desirable properties. One of the most widely-used complete class theorem is the one that states that Bayes rules and simple generalizations of them constitute a complete class (see page 353).

***** Wolfowitz (1951) ϵ-complete classes of decision functions

L-Unbiasedness

Admissibility involves the relationship between the expected values of the loss function with different decision rules at the same distribution in the family being considered. We can also consider the expected values taken at a given point in the distribution space of the loss function of a given decision rule at the given value of the parameter compared with the loss at some other distribution. This leads to the concept of L-unbiasedness.

A decision rule T is L-unbiased for a given loss function L if for all P and $\tilde{P}$,

$$\mathrm{E}_P\big(L(P, T(X))\big) \leq \mathrm{E}_P\big(L(\tilde{P}, T(X))\big). \tag{3.98}$$

The expression on the left of equation (3.98) is the risk of T for given L, but the expression on the right is not a risk. Notice that L-unbiasedness relates to the same rule evaluated under the same expectation at different points in the space of distributions. Admissibility, on the other hand, relates to different rules evaluated at the same point in the space of distributions as the distribution used in the expectation operation. A decision rule may be L-unbiased but not admissible; in the $\mathrm{N}(\mu, 1)$ distribution, for example,

the sample median is L-unbiased under a squared-error loss, but it is not admissible under that loss, while the sample mean is L-unbiased under an absolute-error loss, but it is not admissible.

This is the basis for defining unbiasedness for statistical tests and confidence sets.

Unbiasedness for estimators has a simple definition. For squared-error loss for estimating $g(\theta)$, if T is L-unbiased, then, and only then, T is an unbiased estimator of $g(\theta)$. Of course, in this case, the loss function need not be considered and the requirement is just $\mathrm{E}_\theta(T(X)) = g(\theta)$.

L-Invariance

On page 221 we referred to equivariant estimators in parametric transformation group families (see Section 2.6). We mentioned that associated with the group $\mathcal{G}$ of transformations of the random variable is a group, $\overline{\mathcal{G}}$, of transformations of the parameter and a group of transformations on the estimator, $\mathcal{G}^*$.

In a decision-theoretic approach, the relevance of equivariance depends not only on the family of distributions, but also on the equivariance of the loss function. In the loss function $L(P, T(X))$, the first argument under a transformation can be thought of as a map $P_X \to P_{g(X)}$ or equivalently as a map $P_\theta \to P_{\bar{g}(\theta)}$. The statistical decision procedure $T(X)$ is *L-invariant* for a given loss function L if for each $g \in \mathcal{G}$, there exists a unique $g^* \in \mathcal{G}^*$, such that

$$L(P_X, T(X)) = L(P_{g(X)}, g^*(T(X))), \tag{3.99}$$

or equivalently for each $\bar{g} \in \overline{\mathcal{G}}$,

$$L(P_\theta, T(X)) = L(P_{\bar{g}(\theta)}, g^*(T(X))). \tag{3.100}$$

The g^* in these expressions is the same as in equation (3.22). We will often require that statistical procedures be *equivariant*, in the sense that the quantities involved (the estimators, the confidence sets, and so on) change in a manner that is consistent with changes in the parametrization. The main point of this requirement, however, is to ensure L-invariance, that is, invariance of the loss. We will discuss equivariance of statistical procedures in more detail in Section 3.4.

Uniformly Minimizing the Risk

All discussions of statistical inference are in the context of some family of distributions, and when we speak of a "uniform" property, we mean a property that holds for all members of the family.

If we have the problem of estimating $g(\theta)$ under some given loss function L, it is often the case that for some specific value of θ, say θ_1, one particular

estimator, say T_1, has the smallest expected loss, while for another value of θ, say θ_2, another estimator, say T_2, has a smaller expected loss. Neither T_1 nor T_2 is uniformly optimal.

The risk is a function of the parameter being estimated; therefore, to minimize the risk is not a well-posed problem. A solution is to seek a decision rule that is uniformly best within some restricted class of decision rules.

3.3.2 Approaches to Minimizing the Risk

We use the principle of minimum risk in the following restricted ways. In all cases, the approaches depend, among other things, on a given loss function.

- If there is a sufficient statistic and if the loss function is convex, we use the result of the Rao-Blackwell theorem; that is, we condition any given statistic T_0 on the sufficient statistic, T:

  $$T_{\mathrm{RB}} = \mathrm{E}(T_0|T).$$

  Finding a statistic with a smaller risk by this method is called "Rao-Blackwellization".
  Note that If the loss function is strictly convex and T_0 is not a function of T, then T_0 is inadmissible.
- We may first place a restriction on the statistical procedure and then minimize risk subject to that restriction.
  For example, in estimation problems:
  - require unbiasedness
    In this case, we can often eliminate θ from consideration; that is, we can uniformly minimize the risk.
    In a common situation we define loss as squared-error (because it is unbiased, this means variance), and this yields UMVU.
    Sufficiency and completeness play a major role in UMVUE.
    The information inequality is important in unbiased estimation.
    This approach is great for exponential families.
  - require equivariance
    This must be made more precise (unlike unbiasedness, "equivariance" requires more qualification).
    Equivariance implies independence of the risk from θ; we can uniformly minimize the risk by just minimizing it anywhere.
    This yields UMRE, or just MRE because uniformity is implied.
    This approach is especially useful for group families.
- We may minimize some global property of the risk ("global" over the values of θ).
  For example:
  - minimize maximum risk
    The risk may be unbounded, so obviously in that case, it does not

make sense to attempt to minimize the maximum risk. Even if the risk is unbounded, the maximum risk may not exist, so we consider

$$\sup_{\theta\in\Theta} R(\theta, T(X)). \tag{3.101}$$

The estimator that yields

$$\inf_{T} \sup_{\theta\in\Theta} R(\theta, T(X)) \tag{3.102}$$

is the minimax estimator.

A comment about the supremum may be in order here. We mentioned earlier that in parametric inference, we often consider the closure of the parameter space, $\overline{\Theta}$, and in the maximum likelihood estimator in equation (3.49), for example, that allowed us to consider $\max\{\theta \in \overline{\Theta}\}$. We cannot do this in considering the "maximum" risk in equation (3.101) because we do not know how R behaves over $\overline{\Theta}$. (It could be discontinuous anywhere within $\overline{\Theta}$.)

– minimize "average" risk

How to average? Let $\Lambda(\theta)$ be such that $\int_\Theta \mathrm{d}\Lambda(\theta) = 1$, then average risk is $\int_\Theta R(\theta, T)\mathrm{d}\Lambda(\theta)$.

The estimator that minimizes the average risk wrt $\Lambda(\theta)$, T_Λ, is called the Bayes estimator, and the minimum risk, $\int_\Theta R(\theta, T_\Lambda)\mathrm{d}\Lambda(\theta)$, is called the Bayes risk.

The averaging function allows various interpretations, and it allows the flexibility of incorporating prior knowledge or beliefs. The regions over which $\Lambda(\theta)$ is large will be given more weight; therefore the estimator will be pulled toward those regions.

In formal Bayes procedures we follow the approach indicated in equations (3.3) and (3.4). The distribution Q_0 in equation (3.3) has the PDF $\mathrm{d}\Lambda(\theta)$, which we call the *prior probability density* for θ.

We then form the joint distribution of θ and X, and then the conditional distribution of θ given X, which is the distribution Q_H in equation (3.3) and is called the posterior distribution. The *Bayes estimator* is determined by minimizing the risk, where the expectation is taken with respect to the posterior distribution.

Because the Bayes estimator is determined by the posterior distribution, the Bayes estimator must be a function of a sufficient statistic.

We will discuss Bayesian inference more fully in Chapter 4.

– combinations of global criteria

We could consider various combinations of the global criteria. For example, we may see an estimator that generally minimizes the average risk, but such that its maximum risk is not so large. An intuitively reasonable bound on the maximum risk would bs some excess of the minimum maximum bound. This approach is called *restricted Bayes*, and results in the following constrained optimization problem:

$$\begin{aligned}&\min_T \int R(\theta, T)\mathrm{d}\Lambda(\theta)\\ &\text{s.t.}\quad \sup_{\theta\in\Theta} R(\theta, T(X)) \le (M+\epsilon)\inf_T \sup_{\theta\in\Theta} R(\theta, T(X))\end{aligned}$$

- We may combine various criteria.
  It is often appropriate to combine criteria or to modify them. This often results in "better" estimators. For example, if for $\theta \in \Theta$, $g(\theta) \in [\gamma_1, \gamma_2]$, and $T(X)$ is an estimator of $g(\theta)$ such that $\Pr(T(X) \notin [\gamma_1, \gamma_2]) \neq 0$, then $T^*(X)$ defined as

$$T^*(X) = \begin{cases} T(X) & \text{if } T(X) \in [\gamma_1, \gamma_2] \\ \gamma_1 & \text{if } T(X) < \gamma_1 \\ \gamma_2 & \text{if } T(X) > \gamma_2 \end{cases}$$

  dominates $T(X)$.
- We may focus on asymptotic criteria.
  Sometimes we seek estimators that have good asymptotic properties, such as consistency.

Optimal Point Estimation under Squared-Error Loss

In estimation problems, squared-error loss functions are often the most logical (despite the examples above!). A squared-error loss function is strictly convex, so the useful properties of convex loss functions, such as those relating to the use of sufficient statistics (Rao-Blackwell, for example), hold for squared-error loss functions. Squared-error is of course the loss function in UMVU estimation, and so we use it often.

Example 3.17 UMVUE of binomial parameter
Consider the binomial family of distributions with fixed n and parameter π. Consider the estimator $T(X) = X/n$ for π. We see that $\mathrm{E}(T(X)) = \pi$: hence T is unbiased, and therefore under squared-error loss, the risk is the variance, which is $pi(1-\pi)/n$. The binomial is a Fisher information regular family, and from equation (3.39), we see that the CRLB is $I(\pi)^{-1} = pi(1-\pi)/n$; hence T is a UMVUE. ∎

Squared-error loss functions yield nice properties for linear functions of estimands. If T is an estimator of $g(\theta)$, then an obvious estimator of $ag(\theta)+b$ is $aT + b$. Under squared-error loss, we have the properties stated in the following theorem.

Theorem 3.9 (linearity of optimal estimators under squared-error loss)

If T is $\left\{\begin{array}{c}\text{Bayes}\\ \text{UMVU}\\ \text{minimax}\\ \text{admissible}\end{array}\right\}$ *for $g(\theta)$, then $aT+b$ is* $\left\{\begin{array}{c}\text{Bayes}\\ \text{UMVU}\\ \text{minimax}\\ \text{admissible}\end{array}\right\}$ *for $ag(\theta)+b$,*
where all properties are taken under squared-error loss.

The various pieces of this theorem will be considered in other places where the particular type of estimation is discussed.

If in a Bayesian setup, the prior distribution and the posterior distribution are in the same parametric family, that is, if $\mathcal{Q}$ in equations (3.3) and (3.4) represents a single parametric family, then a squared-error loss yield Bayes estimators for $\mathrm{E}(X)$ that are linear in X. (If a prior distribution on the parameters together with a conditional distribution of the observables yield a posterior in the same parametric family as the prior, the prior is said to be *conjugate* with respect to the conditional distribution of the observables. We will consider various types of priors more fully in Chapter 4.)

Because we use squared-error loss functions so often, we must be careful not to assume certain common properties hold. Other types of loss functions can provide useful counterexamples.

3.3.3 Admissibility

By Definition 3.13, a decision δ^* is admissible if there does not exist a decision δ that dominates δ^*. Because this definition is given as a negative condition, it is often easier to show that a rule is inadmissible, because all that is required to do that is to exhibit another rule that dominates it. In this section we consider some properties of admissibility and ways of identifying admissible or inadmissible rules.

Admissibility of Estimators under Squared-Error Loss

Any property defined in terms of the risk depends on the loss function. As we have seen above, the squared-error loss often results in estimators that have "nice" properties. Here is another one.

Under a squared-error loss function an unbiased estimator is always at least as good as a biased estimator unless the bias has a negative correlation with the unbiased estimator.

Theorem 3.10
Let $\mathrm{E}(T(X)) = g(\theta)$*, and let* $\widetilde{T}(X) = T(X) + B$*, where* $\mathrm{Cov}(T, B) \geq 0$*. Then under squared-error loss, the risk of* $T(X)$ *is uniformly less than the risk of* $\widetilde{T}(X)$*; that is,* $\widetilde{T}(X)$ *is inadmissible.*

Proof.

$$R(g(\theta), \widetilde{T}) = R(g(\theta), T) + \mathrm{V}(B) + \mathrm{Cov}(T, B) \quad \forall\, \theta.$$

∎

Also under a squared-error loss function, an unbiased estimator dominates a biased estimator unless the bias is a function of the parameter.

Theorem 3.11
Let $\mathrm{E}(T(X)) = g(\theta)$*, and let* $\widetilde{T}(X) = T(X) + B$*, where* $B \neq 0$ *a.s. and* B *is independent of* θ*. Then under squared-error loss,* $\widetilde{T}(X)$ *is inadmissible.*

Proof.

$$R(g(\theta), \widetilde{T}) = R(g(\theta), T) + B^2.$$

∎

Now, let us consider linear estimators of $g(\theta) = \mathrm{E}(T(X))$

$$\widetilde{T}(X) = aT(X) + b$$

that generalize the estimators above, except we consider a and b to be constants. We have the following results under squared-error loss.

- If $a = 1$ and $b \neq 0$, then $\widetilde{T}(X)$ is inadmissible by Theorem 3.11.
- If $a > 1$, then $\widetilde{T}(X)$ is inadmissible for any b because

$$R(g(\theta), \widetilde{T}) = a^2 R(g(\theta), T) > R(g(\theta), T).$$

- If $a < 0$, then $\widetilde{T}(X)$ is inadmissible for any b because

$$R(g(\theta), \widetilde{T}) > R(g(\theta), 0)$$

(Exercise 3.9).

Admissibility of Estimators in One-Parameter Exponential Families

In the previous section, we identified conditions that assured the inadmissibility of linear estimators, and later we will see some examples in which we easily establish inadmissibility. It is of course a more interesting problem to identify conditions that assure admissibility. Efforts to do this are much less successful, but we do have a useful result for linear estimators in a one-parameter exponential family with PDF as given in equation (2.15),

$$f(x) = \beta(\theta)\mathrm{e}^{\theta T(x)}. \tag{3.103}$$

Karlin's theorem MS2 Theorem 4.14
use information inequality (3.39).
*** in binomial, show that $\overline{X}$ is admissible **** example

Admissible and Bayes Estimators

There are important and useful connections between admissible estimators and Bayes estimators.

- A unique Bayes estimator is admissible with respect to the same loss function and distribution.
- An admissible estimator is either Bayes or limiting Bayes with respect to the same loss function and distribution.

We will consider these properties in Chapter 4. It is sometimes easy to construct a Bayes estimator, and having done so, if the estimator is unique, we immediately have an admissible estimator.

Inadmissible Estimators

Some estimators that have generally good properties or that are of a standard type may not be admissible. Heuristic methods such as MLE or the method of moments are not developed in the context of decision theory, so it should not be surprising that estimators based ont these methods may not be admissible.

Example 3.18 Inadmissible Method of Moments Estimator
Consider the case of estimating θ in the finite population $\{1, \ldots, \theta\}$. Suppose we sample from this population with replacement, obtaining $X_1, \ldots, X_n$. Because $\mathrm{E}(\overline{X}) = (\theta+1)/2$, the method of moments estimator of θ is $T = 2\overline{X} - 1$. This estimator is inadmissible (for any reasonable loss function including squared-error loss), since $T^* = \max(X_{(n)}, T)$ is always at least as close to θ, and can be closer.

The method of moments estimator in this case is not even a function of a sufficient statistic, so we would not expect it to have good properties. Note also that the MME of θ may produce a value that could never be the true value of θ. (Of course, that is also the case with T^*.) ∎

There are many surprising cases of inadmissibility, as we see in the following examples. We show that a given rule is not admissible by exhibiting a rule that dominates it. It is important to recognize, of course, that the dominating rule may also not be admissible either.

It may be possible to construct a randomized estimator that shows that a given estimator is not admissible. Another way is to form a scalar multiple of a "good" estimator, and show that it dominates the "good" estimator. In Example 3.19 the scaling is a function of the statistic, and in Example 3.20 the scaling is a constant.

Example 3.19 Inadmissible Estimator of the Mean in a Multivariate Normal Distribution
The estimation of the mean of a normal distribution has interesting admissibility properties. It is relatively straightforward to show that $\overline{X}$ is admissible for estimating θ in $\mathrm{N}(\theta, 1)$ under squared-error loss. It can also be shown that $\overline{X}$ is admissible for estimating θ in $\mathrm{N}_2(\theta, I_2)$, and of course, in the simpler case of $n = 1$, X is admissible for estimating θ.

However, for $r > 2$, X is not admissible for estimating θ in $\mathrm{N}_r(\theta, I_r)$!

For $r > 2$, the estimator

$$\hat{\theta}_{\mathrm{J}} = \left(1 - c\frac{r-2}{\|X\|^2}\right) X \tag{3.104}$$

though biased, dominates X. This is called the James-Stein estimator.

Why this is the case for $r > 2$ has to do with the existence of

$$\mathrm{E}\left(\frac{1}{\|X\|^2}\right);$$

see page 188. ∎

Further comments on Example 3.19

The James-Stein estimator is generally shrunk toward 0. This type of adjustment is called Stein shrinkage. Choice of c allows for different amounts of bias and different amounts of reduction in the risk. The regularization parameter in ridge regression is similar to the c in this expression; see Example 5.27.

The fact that shrinkage in the case of the multivariate normal distribution may improve the estimator is related to the outlyingness of data in higher dimensions.

A shrunken estimator is biased, and ordinarily we would not expect a biased estimator to dominate a "good" unbiased one. It should be noted, however, that the bias of the shrunken estimator has a negative correlation with the basic estimator (recall Theorem 3.10).

The fact that the James-Stein estimator dominates the UMVUE, however, does not mean that the James-Stein estimator itself is admissible. Indeed, it is not. The estimator

$$\hat{\theta}_{\mathrm{J}_+} = \min\left(1, c\frac{r-2}{\|X\|^2}\right) X \tag{3.105}$$

dominates the James-Stein estimator under squared-error loss. This is sometimes called the positive-part James-Stein estimator. The positive-part James-Stein estimator, however, is also inadmissible under squared-error loss (see Strawderman (1971) for further discussion). ∎

Consider another example, due to Lehmann, for a general one-parameter exponential family.

Example 3.20 Inadmissible Estimator in a One-Parameter Exponential Family

Let X have the density

$$p_\theta(x) = \beta(\theta)\mathrm{e}^{\theta x}\mathrm{e}^{-|x|},$$

where $\theta \in]-1, 1[$ and $\beta(\theta) = 1-\theta^2$ (so it integrates to 1). Consider a sample of size one, X, and the problem of estimating $g(\theta) = \mathrm{E}_\theta(X)$ with squared-error loss. Now,

$$\begin{aligned} \mathrm{E}_\theta(X) &= -\frac{\beta'(\theta)}{\beta(\theta)} \\ &= \frac{2\theta}{1-\theta^2}, \end{aligned}$$

and

$$\begin{aligned} \mathrm{V}_\theta(X) &= \frac{\mathrm{d}}{\mathrm{d}\theta}\mathrm{E}_\theta(X) \\ &= 2\frac{1+\theta^2}{(1-\theta^2)^2}; \end{aligned}$$

hence, the risk is

$$R(g(\theta), X) = 2\frac{1+\theta^2}{(1-\theta^2)^2}.$$

Now, consider the estimator $T_a = aX$. Its risk under squared-error is

$$\begin{aligned} R(\theta, T_a) &= \mathrm{E}_\theta(L(\theta, T_a)) \\ &= \mathrm{E}_\theta((g(\theta) - T_a)^2) \\ &= 2a^2\frac{1+\theta^2}{(1-\theta^2)^2} + 4(1-a^2)\frac{\theta^2}{(1-\theta^2)^2}. \end{aligned}$$

If $a = 0$, that is, the estimator is the constant 0, the risk is $4\theta^2/(1-\theta^2)^2$, which is smaller than the risk for X for all $\theta \in]-1, 1[$.

The natural sufficient statistic in this one-parameter exponential family is inadmissible for its expectation! ∎

Other Forms of Admissibility

We have defined admissibility in terms of a specific optimality criterion, namely minimum risk. Of course, the risk depends on the loss function, so admissibility depends on the particular loss function.

Although this meaning of admissibility, which requires a decision-theory framework, is by far the most common meaning, we can define admissibility in a similar fashion with respect to any optimality criterion; for example, the estimator $T(X)$ is *Pitman-admissible* for $g(\theta)$ if there does not exist an estimator that is *Pitman-closer* to $g(\theta)$. In Example 3.3 on page 220 we saw that the sample mean even in a univariate normal distribution is not Pitman admissible. The type of estimator used in that example to show that the univariate mean is not Pitman admissible is a shrinkage estimator, just as a shrinkage estimator was used in Example 3.19.

3.3.4 Minimaxity

Instead of uniform optimality properties for decisions restricted to be unbiased or equivariant or optimal average properties, we may just seek to find one with the smallest maximum risk. This is *minimax estimation*.

For a given decision problem, the maximum risk may not exist, so we consider

$$\sup_{\theta\in\Omega} R(\theta, \delta(X)).$$

The decision that yields

$$\inf_{\delta} \sup_{\theta\in\Omega} R(\theta, \delta(X)) \tag{3.106}$$

is the *minimax decision*.

Minimaxity, as with most optimality properties, depends on the loss function.

Minimaxity and Admissibility

There are various connections between minimaxity and admissibility.

Theorem 3.12
An admissible estimator with a constant risk is minimax with respect to the same loss function and distribution.

Proof.
We see that this must be the case because if such an estimator were not minimax, then an estimator with smaller maximum risk would dominate it and hence it would not be admissible. ∎

Minimax and Bayes Estimators

Just as with admissible estimators, there are interesting connections between minimax and Bayes estimators. One of the most important is the fact that a Bayes estimator with a constant risk is minimax with respect to the same loss function and distribution. (This is Theorem 4.4.) Hence, one way of finding a minimax estimator is to find a Bayes estimator with constant risk. For a given loss function, and given distribution of the observable random variable, the minimax estimator is the Bayes estimator for "worst" prior distribution. We will consider this and other properties in Chapter 4.

Minimax Estimators under Squared-Error Loss in Exponential Families

For one-parameter exponential families, under squared-error loss, Theorem 4.14 in MS2 provides a condition for identifying admissible estimators, and hence minimax estimators. The minimax estimator is not always the obvious one.

Often a randomized estimator can be constructed so that it is minimax.

Example 3.21 Risk functions for estimators of the parameter in a binomial distribution
Suppose we have an observation X from a binomial distribution with parameters n and π. The PDF (wrt the counting measure) is

$$p_X(x) = \binom{n}{x}\pi^x(1-\pi)^{n-x}\mathrm{I}_{\{0,1,\ldots,n\}}(x).$$

We wish to estimate π.

The MLE of π is

$$T(X) = X/n. \tag{3.107}$$

We also see that this is an unbiased estimator. Under squared-error loss, the risk is

$$R_T(\pi) = \mathrm{E}\left((X/n-\pi)^2\right) = \pi(1-\pi)/n,$$

which, of course, is just variance.

Let us consider a randomized estimator, for some $0 \leq \alpha \neq 1$,

$$\delta_\alpha(X) = \begin{cases} T & \text{with probability } 1-\alpha \\ 1/2 & \text{with probability } \alpha \end{cases} \tag{3.108}$$

This is a type of shrunken estimator. The motivation to move T toward $1/2$ is that the maximum of the risk of T occurs at $1/2$. By increasing the probability of selecting that value the risk at that point will be reduced, and so perhaps this will reduces the risk in some overall way.

Under squared-error loss, the risk of $\delta_\alpha(X)$ is

$$\begin{aligned} R_{\delta_\alpha}(\pi) &= (1-\alpha)\mathrm{E}\left((X/n-\pi)^2\right) + \alpha\mathrm{E}\left((1/2-\pi)^2\right) \\ &= (1-\alpha)\pi(1-\pi)/n + \alpha(1/2-\pi)^2. \end{aligned}$$

The mass point has a spreading effect and the risk dips smoothly The risk of $\delta_\alpha(X)$ also has a maximum at $\pi = 1/2$, but it is $(1-\alpha)/4n$, compared to $R_T(1/2) = 1/4n$.

We see that for $\alpha = 1/(n+1)$ the risk is constant with respect to π; therefore $\delta_{1/(n+1)}(X)$ is a minimax estimator wrt squared-error loss.

Risk functions are shown in Figure 3.1 for $T(X)$ and for $\delta_{.05}(X)$ and $\delta_{1/(n+1)}(X)$. Notice that neither $\delta_\alpha(X)$ nor $T(X)$ dominates the other. ∎

3.3.5 Summary and Review

We have discussed five general approaches to statistical inference, and we have identified certain desirable properties that a method of inference may have.

A first objective in mathematical statistics is to characterize optimal properties of statistical methods. The setting for statistical inference includes the distribution families that are assumed a priori, the objectives of the statistical

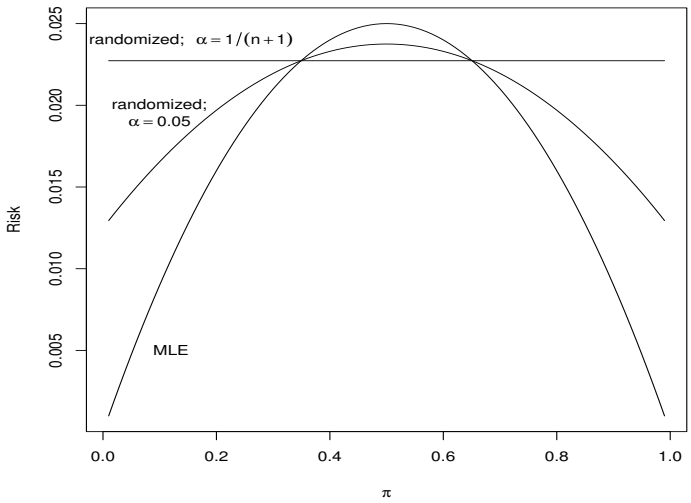


Figure 3.1. Risk Functions for Squared-Error Loss in Binomial with $n = 10$ (Example 3.21).

inference, and the criteria by which the statistical methods to achieve those objectives are to be evaluated.

A second objective in mathematical statistics is to develop techniques for finding optimal methods in a particular setting. We have considered some of these procedures above, and they will be major recurring topics throughout the rest of this book.

Nonexistence of Optimal Methods

There are many criteria by which to evaluate a statistical method. In a given setting **there may not be a statistical procedure that is optimal** with respect to a given criterion.

The criteria by which to evaluate a statistical method include basic things about the nature of the statistics used in the method, such as sufficiency, minimality, and completeness. These properties are independent of the objectives of the procedure and of the particular statistical method used. They depend on the assumed distribution family and the nature of the available sample.

- **sufficiency.** There is always a sufficient statistic.
- **minimal sufficiency.** There is always a minimal sufficient statistic.
- **completeness.** There may not be a complete statistic. This depends on the assumed family of distributions. (See Example 3.4.)

For a given objective in a given family of distributions maximizing the likelihood is a desirable property for a statistical procedure; that is, other things being equal, we would choose a statistical procedure that maximizes the likelihood.

- **maximum likelihood.** There may not be a procedure that maximizes the likelihood. (See Example 6.13.)

For a given objective there are a number of criteria for evaluating a statistical method that is to achieve that objective. Specifically, if the objective is estimation, we can identify three general criteria for evaluating an estimator

- **Pitman closeness.** There may not be an estimator that is Pitman-closest.
- **equivariance.** Equivariance depends first of all on the assumed family of distributions and the group of transformations. There may not be an equivariant estimator in a given setting.
- **expected difference.** There are various functions of the difference $T_n(X) - g(\theta)$, whose expected values may be relevant. Whether or not an expected value of the difference can be minimized depends on the family of distributions, the nature of T_n (and on n itself), and on the nature of $g(\theta)$. For example in a Cauchy distribution, for many functions T_n, $\mathrm{E}(T_n(X))$ does not exist. In this particular case, $\mathrm{E}(T_n(X))$ may exist for n greater than some value k, but not exist for $n < k$.
  - **bias; $\mathrm{E}(T(X) - g(\theta))$.** An unbiased estimator would seem to be desirable. For a given family of distributions and a given $g(\theta)$, there may not be an unbiased estimator. (See Examples 5.2 and 5.3.)
  - **mean absolute error; $\mathrm{E}(|T(X) - g(\theta)|)$.** For a given family of distributions and a given $g(\theta)$, $\mathrm{E}(|T(X) - g(\theta)|)$ may not exist, but if it does there may be no estimator T_n that minimizes it for all θ.
  - **mean squared error; $\mathrm{E}((T(X) - g(\theta))^2)$.** For a given family of distributions and a given $g(\theta)$, $\mathrm{E}((T(X) - g(\theta))^2)$ may not exist, but if it does there may be no estimator T_n that minimizes it for all θ.
  - **conditional expectation; for example, $\mathrm{E}((T(X) - g(\theta))^2 | \mathrm{E}(T(X) - g(\theta)) = 0)$.** For a given family of distributions and a given $g(\theta)$, the conditional distribution may not exist, but if it does there may be no estimator T_n that minimizes the conditional expectation for all θ.

In a decision theory approach, the criteria for evaluating a statistical method revolve around the loss function, which depends on the objectives of the procedure. These criteria generally involve some kind of expectation of the loss function, such as the risk or the expectation of the loss taken with respect to the posterior distribution.

- **minimum risk.** For any nontrivial inference problem, it is generally not possible to minimize the risk uniformly for almost any reasonable loss function. (See page 266.)

- **minimum risk with restrictions.** We often impose the restriction that the procedure be unbiased or be equivariant. Under either of these restrictions there may not be an optimal statistical procedure. As we have already noted, there may not even be a procedure that satisfies the restriction of being unbiased.
- **minimum maximum risk.** The risk may be unbounded, so no minimax procedure can exist.
- **minimum average risk.** Whether or not there can be a procedure that minimizes the average risk clearly depends on the averaging process.

In the frequentist approach to decision theory, admissibility is an important unifying concept. Having defined the concept, we can limit our search for optimal procedures to admissible procedures if they can be identified. From a negative perspective, if we have a procedure that is optimal with respect to other criteria, we generally ask whether or not it is admissible. We may show that the procedure being considered is not admissible by demonstrating a procedure that dominates the procedure in question. Often, when we do this however, the dominating procedure is not admissible either.

In the next section we consider the restriction of equivariance that we have referred to already. This property is relevant only in inference problems that can be formulated in a way that connects the underlying sample space, parameter space, and loss function in a special way.

3.4 Invariant and Equivariant Statistical Procedures

Statistical decisions or actions based on data should not be affected by simple transformations on the data or by reordering of the data, so long as these changes on the data are reflected in the statement of the decision; that is, the actions should be *invariant*. If the action is a yes-no decision, such as in hypothesis testing, it should be completely invariant. If a decision is a point estimate, its value is not unaffected, but it should be *equivariant*, in the sense that it reflects the transformations in a meaningful way so that the loss function should be invariant to the transformations.

Given a decision problem with loss function L, we seek a decision rule that is L-invariant (see page 266).

In the following, we will formalize this equivariance principle by defining appropriate classes of transformations, and then specifying rules that statistical decision functions must satisfy. We identify "reasonable" classes of transformations on the sample space and the corresponding transformations on other components of the statistical decision problem. We will limit consideration to transformations that are one-to-one and onto.

Development of equivariant statistical procedures is based on an algebraic group of transformations (a group in which the operation is composition of functions; see Definition 0.0.2 on page 629 and Section 0.1.11 beginning on

page 754) and a suitable family of probability distributions, such as a "group family" of distributions. (See Section 2.6, beginning on page 178, for further discussion of such families of distributions.)

3.4.1 Formulation of the Basic Problem

We are interested in what happens under a one-to-one transformation of the random variable $g(X)$; in particular, we are interested in a transformation of the parameter $\tilde{g}(\theta)$ such that $P_{g(X)|\tilde{g}(\theta)}$ is a member of the same distributional family. (In this section we will consider only parametric inference; that is, we will consider distributions $P_{X|\theta}$ for $\theta \in \Theta$, but in a more general sense, we can just consider θ to be some index in the distributional family.) We want to identify optimal methods of inference for $P_{X|\theta}$ that will remain optimal for $P_{g(X)|\tilde{g}(\theta)}$.

Whether or not such methods exist depends on the type of transformations, the distributions and parameters of interest, and the form of the loss function. In the following, we will identify the special cases that admit minimum risk equivariant procedures.

The invariance or equivariance of interest is with respect to a given class of transformations. A family of distributions whose probability measures accommodate a group of transformations in a natural way is called a *group family*. The most common class of transformations of interest is the group of linear transformations of the form $\tilde{x} = Ax + c$, and the group families of interest have a certain invariance with respect to a group of linear transformations on the random variable. We call such a group family a *location-scale family* (Definition 2.3). More generally, given a distribution with parameter θ, that distribution together with a group of transformations on θ forms a group family.

Transformations on the Sample Space, the Parameter Space, and the Decision Space

Following Definition 2.4 for an *invariant parametric family of distributions*, we have two transformation groups, $\mathcal{G}$, with elements

$$g : \mathcal{X} \mapsto \mathcal{X}, \quad 1:1 \text{ and onto},$$

and the induced group $\widetilde{\mathcal{G}}$, with elements

$$\tilde{g} : \Theta \mapsto \Theta, \quad 1:1 \text{ and onto},$$

in such a way that for given $g \in \mathcal{G}$, there is a $\tilde{g} \in \widetilde{\mathcal{G}}$ such that for any set A,

$$\Pr_\theta(g(X) \in A) = \Pr_{\tilde{g}(\theta)}(X \in A); \tag{3.109}$$

that is, $\tilde{g}$ *preserves* Θ. The group $\widetilde{\mathcal{G}}$ is *transitive* in the sense defined on page 755.

**** we consider h *** give restrictions on h ***** Given a Borel function h on Θ and the general problem of making inferences about $h(\theta)$ under a given loss function L, if transformations $g \in \mathcal{G}$ and $\tilde{g} \in \widetilde{\mathcal{G}}$ that preserve the probability model are made, we seek a statistical procedure T such that

$$L(h(\theta), T(X)) = L(h(\tilde{g}(\theta)), T(g(X))). \tag{3.110}$$

For a general statistical decision function T, we seek a transformation g^* that yields the same (or appropriately transformed) decision within the transformed distribution using the transformed data. The decision function takes the sample space into the decision space $\mathcal{A}$; that is, $T : \mathcal{X} \mapsto \mathcal{A} \subseteq \mathbb{R}$.

*** give formal definition

*** give examples of orbits

Invariance of the Loss Function

For a given loss function $L(\theta, T(X))$ with transformations g and $\tilde{g}$ applied to the observable random variable and to the parameter, we seek a transformation g^* such that

$$L(\theta, T(X)) = L(\tilde{g}(\theta), g^*(T(X))). \tag{3.111}$$

Such transformations yield invariance of the risk:

$$\mathrm{E}_\theta(L(\theta, T(X))) = \mathrm{E}_{\tilde{g}(\theta)}(L(\tilde{g}(\theta), g^*(T(X)))). \tag{3.112}$$

The question is whether or not such a transformation exists. Its existence clearly depends on the loss function. If equation (3.111) holds, then the transformation g^* is said to be L-invariant with respect to the loss function L.

In most statistical decision problems, we assume a *symmetry* or *invariance* or *equivariance* of the problem before application of any of these transformations, and the problem that results from applying a transformation. For given classes of transformations, we consider loss functions that admit L-invariant transformations; that is, we require that the transformation have the property of L-invariance with respect to the loss function as expressed in equation (3.111). This means that a good statistical procedure, T, for the original problem is good for the transformed problem. Note that this is an *assumption* about the class of meaningful loss functions for this kind of statistical problem.

Example 3.22 Transformations in a Bernoulli distribution
Suppose we have a random sample of size n, $X_1, \ldots, X_n$ from the Bernoulli(π) distribution and we wish to estimate π. In Example 3.17, we found that $T(X) = \sum X_i/n$ is a UMVUE for π; that is, it is optimal under squared-error loss among the class of unbiased estimators.

Now, consider the binomial transformation of Example 2.3. In this transformation, the values assumed by the binary random variables are reversed;

that is, the random variable is transformed as $g(X) = 1 - X$. We see that the transformation $\tilde{g}(\pi) = 1 - \pi$ preserves the parameter space, in the sense of equation (3.109).

Under this new setup, following the same approach that led to the estimator $T(X)$, we see that $T(g(X)) = T(1 - X)$ is an optimal estimator of $\tilde{g}(\pi) = 1 - \pi$ under squared-error loss among the class of unbiased estimators. Hence, in this case, the squared-error loss function allowed us to develop an equivariant procedure.

We note that the estimator $T(g(X)) = g^*(T(X)) = 1 - T(X)$, and we have, as in equation (3.111),

$$L(\pi, T(X)) = L(\tilde{g}(\pi), g^*(T(g(X)))).$$

∎

In the Bernoulli example above, loss functions of various forms would have allowed us to develop an equivariant procedure for estimation of the transformed π. This is not always the case. For some types of transformations g and $\tilde{g}$ on the sample and parameter spaces, we can develop equivariant procedures only if the loss function is of some particular form. For example, in a location family, with transformations of the form $g(X) = X + c$ and $\tilde{g}(\mu) = \mu + c$, in order to develop an equivariant procedure that satisfies equation (3.111) we need a loss function that is a function only of $a - g(\theta)$. **********

Following the same approach as above, we see that in a univariate scale family, with transformations of the form $g(X) = cX$, in order to develop an equivariant procedure, we need a loss function that is a function only of $a/g(\theta)$. In order to develop equivariant procedures for a general location-scale family $P_{(\mu,\Sigma)}$ we need a loss function of the form

$$L((\mu, \Sigma), a) = L_{\text{ls}}(\Sigma^{1/2}(a - \mu)). \tag{3.113}$$

In order to achieve invariance of the loss function for a given group of transformations $\mathcal{G}$, for each $g \in \mathcal{G}$, we need a 1:1 function g^* that maps the decision space onto itself, $g^* : \mathcal{A} \mapsto \mathcal{A}$. The set of all such g^* together with the induced structure is a group, $\mathcal{G}^*$ with elements

$$g^* : \mathcal{A} \mapsto \mathcal{A}, \quad 1:1 \text{ and onto.}$$

The relationship between $\mathcal{G}$ and $\mathcal{G}^*$ is an isomorphism; that is, for $g \in \mathcal{G}$ and $g^* \in \mathcal{G}^*$, there is a function h such that if $g^* = h(g)$, then $h(g_1 \circ g_2) = h(g_1) \circ h(g_2)$.

Invariance of Statistical Procedures

To study invariance of statistical procedures we will now identify three groups of transformations $\mathcal{G}$, $\tilde{\mathcal{G}}$, and $\mathcal{G}^*$, and the relationships among the groups. This

notation is widely used in mathematical statistics, maybe with some slight modifications.

Let $\mathcal{G}$ be a group of transformations that map the probability space onto itself. We write

$$g(X) = \tilde{X}. \tag{3.114}$$

Note that X and $\tilde{X}$ are random variables, so the domain and the range of the mapping are subsets of *probability spaces*; the random variables are based on the same underlying measure, so the probability spaces are the same; the transformation is a member of a transformation group, so the domain and the range are equal and the transformations are one-to-one.

For a given statistical procedure T that yields the action a for an observation X, we have under the various transformations

$$\begin{aligned} g^*(T(g^{-1}(\tilde{X}))) &= g^*(T(X)) && (3.115)\\ &= g^*(a) && (3.116)\\ &= \tilde{a}. && (3.117) \end{aligned}$$

We are interested in a probability space, $(\Omega, \mathcal{F}, \mathcal{P}_\Theta)$, that is invariant to a class of transformations $\mathcal{G}$; that is, one in which $\mathcal{P}_\Theta$ is a group family with respect to $\mathcal{G}$. The induced groups $\overline{\mathcal{G}}$ and $\mathcal{G}^*$ determine the transformations to be applied to the parameter space and the action space.

The basic idea underlying invariance of statistical procedures naturally is invariance of the risk under the given transformations.

We seek a statistical procedure $T(x)$ that is an invariant function under the transformations.

Because if there is a maximal invariant function m (see Definition 0.1.53 on page 755) all invariant functions are dependent on m, our search for optimal invariant procedures can use m. The concept of maximal invariance is similar to the concept of sufficiency. A sufficient statistic may reduce the sample space; a maximal invariant statistic may reduce the parameter space. (Maximal invariant statistics have some technical issues regarding measurability, however; X being measurable does not guarantee $m(X)$ is measurable under the same measure.)

A probability model may be defined in different ways. There may be an equivalence between two different models that is essentially a result of a reparametrization: $\tilde{\theta} = \tilde{g}(\theta)$. A random variable in the one model may be a function of the random variable in the other model: $\widetilde{X} = g(X)$. There are two ways of thinking of estimation under a reparametrization, both in the context of an estimator $T(X)$ of $h(\theta)$, and with the transformations defined above:

- functional, $g^*(T(X))$ estimates $g^*(h(\theta))$;
- formal, $T(g(X))$ estimates $g^*(h(\theta))$.

Functional equivariance is trivial. This is the equivariance we expect under a simple change of units, for example. If X is a random variable that models physical temperatures in some application, it should not make any real difference whether the temperatures are always measured in degrees Celsius or degrees Fahrenheit. The random variable itself does not include units, of course (it is a real number). If the measurements are made in degrees Celsius at a time when X is the random variable used to model the distribution of the data and the estimator $T(X)$ and the estimand $h(\theta)$ relates to X in a linear fashion (if $h(\theta)$ is the mean of X, for example), and later in a similar application the measurements are made in degrees Fahrenheit, applying $g^*(t) = 9t/5 + 32$ to both $T(X)$ and $h(\theta)$ preserves the interpretation of the model.

Formal equivariance, however, is not meaningful unless the problem itself has fundamentally symmetric properties; the family of probability distributions is closed under some group of transformations on the sample space one on the parameter space. In this case, we need a corresponding transformation on the decision space. The statistical procedure is equivariant if the functional equivariance is the same as the formal equivariance; that is,

$$T(g(X)) = g^*(T(X)). \tag{3.118}$$

3.4.2 Optimal Equivariant Statistical Procedures

In the decision-theoretic approach to statistical inference, we generally seek procedures that have minimum risk with respect to a given loss function. As we have seen, there are situations where we cannot obtain this uniformly. By restricting attention to procedures with properties such as L-unbiasedness or L-invariance, however, we may be able to achieve uniformly best procedures within that restricted class. Within the class of unbiased procedures, we seek UMVU estimators and UMPU tests. Likewise, within a collection of equivariant procedures, we seek ones with minimum risk.

The simplest and most interesting transformations are translations and scalings, and the combinations of these two, that is linear transformations. Consequently, the two most common types of invariant inference problems are those that are location invariant (or equivariant) and those that are scale invariant (or equivariant). Because in a linear transformation we scale first, a scale invariant procedure is invariant to location transformations, but a location invariant procedure is not invariant to scale transformations.

In the remainder of this section, we concentrate on problems of point estimation. In Section 7.2.5 beginning on page 525 we discuss equivariant (invariant) test procedures, and in Section 7.9.3 beginning on page 549 we discuss equivariant confidence sets. We discuss equivariance in the context of Bayesian analysis in Section 4.3.2 beginning on page 354.

Equivariant Point Estimation

If the estimand under the untransformed problem is θ, the estimand after the transformations is $\tilde{g}(\theta)$. If $T(X)$ is an estimator of θ, equivariance of the estimator requires that $g^*(T(X)) = T(g(X))$ be an estimator of $\tilde{g}(\theta)$ with the same risk.

The properties of the estimator in the untransformed problem are preserved under the transformations. An estimator that is equivariant except possibly on a set of zero probability is said to be *almost equivariant*.

Within a collection of equivariant estimators, we would choose the one with minimum risk. This is MRE estimation, and the estimator is an MREE. (Some authors call it MRI and MRIE.)

By the definition of "equivariance" in this context, the MRE estimator is UMRE, so the concept of uniformity does not arise as a separate issue here.

Finding an Optimal Equivariant Point Estimator

To find an MREE for a given group of transformations, we

1. identify the necessary form(s) of the loss function for the transformations
2. identify necessary and/or sufficient properties of equivariant estimators
3. identify an equivariant estimator
4. characterize all equivariant estimators in terms of a given one
5. identify the one that minimizes the risk for a given loss function

We must accept certain limitations alluded to above: the statistical inference problem must have be of a special type with respect to the types of transformations, the probability distribution and parameter of interest, and the given loss function.

In the next two sections, we illustrate these steps for estimators with location equivariance and scale equivariance.

Location Equivariant Estimation

In location equivariant estimation, we assume a family of distributions that are location invariant. We write a PDF of a member of this family as $p(x+c)$. The basic transformation is a translation on both the random variable and the location parameter: $\widetilde{X} = X + c$ and $\tilde{\mu} = \mu + c$. The estimand of interest is μ.

A reasonable loss function $\widetilde{L}$ must have the property (3.113), that is, $\widetilde{L}(\mu+c, a+c) = \widetilde{L}(\mu, a)$ for any c, μ and a; hence, $\widetilde{L}(\mu, a)$ is a function only of $(a-\mu)$:

$$\widetilde{L}(\mu, a) = L(a - \mu). \tag{3.119}$$

(To repeat the argument that led to equation (3.113) and to see it in this particular case, let $\mu = -c$, and so we have $\widetilde{L}(0, a) = \widetilde{L}(0, a - \mu)$, and this equality must continue to hold as μ and c move in tandem.)

Now, we consider properties of location equivariant estimators.
The estimator must have the property (3.118), that is,

$$T(x+a) = T(x) + a. \tag{3.120}$$

It is easy to see that if T_0 is a location equivariant estimator and

$$T(x) = T_0(x) + u(x), \tag{3.121}$$

where u is any Borel function that is invariant to translations (that is, $u(x+a) = u(x)$), then $T(x)$ is a location equivariant estimator. (Notice the difference in "invariant" and "equivariant".)

We now show that any location equivariant estimator must be of the form (3.121) and furthermore, we characterize the function u.

Theorem 3.13
Given $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ and let T_0 be a location equivariant estimator.
(i)If $n = 1$, then any location equivariant estimator T must satisfy

$$T(x) = T_0(x) + u, \tag{3.122}$$

where u is constant.
(ii)If $n > 1$, then any location equivariant estimator T must satisfy

$$T(x) = T_0(x) + u(d), \tag{3.123}$$

where

$$d = (x_i - x_n) \quad \text{for } i = 1, \ldots, n-1.$$

Proof.
Part (i) follows from equation (3.120).
Part (ii), $n > 1$: Let T_0 be a location equivariant estimator. Suppose T is a location equivariant estimator, and $\forall x \in \mathbb{R}^n$ let $\tilde{u}(x) = T(x) - T_0(x)$. Because T and T_0 are location equivariant, we have for any $c \in \mathbb{R}$,

$$\begin{aligned} T(x_1, \ldots, x_n) - T_0(x_1, \ldots, x_n) &= T(x_1 + c, \ldots, x_n + c) - T_0(x_1 + c, \ldots, x_n + c) \\ &= \tilde{u}(x_1 + c, \ldots, x_n + c). \end{aligned}$$

Now, let $c = -x_n$. So

$$\tilde{u}(x_1 - x_n, \ldots, x_{n-1} - x_n, 0) = T(x_1, \ldots, x_n) - T_0(x_1, \ldots, x_n)$$

or, with $u(x_1 - x_n, \ldots, x_{n-1} - x_n) = \tilde{u}(x_1 - x_n, \ldots, x_{n-1} - x_n, 0)$ and $d = (x_i - x_n)$,

$$T(x) = T_0(x) + u(d).$$

∎

With this knowledge about the form of any location equivariant estimator, we now seek one with minimum risk. For an estimator based on only one

observation, the problem is trivial and it is just to determine an optimal constant u.

We will assume a random sample of size $n > 1$, we will use d as defined above, and we will assume a distribution with PDF $p(x - c)$. If we have a location equivariant estimator T_0 with finite risk, we determine the MREE (if it exists) as

$$T_*(x) = T_0(x) - u_*(d), \tag{3.124}$$

where $u_*(d)$ minimizes the conditional risk at c

$$\mathrm{E}_c\Big(L(T_0(X) - u(D)) \,|\, D = d\Big). \tag{3.125}$$

For simplicity, we take $c = 0$, and write E_0 for the expectation.

Whether or not such $u_*(d)$ exists, and if so is unique, depends on the form of the loss function (which, in any event, must be of the form of equation (3.119).) In particular, for squared-error loss, which is of this form, we have

$$u_*(d) = \mathrm{E}_0\big(T_0(X) \,|\, d\big). \tag{3.126}$$

Note that for squared-error loss, if a UMVUE exists and is equivariant, it is MRE.

For squared-error loss, a location-equivariant point estimator of the location has a special form, as given in the following theorem.

Theorem 3.14
Given a sample $X_1, \ldots, X_n$ from a location family with joint Lebesgue PDF $p(x_1-\mu, \ldots, x_n-\mu)$, if there is a location-equivariant estimator of μ with finite risk under squared-error loss, then the unique MREE of μ under squared-error loss is

$$T_*(x) = \frac{\int t\, p(X_1 - t, \ldots, X_n - t)\mathrm{d}t}{\int p(X_1 - t, \ldots, X_n - t)\mathrm{d}t}; \tag{3.127}$$

that is, $T_*(x)$ in equation (3.124), with $u_*(x)$ from equation (3.126)), can be written as (3.127).

Proof.
Let $T_0(X)$ be a location-equivariant estimator of μ with finite risk. MS2 theorem 4.5 ∎

The estimator $T_*(x)$ in equation (3.127) is called a *Pitman estimator*.

Note that a location equivariant estimator is not necessarily invariant to scale transformations.

Scale Equivariant Estimation

In scale equivariant estimation, the basic transformation is a multiplication on both the random variable and the a power nonzero power of the scale parameter: $\widetilde{X} = rX$, for $r > 0$, and $\tilde{\sigma} = r^h\sigma^h$. This development parallels that for location equivariant estimation in the previous section.

The estimand of interest is σ^h, for some nonzero h. A reasonable loss function $\widetilde{L}$ must have the property (3.113), $\widetilde{L}(r\sigma, r^h a) = \widetilde{L}(\sigma, a)$, hence,

$$\widetilde{L}(\sigma, a) = L(a/\sigma^h), \tag{3.128}$$

and the estimator must have the property

$$T(rx) = r^h T(x). \tag{3.129}$$

If T_0 is a scale equivariant estimator, then any scale equivariant estimator must be of the form

$$T(x) = \frac{T_0(x)}{u(z)}, \tag{3.130}$$

where

$$z_i = \frac{x_1}{x_n}, \text{ for } i = 1, \ldots, n-1, \quad \text{and } z_n = \frac{x_n}{|x_n|}.$$

If we have a scale equivariant estimator T_0 with finite risk, we determine the MREE (if it exists) as

$$T_*(x) = T_0(x)/u_*(z), \tag{3.131}$$

where $u_*(z)$ minimizes the conditional risk at $r = 1$:

$$\mathrm{E}_1\Big(\gamma(T_0(X)/u(z)) \,|\, z\Big). \tag{3.132}$$

Note that the loss function has a special form. In the scale equivariant estimation problem, there are a couple of special loss functions. One is a squared error of the form

$$L(a/\sigma^h) = \frac{(a - \sigma^h)^2}{\sigma^{2h}}, \tag{3.133}$$

in which case

$$u_*(z) = \frac{\mathrm{E}_1\Big((T_0(x))^2 \,|\, y\Big)}{\mathrm{E}_1\Big(T_0(x) \,|\, y\Big)}, \tag{3.134}$$

and the estimator is a Pitman estimator.

Another special loss functions is of the form

$$L(a/\sigma^h) = a/\sigma^h - \log(a/\sigma^h) - 1, \tag{3.135}$$

called "Stein's loss", in which case

$$u_*(z) = \mathrm{E}_1(T_0(X) \,|\, y). \tag{3.136}$$

Stein's loss has the interesting property that it is the only scale-invariant loss function for which the UMVUE is also the MREE (difficult proof).

A scale equivariant estimator is invariant to location transformations; that is, if T is scale invariant, then $T(x + a) = T(x)$.

Location-Scale Equivariant Estimation

Location-scale equivariance involves the combination of the two separate developments. The basic transformations are location and scale: $\widetilde{X} = bX + c$ and $\tilde{\theta} = b\theta + c$.

The loss function (3.128) for estimation of the scale parameter is invariant to both location and scale transformations, and the estimator of the scale must have the form of (3.130).

In order for the loss function for estimation of the location parameter to be invariant under a location and scale transformation, the loss function must be of the form

$$\widetilde{L}(\mu, a) = L((a - \mu)/\sigma), \tag{3.137}$$

and the location estimator must have the property

$$T(bx + c) = b^r T(x) + c. \tag{3.138}$$

Analysis of these estimators does not involve anything fundamentally different from combinations of the ideas discussed separately for the location and scale cases.

Equivariant Estimation in a Normal Family

MRE estimation has particular relevance to the family of normal distributions, which is a location-scale group family.

Example 3.23 Equivariant Estimation in a Normal Family
Suppose $X_1, X_2, \ldots, X_n$ are iid as $\mathrm{N}(\mu, \sigma^2)$ distribution, and consider the problem of estimation of μ and σ^2.

****

$$T_{\sigma^2}(X) = \frac{1}{n+1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{3.139}$$

**** compare MLE, minimum MSE

$$T_{\mu}(X) = \overline{X} \tag{3.140}$$

∎

The MRE estimator of the location under a convex and even loss function of the form (3.138) and MRE estimator of the scale under a loss of the form (3.130) are independent of each other. Another interesting fact is that in location families that have densities with respect to Lebesgue measure and with finite variance, the risk of a MRE location estimator with scaled squared-error loss is larger in the normal family than in any other location-scale group family.

3.5 Probability Statements in Statistical Inference

In a statistical paradigm in which a parameter characterizing the probability distribution of the observable random variable is itself a random variable, the objective of statistical inference is to use observable data to adjust the assumed probability distribution of the parameter. In this case, the results of the statistical analysis can be summarized by probability statements.

If the parameter is not considered to be a random variable, statements of probability that the parameter has given values do not make sense, except as ways of quantifying "beliefs" about the values.

Within a "objective" paradigm for statistical inference, there are two instances in which statements about probability are associated with the decisions of the inferential methods. In hypothesis testing, under assumptions about the distributions, we base our inferential methods on probabilities of two types of errors. In confidence sets the decisions are associated with probability statements about coverage of the parameters.

In both of these types of inference, the basic set up is the standard one in statistical inference. We have a random sample of independent observations $X_1, \ldots, X_n$ on a random variable X that has a distribution P_θ, some aspects of which are unknown. We assume some family of probability distributions $\mathcal{P}$ such that $P_\theta \in \mathcal{P}$. We begin with some preassigned probability that, following the prescribed method of inference, we will arrive at set of distributions $\mathcal{P}_\theta$ that contain the distribution P_θ. Our objective is to determine such methods, and among a class of such methods, determine ones that have optimal properties with respect to reasonable criteria.

After having completed such a process, it may or may not be appropriate to characterize the relationship of the "true" unknown distribution P_θ to the set of $\mathcal{P}_\theta$ with any statement about "probability". If the particular distribution or some parameter in the distribution is considered to be a (nondegenerate) random variable, we may speak of a probability conditional on the observations used in the inference process. (This is a "posterior" probability.) On the other hand, if the underlying probability model of the observable data is fixed, then either $P_\theta \in \mathcal{P}_\theta$ with probability 1, or else $P_\theta \notin \mathcal{P}_\theta$ with probability 1.

In these types of statistical inference, as we will describe below, we use the terms "significance level", "size", "confidence level", and "confidence coefficient" to describe our findings.

3.5.1 Tests of Hypotheses

Given a set of data, X, and a family of possible distributions that gave rise to the data, $\mathcal{P}$, a common objective of statistical inference is to specify a particular member or subclass of $\mathcal{P}$ that "likely" generated X. For example, if $\mathcal{P} = \{\mathrm{N}(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+\}$, given $X = x$, we may choose $\mathrm{N}(\bar{x}, s^2)$ as a good candidate for the population from which the data arose. This choice is based on statistical estimators that we know to be "good" ones.

In another kind of statistical inference, given a set of data X and a family of distributions $\mathcal{P}$, we are to decide whether the data "likely" came from some hypothesized subfamily $\mathcal{P}_0$ of $\mathcal{P}$. Our possible decisions are "yes" or "no". Rather than a general "no", a specific alternative may be hypothesized.

This kind of statistical inference is called "testing statistical hypotheses". We will discuss this topic more fully in Chapter 7. In Section 4.5 we discuss testing from a Bayesian perspective. Here, we just introduce some terms and consider some simple cases.

Statistical Hypotheses

The hypotheses concern a specific member $P \in \mathcal{P}$. This is the distribution that generated the observed data.

We have a *null hypothesis*

$$H_0 : \quad P \in \mathcal{P}_0 \tag{3.141}$$

and an *alternative hypothesis*

$$H_1 : \quad P \in \mathcal{P}_1, \tag{3.142}$$

where $\mathcal{P}_0 \subseteq \mathcal{P}$, $\mathcal{P}_1 \subseteq \mathcal{P}$, and $\mathcal{P}_0 \cap \mathcal{P}_1 = \emptyset$. If $\mathcal{P}_0 \cup \mathcal{P}_1 = \mathcal{P}$, the alternative hypothesis is effectively "everything else".

In the paradigm of equations (3.1) and (3.2), in which we characterize statistical inference as beginning with a family of probability distributions $\mathcal{P} = \{P_\theta \mid \theta \in \Theta\}$ and, using observed data, deciding that the family is $\mathcal{P}_H$, where $\mathcal{P}_H \subseteq \mathcal{P}$, the problem of statistical hypothesis testing can be described as beginning with $\mathcal{P} = \mathcal{P}_0 \cup \mathcal{P}_1$, and deciding either that $\mathcal{P} = \mathcal{P}_0$ or $\mathcal{P} = \mathcal{P}_1$.

In a Bayesian setup of the canonical problem in statistical inference as described in equations (3.3) and (3.4), the problem of statistical hypothesis testing can be described as beginning with $\mathcal{P} = \{P_\Theta \mid \Theta \sim Q_0 \in \mathcal{Q}\}$ where Q_0 is some prior distribution, and then deciding that the family of probability distributions giving rise to the observed data is $\mathcal{P}_H = \{P_\Theta \mid \Theta \sim Q_H \in \mathcal{Q}\}$.

An hypothesis that specifies exactly one distribution is called a *simple hypothesis*; otherwise it is called a *composite hypothesis*. H_0 above is a simple hypothesis if there is only one distribution in $\mathcal{P}_0$.

If the family of distributions is associated with a parameter space Θ, we may equivalently describe the tests as

$$H_0 : \theta \in \Theta_0 \qquad \text{versus} \qquad H_1 : \theta \in \Theta_1.$$

An hypothesis $H : \theta \in \Theta_H$ in which $\#\Theta_H = 1$ is a simple hypothesis; if $\#\Theta_H > 1$ it is a composite hypothesis. Of course we are often interested in the case where $\Theta = \Theta_0 \cup \Theta_1$. An hypothesis of the form $H_0 : \theta = \theta_0$ is a simple hypothesis, while $H_i : \theta \geq \theta_0$ is a composite hypothesis.

Test Statistics and Critical Regions

A straightforward way of performing the test involves use of a test statistic, $T(X)$, computed from a random sample of data, with which we associated a rejection region C, and if $T(X) \in C$, we reject H_0 in favor of H_1. Now, if H_0 is true and we reject it, we have made an error. So a reasonable procedure is to choose C such that if the null hypothesis is true, for some preassigned (small) value, α,

$$\Pr(T(X) \in C | H_0) \leq \alpha. \tag{3.143}$$

We call this bound a *significance level* of the test.

Although the term "significance level" is widely used, the fact that we have defined it as a bound means that it is not very useful (although the definition in equation (3.143) is the standard one). The LUB is clearly the measure of interest. We call

$$\sup_{P \in H_0} \Pr(T(X) \in C | P) \tag{3.144}$$

the *size of the test*.

We seek a statistic $T(X)$ such that $\Pr(T(X) \in C)$ is large if the null hypothesis is not true. Thus, C is a region of more "extreme" values of the test statistic if the null hypothesis is true. A statistical test in this kind of scenario is called a "significance test".

If $T(X) \in C$, the null hypothesis is rejected. The rejection region is also called the *critical region*. The complement of the rejection region is called the acceptance region.

It is desirable that the test have a high probability of rejecting the null hypothesis if indeed the null hypothesis is not true.

p-Values

A procedure for testing that is mechanically equivalent to this is to compute the realization of the test statistic $T(X)$, say t, and then to determine the probability that $T(X)$ is more extreme than t. In this approach, the realized value of the test statistic determines a region C_t of more extreme values. The probability that the test statistic is in C_t if the null hypothesis is true, $\Pr(T \in C_t)$, is called the "p-value" or "observed significance level" of the realized test statistic.

In this framework we are testing one hypothesis versus another hypothesis. The two hypotheses are not treated symmetrically, however. We are still directly testing the null hypothesis. This asymmetry allows us to focus on two kinds of losses that we might incur. The losses relate to the two kinds of errors that we might make.

Test Rules

Instead of thinking of a test statistic T and a rejection region C, as above, we can formulate the testing procedure in a slightly different way. We can think of the test as a decision rule, $\delta(X)$, which is a statistic that relates more directly to the decision about the hypothesis. We sometimes refer to the statistic $\delta(X)$ as "the test", because its value is directly related to the outcome of the test; that is, there is no separately defined rejection region.

A *nonrandomized test procedure* is a rule $\delta(X)$ that assigns two decisions to two disjoint subsets, C_0 and C_1, of the range of $T(X)$. In general, we require $C_0 \cup C_1$ be the support of $T(X)$. We equate those two decisions with the real numbers d_0 and d_1, so $\delta(X)$ is a real-valued function,

$$\delta(x) = \begin{cases} d_0 \text{ for } T(x) \in C_0 \\ d_1 \text{ for } T(x) \in C_1. \end{cases} \tag{3.145}$$

For simplicity, we choose $d_0 = 0$ and $d_1 = 1$. Note for $i = 0, 1$,

$$\Pr(\delta(X) = i) = \Pr(T(X) \in C_i). \tag{3.146}$$

As above, we call C_1 the *critical region*, and generally denote it by just C.

If $\delta(X)$ takes the value 0, the decision is not to reject; if $\delta(X)$ takes the value 1, the decision is to reject. If the range of $\delta(X)$ is $\{0, 1\}$, the test is a nonrandomized test. Sometimes, however, it is useful to expand the range of $\delta(X)$ to be $[0, 1]$, where we can interpret a value of $\delta(X)$ as the probability that the null hypothesis is rejected. If it is not the case that $\delta(X)$ equals 0 or 1 a.s., we call the test a *randomized test*.

Testing as an Estimation Problem

In the general setup above, we can define an indicator function $\mathrm{I}_{\Theta_0}(\theta)$. The testing problem is equivalent to the problem of estimating $\mathrm{I}_{\Theta_0}(\theta)$. Let us use a statistic $S(X)$ as an estimator of $\mathrm{I}_{\Theta_0}(\theta)$. The estimand is in $\{0, 1\}$, and so $S(X)$ should be in $\{0, 1\}$, or at least in $[0, 1]$.

Notice the relationship of $S(X)$ to $\delta(X)$. For the estimation approach using $S(X)$ to be equivalent to use of the test rule $\delta(X)$, it must be the case that

$$S(X) = 1 \Longleftrightarrow \delta(X) = 0 \quad \text{(i.e., do not reject)} \tag{3.147}$$

and

$$S(X) = 0 \Longleftrightarrow \delta(X) = 1 \quad \text{(i.e., reject)} \tag{3.148}$$

Following a decision-theoretic approach to the estimation problem, we define a loss function. In the a simple framework for testing, the loss function is 0-1. Under this loss, using $S(X) = s$ as the rule for the test, we have

$$L(\theta, s) = \begin{cases} 0 \text{ if } s = \mathrm{I}_{\Theta_0}(\theta) \\ 1 \text{ otherwise.} \end{cases} \tag{3.149}$$

Power of the Test

We now can focus on the test under either hypothesis (that is, under either subset of the family of distributions) in a unified fashion. We define the *power function* of the test, for any given $P \in \mathcal{P}$ as

$$\beta(\delta, P) = \mathrm{E}_P(\delta(X)). \tag{3.150}$$

We also often use the notation $\beta_\delta(P)$ instead of $\beta(\delta, P)$. In general, the probability of rejection of the null hypothesis is called the power of the test.

An obvious way of defining optimality for tests is in terms of the power for distributions in the class of the alternative hypothesis; that is, we seek "most powerful" tests.

Errors

If $P \in \mathcal{P}_0$ and $\delta(X) = 1$, we make an error; that is, we reject a true hypothesis. We call that a "type I error". For a randomized test, we have the possibility of making a type I error if $\delta(X) > 0$. In general, if $P \in \mathcal{P}_0$, $\beta_\delta(P)$ is the probability of a type I error. Conversely, if $P \in \mathcal{P}_1$, then $1 - \beta_\delta(P)$ is the probability of a "type II error", that is failing to reject a false hypothesis.

Testing as a Decision Problem

For a statistical hypothesis as described above with $\delta(x)$ as in equation (3.145), and $d_0 = 0$ and $d_1 = 1$, write

$$\phi(x) = \Pr(\delta(X) = 1 \mid X = x). \tag{3.151}$$

Notice that this is the same as the power, except ϕ here is a function of the observations, while we think of the power as a function of the true distribution. Assuming only the two outcomes, we have

$$1 - \phi(x) = \Pr(\delta(X) = 0 \mid X = x). \tag{3.152}$$

For this decision problem, an obvious choice of a loss function is the 0-1 loss function:

$$\begin{array}{ll} L(P, i) = 0 & \text{if } H_i \\ L(P, i) = 1 & \text{otherwise.} \end{array} \tag{3.153}$$

It may be useful to consider a procedure with more than just two outcomes; in particular, a third outcome, γ, may make sense. In an application in analysis of data, this decision may suggest collecting more data; that is, it may correspond to "no decision", or, usually only for theoretical analyses, it may suggest that a decision be made randomly. We will, at least in the beginning, however, restrict our attention to procedures with just two outcomes.

For the two decisions and two state of nature case, there are four possibilities:

- the test yields 0 and H_0 is true (correct decision);
- the test yields 1 and H_1 is true (correct decision);
- the test yields 1 and H_0 is true (type I error); and
- the test yields 0 and H_1 is true (type II error).

We obviously want a test procedure that minimizes the probability of either type of error. It is clear that we can easily decrease the probability of one (if its probability is positive) at the cost of increasing the probability of the other.

We do not treat H_0 and H_1 symmetrically; H_0 is the *hypothesis* to be tested and H_1 is the *alternative*. This distinction is important in developing a practical methodology of testing.

We adopt the following approach for choosing δ (under the given assumptions on X, and the notation above):

1. Choose $\alpha \in]0, 1[$ and require that $\delta(X)$ be such that
$$\Pr(\delta(X) = 1 \mid H_0) \leq \alpha.$$
α is called the *level of significance*.
2. Subject to this, find $\delta(X)$ so as to minimize
$$\Pr(\delta(X) = 0 \mid H_1).$$

The definition of significance level is not as ambiguous as it may appear at first glance.

One *chooses* α; that is the level of significance.

For some $\tilde{\alpha} > \alpha$, although $\Pr(\delta(X) = 1 \mid \theta \in \Theta_0) \leq \tilde{\alpha}$, we would not say that $\tilde{\alpha}$ is the level (or a level) of significance.

Notice that the restriction on the type I error in the first step applies $\forall P \in H_0$. If the size is less than the level of significance, the test is said to be *conservative*, and in that case, we often refer to α as the "nominal size".

Approximate Tests

If the distribution of the test statistic T or δ under the null hypothesis is known, the critical region or the p-value can be determined. If the distribution is not known, some other approach must be used. A common method is to use some approximation to the distribution. The objective is to approximate a quantile of T under the null hypothesis. In asymptotic inference, the approximation is often based on an asymptotic distribution of the test statistic.

In computational inference, a Monte Carlo test may be used. In Monte Carlo tests the quantile of T is estimated by simulation of the distribution.

Unbiased Tests

A test δ of $H_0 : P \in \mathcal{P}_0$ versus $H_1 : P \in \mathcal{P}_1$ is said to be *unbiased at level* α if the power function satisfies

$$\begin{aligned} \beta_\delta(P) &\leq \alpha \quad \text{for } P \in \mathcal{P}_0 \\ \beta_\delta(P) &\geq \alpha \quad \text{for } P \in \mathcal{P}_1 \end{aligned}$$

Uniformly Best Tests

The risk or the expected error in a test depends on the specific distribution within the family of distributions assumed. We may seek a test that has minimum expected errors of both types, or, in a practical approach to this objective, we may cap the probability of a type I error and seek the most powerful test for distributions within the class of the alternative hypothesis.

As we have seen in the estimation problem, optimality generally depends on the specific distribution, and it may not be possible to achieve it uniformly; that is, for all distributions within a given family.

We may then take the approach mentioned on page 267 for estimation and restrict the allowable tests in some way. We may require that the tests be unbiased, for example. That approach leads us to seek a UMPU test, that is, a uniformly most powerful unbiased test. Alternatively, as we mentioned before, we may seek a test that is optimal over the full set of distributions $\mathcal{P}$ by some global measure of optimality.

3.5.2 Confidence Sets

In a problem of statistical inference for a family of distributions $\mathcal{P}$, given a random sample X, a level $1-\alpha$ *confidence set*, or *confidence set* (the terms are synonymous), is a is a random subset of $\mathcal{P}$, $\mathcal{P}_S$, such that

$$\Pr_P\left(\mathcal{P}_S \ni P\right) \geq 1-\alpha \quad \forall P \in \mathcal{P}. \tag{3.154}$$

More precisely, we call $\mathcal{P}_S$ a *random family of level* $1-\alpha$ *confidence sets*. This definition obviously leaves many issues to be examined because of the $\geq$ relationship. A family of $1-\alpha_1$ confidence sets is also a family of $1-\alpha_2$ confidence set for $\alpha_2 \geq \alpha_1$; and if $\mathcal{P}_S$ is a level $1-\alpha$ confidence set, then $\mathcal{P}_{\widetilde{S}}$ is also a level $1-\alpha$ confidence set if $\mathcal{P}_{\widetilde{S}} \supset \mathcal{P}_S$.

The "S" in this notation refers to a random sample of X, and the notation $\mathcal{P}_S$ is intended to imply that a random set is being indicated, in contrast to the notation $\mathcal{P}_H$ used above to refer to an hypothesized set. The set $\mathcal{P}_S$ is determined by the random sample, while the set $\mathcal{P}_H$ is determined a priori. The source of randomness also accounts for my preferred notation, $\mathcal{P}_S \ni P$,

which can be thought of a referring to the random event in which the set $\mathcal{P}_S$ includes the element P.

As with the term "significance" in the hypothesis testing problem, the standard usage of the term "confidence" is subject to a certain amount of ambiguity. In hypothesis testing, we use "level of significance" and "size" of the tests. (Recall that, adding to the confusion" we also use "significance level" of a test statistic to refer to a minimal size test that would reject the null hypothesis; that is, to refer to a p-value.) In setting confidence regions, we refer to "confidence level" and "confidence coefficient". We call

$$\inf_{P \in \mathcal{P}} \Pr_P(\mathcal{P}_S \ni P) \tag{3.155}$$

the *confidence coefficient* of $\mathcal{P}_S$.

The confidence coefficient is also called the *coverage probability*.

In a parametric setting, we can equivalently define a random family Θ_S of $1-\alpha$ confidence regions (sets) for the parameter space Θ by

$$\Pr_\theta(\Theta_S \ni \theta) \geq 1-\alpha \quad \forall \theta \in \Theta.$$

A realization of a confidence set, say Θ_s, is also called a confidence set. Although it may seem natural to state that the "probability that θ is in $A(x)$ is $1-\alpha$", this statement can be misleading unless a certain underlying probability structure is assumed.

We will introduce and discuss other terms in Chapter 7. In Chapter 4 we discuss confidence sets from a Bayesian perspective. Here, we just define the term and consider some simple cases.

Pivot Functions

For forming confidence sets, we often can use a function of the sample that also involves the parameter of interest, $g(T, \theta)$. The confidence set is then formed by separating the parameter from the sample values.

A class of functions that are particularly useful for forming confidence sets are called *pivotal* values, or pivotal functions. A function $g(T, \theta)$ is said to be a pivotal function if its distribution does not depend on any unknown parameters. This allows exact confidence intervals to be formed for the parameter θ.

Confidence Intervals

Our usual notion of a confidence leads to the definition of a $1-\alpha$ confidence interval for the (scalar) parameter θ as the random interval $[T_L, T_U]$, that has the property

$$\Pr(T_L \leq \theta \leq T_U) \geq 1-\alpha. \tag{3.156}$$

This is also called a $(1-\alpha)100\%$ confidence interval. The interval $[T_L, T_U]$ is not uniquely determined.

The concept extends easily to vector-valued parameters. Rather than taking vectors T_L and T_U, however, we generally define an ellipsoidal region, whose shape is determined by the covariances of the estimators.

A realization of the random interval, say $[t_L, t_U]$, is also called a confidence interval.

In practice, the interval is usually specified with respect to an estimator of θ, T. If we know the sampling distribution of $T-\theta$, we may determine c_1 and c_2 such that

$$\Pr\left(c_1 \leq T-\theta \leq c_2\right) = 1-\alpha;$$

and hence

$$\Pr\left(T-c_2 \leq \theta \leq T-c_1\right) = 1-\alpha.$$

If either T_L or T_U is infinite or corresponds to a bound on acceptable values of θ, the confidence interval is one-sided. Suppose $\Theta = (a, b)$, where a or b may be infinite. In equation (3.156), if $T_L = a$, then T_U is called an *upper confidence bound*, and if $T_U = b$, then T_L is called a *lower confidence bound*. (It is better not to use the terms "upper confidence interval" or "lower confidence interval", because of the possible ambiguity in these terms.)

For two-sided confidence intervals, we may seek to make the probability on either side of T to be equal, to make $c_1 = -c_2$, and/or to minimize $|c_1|$ or $|c_2|$. This is similar in spirit to seeking an estimator with small variance.

We can use a pivot function $g(T, \theta)$ to form confidence intervals for the parameter θ. We first form

$$\Pr\Big(g_{(\alpha/2)} \leq g(T,\theta) \leq g_{(1-\alpha/2)}\Big) = 1-\alpha,$$

where $g_{(\alpha/2)}$ and $g_{(1-\alpha/2)}$ are quantiles of the distribution of $g(T, \theta)$; that is,

$$\Pr(g(T,\theta) \leq g_{(\pi)}) = \pi.$$

If, as in the case considered above, $g(T, \theta) = T - \theta$, the resulting confidence interval has the form

$$\Pr\Big(T-g_{(1-\alpha/2)} \leq \theta \leq T-g_{(\alpha/2)}\Big) = 1-\alpha.$$

Example 3.24 Confidence Interval for Mean of a Normal Distribution

Suppose $Y_1, Y_2, \ldots, Y_n$ are iid as $\mathrm{N}(\mu, \sigma^2)$ distribution, and $\overline{Y}$ is the sample mean. The quantity

$$g(\overline{Y}, \mu) = \frac{\sqrt{n(n-1)}\,(\overline{Y}-\mu)}{\sqrt{\sum\left(Y_i-\overline{Y}\right)^2}}$$

has a Student's t distribution with $n-1$ degrees of freedom, no matter what is the value of σ^2. This is one of the most commonly-used pivotal values.

The pivotal value can be used to form a confidence value for θ by first writing

$$\Pr\left(t_{(\alpha/2)} \leq g(\overline{Y},\mu) \leq t_{(1-\alpha/2)}\right) = 1-\alpha,$$

where $t_{(\pi)}$ is a percentile from the Student's t distribution. Then, after making substitutions for $g(\overline{Y},\mu)$, we form the familiar confidence interval for μ:

$$\left(\overline{Y} - t_{(1-\alpha/2)}\, S/\sqrt{n}, \quad \overline{Y} - t_{(\alpha/2)}\, S/\sqrt{n}\right),$$

where S^2 is the usual sample variance, $\sum(Y_i-\overline{Y})^2/(n-1)$.

(Note the notation: $t_{(\pi)}$, or for clarity, $t_{\nu,(\pi)}$ is the π quantile of a Student's t distribution. That means that

$$\Pr(Y \leq t_{\nu,(\pi)}) = \pi.$$

Other authors sometimes use a similar notation to mean the $1-\pi$ quantile and other times to mean the π quantiles; that is, the same authors use it both ways. I always use the notation in the way I indicate above. The reasons for the different symbols go back to the fact that $t_{\nu,(\pi)} = -t_{\nu,(1-\pi)}$, as for any distribution that is symmetric about 0.) ∎

Other similar pivotal functions have F distributions. For example, consider the usual linear regression model in which the n-vector random variable Y has a $\mathrm{N}_n(X\beta,\sigma^2 I)$ distribution, that is,

$$Y \sim \mathrm{N}_n(X\beta,\sigma^2 I), \tag{3.157}$$

where X is an $n\times m$ known matrix, and the m-vector β and the scalar σ^2 are unknown. A pivotal value useful in making inferences about β is

$$g(\widehat{\beta},\beta) = \frac{\left(X(\widehat{\beta}-\beta)\right)^{\mathrm{T}} X(\widehat{\beta}-\beta)/m}{(Y-X\widehat{\beta})^{\mathrm{T}}(Y-X\widehat{\beta})/(n-m)},$$

where

$$\widehat{\beta} = (X^{\mathrm{T}}X)^{+}X^{\mathrm{T}}Y.$$

The random variable $g(\widehat{\beta},\beta)$ for any finite value of σ^2 has an F distribution with m and $n-m$ degrees of freedom.

For a given parameter and family of distributions there may be multiple pivotal values. For purposes of statistical inference, such considerations as unbiasedness and minimum variance may guide the choice of a pivotal value to use. Alternatively, it may not be possible to identify a pivotal quantity for a particular parameter. In that case, we may seek an approximate pivot. A function is asymptotically pivotal if a sequence of linear transformations of the function is pivotal in the limit as $n\to\infty$.

If the distribution of T is known, c_1 and c_2 can be determined. If the distribution of T is not known, some other approach must be used. A common method is to use some numerical approximation to the distribution. Another method is to use bootstrap resampling.

Optimal Confidence Sets

We seek confidence sets that are "small" or "tight" in some way. We want the region of the parameter space that is excluded by the confidence set to be large; that is, we want the probability that the confidence set exclude parameters that are not supported by the observational evidence to be large. This is called "accuracy". We see most accurate confidence sets.

As with point estimation and tests of hypotheses, the risk in determining a confidence set depends on the specific distribution within the family of distributions assumed. We, therefore, seek *uniformly most accurate* confidence sets.

As in other cases where we seek uniform optimality, such procedures may not exist. We, therefore, may then take a similar approach for determining confidence sets, and restrict the allowable regions in some way. We may require that the confidence sets be unbiased, for example.

Unbiased Confidence Sets

A family of confidence sets Θ_S for θ is said to be *unbiased* (without regard to the level) if

$$\Pr_{\theta_0}(\Theta_S \ni \theta_1) \leq \Pr_{\theta_0}(\Theta_S \ni \theta_0) \; \forall \theta_0, \theta_1 \in \Theta. \tag{3.158}$$

Prediction Sets and Tolerance Sets

We often want to identify a set in which a future observation on a random variable has a high probability of occurring. This kind of set is called a *prediction set*. For example, we may assume a given sample $X_1, \ldots, X_n$ is from a $\mathrm{N}(\mu, \sigma^2)$ and we wish to determine a measurable set $S(X)$ such that for a future observation X_{n+1}

$$\inf_{P \in \mathcal{P}} \Pr_P(X_{n+1} \in S(X)) \geq 1 - \alpha. \tag{3.159}$$

More generally, instead of X_{n+1}, we could define a prediction interval for any random variable V.

The difference in this and a confidence set for μ is that there is an additional source of variation. The prediction set will be larger, so as to account for this extra variation.

We may want to separate the statements about V and $S(X)$. A *tolerance set* attempts to do this.

Given a sample X, a measurable set $S(X)$, and numbers δ and α in $]0,1[$, if

$$\inf_{P\in\mathcal{P}} \Pr_P(\Pr_P(V \in S(X)|X) \geq \delta) \geq 1-\alpha, \tag{3.160}$$

then $S(X)$ is called a δ-tolerance set for V with confidence level $1-\alpha$.

Approximate Confidence Sets

In some cases, we have a tractable probability model, so that we can determine confidence sets with exact levels of significance. In other cases the problem is not tractable analytically, so we must resort to approximations, which may be base on asymptotic distributions, or to estimates, which may be made using simulations.

Asymptotic inference uses asymptotic approximations. Computational inference uses probabilities estimated by simulation of an assumed or hypothesized data generating process or by resampling of an observed sample.

3.6 Variance Estimation

Statistical inferences that involve or are derived from statements of probability, such as hypothesis testing and determining confidence sets, require knowledge of the distribution of the statistic that is used. Often we know or can work out that distribution exactly, given the assumptions in the underlying probability model. In other cases we use approximate distributions. In either case, we are often faced with the problem of estimating the variance of a statistic.

In this section we first restrict our attention to the case in which the statistic of interest is a scalar; that is, the case in which the variance itself is a scalar. We describe two general methods, the jackknife and the bootstrap, based on resampling. We then consider the more general problem of estimating the variance-covariance matrix for a vector statistic. In either case, the first issue to address is the meaning of consistency of a variance-covariance estimator, which we will consider in a general way in Section 3.8.1, and define specifically in Definition 3.18. The jackknife and bootstrap can be used to estimate a variance-covariance matrix, and we also consider a "substitution" estimator.

3.6.1 Jackknife Methods

Jackknife methods make use of systematic partitions of a dataset to estimate properties of an estimator computed from the full sample.

Suppose that we have a random sample, $Y_1, \ldots, Y_n$, from which we compute a statistic T as an estimator of a parameter θ in the population from

which the sample was drawn. In the jackknife method, we compute the statistic T using only a subset of size $n-d$ of the given dataset; that is, we delete a set of size d.

There are of course

$$\mathrm{C}_d^n = \binom{n}{d}$$

such sets.

Let $T_{(-j)}$ denote the estimator computed from the sample with the j^{th} set of observations removed; that is, $T_{(-j)}$ is based on a sample of size $n-d$. The estimator $T_{(-j)}$ has properties similar to those of T. For example, if T is unbiased, so is $T_{(-j)}$. If T is not unbiased, neither is $T_{(-j)}$; its bias, however, is likely to be different.

The mean of the $T_{(-j)}$,

$$\overline{T}_{(\bullet)} = \frac{1}{\mathrm{C}_d^n} \sum_{j=1}^{\mathrm{C}_d^n} T_{(-j)}, \tag{3.161}$$

can be used as an estimator of θ. The $T_{(-j)}$ may also provide some information about the estimator T from the full sample.

For the case in which T is a linear functional of the ECDF, then $\overline{T}_{(\bullet)} = T$, so the systematic partitioning of a random sample will not provide any additional information.

Consider the weighted differences in the estimate for the full sample and the reduced samples:

$$T_j^* = nT - (n-d)T_{(-j)}. \tag{3.162}$$

The T_j^* are called "pseudovalues". (If T is a linear functional of the ECDF and $d = 1$, then $T_j^* = T(x_j)$; that is, it is the estimator computed from the single observation, x_j.)

We call the mean of the pseudovalues the "jackknifed" T and denote it as $\mathrm{J}(T)$:

$$\begin{aligned} \mathrm{J}(T) &= \frac{1}{\mathrm{C}_d^n} \sum_{j=1}^{\mathrm{C}_d^n} T_j^* \\ &= \overline{T}^*. \end{aligned} \tag{3.163}$$

In most applications of the jackknife, it is common to take $d = 1$, in which case $\mathrm{C}_d^n = n$. The term "jackknife" is often reserved to refer to the case of $d = 1$, and if $d > 1$, the term "delete d jackknife" is used. In the case of $d = 1$, we can also write $\mathrm{J}(T)$ as

$$\mathrm{J}(T) = T + (n-1)\left(T - \overline{T}_{(\bullet)}\right)$$

or

$$\mathrm{J}(T) = nT - (n-1)\overline{T}_{(\bullet)}. \tag{3.164}$$

It has been shown that $d = 1$ has some desirable properties under certain assumptions about the population (see Rao and Webster (1966)). On the other hand, in many cases consistency requires that $d \to \infty$, although at a rate substantially less than n, that is, such that $n - d \to \infty$.

Notice that the number of arithmetic operations required to compute a jackknife statistic can be large. When $d = 1$, a naive approach requires a number of computations of $\mathrm{O}(n^2)$, although in most cases computations can be reduced to $\mathrm{O}(n)$. In general, the order of the number of computations may be in $\mathrm{O}(n^{d+1})$. Even if the exponent of n can be reduced by clever updating computations, a delete-d jackknife can require very intensive computations. Instead of evaluating the statistic over all C_d^n subsets, in practice, we often use an average of the statistic computed only over a random sampling of the subsets.

Jackknife Variance Estimator

Although the pseudovalues are not independent (except when T is a linear functional), we treat them as if they were independent, and use $\mathrm{V}(\mathrm{J}(T))$ as an estimator of the variance of T, $\mathrm{V}(T)$. The intuition behind this is simple: a small variation in the pseudovalues indicates a small variation in the estimator. The sample variance of the mean of the pseudovalues can be used as an estimator of $\mathrm{V}(T)$:

$$\widehat{\mathrm{V}(T)}_{\mathrm{J}} = \frac{\sum_{j=1}^{\mathrm{C}_d^n}\left(T_j^* - \overline{T}^*\right)^2}{r(r-1)}. \tag{3.165}$$

Notice that when T is the mean and $d = 1$, this is the standard variance estimator.

From expression (3.165), it may seem more natural to take $\widehat{\mathrm{V}(T)}_{\mathrm{J}}$ as an estimator of the variance of $\mathrm{J}(T)$, and indeed it often is.

A variant of this expression for the variance estimator uses the original estimator T:

$$\frac{\sum_{j=1}^{\mathrm{C}_d^n}(T_j^* - T)^2}{r(r-1)}. \tag{3.166}$$

How good a variance estimator is depends on the estimator T and on the underlying distribution. Monte Carlo studies indicate that $\widehat{\mathrm{V}(T)}_{\mathrm{J}}$ is often conservative; that is, it often overestimates the variance.

The alternate expression (3.166) is greater than or equal to $\widehat{\mathrm{V}(T)}_{\mathrm{J}}$, as is easily seen (exercise); hence, it may be an even more conservative estimator.

3.6.2 Bootstrap Methods

From a given sample $y_1, \ldots, y_n$, suppose that we have an estimator $T(y)$. The estimator T^* computed as the same function T, using a bootstrap sample (that is, $T^* = T(y^*)$), is a *bootstrap observation* of T.

The bootstrap estimate of some function of the estimator T is a plug-in estimate that uses the empirical distribution P_n in place of P. This is the bootstrap principle, and this bootstrap estimate is called the *ideal bootstrap*.

For the variance of T, for example, the ideal bootstrap estimator is the variance $\mathrm{V}(T^*)$. This variance, in turn, can be estimated from bootstrap samples. The bootstrap estimate of the variance, then, is the sample variance of T^* based on the m samples of size n taken from P_n:

$$\widehat{\mathrm{V}}(T) = \widehat{\mathrm{V}}(T^*) \tag{3.167}$$

$$= \frac{1}{m-1}\sum(T^{*j} - \overline{T}^*)^2, \tag{3.168}$$

where T^{*j} is the j^{th} bootstrap observation of T. This, of course, can be computed by Monte Carlo methods by generating m bootstrap samples and computing T^{*j} for each.

If the estimator of interest is the sample mean, for example, the bootstrap estimate of the variance is $\widehat{\mathrm{V}}(Y)/n$, where $\widehat{\mathrm{V}}(Y)$ is an estimate of the variance of the underlying population. (This is true no matter what the underlying distribution is, as long as the variance exists.) The bootstrap procedure does not help in this situation.

3.6.3 Substitution Methods

The jackknife and bootstrap can be used to estimate a variance-covariance estimator. Another useful type of estimator is called a substitution estimator or sandwich estimator.

The idea in the "substitution method" for estimating V_n is to arrive at an expression for V_n that involves a simpler variance along with quantities that are known functions of the sample. Often that simpler variance can be estimated by an estimator with known desirable properties. An estimator of V_n in which the simpler estimator and the known sample functions are used is called a substitution estimator. A simple example is the estimator of the variance of $\widehat{\beta}$ in a linear regression following the model (3.157) on page 299. The variance-covariance matrix is $(X^{\mathrm{T}}X)^{-1}\sigma^2$. A substitution estimator is one in which the regression MSE is substituted for σ^2.

The so-called "sandwich estimators" are often substitution estimators.

$$(Z^{\mathrm{T}}Z)^{-1}V(Z^{\mathrm{T}}Z)^{-1}$$

V is some variance-covariance estimator that probably includes a scalar multiple of $\widehat{\sigma}^2$.

3.7 Applications

3.7.1 Inference in Linear Models

3.7.2 Inference in Finite Populations

One of the most important areas of application of statistics is in sampling and making inferences about a finite set of objects, such as all inhabitants of a given country. Finite population sampling or, as it is often called, "survey sampling" is the process of designing and collecting a sample of some characteristic of the members of the finite set. There are various issues and considerations in finite population sampling that rarely arise in other areas of statistical inference.

A *finite population* is some finite set $\mathcal{P} = \{(1, y_1), \ldots, (N, y_N)\}$, where the y_i are real numbers associated with the set of objects of interest. (Note that here we use "population" in a different way from the use of the term as a probability measure.) Finite population sampling is the collection of a sample $\mathcal{S} = \{(L_1, X_1), \ldots, (L_n, X_n)\}$ where $X_i = y_j$, for some j. (In general, the set $X = \{X_1, \ldots, X_n\}$ may be a multiset.) In discussions of sampling it is common to use n to denote the size of the sample and N to denote the size of the population. Another common notation used in sampling is Y to denote the population total, $Y = \sum_{i=1}^{N} y_i$. The objective of course is to make inferences about the population, such as to estimate the total Y.

From a parametric point of view, the parameter that characterizes the population is $\theta = (y_1, \ldots, y_N)$, and the parameter space, Θ, is the subspace of $\mathbb{R}^N$ containing all possible values of the y_i.

There are various ways of collecting the sample. A simple random sample with replacement is a sample in which the X_i are iid. A related concept is a simple random sample without replacement, in which the $X_i = y_j$ are constrained so that a given value of j can occur once at most.

A common method of collecting a sample is to select elements from the finite population with different probabilities. If $\pi_i > 0$ for all i is the probability that y_i is included in the sample, and if

$$\mathcal{L}_S = \{i : y_i \text{ is included in the sample}\}$$

then clearly

$$\widehat{Y} = \sum_{i \in \mathcal{L}_S} \frac{y_i}{\pi_i}$$

is an unbiased estimator of the population total.

The variance of this estimator depends on the π_i as well as π_i, where π_{ij} is the probability that both y_i and y_j are included in the sample.

Much of the theory for finite population inference depends on how a probability distribution is used. As we have implied above the probability distribution used in inference arises from the random selection of the sample itself. This is called a "design based" approach. Other approaches to statistical inference in finite populations begin by modeling the population as a realization

$(y_1, \ldots, y_N)$ of a random vector $(Y_1, \ldots, Y_N)$. This is called a superpopulation model. In the context of a superpopulation, probability distributions can be assumed for each Y_i, possibly with associated covariates. This may lead to a "model based" approach to statistical inference in finite populations.

Just as with statistical procedures in other settings, a superpopulation model also allows us to investigate asymptotic properties of statistics.

We will discuss some topics of inference in finite populations further in Section 5.5.2.

3.8 Asymptotic Inference

In the standard problem in statistical inference, we are given some family of probability distributions, we take random observations on a random variable, and we use some function of the random sample to estimate some aspect of the underlying probability distribution or to test some statement about the probability distribution.

The approach to statistical inference that we would like to follow is to identify a reasonable statistic to use as an estimator or a test statistic, then work out its distribution under the given assumptions and under any null hypothesis, and, knowing that distribution, assess its goodness for the particular application and determine levels of confidence to associate with our inference. In many of interesting problems in statistical inference we cannot do this, usually because the distributions are not tractable.

It is often easy, however, to determine the limiting distribution of a statistic. In that case, we can base an approximate inference on the *asymptotic* properties of the statistic. This is *asymptotic inference*.

The Basic Setup and Notation

As usual in statistical inference, we have a family of probability distributions $\mathcal{P} = \{P_\theta\}$, where θ may be some parameter in a real-valued parameter space Θ ("parametric inference"), or θ may just be some index in an index set $\mathcal{I}$ to distinguish one distribution, P_{θ_1}, from another, P_{θ_2} ("nonparametric inference"). The parameter or the index is not observable; however, we assume $P_{\theta_1} \neq P_{\theta_2}$ if $\theta_1 \neq \theta_2$ ("identifiability").

We have an observable random variable X. We have a random sample, $X_1, \ldots, X_n$, which we may also denote by X; that is, we may use X not just as the random variable (that is, a Borel function on the sample space) but also as the sample: $X = X_1, \ldots, X_n$.

Both θ and X may be vectors. (Recall that I use "real-valued" to mean either a scalar (that is, an element in $\mathbb{R}$) or a real-valued vector (that is, an element in $\mathbb{R}^k$, where k is a positive integer possibly larger than 1).)

The canonical problem in parametric inference is to estimate $g(\theta)$ or to test an hypothesis concerning $g(\theta)$, where g is some real-valued measurable function.

We denote the statistic (possibly an estimator or a test statistic) as $T_n(X)$, or just T_n. We also use the same symbol to denote the sequence of statistics, although to emphasize the sequence, as opposed to the n^{th} term in the sequence, we may write $\{T_n\}$.

We will often be interested in weak convergence, and in that case the order of convergence $\mathrm{O}(f(n))$ as in Example 1.23 on page 85 will be of interest.

3.8.1 Consistency

Consistency is a general term used for various types of asymptotic convergence and has different meanings for different statistical procedures. Unless it is clear from the context, we must qualify the word "consistency" with the type of convergence and with the type of inference. We speak of consistent point estimators and consistent tests of hypotheses.

In this section we will discuss consistency of point estimators. This relates to the convergence of the sequence of estimators $T_n(X)$ to the estimand $g(\theta)$, and these types correspond directly to those discussed in Section 1.3.3. We will consider consistency of hypothesis tests and the related concepts of asymptotic correctness and accuracy in Chapter 7.

Convergence is defined with respect to a distribution. In a problem of statistical inference we do not know the distribution, only the distributional family, $\mathcal{P}$. To speak of consistency, therefore, we require that the convergence be with respect to every distribution in $\mathcal{P}$.

The three most common kinds of consistency for point estimators are weak consistency, strong consistency, and L_r-consistency.

Definition 3.14 (weak consistency)
$T_n(X)$ is said to be *weakly consistent* for $g(\theta)$ iff

$$T_n(X) \xrightarrow{\mathrm{p}} g(\theta) \quad \text{wrt any } P \in \mathcal{P}.$$

∎

This kind of consistency involves a *weak convergence*. It is often what is meant when we refer to "consistency" without a qualifier. Whenever the asymptotic expectation of a sequence of estimators is known, consistency is usually proved by use of a Chebyshev-type inequality.

Definition 3.15 (strong (or a.s.) consistency)
$T_n(X)$ is said to be *strongly consistent* for $g(\theta)$ iff

$$T_n(X) \xrightarrow{\mathrm{a.s.}} g(\theta) \quad \text{wrt any } P \in \mathcal{P}.$$

∎

Definition 3.16 (L_r-consistency)
$T_n(X)$ is said to be L_r-*consistent* for $g(\theta)$ iff

$$T_n(X) \xrightarrow{L_r} g(\theta) \quad \text{wrt any } P \in \mathcal{P}.$$

∎

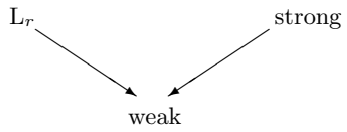
L_r-convergence applies to convergence in expectation:

$$\lim_{n\to\infty} \mathrm{E}(\|T_n(X) - g(\theta)\|_r^r) = 0.$$

For $r = 1$, L_r-consistency is called *consistency in mean*. For $r = 2$, L_r-consistency is called *consistency in mean squared error*.

The term "consistency" or "consistent" is often used without a qualifier. While this may be considered bad practice, it is fairly common. In certain contexts, we often refer to weak consistency as just "consistency", without the qualifier. "Consistency" or "consistent" without a qualifier, however, often means consistency or consistent in mean squared error.

There are relationships among these types of consistency that are similar to those among the types of convergence. As in Figure 1.3 on page 81, we have Recall Example 1.21 to see that a.s. consistency does not imply consistency



in mean squared error.

For any convergent sequence, the rate of convergence is of interest. We quantify this rate of convergence in terms of another sequence, which is often some nice function of n.

Definition 3.17 (a_n-consistency)
Given a sequence of positive constants $\{a_n\}$ with $\lim_{n\to\infty} a_n = \infty$, $T_n(X)$ is said to be a_n-*consistent* for $g(\theta)$ iff $a_n(T_n(X) - g(\theta)) \in \mathrm{O_P}(1)$ wrt any $P \in \mathcal{P}$, that is,

$$\forall \epsilon > 0 \; \exists \text{ constant } C_\epsilon > 0 \ni \sup_n \Pr(a_n \|T_n(X) - g(\theta)\| \geq C_\epsilon) < \epsilon.$$

∎

Notice that this is a kind of weak consistency. Although it is common to include the a_n sequence as a scale factor, if $T_n(X)$ is a_n-*consistent* for $g(\theta)$ then we could write $(T_n(X) - g(\theta)) \in \mathrm{O_P}(a_n^{-1})$. a_n-consistency plays a major role in asymptotic inference.

The most common kind of a_n-consistency that we consider is $\sqrt{n}$-consistency; that is, $a_n = \sqrt{n}$, as in Example 1.23 on page 85, where we saw that $\overline{X}_n - \mu \in \mathrm{O_P}(n^{-1/2})$.

We are interested the limiting behavior of such properties of statistics as the variance, the bias, and the mean squared error. We often express sequences of these properties in terms of big O (not in probability) of a convergent sequence. The variance and mean squared error are often in $\mathrm{O}(n^{-1})$, as for the variance of $\overline{X}_n$, for example. A sequence of estimators whose variance or mean squared error is in $\mathrm{O}(n^{-r})$ is a better sequence of estimators than one whose mean squared error is in $\mathrm{O}(n^{-s})$ if $r > s$.

Quantities such as the variance, the bias, and the mean squared error are defined in terms of expectations, so firstly, we need to be precise in our meaning of asymptotic expectation. In the following we will distinguish asymptotic expectation from limiting expectation. A related term is "approximate" expectation, but this term is sometimes used in different ways. Some authors use the term "approximately unbiased" in reference to a limiting expectation. Other authors and I prefer the term "unbiased in the limit" to refer to this property. This property is different from asymptotically unbiased, as we will see.

Consistency of a Sequence to a Sequence

In some cases, rather than a fixed estimand $g(\theta)$, we are interested in a sequence of estimands $g_n(\theta)$. In such cases, it may not be adequate just to consider $|T_n - g_n(\theta)|$. This would not be meaningful if, for example, $g_n(\theta) \to 0$. This kind of situation occurs, for example, when $g_n(\theta)$ is the variance of the mean of a sample of size n from a population with finite variance. In such cases we could define any of the types of consistency described above using the appropriate type of convergence in this expression,

$$|T_n/g_n(\theta) - 1| \to 0. \tag{3.169}$$

A common situation is one in which $g_n(\theta)$ is a sequence of variance-covariance matrices, say Σ_n. Because, for nondegenerate distributions, these are positive definite matrices, we may restrict our attention to a sequence of positive definite estimators, say V_n. For this case, we give a special definition for consistent estimators of the sequence of variance-covariance matrices.

Definition 3.18 (consistent estimators of variance-covariance matrices)

Let $\{\Sigma_n\}$ be a sequence of $k \times k$ positive definite matrices and V_n be a positive definite matrix estimator of Σ_n for each n. Then V_n is said to be consistent for Σ_n if

$$\left\| \Sigma_n^{-1/2} V_n \Sigma_n^{-1/2} - I_k \right\| \xrightarrow{\mathrm{P}} 0. \tag{3.170}$$

Also V_n is said to be strongly consistent for Σ_n if

$$\left\| \Sigma_n^{-1/2} V_n \Sigma_n^{-1/2} - I_k \right\| \xrightarrow{\text{a.s.}} 0. \tag{3.171}$$

∎

Note the similarity of these expressions to expression (3.169). In many cases of interest $\|\Sigma_n\| \to 0$, so these expressions are not the same as $\|V_n - \Sigma_n\| \to 0$.

Theorem 3.15
Assume the conditions of Definition 3.18. Equation (3.170) holds iff for every sequence $\{l_n\} \in \mathbb{R}^k$,

$$l_n^{\mathrm{T}} V_n l_n \xrightarrow{\text{p}} 1. \tag{3.172}$$

Proof. Exercise. ∎

3.8.2 Asymptotic Expectation

Asymptotic inference is based on the asymptotic distribution of a statistic, T_n. Recall that the asymptotic distribution is defined in terms of the convergence of the CDFs at each point of continuity t of the CDF of X, F: $\lim_{n\to\infty} F_n(t) = F(t)$, and an expectation can be defined in terms of a CDF. The properties of the asymptotic distribution, such as its mean or variance, are the asymptotic values of the corresponding properties of T_n.

Because $\{T_n\}$ may converge to a degenerate random variable, it may be more useful to consider more meaningful sequences of the form $\{a_n(X_n - c)\}$ as in Sections 1.3.7 and 1.3.8. Even if T_n is a normalized statistic, such as $\overline{X}$, with variance of the form σ^2/n, the limiting values of various properties of T_n may not be very useful. We need an "asymptotic variance" different from $\lim_{n\to\infty} \sigma^2/n$. Hence, we defined "an asymptotic expectation" in Definition 1.43 in terms of the expectation in the asymptotic distribution.

We refer to $\lim_{n\to\infty} \mathrm{E}(T_n)$ as the *limiting expectation*. It is important to recognize the difference in limiting expectation and asymptotic expectation. (These two terms are not always used in this precise way, so the student must be careful to understand the meaning of the terms in their context.)

Notice that this definition of asymptotic expectation may allow us to address more general situations. For example, we may consider the asymptotic variance of a sequence of estimators $\sqrt{n}T_n(X)$. The asymptotic variance may be of the form $\mathrm{V}(T/n)$ (which we should not be tempted to say is just 0, because $n \to \infty$).

Notation and Terminology

The definition of asymptotic expectation in Definition 1.43 is a standard one. Terminology related to this definition, however, is not always standard. To illustrate, we consider a result in a common situation: $\sqrt{n}(\overline{X} - \mu) \xrightarrow{\text{d}} \mathrm{N}(0, \sigma^2)$. By the definition, we would say that the asymptotic variance of $\overline{X}$ is σ^2/n;

whereas, often in the literature (see, for example, TPE2, page 436, following equation (1.25)), σ^2 would be called the asymptotic variance. This confusion in the notation is not really very important, but the reader should be aware of it. To me, it just seems much more logical to follow Definition 1.43 and call σ^2/n the asymptotic variance. (See Jiang (2010), page 19, for additional comments on the terminology.)

3.8.3 Asymptotic Properties and Limiting Properties

After defining asymptotic expectation, we noted an alternative approach based on a limit of the expectations, which we distinguished by calling it the limiting expectation. These two types of concepts persist in properties of interest that are defined in terms of expectations, such as bias and variance and their combination, the mean squared error.

One is based on the asymptotic distribution and the other is based on limiting moments. Although in some cases they may be the same, in general they are different, as we will see.

Asymptotic Bias and Limiting Bias

Now consider a sequence of estimators $\{T_n(X)\}$ for $g(\theta)$, with $\mathrm{E}(|T_n|) < \infty$, in the family of distributions $\mathcal{P} = \{P_\theta\}$. We define the *limiting bias* of $\{T_n\}$ within the family $\mathcal{P}$ to be $\lim_{n\to\infty} \mathrm{E}(T_n) - g(\theta)$.

Suppose $T_n(X) \xrightarrow{\mathrm{d}} T$ and $\mathrm{E}(|T|) < \infty$. The limiting bias of $\{T_n\}$ within the family $\mathcal{P}$ is $\mathrm{E}(T) - g(\theta)$.

Notice that the bias may be a function of θ; that is, it may depend on the specific distribution within $\mathcal{P}$.

If the limiting bias is 0 for any distribution within $\mathcal{P}$, we say $\{T_n(X)\}$ is *unbiased for $g(\theta)$ in the limit*.

It is clear that if $T_n(X)$ is unbiased for $g(\theta)$ for all n, then $\{T_n(X)\}$ is unbiased for $g(\theta)$ in the limit.

We can easily construct an estimator that is biased in any finite sample, but is unbiased in the limit. Suppose we want an estimator of the mean μ (which we assume is finite). Let

$$T_n = \overline{X}_n + \frac{c}{n},$$

for some $c \neq 0$. Now, the bias for any n is c/n. The limiting bias of T_n for μ, however, is 0, and since this does not depend on μ, we say it is unbiased in the limit.

To carry this further, suppose $X_1, \ldots, X_n \overset{\mathrm{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$, and with

$$T_n = \overline{X}_n + \frac{c}{n}$$

as above, form $\sqrt{n}(T_n - \mu) = \sqrt{n}(\overline{X}_n - \mu) + c/\sqrt{n}$. We know $\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{\mathrm{d}} \mathrm{N}(0, \sigma^2)$ and $c/\sqrt{n} \to 0$, so by Slutsky's theorem,

$$\sqrt{n}(T_n - \mu) \xrightarrow{\mathrm{d}} \mathrm{N}(0, \sigma^2).$$

Hence, the limiting bias of T_n for μ is also 0, and since this does not depend on μ, we say it is unbiased in the limit.

We define the *asymptotic bias* in terms of the asymptotic expectation of $\{T_n\}$ given a sequence of positive constants $\{a_n\}$ with $\lim_{n\to\infty} a_n = \infty$ or with $\lim_{n\to\infty} a_n = a > 0$, and such that $a_n T_n(X) \xrightarrow{\mathrm{d}} T$. An asymptotic bias of $\{T_n\}$ is $\mathrm{E}(T - g(\theta))/a_n$.

If $\mathrm{E}(T - g(\theta))/a_n = 0$, we say $\{T_n(X)\}$ is *asymptotically unbiased for* $g(\theta)$.

It is clear that if $T_n(X)$ is unbiased for $g(\theta)$ for any n, then $\{T_n(X)\}$ is asymptotically unbiased for $g(\theta)$.

Example 3.25 unbiased in limit but asymptotically biased
To illustrate the difference in asymptotic bias and limiting bias, consider $X_1, \ldots, X_n \overset{\mathrm{iid}}{\sim} \mathrm{U}(0, \theta)$, and the estimator $X_{(n)}$ (which we know to be sufficient for $g(\theta) = \theta$). We can work out the asymptotic distribution of $n(\theta - X_{(n)})$ to be exponential with parameter θ. (The distributions of the order statistics from the uniform distribution are betas. These distributions are interesting and you should become familiar with them.) Hence, $X_{(n)}$ is asymptotically biased. We see, however, that the limiting bias is $\lim_{n\to\infty} \mathrm{E}(X_{(n)} - \theta) = \frac{n-1}{n}\theta - \theta = 0$; that is, $X_{(n)}$ is unbiased in the limit. ∎

Notice the role that the sequence $\{a_n\}$ plays. This would allow us to construct a sequence that is biased in the limit, but is asymptotically unbiased.

Consistency

There are also, of course, relationships between consistency and limiting bias. Unbiasedness in the limit implies consistency in mean.

Example 3.26 consistency and limiting and asymptotic bias
Consider $X_1, \ldots, X_n \overset{\mathrm{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$, and an estimator of the mean

$$S_n = \overline{X}_n + \frac{c}{\sqrt{n}},$$

for some $c \neq 0$. (Notice this estimator is slightly different from T_n above.) As above, we see that this is unbiased in the limit (consistent in the mean), and furthermore, we have the mean squared error

$$\begin{aligned} \mathrm{MSE}(S_n, \mu) &= \mathrm{E}((S_n - \mu)^2) \\ &= \frac{\sigma^2}{n} + \left(\frac{c}{\sqrt{n}}\right)^2 \end{aligned}$$

tending to 0, hence we see that this is consistent in mean squared error. However, $\sqrt{n}(S_n - \mu) = \sqrt{n}(\overline{X}_n - \mu) + c$ has limiting distribution $\mathrm{N}(c, \sigma^2)$; hence S_n is asymptotically biased. ∎

We also note that an estimator can be asymptotically unbiased but not consistent in mean squared error. In Example 3.26, we immediately see that X_1 is asymptotically unbiased for μ, but it is not consistent in mean squared error for μ.

Another interesting example arises from a distribution with slightly heavier tails than the normal.

Example 3.27 consistency and limiting bias
Consider the double exponential distribution with $\theta = 1$, and the estimator of the mean

$$R_n(X) = \frac{X_{(n)} + X_{(1)}}{2}.$$

(This is the mid-range.) We can see that R_n is unbiased for any finite sample size (and hence, is unbiased in the limit); however, we can show that

$$\mathrm{V}(R_n) = \frac{\pi^2}{12},$$

and, hence, R_n is not consistent in mean squared error. ∎

Asymptotic Variance, Limiting Variance, and Efficiency

We define the asymptotic variance and the limiting variance in similar ways as in defining the asymptotic bias and limiting bias, and we also note that they are different from each other. We also define asymptotic mean squared error and the limiting mean squared error in a similar fashion. The limiting mean squared error is of course related to consistency in mean squared error.

Our interest in *asymptotic* (not "limiting") variance or mean squared error is as they relate to optimal properties of estimators. The "efficiency" of an estimator is related to its mean squared error.

Usually, rather than consider efficiency in a absolute sense, we use it to compare two estimators, and so speak of the *relative efficiency*. When we restrict our attention to unbiased estimators, the mean-squared error is just the variance, and in that case we use the phrase *efficient* or *Fisher efficient* (Definition 3.8) to refer to an estimator that attains its Cramér-Rao lower bound (the right-hand side of inequality (B.25) on page 854.)

As before, assume a family of distributions $\mathcal{P}$, a sequence of estimators $\{T_n\}$ of $g(\theta$, and a sequence of positive constants $\{a_n\}$ with $\lim_{n\to\infty} a_n = \infty$ or with $\lim_{n\to\infty} a_n = a > 0$, and such that $a_n T_n(X) \xrightarrow{\mathrm{d}} T$ and $0 < \mathrm{E}(T) < \infty$. We define the asymptotic mean squared error of $\{T_n\}$ for estimating $g(\theta)$ wrt $\mathcal{P}$ as an asymptotic expectation of $(T_n - g(\theta))^2$; that is, $\mathrm{E}((T - g(\theta))^2)/a_n$, which we denote as $\mathrm{AMSE}(T_n, g(\theta), \mathcal{P})$.

For comparing two estimators, we may use the *asymptotic relative efficiency*. The asymptotic relative efficiency of the estimators S_n and T_n for $g(\theta)$ wrt $\mathcal{P}$ is defined as

$$\text{ARE}(S_n, T_n) = \text{AMSE}(S_n, g(\theta), \mathcal{P})/\text{AMSE}(T_n, g(\theta), \mathcal{P}). \tag{3.173}$$

The ARE is essentially a scalar concept; for vectors, we usually do one at a time, ignoring covariances.

Asymptotic Significance

For use of asymptotic approximations for confidence sets and hypothesis testing, we need a concept of asymptotic significance. As in the case of exact significance, the concepts in confidence sets and hypothesis tests are essentially the same.

We assume a family of distributions $\mathcal{P}$, a sequence of statistics $\{T_n\}$, and a sequence of tests $\{\delta(X_n)\}$ based on the iid random variables $X_1, \ldots, X_n$. The test statistic $\delta(\cdot)$ is defined in terms the decisions; it takes the value 1 for the case of deciding to reject H_0 and conclude H_1, and the value 0 for the case of deciding not to reject H_0.

Asymptotic Properties of Tests

In hypothesis testing, the standard setup is that we have an observable random variable with a distribution in the family $\mathcal{P}$. Our hypotheses concern a specific member $P \in \mathcal{P}$. We have a null hypothesis

$$H_0 : \quad P \in \mathcal{P}_0$$

and an alternative hypothesis

$$H_1 : \quad P \in \mathcal{P}_1,$$

where $\mathcal{P}_0 \subseteq \mathcal{P}$, $\mathcal{P}_1 \subseteq \mathcal{P}$, and $\mathcal{P}_0 \cap \mathcal{P}_1 = \emptyset$.

Definition 3.19 (limiting size)
Letting $\beta(\delta(X_n), P)$ be the power function,

$$\beta(\delta(X_n), P) = \Pr_P(\delta(X_n) = 1).$$

We define

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}_0} \beta(\delta(X_n), P), \tag{3.174}$$

if it exists, as the *limiting size* of the test. ∎

Definition 3.20 (asymptotic significance)
Given the power function $\beta(\delta(X_n), P)$. If

$$\limsup_n \beta(\delta(X_n), P) \leq \alpha \, \forall P \in \mathcal{P}_0, \tag{3.175}$$

then α is called an *asymptotic significance level* of the test. ∎

Definition 3.21 (consistency)
The sequence of tests $\{\delta(X_n)\}$ is *consistent* for the test $P \in \mathcal{P}_0$ versus $P \in \mathcal{P}_1$ iff

$$\lim_{n\to\infty} (1 - \beta(\delta(X_n), P)) = 0 \, \forall P \in \mathcal{P}_1. \tag{3.176}$$

∎

Definition 3.22 (uniform consistency)
The sequence of tests $\{\delta_n\}$ is *uniformly consistent* iff

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}_1} (1 - \beta(\delta_n, P)) = 0.$$

∎

Asymptotic Properties of Confidence Sets

Let $C(X)$ be a confidence set for $g(\theta)$.

Definition 3.23 (asymptotic significance level)
If

$$\liminf_n \Pr(C(X) \ni g(\theta)) \geq 1 - \alpha \, \forall P \in \mathcal{P}, \tag{3.177}$$

then $1 - \alpha$ is an *asymptotic significance level* of $C(X)$. ∎

Definition 3.24 (limiting confidence coefficient)
If

$$\lim_{n\to\infty} \inf_{P\in\mathcal{P}} \Pr(C(X) \ni g(\theta)) \tag{3.178}$$

exists, then it is the *limiting confidence coefficient* of $C(X)$. ∎

"The" Asymptotic Distribution

In determining asymptotic confidence sets or asymptotic relative efficiencies, we need expressions that do not depend on unknown parameters. This fact determines which asymptotic expectations are useful.

The asymptotic expectation of some sequence of statistics, or of pivotal quantities, is determined by the sequence $\{a_n\}$ (used above in the definitions).

In the univariate delta method, for example, we find a quantity $a_n(g(X_n) - g(c))$ that converges in distribution to $\mathrm{N}(0, v)$, where v does not depend on

an unknown parameter. In that case, we can set a confidence interval based on the approximate distribution of $g(X_n)$ as $\mathrm{N}(g(c), v/a_n^2)$.

To speak of the asymptotic distribution of $a_n(g(X_n) - g(c))$ is clear; but to refer to "the" asymptotic distribution of $g(X_n)$ is somewhat less so.

Because it is the useful approximate distribution resulting from asymptotic expectations, we often say that "the asymptotic distribution" of $g(X_n)$ is $\mathrm{N}(g(c), v/a_n^2)$. You should recognize that "the" in this statement is somewhat arbitrary. It might be better to call it "the asymptotically approximate distribution that I'm going to use in this application".

Again, we should distinguish "asymptotic" from "limiting".

In the example of the delta method above, it is likely that

$$g(X_n) \xrightarrow{\mathrm{d}} g(c);$$

that is, $g(X_n)$ converges in distribution to the constant $g(c)$; or the limiting distribution of $g(X_n)$ is degenerate at $g(c)$. "The" asymptotic variance is 0.

**** discuss expansion of statistical functionals *** refer to Serfling

This would not be very useful in asymptotic inference. We therefore seek "an" asymptotic variance that is more useful. In asymptotic estimation using $g(X_n)$, we begin with an expression of the form $a_n(g(X_n) - g(c))$ that has a limiting distribution of the desired form (usually that means such that the variance does not involve any unknown parameters and it does not involve n). If this distribution is in a location-scale family, then we make the appropriate linear transformation (which probably results in a variance that does involve n).

We then often refer to this as the asymptotic distribution of $g(X_n)$. Sometimes, as mentioned above, however, the limiting distribution of $g(X_n)$ is degenerate.

This is not to imply that asymptotic expectations are entirely arbitrary. Proposition 2.3 in MS2 shows that there is a certain uniqueness in the asymptotic expectation. This proposition involves three cases regarding whether the expectation of $g(X_n)$ (without the a_n sequence) is 0. In the example above, we have a degenerate distribution, and hence the asymptotic expectation that defines the asymptotic variance is 0.

3.8.4 Properties of Estimators of a Variance Matrix

If the statistic is a vector, we need an estimator of the variance-covariance matrix. Because a variance-covariance matrix is positive definite, it is reasonable to consider only estimators that are positive definite a.s.

We have defined what it means for such an estimator to be consistent (Definition 3.18 on page 309).

Theorem 3.16
conditions for the consistency of substitution estimators.

Proof.
****

Theorem 3.17
Given a sequence of estimators $\{T_n\}$ *of* $\{g_n(\theta)\}$ *with variance-covariance matrices* $\{\Sigma_n\}$*, if*

$$\Sigma_n^{-1/2}(T_n - g_n(\theta)) \xrightarrow{\mathrm{d}} \mathrm{N}(0, I_k),$$

and if V_n *is consistent for* Σ_n*, then*

$$V_n^{-1/2}(T_n - g_n(\theta)) \xrightarrow{\mathrm{d}} \mathrm{N}(0, I_k). \tag{3.179}$$

Proof.
****

Notes and Further Reading

The general problem of statistical inference, that is, the use of observed data for which we have a family of probability distributions to provide information about those probability distributions, is an "inverse problem". Nineteenth and twentieth century scientists who made inferences about probability models referred to the problem as one of "inverse probability". Statisticians in the early twentieth century also used this term. Although the maximum likelihood approach could be thought of as a method of inverse probability, R. A. Fisher, who developed likelihood methods, made a distinction between the methods and "inverse probability" as a general term fell into disuse.

Foundations

Although statistical science has been very successful in addressing real-world problems, there are some issues at the foundations of statistics that remain somewhat controversial. One of these issues is the incorporation of subjectivity in statistical analysis, and another is the relevance of certain principles, such as "conditionality" and sufficiency in statistical inference.

In Chapter 1 I took the view that probability theory is an area of pure mathematics; hence, given a consistent axiomatic framework, "beliefs" are irrelevant. Distinctions between "objectivists" and "subjectivists" have no place in probability theory.

In statistics, however, this is a different matter. Instead of a vague notion of "subjective probability", we may explicitly incorporate the subjectivity in our decisions, that is, in our statistical inference. Press and Tanur (2001) argue that scientists have never behaved fully objectively, but rather, some of the greatest scientific minds have relied on intuition, hunches, and personal beliefs to understand empirical data. These subjective influences have often aided

in the greatest scientific achievements. Press and Tanur (2001) also express the view that science will advance more rapidly if the methods of Bayesian statistical analysis, which may incorporate subjective beliefs, are used more widely. In this chapter, we have laid a framework for Bayesian approaches, and in Chapter 4, we will discuss it more fully.

Samaniego (2010) provides an interesting comparison of the general Bayesian and frequentist approaches to estimation.

Evidence and Decision Making

The "conditionality" principle was formulated by Birnbaum (1962) as a connection between the sufficiency principle and the likelihood principle. The context of the conditionality principle is a set of possible experiments (designs, or data-generating processes) for obtaining data for statistical inference. The conditionality principle basically states that if a particular experiment is selected randomly (that is, independent of any observations), then only the experiment actually performed is relevant. Berger and Wolpert (1988) discuss these principles and their relationships to each other.

The likelihood principle, alluded to in Example 3.12 on page 237 and stated on page 245, is perhaps the principle that brings into sharpest contrast some different fundamental approaches to statistical inference. To many statisticians, general statistical principles that focus on the *observed data* rather than on the *data-generating process* often miss salient points about the process. Lucien Le Cam (in Berger and Wolpert (1988), page 185.1) expressed a common opinion among statisticians, "One should keep an open mind and be a bit 'unprincipled'."

Royall (1997) evidence versus hypothesis testing ********

Information

Fisher information is the most familiar kind of information to most statisticians, but information that is related to entropy (see Section 1.1.5) is often used as a basis for statistical inference. Soofi (1994) discusses some subtleties in the different definitions of information that are ultimately based on Shannon information.

General Approaches

While I have classified the general approaches to statistical inference into five groups, there are obviously other classifications, and, in any event, there are overlaps among the classes; for example, approaches based on empirical likelihood follow ideas from both likelihood and ECDF methods. In subsequent chapters, we will have more to say about each of these approaches, especially a decision-theoretic approach and use of a likelihood function.

Least-squares estimation was first studied systematically by C. F. Gauss in the early 1800's in the context of curve fitting. The name of the Gauss-Markov theorem reminds us of his work in this area.

Karl Pearson around the beginning of the twentieth century promoted estimation based on fitting moments. These methods are the earliest and simplest of the general plug-in methods. This class of methods is also called "analog" estimation (see Manski (1988) for a general discussion in a specific area of application). The good statistical properties of the ECDF lend appeal to plug-in methods. The strong uniform convergence given in the Glivenko-Cantelli theorem is another illustration of Littlewood's third principle regarding the "nearly" uniform convergence of pointwise convergent sequences.

R. A. Fisher developed and studied maximum likelihood methods in the 1920's. These are probably the most widely-used statistical methods across a broad range of applications.

Ideas and approaches developed by engineers and physical scientists lead to statistical methods characterized by maximum entropy. Much of this work dates back to Claude Shannon in the 1930's. E. T. Jaynes in the 1950's formalized the approach and incorporated it in a Bayesian framework. His posthumous book edited by G. Larry Bretthorst (Jaynes, 2003) is a very interesting discussion of a view toward probability that leads to a Bayesian maximum entropy principle for statistical inference. We will discuss the maximum entropy principle in more detail in the context of Bayesian priors in Section 4.2.5, beginning on page 346. Wu (1997) expands on the use of the maximum entropy principle in various areas of application. Pardo (2005) gives an extensive overview of the use of functionals from information theory in statistical inference. In some disciplines, such as electrical engineering, this approach seems to arrive very naturally.

Pitman's measure of closeness was introduced in 1937. The idea did not receive much attention until the article by Rao (1981), in which was given the definition we have used, which is slightly different from Pitman's. Pitman's original article was reproduced in a special issue of *Communications in Statistics* (Pitman, 1991) devoted to the topic of Pitman closeness. The lack of transitivity of Pitman's closeness follows from Arrow's "impossibility theorem", and is a natural occurrence in paired comparisons (see David (1988)). The example on page 220 is called a "cyclical triad".

David and Salem had considered estimators similar to (3.20) for a normal mean in 1973, and in David and Salem (1991) they generalized these shrunken estimators to estimators of the means that are Pitman-closer than the sample mean in a broad class of location families.

The basic ideas of the "decision theory" approach, such as risk and admissibility, were organized and put forth by Wald (1950). Wald showed that many of the classical statistical methods, such as hypothesis testing and even design of experiments, could be developed within the context of decision theory. Wald also related many of the basic ideas of decision theory to game theory for two-person games, and Blackwell and Girshick (1954) and Ferguson

(1967) expanded on these relations. In a two-person game, one player, "nature", chooses action "θ" and the other player, "the statistician", chooses action "a" and the elements of the payoff matrix are the values of the loss function evaluated at θ and a.

Complete Class Theorems

Wald, starting in Wald (1939) and especially in Wald (1947a), gave the first characterizations of a class of decision rules that are complete or are essentially complete. These theorems are collected in Wald (1950). See also Kiefer (1953), who proved some additional properties of complete classes, and Le Cam (1955), who relaxed some of the assumptions in the theorems.

Estimating Functions and Generalized Estimating Equations

The idea of an estimating function is quite old; a simple instance is in the method of moments. A systematic study of estimating functions and their efficiency was begun independently by Godambe (1960) and Durbin (1960). Small and Wang (2003) provide a summary of estimating functions and their applications. Estimating functions also play a prominent role in quasi-likelihood methods, see Heyde (1997). We will discuss this further in Chapter 6.

Unbiasedness

The concept of unbiasedness in point estimation goes back to Gauss in the early nineteenth century, who wrote of fitted points with no "systematic error". Although nowadays unbiasedness is most often encountered in the context of point estimation, the term "unbiased" was actually first used by statisticians to refer to tests (Neyman and Pearson, 1936, cited in Lehmann (1951)), then used to refer to confidence sets (Neyman, 1937, cited in Lehmann (1951)), and later introduced to refer to point estimators (David and Neyman, 1938, cited in Lehmann (1951)). See Halmos (1946) and Lehmann (1951) for general discussions, and see page 296 for unbiased tests and page 300 for unbiased confidence sets. The idea of unbiasedness of an estimating function was introduced by Kendall (1951).

In a decision-theoretic framework, L-unbiasedness provides an underlying unifying concept.

Equivariant and Invariant Statistical Procedures

Equivariant and invariant statistical models have a heuristic appeal in applications. The basic ideas of invariance and the implications for statistical inference are covered in some detail in the lectures of Eaton (1989).

Approximations and Asymptotic Inference

In many cases of interest we cannot work out the distribution of a particular statistic. There are two ways that we can proceed. One is to use computer simulation to *estimate* properties of our statistic. This approach is called *computational inference* (see Gentle (2009)). The other approach is to make some approximations, either of the underlying assumptions or for the unknown distribution.

Some approximations are just based on known similarities between two distributions. The most common kind of approximation, however, is based on the asymptotic or "large-sample" properties of the statistic. This approach of asymptotic inference, as discussed in Section 3.8, is generally quite straightforward and so it has widespread usage.

It is often difficult to know how the asymptotic properties relate to the properties for any given finite sample. The books by Barndorff-Nielson and Cox (1994), DasGupta (2008), Jiang (2010), Lehmann (1999), Serfling (1980), and van der Vaart (1998) provide extensive coverage of asymptotic inference.

Variance Estimation

A sandwich-type estimator was introduced introduced by Eiker (1963) for estimation of the variance-covariance matrix of the least-squares estimator of the coefficient vector in linear regression in the case where the errors are uncorrelated, but possibly have different distributions. Huber (1967) used a similar kind of estimator as a robust estimator. White (1980) introduced a similar estimator for heteroscedastic situations in economics. The term "sandwich estimator" was introduced in the context of estimation of the variance-covariance matrix for the solution of a generalized estimation equation, and it is widely used in that type of problem.

Subsampling and Resampling

The idea of the jackknife goes back to Quenouille in 1949. The ordinary standard (delete-1) jackknife was popularized by John Tukey for both bias correction and variance estimation. (Tukey, of course, gave it the poetic name.) It is currently widely-used, especially in sample surveys. The delete-d ($d > 1$) jackknife was introduced and studied by Wu (1986). Shao and Tu (1995) provide an extensive discussion of the jackknife.

The theory and methods of the bootstrap were largely developed by Efron, and Efron and Tibshirani (1993) introduce the principles and discuss many extensions and applications.

Predictive Inference and Algorithmic Statistical Models

Geisser (1993) argues that predictive inference is a more natural problem that the ordinary objective in statistical inference of identifying a subfamily

of distributions that characterizes the data-generating process that gave rise to the observed data as well as future observations from that process.

Breiman (2001) emphasizes the role of algorithmic models when the objective is prediction instead of a simple descriptive model with a primary aim of aiding understanding.

Exercises

3.1. Show that the estimator (3.20) is Pitman-closer to μ than is $\overline{X}$.

3.2. a) Suppose $T_1(X)$ and $T_2(X)$ have continuous symmetric PDFs $p_1(t-\theta)$ and $p_2(t-\theta)$ (that is, their distributions are both location families). Suppose further that $p_1(0) > p_2(0)$. Show that for some $\epsilon > 0$

$$\Pr(|T_1 - \theta| < \epsilon) > \Pr(|T_2 - \theta| < \epsilon).$$

b) Is T_1 in question 3.2a Pitman-closer to θ than T_2? Tell why or why not.

c) Now suppose X_1 and X_2 are iid with a continuous symmetric PDF p. Let $T_1(X) = X_1$ and $T_2(X) = (X_1 + X_2)/2$. Show that if

$$2\int (p(x))^2 \mathrm{d}x < p(0),$$

then for some ϵ

$$\Pr(|T_1 - \theta| < \epsilon) > \Pr(|T_2 - \theta| < \epsilon).$$

3.3. a) Prove that if $T(X)$ is a sufficient statistic for θ and if $Y(X)$ is distributed independently of $T(X)$, then the distribution of Y does not depend on θ.

b) Does a converse hold? State and prove, or show why not.

3.4. Let T be a sufficient statistic for P. Prove the statements made on page 222 about functions and sufficient statistics; specifically,

a) show by example that $W = f(T)$ for some function f is not necessarily a sufficient statistic for P; however

b) if $T = g(S)$, where g is a measurable function and S is a statistic, then S is sufficient for P.

3.5. Show that $T_1(X)$ and $T_2(X)$ in Example 3.6 on page 227 are both sufficient and complete for θ.

3.6. Show that $T(X)$ in Example 3.7 on page 227 is sufficient and complete for θ.

3.7. Work out the information matrix for $\theta = (\mu, \sigma)$ in the $\mathrm{N}(\mu, \sigma^2)$ family using

a) the expectation of the product of first derivatives with respect to the parameters

b) the expectation of the second derivatives with respect to the parameters
c) the integrals of the derivatives with respect to the variable (which is an expectation).

3.8. Determine the Hessian of the log-likelihood in Example 3.13, and show that at the stationary point it is negative definite. (The off-diagonals are 0.)

3.9. Suppose $T(X)$ has finite first and second moments, and let $g(\theta) = \mathrm{E}(T(X))$. Show that for the estimator $\widetilde{T}(X) = aT(X) + b$ when $a < 0$,

$$R(g(\theta), \widetilde{T}) > R(g(\theta), 0);$$

that is, $\widetilde{T}(X)$ is inadmissible for any b because it is dominated by the estimator $T(X) \equiv 0$.

3.10. What can you say about a point estimator that is admissible under an absolute-error loss? (Compare Theorem 3.10.)

3.11. Show that the expression (3.166) is greater than or equal to $\widehat{\mathrm{V}(T)}_{\mathrm{J}}$.

3.12. Assume a random sample of size $n > 2$ from a distribution with PDF of the form

$$p(x;\theta) = \frac{f(x)}{h(\theta)}\mathrm{I}_{(0,\,\theta)}(x).$$

a) Show that $X_{(n)}$ is not unbiased for θ.
b) Show that $T = 2X_{(n)} - X_{(n-1)}$ is unbiased to order $\mathrm{O}(n^{-1})$.
c) Show that the asymptotic risk with squared error loss for T and $X_{(n)}$ are the same to $\mathrm{O}(n^{-2})$.

3.13. Prove Theorem 3.15.

4

Bayesian Inference

We have used an urn process to illustrate several aspects of probability and sampling. An urn that contains balls of different colors can be used to illustrate a primitive notion of probability – "What is the probability of drawing a red ball?" – that can be integrated into our axiomatic development of probability (as a set measure). Almost 250 years ago Pierre-Simon Laplace, the French mathematician and astronomer, considered the urn problem and asked a very different question: "Given that there are n balls in the urn, some of which are red, if the first ball drawn is red, what is the probability that the proportion of red balls, P, is p_0 (some constant)?" While this form of question may be quite natural to a layman, it is not consistent with our notion of probability. There is a fixed number of red balls in the urn; the proportion P is either p_0 or it is not.

Even if we adhere to our definitions of "probability", we should be able to rephrase this question into one for which statistical decision theory should provide an answer. We might feel more comfortable, however, using different words, and maybe even asking about "subjective probability" or "belief" about the proportion of red balls in the urn. Laplace went on to answer the question in a manner that we will identify later as a systematic approach to such problems:

$$\begin{aligned}\Pr(P = p_0|\text{first ball red}) &= \frac{p_0/(n-2)}{\sum_{p=2/n}^{(n-1)/n} p/(n-2)} \\ &= \frac{p_0}{n(n-1)/2-1}.\end{aligned}$$

Another question that Laplace addressed concerned the probability π that a human birth would be male. From the point of view that this is a random process, the word "probability" in this context is consistent with our understanding of the word. Laplace, however, went on to pose the question, "What is the probability that the probability π is less than or equal to one half?" Whether he felt it was relevant or not, he did not remark on the differences

in the meanings of "probability" in the question. Laplace's answer to this question, based on the observed numbers of male and female births in Paris during a period in the second half of the eighteenth century, is the same as we would get using Bayes theorem with a uniform prior on π. We will later consider other examples of this genre, so we will not elaborate on this one here.

These two examples have two main points: how the word "probability" is used, and the statistical framework in the analysis. The setup and the analysis used in these problems are of the type called Bayesian inference.

In this chapter, we will consider the general framework of Bayesian inference. In most of the discussion in this chapter we will assume a parametric model; that is, we assume that the observable random variable of interest has a distribution in the family $\mathcal{P} = \{P_\Theta : \Theta \in \Theta \subseteq \mathbb{R}^k\}$.

4.1 The Bayesian Paradigm

The field of Bayesian statistics has become a mainstream part of statistical inference. Bayesian methods allow us easily to incorporate prior information into the inference. One of the major ways in which Bayesian methods differ from other statistical methods, however, is in the basic definition of the problem. In the standard paradigm of parametric statistical inference, as expressed in equations (3.1) and (3.2) of Chapter 3, the objective of the inference is to make a decision about the values of the parameter. In Bayesian statistics, the parameter, that is, the index of the underlying distribution, is viewed as a random variable, so the canonical problem in Bayesian statistics is somewhat different.

Although we now have two random variables, Θ and X, we must not confuse them. Realizations of the random variable X are observable; realizations of Θ are not directly observable. Realizations of Θ determine aspects of the distribution of X.

We still address the fundamental problem in statistics: beginning with a given distributional family $\mathcal{P} = \{P_\Theta : \Theta \in \Theta\}$, we use observable realizations of X to make inferences about how Θ ranges over Θ. Instead of the formulation of the problem in terms of equations (3.1) and (3.2), a formulation in terms of equations (3.3) and (3.4) on page 207 may be more appropriate. We begin with

$$\mathcal{P} = \{P_\Theta \mid \Theta \sim Q_0 \in \mathcal{Q}\}, \tag{4.1}$$

where Θ is a random variable and Q_0 is a "prior distribution". Using observed data from a distribution P_Θ that depends on Θ and relations among joint, marginal, and conditional distributions, we arrive at the class of populations

$$\mathcal{P}_H = \{P_\Theta \mid \Theta \sim Q_H \in \mathcal{Q}\}, \tag{4.2}$$

where Q_H is some "posterior distribution" conditional on the observations.

For many families of distributions of the observable random variable, there are corresponding families of prior distributions that yield a parametric family of posterior distributions that is the same as the family of priors. This means that $\mathcal{Q}$ in equations (4.1) and (4.2) represents a single parametric family. We call a member of such a family of priors a *conjugate prior* with respect to the conditional distributional family of the observables. Clearly, we can always define a family of conjugate priors. A trivial example is when $\mathcal{Q}$ is the family of all distributions on Θ. The concept of conjugate priors, however, becomes relevant when the family $\mathcal{Q}$ is fairly restrictive, especially as a parametric family.

In the sense that Q_0 allows direct incorporation of prior beliefs or subjective evidence, statistical inference following this paradigm is sometimes referred to as subjective inference.

In Bayesian inference, as usual, we assume that P_Θ, Q_0, and Q_H are dominated by some σ-finite measure or by a mixture of σ-finite measures.

Notation

As we proceed, even in the simplest cases, there are five CDFs and PDFs that are involved in this approach. Rather than introduce specialized notation, we will use the same simple and consistent notation for these that we have used previously. The notation makes use of upper and lower case fonts of the same letter together with subscripts as symbols for the CDFs and PDFs of particular random variables. Thus, for example, "$f_{X|\theta}$" represents the PDF of the observable random variable whose distribution is conditional on the random variable Θ, and "$f_{\Theta|x}$" represents the PDF of the unobservable random parameter whose distribution is conditional on the observations, that is, the "posterior PDF". The corresponding CDFs would be represented as "$F_{X|\theta}$" and "$F_{\Theta|x}$". The visual clutter of the subscripts is an overhead that is more than compensated for by the consistency of the notation.

Occasionally, especially in the context of computations as in Section 4.7, we will use a specialized and simple notation: $[X, Y]$, $[X|Y]$, and $[X]$ represent the joint, conditional, and marginal densities, respectively.

Bayesian Inference

The inference in going from the family in equations (4.1) to the family in equations (4.2) is just based on our probability models. Once the models are in place, this inference does not depend on any loss function. Bayesian inference for making decisions, as for the specific questions raised in the two examples from Laplace above, however, is generally set in a decision theory framework.

In the decision-theoretic approach to statistical inference, we first quantify the loss relative to the true state of nature in any decision that we make. There is an obvious gap between the theory and the practice. It is easy to write some

loss function as we discussed on page 260 and the following pages, but whether these correspond to reality is another question.

Once we are willing to accept a particular loss function as a quantification of our reality, we still have a problem because the loss function involves the true state of nature, which we do not know. As we discussed in Section 3.3, we use an expectation of the loss function, under some less restrictive assumptions about the true state of nature.

Our objective is to develop methods of statistical inference that minimize our losses or expected losses. Some of the issues in defining this goal more precisely were discussed in Section 3.3.2. A more fundamental question arises as to how to define the expected loss.

Functionals of the Loss Function

In Section 3.3, given an assumed family of distributions of the observable $\{P_\theta\}$, we defined the risk,

$$R(P_\theta, T) = \mathrm{E}_{P_\theta}\big(L(P_\theta, T)\big), \tag{4.3}$$

and took this as the basis of the decision-theoretic approach.

The risk is a function of two elements, T and P_θ or just θ in a parametric setting. For a given T we often write the risk as $R_T(\theta)$. We generally wish to choose T so as minimize the risk in (4.3). The risk is the basis for defining admissibility, which from some perspectives is one of the most important properties of a statistical procedure.

Various considerations, as discussed on page 266 and the following pages, led us to use a distribution with PDF $\mathrm{d}F_\Theta(\theta)$ to average the risk, yielding

$$r(F_\Theta, T) = \int_\Theta R(P_\theta, T)\mathrm{d}F_\Theta(\theta) \tag{4.4}$$

$$= \int_\Theta \int_{\mathcal{X}} L(\theta, T(x))\mathrm{d}F_{X|\theta}(x)\mathrm{d}F_\Theta(\theta), \tag{4.5}$$

which is called the *Bayes risk*. (The term "Bayes risk" is sometimes used to refer to the minimum of $r(F_\Theta, T(X))$ with respect to T.) We denote the Bayes risk in various ways, $r(F_\Theta, T(X))$, $r_T(F_\Theta)$, and so on. While the risk is a function of θ or of the distribution family P_θ, the Bayes risk is a function of the distribution F_Θ.

The expectation in (4.3) is taken with respect to the distribution $[X|\Theta]$, and thus is a "conditional risk". Because it is an expected value with respect to the observable random variable, it is a "frequentist risk".

Because admissibility is defined in terms of this risk, "Inadmissibility is a frequentist concept, and hence its relevance to a Bayesian can be debated" (Berger (1985), page 257). We, however, will continue to consider the admissibility of statistical procedures, and even consider it to be a relevant fact that most Bayes procedures are admissible (Theorem 4.3).

A quantity that is of more interest in the Bayesian paradigm is a different functional of the loss function. It is an expectation taken with respect to the distribution of the parameter, $[\Theta]$ or $[\Theta|X]$. This is called the *Bayesian expected loss*. The Bayesian expected loss is defined "at the time of decision making". If no statistical analysis is involved, but rather the Bayesian expected loss is merely a probabilistic construct, it is

$$\rho(F_\Theta, T) = \int_\Theta L(\theta, T) \mathrm{d}F_\Theta(\theta).$$

In Bayesian statistical analysis, the expected loss is defined in terms of the conditional distribution:

$$\rho(F_\Theta, T) = \int_\Theta L(\theta, T) \mathrm{d}F_{\Theta|x}(\theta). \tag{4.6}$$

In summary, given a loss function $L(\theta, T)$, where T is a function of a random variable X, a conditional distribution $[X|\Theta]$ and a marginal distribution $[\Theta]$, we have the three quantities that are functionals of L.

- Risk:
$$R_T(\theta) = \mathrm{E}_{X|\theta}(L(\theta, T)). \tag{4.7}$$
- Bayes risk:
$$r_T(\Theta) = \mathrm{E}_\Theta\left(\mathrm{E}_{X|\theta}(L(\theta, T))\right) \tag{4.8}$$
$$= \mathrm{E}_X\left(\mathrm{E}_{\Theta|X}(L(\theta, T))\right). \tag{4.9}$$
- Bayes expected loss:
$$\rho_T(\Theta) = \mathrm{E}_{\Theta|X}(L(\theta, T)). \tag{4.10}$$

Any expectation taken wrt $[X|\Theta]$ or $[X]$ is a "frequentist" concept.

Bayes Actions

If $F_\Theta(\theta)$ in equation (4.5) is a CDF, the rule or action that minimizes the conditional expectation with respect to the distribution with that CDF is called the *Bayes action* or the *Bayes rule*. We often denote the Bayes rule wrt F_Θ as $\delta_{F_\Theta}(X)$.

The risk that is achieved by the Bayes rule, that is, the minimum average risk, is

$$\int_\Theta R(\theta, \delta_{F_\Theta}(X)) \mathrm{d}F_\Theta(\theta).$$

(As noted above, sometimes this minimum value is called the Bayes risk.)

The averaging function allows various interpretations, and it allows the flexibility of incorporating prior knowledge or beliefs. The regions over which

$\mathrm{d}F_\Theta(\theta)$ is large will be given more weight; therefore an estimator will be pulled toward those regions.

In formal Bayes procedures, $\mathrm{d}F_\Theta(\theta)$ is normalized so that $\int_\Theta \mathrm{d}F_\Theta(\theta) = 1$, we call $\mathrm{d}F_\Theta(\theta)$ the *prior probability density* for θ. The prior distribution of Θ may also depend on parameters, which are called "hyperparameters".

In an exercise in simple probabilistic reasoning without any data, we may set up a quantity similar to $r(F_\Theta, T)$ in expression (4.5). In that case instead of an expected value $R(P_\theta, T)$, we have the loss function $L(P_\theta, a)$ for a specific action a. Using this setup is not statistical inference, of course, but it may be useful in analyzing expected losses under an assumed probability model F_Θ. This situation is referred to as a "no-data problem".

To continue with the statistical inference, we next form the joint distribution of Θ and X, and then the conditional distribution of Θ given X, called the *posterior distribution*. The Bayes rule is determined by minimizing the risk, where the expectation is taken with respect to the posterior distribution. Because the Bayes rule is determined by the posterior distribution, the Bayes rule must be a function of a sufficient statistic.

In some cases, we wish to determine a rule similar to a Bayes rule that minimizes the average risk in equation (4.5) even though $F_\Theta(\theta)$ is not a CDF and $\mathrm{d}F_\Theta(\theta)$ cannot be normalized to integrate to 1. If the integral in equation (4.5) exists, we can proceed to determine an action that minimizes it without actually determining a posterior distribution. In that case, we say that the prior distribution is an *improper prior*; and of course $\mathrm{d}F_\Theta(\theta)$ is not a PDF, but it serves the same purpose as a PDF in the sense that it is a prior weighting function. We will consider this situation on page 345 and give an example on page 359.

Probability Statements in Statistical Inference

Some methods of statistical inference are based on probabilities of a statistic taking on certain values given a specific member of a family of probability distributions; that is, perhaps, given a value of a parameter. The two main statistical methods that rely on statements of probability are hypothesis testing and determining confidence sets. In these methods we assume a model P_Θ for the state of nature and then consider probabilities of the form $\Pr(T(X) = 1|\Theta = \theta)$ or $\Pr(T(X) \ni \Theta|\Theta = \theta)$. The proper interpretation of a confidence set, for example, is "[... given the assumptions, etc. ...] the probability that a random region formed in this manner includes true the value of the parameter is ..."

This kind of probability statement is somewhat awkward for use in interpreting the results of a statistical analysis.

Instead of a statement about $\Pr(\delta(X)|\theta)$, many people would prefer a statement about $\Pr(\Theta \in T(X)|X = x)$, that is,

$$\Pr(\Theta \in T(x))$$

even if they don't think of Θ as a random variable. In the Bayesian approach to testing and determining confidence sets, we do think of the parameter as a random variable and so we can make statements about the probability of the parameter taking on certain values.

If the parameter is a random variable, especially if it is a continuous random variable, point estimation of the parameter or a test of an hypothesis that a parameter takes a specific value when the parameter is modeled as a continuous random variable does not make much sense. The idea of a point estimator that formally minimizes the Bayes risk, however, remains viable. Going beyond point estimation, the Bayesian paradigm provides a solid theoretical infrastructure for other aspects of statistical inference, such as confidence intervals and tests of hypotheses. The parameter random variable is different in a fundamental way from the other random variable in the estimation problem: the parameter random variable is not observable; the other random variable is — that is, we can observe and record realizations of this random variable of interest, and those observations constitute the sample, which is the basis for the statistical inference.

4.2 Bayesian Analysis

The starting point in ordinary Bayesian inference is the conditional distribution of the observable random variable. (In a frequentist approach, this is just the distribution — not the "conditional" distribution.)

The prior density represents a probability distribution of the parameter assumed a priori, that is, without the information provided by a random sample. Bayesian inference depends on the conditional distribution of the parameter, given data from the random variable of interest.

4.2.1 Theoretical Underpinnings

The relationships among the conditional, marginal, and joint distributions can be stated formally in the "Bayes formula". The simple relationship of probabilities of events as in equations (1.230) and (1.231) allows us to express a conditional probability in terms of the two marginal probabilities and the conditional probability with the conditioning reversed;

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}. \tag{4.11}$$

This relationship expresses the basic approach in Bayesian statistical analysis. Instead of probabilities of discrete events, however, we wish to utilize relationships among probability densities.

We consider the random variable X with range $\mathcal{X} \subseteq \mathbb{R}^d$ and Θ with range $\Theta \subseteq \mathbb{R}^k$. We consider the product space $\mathcal{X} \times \Theta$ together with the product

σ-field $\sigma(\mathcal{B}(\mathcal{X}) \times \mathcal{B}(\Theta))$ where $\mathcal{B}(\mathcal{X})$ and $\mathcal{B}(\Theta)$ are the Borel σ-fields over the respective ranges. We will consider the family of distributions $\mathcal{P}$ with probability measures dominated by a σ-finite measure ν, and characterized by the PDFs $f_{X|\theta}(x)$ with $\theta \in \Theta$.

Theorem 4.1 (Bayes theorem)
Assume the density, $f_{X|\theta}(x) = \mathrm{d}F_{X|\theta}(x)/\mathrm{d}\nu$ is Borel on the product measure space, $(\mathcal{X} \times \Theta, \sigma(\mathcal{B}(\mathcal{X}) \times \mathcal{B}(\Theta))$. Let F_Θ be a (prior) CDF on Θ. We have that the posterior distribution $F_{\Theta|x}$ is dominated by the probability measure associated with F_Θ, and if $f_X(x) = \int_\Theta f_{X|\theta}(x)\mathrm{d}F_\Theta > 0$ a.e. ν,

$$\frac{\mathrm{d}F_{\Theta|x}}{\mathrm{d}F_\Theta} = \frac{f_{X|\theta}(x)}{f_X(x)}. \tag{4.12}$$

Proof. We first show that $f_X(x) < \infty$ a.e., by directly integrating it using Fubini's theorem:

$$\begin{aligned}
\int_{\mathcal{X}} f_X(x)\mathrm{d}\nu &= \int_{\mathcal{X}} \int_\Theta f_{X|\theta}(x)\mathrm{d}F_\Theta \mathrm{d}\nu \\
&= \int_\Theta \int_{\mathcal{X}} f_{X|\theta}(x)\mathrm{d}\nu \mathrm{d}F_\Theta \\
&= \int_\Theta \mathrm{d}F_\Theta \\
&= 1.
\end{aligned}$$

Thus, $f_X(x)$ is a proper PDF and is finite a.e. ν.

Now for $x \in \mathcal{X}$ and $B \in \mathcal{B}(\Theta)$, let

$$P(B, x) = \frac{1}{f_X(x)} \int_B f_{X|\theta}(x)\mathrm{d}F_\Theta.$$

Because $P(B, x) \geq 0$ for all B and $P(\Theta, x) = 1$, $P(B, x) > 0$ is a probability measure on $\mathcal{B}(\Theta)$.

Furthermore, by Fubini's theorem

$$\int_\Theta \int_{\mathcal{X}} P(B, x)\mathrm{d}\nu \mathrm{d}F_\Theta = \int_{\mathcal{X}} \int_\Theta P(B, x)\mathrm{d}F_\Theta \mathrm{d}\nu$$

and so $P(B, x)$ is measurable ν.

Now, for any $A \in \sigma(X)$ and $B \in \mathcal{B}(\Theta)$, again using Fubini's theorem, we have

$$\begin{aligned}
\int_{A\times\Theta} \mathrm{I}_B(\theta)\mathrm{d}F_{X,\Theta} &= \int_A \int_B f_{X|\theta}(x)\mathrm{d}F_\Theta \mathrm{d}\nu \\
&= \int_A \left(\int_B \frac{f_{X|\theta}(x)}{f_X(x)} \mathrm{d}F_\Theta \right) \left(\int_\Theta f_X(x)\mathrm{d}F_\Theta \right) \mathrm{d}\nu \\
&= \int_\Theta \int_A \left(\int_B \frac{f_{X|\theta}(x)}{f_X(x)} \mathrm{d}F_\Theta \right) f_{X|\theta}(x)\mathrm{d}\nu \mathrm{d}F_\Theta \\
&= \int_{A\times\Theta} P(B, x)\mathrm{d}F_{X,\Theta}.
\end{aligned}$$

Hence $P(B, x) = \Pr(\Theta \in B | X = x)$; that is, it is dominated by the probability measure associated with F_Θ and it is the resulting conditional probability measure of Θ given $X = x$. Its associated CDF is $F_{\Theta|x}$. ∎

Furthermore, if λ is a σ-finite measure that dominates the probability measure associated with F_Θ (in a slight abuse of the notation, we write $F_\Theta \ll \lambda$), and $f_\Theta(\theta) = \frac{\mathrm{d}F_\Theta}{\mathrm{d}\lambda}$, then the chain rule applied to expression (4.12) yields

$$\frac{\mathrm{d}F_{\Theta|x}}{\mathrm{d}\lambda} = \frac{f_{X|\theta}(x) f_\Theta(\theta)}{f_X(x)}. \tag{4.13}$$

The Consequences of Exchangeability

In Example 3.12 on page 237, we encountered a data-generating process with an underlying Bernoulli distribution that presented us with a quandary. The analysis required us to use knowledge of the data-generating process (specifically, the stopping rule). An alternative approach using only the assumption that the Bernoulli observations are exchangeable allows us to ignore the stopping rule. This approach is based on de Finetti's representation theorem (Theorem 1.30 on page 75).

We now reconsider the problem discussed in Example 3.12.

Example 4.1 Sampling in a Bernoulli distribution
We assume an exchangeable sample of size n, $X_1, \ldots, X_n$, from the Bernoulli(π). Suppose that k of the observations in the sample have a value of 1, and the other $n - k$ have a value of 0. Given only this information, we ask what is the probability that a new observation X_{n+1} has a value of 1.

Let $\{X_i\}_{i=1}^\infty$ be an infinite sequence of binary random variables such that for any n, $\{X_i\}_{i=1}^n$ is exchangeable. Then there is a unique probability measure P on $[0, 1]$ such that for each fixed sequence of zeros and ones $\{e_i\}_{i=1}^n$,

$$\Pr(X_1 = e_1, \ldots, X_n = e_n) = \int_0^1 \pi^k (1 - \pi)^{n-k} \mathrm{d}\mu(\pi),$$

where $k = \sum_{i=1}^n e_i$.

In Example 3.12, we considered a variation of this problem in which the sample size n was random. If we have an exchangeable sequence $X_1, X_2, \ldots$ and we choose a finite value of n to form a set $X_1, \ldots, X_n$, we may ask whether the set is exchangeable. The sample is exchangeable so long as the chosen value of n has nothing to do with the X_is. Suppose, however, the value of n is chosen conditionally such that $X_n = 1$. The sample is no longer exchangeable, and the argument above no longer holds. The Bayesian analysis remains the same, however, as we see in Example 4.3. ∎

We will consider other aspects of this problem again in Example 6.1 on page 447 and in Example 7.12 on page 539.

In most cases, a random sample is a set of iid random variables; in the Bayesian framework, we only assume that the $X_1, \ldots, X_n$ are exchangeable and that they are conditionally independent given θ.

4.2.2 Regularity Conditions for Bayesian Analyses

Many interesting properties of statistical procedures depend on common sets of assumptions, for example, the Fisher information regularity conditions. For some properties of a Bayesian procedure, or of the posterior distribution itself, there is a standard set of regularity conditions, often called the Walker regularity conditions, after Walker (1969), who assumed them in the proofs of various asymptotic properties of the posterior distribution. The regularity conditions apply to the parameter space Θ, the prior PDF $f_\Theta(\theta)$, the conditional PDF of the observables $f_{X|\theta}(x)$, and to the support $\mathcal{X} = \{x : f_{X|\theta}(x) > 0\}$. All elements are real, and μ is Lebesgue measure.

The Walker regularity conditions are grouped into three sets:

A1. Θ is closed.
When a general family of distributions is assumed for the observables, this condition may allow for a distribution that is not in that family (for example, in a Bernoulli(π), the parameter is not allowed to take the values 0 and 1), but the convenience of this condition in certain situations more than pays for this anomaly, which occurs with 0 probability anyway.

A2. $\mathcal{X}$ does not depend on θ.
This is also one of the FI regularity conditions.

A3. For $\theta_1 \neq \theta_2 \in \Theta$, $\mu(\{x : f_{X|\theta}(x) \neq f_{X|\theta}(x)\}) > 0$.
This is identifiability; see equation (1.25). Without it parametric inference does not make much sense.

A4. Given $x \in \mathcal{X}$ and $\theta_1 \in \Theta$ and δ a sufficiently small real positive number, then $\forall \theta \ni \|\theta - \theta_1\| < \delta$,

$$|\log(f_{X|\theta}(x)) - \log(f_{X|\theta_1}(x))| < H_\delta(x, \theta_1)$$

where $H_\delta(x, \theta_1)$ is a measurable function of x and θ_1 such that

$$\lim_{\delta \to 0_+} H_\delta(x, \theta_1) = 0$$

and, $\forall \tilde{\theta} \in \Theta$,

$$\lim_{\delta \to 0_+} \left(\int_{\mathcal{X}} H_\delta(x, \theta_1) f_{X|\tilde{\theta}}(x) \mathrm{d}x \right) = 0.$$

This is a continuity condition.

A5. If Θ is not bounded, then for any $\tilde{\theta} \in \Theta$ and a sufficiently large real number Δ,

$$\|\theta\| > \Delta \implies \log(f_{X|\theta}(x)) - \log(f_{X|\tilde{\theta}}(x)) < K_\Delta(x, \tilde{\theta}),$$

where $K_\Delta(x, \tilde{\theta})$ is a measurable function of x and $\tilde{\theta}$ such that

$$\lim_{\delta \to 0_+} \left(\int_{\mathcal{X}} K_\Delta(x, \tilde{\theta}) f_{X|\tilde{\theta}}(x) \mathrm{d}x \right) < 0.$$

B1. $\forall\theta_0 \in \Theta^\circ$, $\log(f_{X|\theta}(x))$ is twice differentiable wrt θ in some neighborhood of θ_0.

B2. $\forall\theta_0 \in \Theta^\circ$,

$$0 \prec \int_{\mathcal{X}} \left(\left.\frac{\partial \log(f_{X|\theta}(x))}{\partial\theta}\right|_{\theta=\theta_0} \right) \left(\left.\frac{\partial \log(f_{X|\theta}(x))}{\partial\theta}\right|_{\theta=\theta_0} \right)^{\mathrm{T}} f_{X|\theta_0}(x)\mathrm{d}x < \infty.$$

B3. $\forall\theta_0 \in \Theta^\circ$,

$$\int_{\mathcal{X}} \left.\frac{\partial f_{X|\theta}(x)}{\partial\theta}\right|_{\theta=\theta_0} \mathrm{d}x = 0$$

and

$$\int_{\mathcal{X}} \left.\frac{\partial^2 f_{X|\theta}(x))}{\partial\theta(\partial\theta)^{\mathrm{T}}}\right|_{\theta=\theta_0} \mathrm{d}x = 0.$$

(Note that in the first condition, the integrand may be a vector, and in the second, it may be a matrix.)

B4. Given $\theta_0 \in \Theta^\circ$ and δ a sufficiently small real positive number, $\forall\theta \in \Theta \ni \|\theta - \theta_0\| < \delta$ and for each i and j,

$$\left| \frac{\partial^2 \log(f_{X|\theta}(x))}{\partial\theta_i\partial\theta_j} - \left.\frac{\partial^2 \log(f_{X|\theta}(x))}{\partial\theta_i\partial\theta_j}\right|_{\theta=\theta_0} \right| < M_\delta(x,\theta_0)$$

where $M_\delta(x,\theta_0)$ is a measurable function of x and θ_0 such that

$$\lim_{\delta\to 0_+} \left(\int_{\mathcal{X}} M_\delta(x,\theta_0) f_{X|\theta_0}(x)\mathrm{d}x \right) < 0.$$

This is also a continuity condition.

C1. $\forall\theta_0 \in \Theta^\circ$, $f_\Theta(\theta)$ is continuous at θ_0 and $f_\Theta(\theta_0) > 0$.

4.2.3 Steps in a Bayesian Analysis

We can summarize the approach in a Bayesian statistical analysis as beginning with these steps:

1. Identify the conditional distribution of the observable random variable; assuming the density exists, call it

$$f_{X|\theta}(x). \tag{4.14}$$

   This is the PDF of the distribution Q_0 in equation (4.1).
2. Identify the prior (marginal) distribution of the parameter; assuming the density exists, call it

$$f_\Theta(\theta). \tag{4.15}$$

   This density may have parameters also, of course. Those parameters are called "hyperparameters", as we have mentioned.

3. Identify the joint distribution; if densities exist, it is

$$f_{X,\Theta}(x,\theta) = f_{X|\theta}(x) f_\Theta(\theta). \tag{4.16}$$

4. Determine the marginal distribution of the observable; if densities exist, it is

$$f_X(x) = \int_\Theta f_{X,\Theta}(x,\theta) \mathrm{d}\theta. \tag{4.17}$$

   This marginal distribution is also called the *prior predictive distribution*. In slightly different notation, we can also write it as

$$f_X(x) = \int_\Theta f_{X|\theta}(x) f_\Theta(\theta) \mathrm{d}\theta.$$

5. Determine the posterior conditional distribution of the parameter given the observable random variable. If densities exist, it is

$$f_{\Theta|x}(\theta) = f_{X,\Theta}(x,\theta)/f_X(x). \tag{4.18}$$

   This is the PDF of the distribution Q_H in equation (4.2), which is often just called the "posterior". The posterior conditional distribution is then the basis for whatever decisions are to be made.
6. Assess the posterior conditional distribution in the light of prior beliefs. This is called a *sensitivity analysis*. Repeat the steps above as appropriate.

These first steps in a Bayesian analysis involve identifying the components in the equation

$$f_{X|\theta}(x) f_\Theta(\theta) \;=\; f_{X,\Theta}(x,\theta) \;=\; f_{\Theta|x}(\theta) f_X(x). \tag{4.19}$$

Although we have written the PDFs above in terms of single random variables (any of which of course could be vectors), in applications we assume we have multiple observations on X. In place of $f_{X|\theta}(x|\theta)$ in (4.14), for example, we would have the joint density of the iid random variables $X_1, \ldots, X_n$, or $\prod f_{X|\theta}(x_i|\theta)$. The other PDFs would be similarly modified.

Given a posterior based on the random sample $X_1, \ldots, X_n$, we can form a *posterior predictive distribution* for $X_{n+1}, \ldots, X_{n+k}$:

$$\begin{aligned} f_{X_{n+1},\ldots,X_{n+k}|x_1,\ldots,x_n}(x_{n+1},\ldots,x_{n+k}) = \\ \int_\Theta f_{X,\Theta}(x_{n+i},\theta) \mathrm{d}F_{\Theta|x_1,\ldots,x_n}(\theta). \end{aligned} \tag{4.20}$$

Rather than determining the densities in equations (4.14) through (4.18) it is generally sufficient to determine kernel functions. That is, we write the densities as

$$f_D(z) \propto g(z). \tag{4.21}$$

This means that we can avoid computation of the normalizing constant, or partition functions. This is especially important for the densities $f_{\Theta|x}$ and f_X, where in some cases these involve integrals that are difficult to evaluate.

There are other shortcuts we can take in forming all of the relevant expressions for a Bayesian analysis. In the next few examples we will go through all of the gory details. Before considering specific examples, however, let's consider some relationships among the various densities.

Conjugate Priors

If the conditional posterior distribution for a given conditional distribution of the observable is in the same family of distributions as the marginal prior distribution, we say that the prior distribution is a *conjugate prior* with respect to the family of distributions of the observable. In equation (4.19), for example, $f_\Theta(\theta)$ is a conjugate prior for $f_{X|\theta}(x)$ if $f_{\Theta|x}(\theta)$ is in the same family of distributions as $f_\Theta(\theta)$.

Conjugate priors often have attractive properties for a Bayesian analysis, as we will see in the examples.

For a PDF in the exponential class, written in the form of equation (2.7),

$$f_{X|\theta}(x) = \exp\left((\eta(\theta))^{\mathrm{T}} T(x) - \xi(\theta)\right) h(x), \tag{4.22}$$

the general form of a conjugate prior is

$$f_\Theta(\theta) = c(\mu, \Sigma) \exp\left(|\Sigma|(\eta(\theta))^{\mathrm{T}} \mu - |\Sigma|\xi(\theta)\right), \tag{4.23}$$

where $c(\mu, \Sigma)$ is a constant with respect to the hyperparameters μ and Σ, which can be thought of as a mean and variance-covariance (Exercise 4.2).

In Table 4.1, I show some conjugate prior distributions for various single-parameter distributions of observables. See Appendix A for meanings of the parameters. In the table, a parameter with a subscript of 0, for example, θ_0 is assumed to be known (that is, not a parameter).

I assume a sample $x_1, \ldots, x_n$, and I use t to represent $T(x) = \sum x_i$.

Examples

Example 4.2 Binomial with Beta Prior

The Bayesian approach can be represented nicely by a problem in which we model the conditional distribution of an observable random variable X as a binomial(n, π) distribution, conditional on π, of course. (Recall from Example 4.1 that if we wish to view the binomial as a sum of Bernoullis, the Bernoullis must at least be exchangeable.)

Suppose we assume that π comes from a beta(α, β) prior distribution; that is, we consider a random variable Π that has beta distribution.

We work out the density functions in the following order:

The conditional distribution of X given π has density (probability function)

Table 4.1. Univariate Conjugate Prior Distributions

observable	prior	posterior
Bernoulli(π)	beta(α, β)	beta($\alpha + t,\ \beta + n - t$)
geometric(π)	beta(α, β)	beta($\alpha + n,\ \beta - n + t$)
Poisson(θ)	gamma(α, β)	gamma($\alpha + t,\ \beta/(n\beta + 1)$)
normal(μ, σ_0^2)	normal(ν, τ^2)	normal$\left(\frac{\nu\sigma_0^2+\tau^2 t}{\sigma_0^2+n\tau^2},\ \frac{\tau^2\sigma_0^2}{\sigma_0^2+n\tau^2}\right)$
normal($\mu_0, 1/\theta$)	inverted gamma(α, β)	inverted gamma $\left(\alpha + n/2,\ \left(1/\beta + \frac{1}{2}\sum(x_i - \mu_0)^2\right)^{-1}\right)$
uniform(θ_0, θ)	Pareto(α, γ)	Pareto($\alpha + n,\ \max(\gamma, x_1, \ldots, x_n)$)
exponential(θ)	inverted gamma(α, β)	inverted gamma $\left(\alpha + n,\ (1/\beta + t)^{-1}\right)$

$$f_{X|\pi}(x) = \binom{n}{x} \pi^x (1-\pi)^{n-x} \mathrm{I}_{\{0,1,\ldots,n\}}(x). \tag{4.24}$$

The marginal (prior) distribution of Π has density

$$f_\Pi(\pi) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \pi^{\alpha-1}(1-\pi)^{\beta-1} \mathrm{I}_{]0,1[}(\pi). \tag{4.25}$$

Suppose the hyperparameters in the beta prior are taken to be $\alpha = 3$ and $\beta = 5$. The prior, that is, the marginal distribution of Π, is as shown in the upper left of Figure 4.1.

How one might decide that $\alpha = 3$ and $\beta = 5$ are appropriate may depend on some prior beliefs or knowledge about the general range of π in the binomial distribution in this particular setting. We will consider this issue in Section 4.2.5.

The joint distribution of X and Π has density

$$f_{X,\Pi}(x,\pi) = \binom{n}{x} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \pi^{x+\alpha-1}(1-\pi)^{n-x+\beta-1} \mathrm{I}_{\{0,1,\ldots,n\}\times]0,1[}(x,\pi). \tag{4.26}$$

Integrating out π, we get the marginal distribution of X to be beta-binomial, with density

$$f_X(x) = \binom{n}{x} \frac{\Gamma(\alpha+\beta)\Gamma(x+\alpha)\Gamma(n-x+\beta)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(n+\alpha+\beta)} \mathrm{I}_{\{0,1,\ldots,n\}}(x). \tag{4.27}$$

Finally, the conditional distribution of Π given x (the posterior) has density,

$$f_{\Pi|x}(\pi) = \frac{\Gamma(n+\alpha+\beta)}{\Gamma(x+\alpha)\Gamma(n-x+\beta)} \pi^{x+\alpha-1}(1-\pi)^{n-x+\beta-1} \mathrm{I}_{]0,1[}(\pi). \tag{4.28}$$

We note that this is a beta distribution; hence, the beta is a conjugate prior for the binomial. If the parameters of the beta prior are α and β, given one observation x, the posterior is a beta with parameters $x + \alpha$ and $n - x + \beta$.

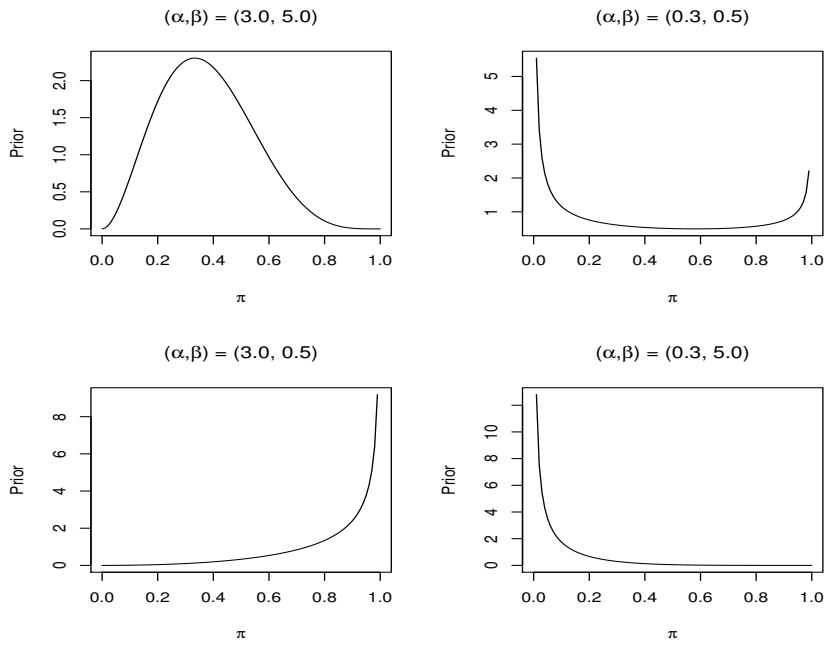


Figure 4.1. Priors with Different Values of α and β

Finally, we need to assess the posterior conditional distribution in the light of prior beliefs. Consider the possible effects of one observation when n is 10. Suppose first that we observe $x = 2$. With the beta(3,5) prior, we get the posterior conditional distribution of Π, as a beta with parameters $x + \alpha = 5$ and $n - x + \beta = 13$. Secondly, suppose that we observe $x = 6$. With the beta(3,5) prior, we get the posterior conditional distribution of Π, as a beta with parameters $x + \alpha = 9$ and $n - x + \beta = 9$. The posterior densities are shown in top panel of Figure 4.2. Compare them with the prior density for beta(3,5) in Figure 4.1.

Now, consider the beta(3,0.5) prior, and first suppose first that we observe $x = 2$. The posterior conditional distribution of Π, is a beta with parameters $x + \alpha = 5$ and $n - x + \beta = 8.5$. Secondly, suppose that we observe $x = 6$, and so with the beta(3,0.5) prior, we get the posterior conditional distribution of Π, as a beta with parameters $x + \alpha = 9$ and $n - x + \beta = 4.5$. The posterior densities are shown in lower panel of Figure 4.2. Compare them with the prior density for beta(3,0.5) in Figure 4.1. We can assess the possible posterior conditional distributions in the light of prior beliefs, and how we might expect to modify those prior beliefs after observing specific values of x. This is called a *sensitivity analysis*.

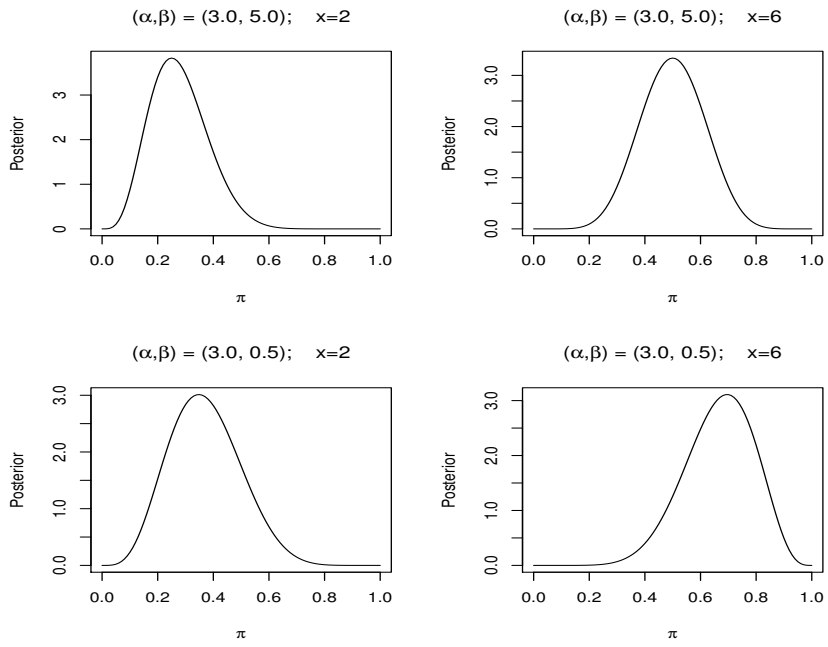


Figure 4.2. Posteriors Resulting from Two Different Priors (Upper and Lower Panels) after Observing $x = 2$ or $x = 6$ (Left and Right Panels)

The posterior distribution of Π represents everything that we know about this parameter that controls the distribution of X, and so, in a sense, our statistical inference is complete. We may, however, wish to make more traditional inferences about the parameter. For example, we may wish to estimate the parameter, test hypotheses concerning it, or determine confidence sets for it. We will return to this problem in Examples 4.6, 4.15, and 4.18.

Example 4.3 (Continuation of Examples 3.12 and 4.1) Sampling in a Bernoulli distribution; Negative binomial with a beta prior
Again consider the problem of statistical inference about π in the family of Bernoulli distributions. We have discussed two data-generating processes: one, take a random sample of size n, $X_1, \ldots, X_n$ from the Bernoulli(π) and count the number of 1's; and two, take a sequential sample, $X_1, X_2, \ldots$, until a fixed number t of 1's have occurred. The first process leads to a binomial distribution and the second leads to a negative binomial distribution. In Example 4.2, we considered a Bayesian approach to inference on π using a binomial distribution with a beta prior. Now consider inference on π using a negative binomial distribution again with a beta prior.

We go through the same steps as in equations (4.24) through (4.28). In place of the conditional distribution of the observable binomial variable X, whose density was given in equation (4.24), we have the conditional distribution of the observable negative binomial variable N, whose density is

$$p_N(n\,;\,t,\pi) = \binom{n-1}{t-1}\pi^t(1-\pi)^{n-t}, \quad n = t, t+1, \ldots. \tag{4.29}$$

(Recall that this is one form of the negative binomial distribution probability function.)

Again starting with a beta(α, β) prior distribution on the random variable Π, and going through the standard steps of forming the joint distribution and the marginal of the observable N, we arrive at the conditional distribution of Π, given $N = n$. The PDF is

$$f_{\Pi|x}(\pi) = \frac{\Gamma(n+\alpha+\beta)}{\Gamma(x+\alpha)\Gamma(n-x+\beta)}\pi^{x+\alpha-1}(1-\pi)^{n-x+\beta-1}\mathrm{I}_{]0,1[}(\pi). \tag{4.30}$$

This is the same posterior distribution as in Example 4.2, and since this represents everything that we know about this parameter that controls the distribution of N, if the data from the two different experiments are the same, as in Example 3.12, then any inference about Π would be the same. Thus, the Bayesian approach conforms to the likelihood principle. ∎

In the next two examples we consider inference in the context of a normal distribution. The first example is very simple because we assume only one unknown parameter.

Example 4.4 Normal with Known Variance and a Normal Prior on the Mean

Suppose we assume the observable random variable X has a $\mathrm{N}(\mu, \sigma^2)$ distribution in which σ^2 is known.

The parameter of interest, μ, is assumed to be a realization of an unobservable random variable $M \in \mathbb{R}$.

Let us use a prior distribution for M that is $\mathrm{N}(\mu_0, \sigma_0^2)$, where the hyperparameters are chosen to reflect prior information or beliefs about the mean of X.

Let us assume that we have one observation on $X = x$. We easily go through the standard steps. First, we get the joint PDF,

$$f_{X,M}(x,\mu) = \frac{1}{2\pi\sigma\sigma_0}\mathrm{e}^{-\frac{1}{2}(x-\mu)^2/\sigma^2 - \frac{1}{2}(\mu-\mu_0)^2/\sigma_0^2}.$$

To get the marginal PDF, we expand the quadratics in the exponent and collect terms in μ, which we want to integrate out. This is a standard operation (see page 684), but nevertheless it is tedious and we'll write it out this once:

$$\begin{aligned}\frac{(x-\mu)^2}{\sigma^2}+\frac{(\mu-\mu_0)^2}{\sigma_0^2} &= \frac{x^2}{\sigma^2}+\frac{\mu_0^2}{\sigma_0^2}+\frac{\sigma^2+\sigma_0^2}{\sigma^2\sigma_0^2}\mu^2-2\frac{\sigma^2\mu_0+\sigma_0^2x}{\sigma^2\sigma_0^2}\mu \qquad (4.31)\\ &= \frac{x^2}{\sigma^2}+\frac{\mu_0^2}{\sigma_0^2}+\left(\mu^2-2\frac{\sigma^2\mu_0+\sigma_0^2x}{\sigma^2+\sigma_0^2}\mu\right)/\left(\frac{\sigma^2\sigma_0^2}{\sigma^2+\sigma_0^2}\right)\\ &= \frac{x^2}{\sigma^2}+\frac{\mu_0^2}{\sigma_0^2}-\frac{(\sigma^2\mu_0+\sigma_0^2x)^2}{\sigma^2\sigma_0^2(\sigma^2+\sigma_0^2)}\\ &\quad+\left(\mu-\frac{\sigma^2\mu_0+\sigma_0^2x}{\sigma^2+\sigma_0^2}\right)^2/\left(\frac{\sigma^2\sigma_0^2}{\sigma^2+\sigma_0^2}\right)\end{aligned}$$

The last quadratic in the expression above corresponds to the exponential in a normal distribution with a variance of $\sigma^2\sigma_0^2/(\sigma^2+\sigma_0^2)$, so we adjust the joint PDF so we can integrate out the μ, leaving

$$f_X(x)=\frac{1}{\sqrt{2\pi}\sigma\sigma_0}\sqrt{\frac{\sigma^2+\sigma_0^2}{\sigma^2\sigma_0^2}}\exp\left(-\frac{1}{2}\left(\frac{x^2}{\sigma^2}+\frac{\mu_0^2}{\sigma_0^2}-\frac{(\sigma^2\mu_0+\sigma_0^2x)^2}{\sigma^2\sigma_0^2(\sigma^2+\sigma_0^2)}\right)\right).$$

Combining the exponential in this expression with (4.31), we get the exponential in the conditional posterior PDF, again ignoring the $-1/2$ while factoring out $\sigma^2\sigma_0^2/(\sigma^2+\sigma_0^2)$, as

$$\left(\mu^2-2\frac{\sigma^2\mu_0+\sigma_0^2x}{\sigma^2+\sigma_0^2}\mu+\frac{(\sigma^2\mu_0+\sigma_0^2x)^2}{(\sigma^2+\sigma_0^2)^2}\right)/\frac{\sigma^2\sigma_0^2}{\sigma^2+\sigma_0^2}$$

Finally, we get the conditional posterior PDF,

$$f_{M|x}(\mu)=\frac{1}{\sqrt{2\pi}}\sqrt{\frac{\sigma^2\sigma_0^2}{\sigma^2+\sigma_0^2}}\exp\left(-\frac{1}{2}\left(\mu-\frac{\sigma^2\mu_0+\sigma_0^2x}{\sigma^2+\sigma_0^2}\right)^2/\frac{\sigma^2\sigma_0^2}{\sigma^2+\sigma_0^2}\right),$$

and so we see that the posterior is a normal distribution with a mean that is a weighted average of the prior mean and the observation x, and a variance that is likewise a weighted average of the prior variance and the known variance of the observable X. ∎

This example, although quite simple, indicates that there can be many tedious manipulations. It also illustrates why it is easier to work with PDFs without normalizing constants.

We will now consider a more interesting example, in which neither μ nor σ^2 is known. We will also assume that we have multiple observations.

Example 4.5 Normal with Inverted Chi-Squared and Conditional Normal Priors

Suppose we assume the observable random variable X has a $\mathrm{N}(\mu,\sigma^2)$, and we wish to make inferences on μ and σ^2. Let us assume that μ is a realization of an unobservable random variable $M\in\mathbb{R}$ and σ^2 is a realization of an unobservable random variable $\Sigma^2\in\mathbb{R}_+$.

We construct a prior family by first defining a marginal prior on Σ^2 and then a conditional prior on $M|\sigma^2$. From consideration of the case of known variance, we choose an inverted chi-squared distribution for the prior on Σ^2:

$$f_{\Sigma^2}(\sigma^2) \propto \frac{1}{\sigma}\sigma^{-(\nu_0/2+1)}\mathrm{e}^{(\nu_0\sigma_0^2)/(2\sigma^2)}$$

where we identify the parameters ν_0 and σ_0^2 as the degrees of freedom and the scale for Σ^2.

Given σ^2, let us choose a normal distribution for the conditional prior of $M|\sigma^2$. It is convenient to express the variance of $M|\sigma^2$ as a scaling of σ^2. Let this variance be σ^2/κ_0. Now combining the prior of Σ^2 with this conditional prior of $M|\sigma^2$, we have the joint prior PDF

$$f_{M,\Sigma^2}(\mu,\sigma^2) \propto \frac{1}{\sigma}(\sigma^2)^{-(\nu_0/2+1)}\exp\left(-\frac{1}{2\sigma^2}\left(\nu_0\sigma_0^2+\kappa_0(\mu-\mu_0)^2\right)\right).$$

We assume a sample $X_1,\ldots,X_n$, with the standard statistics $\overline{X}$ and S^2. We next form the joint density of (X,M,Σ^2), then the marginal of X, and finally the joint posterior of $(M,\Sigma^2|x)$. This latter is

$$\begin{aligned} f_{M,\Sigma^2|x}(\mu,\sigma^2;x) \propto \frac{1}{\sigma}(\sigma^2)^{-(\nu_0/2+1)}\exp\left(-\frac{1}{2\sigma^2}\left(\nu_0\sigma_0^2+\kappa_0(\mu-\mu_0)^2\right)\right) \\ \times(\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\left((n-1)s^2+n(\bar{x}-\mu)^2\right)\right) \end{aligned} \tag{4.32}$$

We would also like to get the conditional marginal posterior of Σ^2 given x. Then, corresponding to our conditional prior of M given σ^2, we would like to get the conditional posterior of M given σ^2 and x. This involves much tedious algebra, completing squares and rearranging terms, but once the expressions are simplified, we find that

$$M|\sigma^2,x \sim \mathrm{N}\left(\frac{\kappa_0\mu_0+n\bar{x}}{\kappa_0+n},\frac{\sigma^2}{\kappa_0+n}\right) \tag{4.33}$$

and

$$\frac{(\nu_0\sigma_0^2+(n-1)s^2+n\kappa_0(\bar{x}-\mu_0)^2/(\kappa_0+n))/2}{\Sigma^2}\,|x \sim \chi^2(\nu_0+n). \tag{4.34}$$

∎

Sufficient Statistics and the Posterior Distribution

Suppose that in the conditional distribution of the observable X for given θ, there is a sufficient statistic T for θ. In that case, the PDF in equation (4.14) can be written as

$$f_{X|\theta}(x) = g(t|\theta)h(x), \tag{4.35}$$

where $t = T(x)$.

Then, following the steps on page 335 leading up to the posterior PDF in equation (4.18), we have the posterior

$$f_{\Theta|x}(\theta) \propto p(t, \theta), \tag{4.36}$$

that is, the posterior distribution depends on X only through T. Notice that the analysis in Example 4.5 was performed using the sufficient statistics $\overline{X}$ and S^2.

All Bayesian inference, therefore, can be based on sufficient statistics.

Use of the Posterior Distribution

The conditional posterior distribution of the parameter contains all of the relevant information about the parameter, based on any prior information (or beliefs) incorporated in the prior distribution, as well as the information contained in the observations under the assumption that their distribution is known conditional on the value of the parameter following the model used in the analysis. Hence, in Example 4.2, the posterior beta distribution of Π tells us everything we might want to know about that parameter, given our prior assumptions and the data observed.

There are different ways we might interpret the analysis. Under the view that the parameter is a random variable, a narrow interpretation of the analysis outlined above is that it addresses the specific value of the random variable that was operative at the time that the data were observed. This interpretation emphasizes the changing nature of the phenomenon being studied. In any specific situation, a given realization of the random parameter governs a data-generating process in that specific instance. A less ephemeral interpretation of the analysis is that the analysis provides a more general inference about the *distribution* of the random parameter.

4.2.4 Bayesian Inference

Although as we pointed out above, once we have the conditional posterior distribution for the parameter, we have all of the information about the parameter. We may, however, wish to make more traditional inferences about the random parameter; that is, we may want to estimate it, test hypotheses concerning it, or set confidence sets for it. Since the parameter is a random variable, the meaning of such inferences requires some interpretation. One simple interpretation is that the inference is about the specific value of the parameter when the data were observed.

We can base the inference on simple heuristics relating to the posterior distribution. For example, in a manner similar to the heuristic that leads to a maximum likelihood estimator, we may consider the mode of the posterior

as the most likely value. The mode is sometimes called the *maximum a posterior probability (MAP)* point estimator. Similar heuristics lead to a type of confidence set based on the probabilities of various regions in the support of the posterior distribution (see Sections 4.6.1 and 4.6.2).

Bayes Actions

While we can arrive at the conditional posterior distribution for the parameter without any reference to a loss function, in order to make specific inferences about the parameter in a formal Bayesian decision-theoretic approach, we need to define a loss function, then formulate the conditional risk for a given inference procedure $\delta(X)$,

$$R(F_{X|\theta}, \delta(X)) = \int_{\mathcal{X}} L(\theta, T(x)) \mathrm{d}F_{X|\theta}(x), \tag{4.37}$$

and finally determine the posterior average risk,

$$r(F_{\Theta}, \delta(X)) = \int_{\Theta} R(F_{X|\theta}, \delta(X)) \mathrm{d}F_{\Theta}(\theta). \tag{4.38}$$

The procedure that minimizes the average risk is called the *Bayes action*. If the action is estimation, the procedure is called the *Bayes estimator*. Although we view the MAP estimator from a heuristic standpoint, it can be shown to be a limit of Bayes estimators under the 0-1 loss function.

We will discuss Bayesian estimation in Section 4.3, Bayesian hypothesis testing in Section 4.5, and Bayesian confidence intervals in Section 4.6. We will continue considering the problem of inferences on π in the binomial distribution that we first addressed in Example 4.2. In that example, there was no loss function and we stopped with the posterior PDF. In Examples 4.6 (page 355), 4.15 (page 365), and 4.18 (page 374), we will consider Bayes actions relating to π.

Generalized Bayes Actions

We often seek a Bayes action even though we do not want to go through any formal steps of identifying a posterior distribution. In some cases, the prior may not actually be a PDF (or proportional to a PDF); that is, the integral of the prior may not be finite. The prior is said to be *improper*. If the prior is improper, the posterior may or may not be improper. If the posterior is not a PDF or proportional to a PDF, then we must be careful in any interpretation we may make of the posterior. We may, however, be able to identify a rule or action in the usual way that we determine a Bayes action when the posterior is proper.

A Bayes action is one that minimizes the risk in equation (4.5) if the weighting function $\mathrm{d}F_{\Theta}(\theta)$ is a PDF. If $\mathrm{d}F_{\Theta}(\theta)$ is not a PDF, that is, if the prior is improper, so long as the integral

$$r(F_\Theta, T) = \int_\Theta \int_{\mathcal{X}} L(\theta, T(x)) \mathrm{d}F_{X|\theta}(x) \mathrm{d}F_\Theta(\theta). \tag{4.39}$$

exists, we call a rule that minimizes it a *generalized Bayes action*.

In Example 4.9 on page 359, we determine the Bayes estimator of the mean in a normal distribution when the prior is uniform over the real line, which obviously is an improper prior.

Limits of Bayes Actions

Another variation on Bayes actions is the limit of the Bayes action as some hyperparameters approach fixed limits. In this case, we have a sequence of prior PDFs, $\{F_\Theta^{(k)} : k = 1, 2, \ldots\}$, and consider the sequence of Bayes actions, $\delta^{(k)}(X)$ that result from these priors. The limit $\lim_{k\to\infty} \delta^{(k)}(X)$ is called a *limiting Bayes action*. The limiting prior, $\lim_{k\to\infty} F_\Theta^{(k)}$, may not be a PDF.

In Example 4.6 on page 356, we determine a limiting Bayes action, given in equation (4.47).

4.2.5 Choosing Prior Distributions

It is important to choose a reasonable prior distribution that reflects prior information or beliefs about the phenomenon of interest.

Families of Prior Distributions

Various families of prior distributions can provide both flexibility in representing prior beliefs and computational simplicity. Conjugate priors, as in Examples 4.2 through 4.5, are often very easy to work with and to interpret. In many cases, a family of conjugate priors would seem to range over most reasonable possibilities, as shown in Figure 4.1 for priors of the binomial parameter.

Generalized distributions or mixtures of common distributions, as discussed in Section 2.10, may correspond to prior beliefs. The ideas of choosing a distribution that matches general assumed properties such as the shape of the distribution or that corresponds to fixed quantiles discussed in Section 2.10.4 may also lead to reasonable prior distributions.

Within a given family of prior distributions, it may be useful to consider ones that are optimal in some way. For example, in testing composite hypotheses we may seek a "worst case" for rejecting or accepting the hypotheses. This leads to consideration of a "least favorable prior distribution". We may also wish to use a prior that reflects an almost complete lack of prior information.. This leads to consideration of "noninformative priors", or priors with maximum entropy within a given class.

If the priors are restricted to a particular class of distributions, say Γ, we may seek an action whose worst risk with respect to any prior in Γ is minimized, that is, we may see an action δ that yields

$$\inf_{\delta} \sup_{P \in \Gamma} r(P, \delta). f_{\Theta}(\theta) \propto |I(\theta)|^{1/2}. \tag{4.40}$$

Such an action is said to be Γ-minimax, usually written as *gamma-minimax*. Clearly, any minimax Bayes action is gamma-minimax Bayes with respect to the same loss function.

Assessing the Problem Formulation

In any statistical analysis, the formulation of a model is important. In Example 4.2 above, we must consider whether or not it is reasonable to assume that the observable data follows some kind of binomial distribution. From first principles, this means that we are willing to assume that there is a set of n independent outcomes that may be 0 or 1, in each case with a constant probability π.

In a Bayesian analysis, not only must we think about whether or not the assumed distribution of the observable is reasonable, we must also consider the assumed prior distribution. Various possibilities for a beta prior for the binomial distribution in Example 4.2 are shown in Figure 4.1. If we know something about the data-generating process, we may conclude that some general shape of the prior is more appropriate than another. Often a scientist who may not know much about statistics, but is familiar with the the data-generating process, will have some general beliefs about what values π is more likely to have. The scientist, for example, may be able to state that there is a "50% chance that π is between 0.2 and 0.6". In that case, the hyperparameters of a beta could be determined that would yield that probability. The process of developing a prior by discussions with a subject matter expert is called "elicitation". As mentioned above, a generalized distribution that corresponds to reasonable quantiles of prior beliefs may be useful.

In the Bayesian approach taken in Example 4.2, we assume that while the n observations were being collected, some random variable Π had a fixed value of π. We are interested both in that value and in the conditional distribution of the random variable Π, given what we have observed. For particular choices of hyperparameters characterizing the prior distribution on Π, we obtain the posterior distributions shown in Figure 4.2. Do these seem reasonable?

In deciding whether the prior is appropriate, sometimes it is worthwhile to consider the effects of various possible outcomes of the experiment. The issue is whether the posterior conditional distribution conforms to how the observed data should change our prior beliefs.

This sensitivity analysis can be done without actually taking any observations because we can determine the posterior density that would result from the given prior density. In Figure 4.3, we plot the posterior distribution of

Π based on a beta$(3, 5)$ prior given various values of that we might have observed.

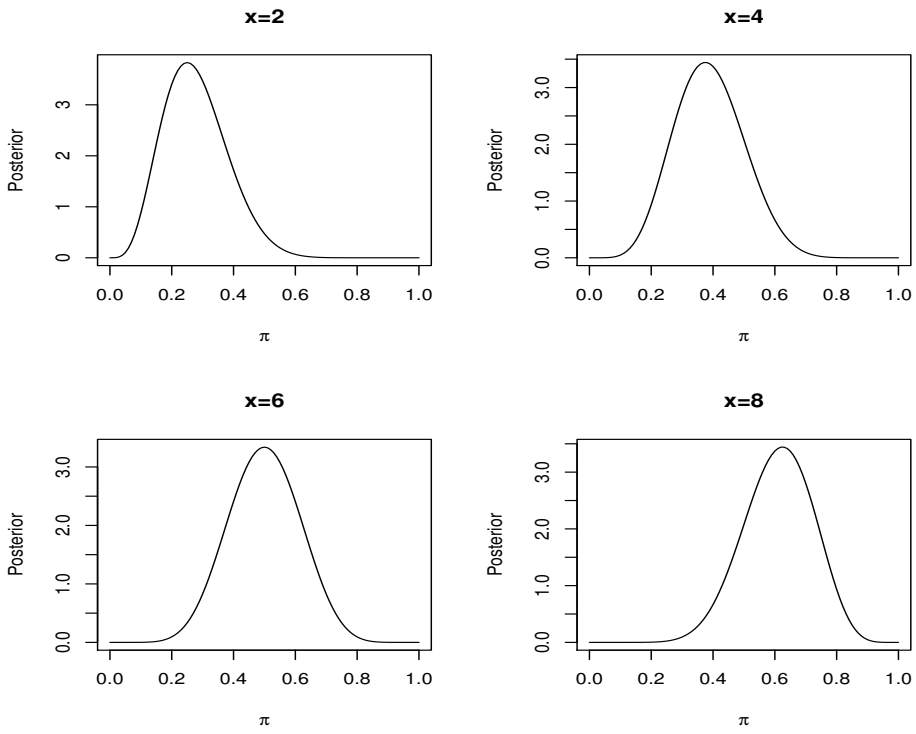


Figure 4.3. Posteriors Resulting from a Beta(3,5) Prior after Various Possible Observations

Assessing the effect on the posterior of various possible observations may give us some feeling of confidence in our choice of a prior distribution.

*** Bayesian robustness

Choice of Hyperparameters

Usually in a Bayesian analysis, it is instructive to consider various priors and particularly various hyperparameters in some detail.

Of course, in most cases, we must also take into account the loss function. Recall the effects in this problem of different hyperparameter values on the point estimation problem (that is, the choice of the Bayes action to minimize the posterior risk) when the loss is squared error.

We might also consider what might be the effect of different hyperparameters. There are several possibilities we could consider. Let's just look at one possibility, which happens to be bimodal, as shown in the upper right of Figure 4.1. In this case, we have chosen $\alpha = 0.3$ and $\beta = 0.5$. This would correspond to a general prior belief that π is probably either close to 0 or close to 1.

Now, again we might consider the effect of various observations on our belief about Π. We get the posteriors shown in Figure 4.4 for various possible values of the observations.

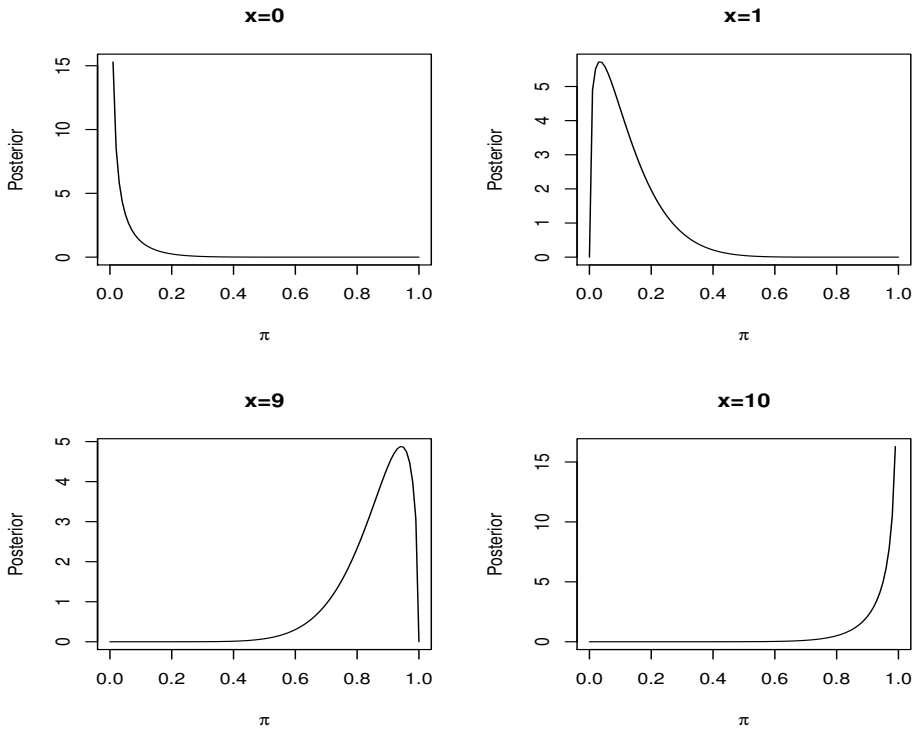


Figure 4.4. Posteriors from the Bimodal Beta(0.3,0.5) Prior

Compare the posteriors in Figure 4.4 with the prior in Figure 4.1. In each case in this example we see that our posterior belief is unimodal instead of bimodal, as our prior belief had been. Although in general a posterior may be multimodal, in the case of a binomial(n, π) distribution with a beta(α, β) prior, the posterior is unimodal, because as we have seen, the posterior is beta with parameters $x + \alpha$ and $n - x + \beta$, both of which cannot be less than 1.

Objective Priors

The extent to which the observed data updates the prior distribution into a posterior distribution depends to some extent on the nature of the prior distribution. Whenever there is little knowledge or only weak belief about the distribution of the parameter of interest, we may wish to define a prior that will allow the "data to speak for themselves". We call such priors "objective", "vague", "flat", "diffuse", or "noninformative". (Although one or the other of these terms may be more appropriate in a given setting, there is no technical difference.) Such priors are often improper.

An example of a noninformative prior is one in which $\mathrm{d}F_\Theta(\theta) = \mathrm{d}\nu$ where ν is Lebesgue measure and Θ is the reals, or some unbounded interval subset of the reals. This is a noninformative prior, in the sense that it gives equal weights to equal-length intervals for Θ. Such a prior is obviously improper.

Another type of noninformative prior is Jeffreys's noninformative prior. This prior is proportional to $\sqrt{\det(I(\theta))}$, where $\det(I(\theta))$ or $|I(\theta)|$ is the determinant of the Fisher information matrix; that is,

$$f_\Theta(\theta) \propto |I(\theta)|^{1/2}. \tag{4.41}$$

The idea is that such a prior is invariant to the parametrization. We can see this by considering the reparametrization $\tilde{\theta} = \bar{g}(\theta)$, where $\bar{g}$ is a 1:1 differentiable function that maps the parameter space onto itself in such a way that the underlying conditional probability distribution of X is unchanged when X is transformed appropriately to $\widetilde{X}$ (see page 178 in Chapter 1). We see this by writing

$$\begin{aligned} I(\tilde{\theta}) &= -\mathrm{E}\left(\frac{\partial^2 \log f_{\widetilde{X}|\tilde{\theta}}(\widetilde{X})}{\partial^2 \tilde{\theta}}\right) \\ &= -\mathrm{E}\left(\frac{\partial^2 \log f_{X|\tilde{\theta}}(X)}{\partial^2 \theta}\left|\frac{\partial^2 \theta}{\partial \tilde{\theta}}\right|^2\right) \\ &= I(\theta)\left|\frac{\partial^2 \theta}{\partial \tilde{\theta}}\right|^2, \end{aligned}$$

and so

$$f_{\widetilde{\Theta}}(\tilde{\theta}) \propto |I(\theta)|^{1/2}, \tag{4.42}$$

where the additional constant of proportionality is the Jacobian of the inverse transformation.

If Θ is the reals, or some unbounded interval subset of the reals, Jeffreys's noninformative prior is improper. If the support of Θ is finite, Jeffreys's noninformative prior is generally proper; see equation (4.49) on page 357.

In a variation of Jeffreys's noninformative prior when $\theta = (\theta_1, \theta_2)$, where θ_2 is a nuisance parameter, we first define $f_{\Theta_2|\Theta_1}(\theta_2)$ as the Jeffreys's prior associated with $f_{X|\theta}$ where θ_1 is fixed and then using $f_{\Theta_2|\Theta_1}(\theta_2)$ derive the

marginal conditional distribution $f_{X|\theta_1}$ if it exists. Finally, we compute the prior on Θ_1 as the Jeffreys's prior using $f_{X|\theta_1}$. Such a prior, if it exists, that is, if $f_{X|\theta_1}$ exists, is called a *reference noninformative prior*.

Maximum Entropy

Entropy can be thought of as the inverse of "information" (see Section 1.1.5), hence large entropy provides less information in this heuristic sense. This provides another general type of "noninformative", or at least, "less informative" prior distribution.

For some family of distributions with finite variance whose support is $\mathbb{R}$, the larger the variance, the larger the entropy We may, however, seek a prior distribution that has maximum entropy for given mean and variance. Of all such distributions dominated by Lebesgue measure, the normal family attains this maximum. (Showing this is Exercise 1.88.)

The appropriate support of the prior, of course, depends on the nature of the distribution of the observables. For a binomial or negative binomial distribution conditional on the parameter π, the prior should have support $]0,1[$. A very nice class of priors for this problem, as we have seen, is the family of beta(α,β) distributions. The entropy of a beta(α,β) distribution is

$$\log\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right) - (\alpha-1)(\psi(\alpha)-\psi(\alpha+\beta)) - (\beta-1)(\psi(\beta)-\psi(\alpha)),$$

and its maximum occurs at $\alpha=\beta=1$ (see Exercise 1.32b).

Empirical Priors

The hyperparameters in a prior distribution are often based on estimates from prior(!) samples of similar data. This seems reasonable, of course, if the prior distribution is to reflect rational beliefs.

The use of data to provide values of hyperparameters can even occur with the same data that is to be used in the Bayesian inference about the conditional distribution, as we will discuss in Section 4.2.6 below.

Hierarchical Priors

In the basic Bayes setup of (4.1),

$$\mathcal{P} = \{P_\Theta \mid \Theta \sim Q_\Xi \in \mathcal{Q}\},$$

we can consider a Bayesian model for the priors,

$$\mathcal{Q} = \{Q_\Xi \mid \Xi \sim R_0 \in \mathcal{R}\}. \tag{4.43}$$

We follow the same analytic process as in Section 4.2.3, except that the distributions have another layer of conditioning. We begin with a distribution

of the observable random that is conditional on both the basic parameters and the hyperparameters, that is, we have the density $f_{X|\theta\xi}(x)$ instead of just $f_{X|\theta}(x)$; and the prior distribution is now considered a conditional distribution. The prior PDF is now $f_{\Theta|\xi}(\theta)$ instead of $f_{\Theta}(\theta)$. Now, after introducing a prior for Ξ with PDF $f_{\Xi}(\xi)$, we proceed to manipulate the densities to arrive at a posterior conditional distribution for the parameter of interest, Θ.

The computations associated with producing the posterior distribution in a hierarchical model are likely to be more difficult than those in a simpler Bayesian model.

4.2.6 Empirical Bayes Procedures

In the basic Bayes setup of (4.1),

$$\mathcal{P} = \{P_\Theta \mid \Theta \sim Q_\xi \in \mathcal{Q}\},$$

we might consider a "frequentist" model of the priors:

$$\mathcal{Q} = \{Q_\xi \mid \xi \in \Xi\}. \tag{4.44}$$

Now, we determine the marginal of X conditional on ξ (there's no θ); that is, instead of the PDF $f_X(x)$ in Section 4.2.3, we have the PDF $f_{X|\xi}(x)$. Given the data on X, we can estimate ξ using traditional statistical methods. Because we have the PDF, a commonly-used method of estimation of ξ is maximum likelihood.

One way of interpreting this setup is that the observable random variable X is a actually a set of random variables, and the k^{th} observation has PDF $f_{X_k|\theta_k}(x)$, where the θ_k is a realization of a random variable Θ with distribution that depends on ξ.

Most people who subscribe to the Bayesian paradigm for statistical analysis eschew empirical Bayes procedures.

4.3 Bayes Rules

To determine a Bayes action, we begin with the standard steps in Section 4.2.3. A loss function was not used in deriving the posterior distribution, but to get a Bayes action, we must use a loss function. Given the loss function, we focus on the posterior risk.

The Risk Function

**** modify this to be more similar to the introductory discussion of Bayes risk and expected loss

After getting the posterior conditional distribution of the parameter given the observable random variable, for a given loss function L, we determine the

estimator δ that minimizes the posterior risk, which is the expected value of the loss wrt the posterior distribution on the parameter:

$$\arg\min_{\delta} \int_{\Theta} L(\theta, \delta(x)) f_{\Theta|x}(\theta) \mathrm{d}\theta.$$

The action that minimizes the posterior risk the Bayes rule.

The Bayes rule is determined by

- the conditional distribution of the observable
- the prior distribution of the parameter
- the nature of the decision; for example, if the decision is an estimator, then the function of the parameter to be estimated, that is, the estimand
- the loss function

The expected loss with respect to the posterior distribution of the parameter is the objective to be minimized.

The Complete Class Theorem

One of the most important characteristics of Bayesian estimation is that all "good" estimators are either Bayes or limiting Bayes; that is, the class of Bayes and limiting Bayes estimators is a complete class of decision rules (see page 265).

Theorem 4.2 (admissible estimators are Bayes)
An admissible estimator is either Bayes or limiting Bayes.

Proof. ∎

4.3.1 Properties of Bayes Rules

For any loss function we have the following relations with admissibility and minimaxity. First, despite the fact that admissibility is not really relevant in the Bayesian paradigm (see page 328), if a Bayes rule is unique it is admissible.

Theorem 4.3 (admissibility of unique Bayes rule)
Suppose that $\delta(X)$ is a unique Bayes rule in a decision problem. Then $\delta(X)$ is admissible in that decision problem.

Proof. Suppose that $\tilde{\delta}(X)$ is a Bayes rule.**** ∎

Theorem 3.12 states that an admissible estimator with a constant risk is minimax with respect to the same loss function and distribution. The same statement is true for a Bayes estimator with a constant risk:

Theorem 4.4 (Bayes rule or limiting Bayes rule is minimax if it has constant risk)

A Bayes rule or limiting Bayes rule with a constant risk is minimax with respect to the same loss function and distribution.

Proof. ∎

***** prove these

4.3.2 Equivariant Bayes Rules

We discussed the general problem of invariance and equivariance of statistical procedures in Section 3.4. We now consider these concepts in the context of Bayesian inference.

We say that a prior distribution Q for Θ is *invariant* with respect to $\widetilde{\mathcal{G}}$ if for $\forall \tilde{g} \in \widetilde{\mathcal{G}}$, the distribution of $\tilde{g}(\Theta)$ is also Q.

That is, for all measurable B,

$$\mathrm{E}_Q(I_B(\tilde{g}(\Theta))) = \mathrm{E}_Q(I_B(\Theta))$$

We define a σ-field $\mathcal{L}$ over the set of functions in a group $\mathcal{G}$, and then for a measurable set of transformations B, we consider right compositions Bh (for $h \in \mathcal{G}$, this is $\{gh : g \in B\}$), and left compositions gB.

Definition 0.1.18

If $\lambda(Bh) = \lambda(B)$ for all $B \in \mathcal{L}$ and $h \in \mathcal{G}$, λ is said to be right Haar invariant, and if $\lambda(gB) = \lambda(B)$ for all $B \in \mathcal{L}$ and $h \in \mathcal{G}$, λ is said to be left Haar invariant.

**** Relevance: relation to Jeffrey's noninformative prior.

4.3.3 Bayes Estimators with Squared-Error Loss Functions

The Bayes estimator depends on the loss function as well as the prior distribution. As in many cases, if the loss function is squared-error, the optimal procedure has some useful properties. For Bayes estimators with squared-error loss functions, we have the following properties.

Theorem 4.5
Under squared-error loss, the Bayes estimator is the posterior mean; that is, the expected value of the estimand, where the expected value is taken wrt the posterior conditional distribution.

Proof. Exercise. ∎

Theorem 4.6
Squared-error loss and a conjugate prior yield Bayes estimators for $\mathrm{E}(X)$ *that are linear in* X.

Proof. Exercise. ∎

Theorem 4.7
Under squared-error loss, if T *is the Bayes estimator for* $g(\theta)$*, then* $aT + b$ *is the Bayes estimator for* $ag(\theta) + b$ *for constants* a *and* b.

Proof. Exercise. ∎

Lemma 4.8.1
If T *is a Bayes estimator under squared-error loss, and if* T *is unbiased, then the Bayes risk* $r_T(P_\Theta) = 0$.

Proof. For the Bayes estimator $T(X)$ with $\mathrm{E}(T(X)|\theta) = g(\theta)$, we have

$$\mathrm{E}(g(\theta)T(X)) = \mathrm{E}(g(\theta)\mathrm{E}(T(X)|\theta)) = \mathrm{E}((g(\theta))^2).$$

Alternatively,

$$\mathrm{E}(g(\theta)T(X)) = \mathrm{E}(T(X)\mathrm{E}(g(\theta)|X)) = \mathrm{E}((T(X))^2).$$

Then we have

$$\begin{aligned} r_T(P_\Theta) &= \mathrm{E}((T(X) - g(\theta)|X))^2) \\ &= \mathrm{E}((T(X))^2) + \mathrm{E}((g(\theta))^2) - 2\mathrm{E}(g(\theta)T(X)) \\ &= 0. \end{aligned}$$

∎

Hence, by the condition of equation (3.84) we have the following theorem.

Theorem 4.8
Suppose T is an unbiased estimator. Then T is not a Bayes estimator under squared-error loss.

Examples

There are two standard examples of Bayesian analyses that serve as models for Bayes estimation under squared-error loss. These examples, Example 4.6 and 4.8, should be in the student's bag of easy pieces. In both of these examples, the prior is in a parametric conjugate family.

In this section, we also consider estimation using an improper prior (Example 4.9).

Example 4.6 (Continuation of Example 4.2) Estimation of the Binomial Parameter with Beta Prior and a Squared-Error Loss
We return to the problem in which we model the conditional distribution of an observable random variable X as a binomial(n, π) distribution, conditional on π, of course. Suppose we assume π comes from a beta(α, β) prior distribution; that is, we consider a random variable Π that has beta distribution. We wish to estimate Π.

Let us choose the loss to be squared-error. In this case we know the risk is minimized by choosing the estimate as $\delta(x) = \mathrm{E}(\Pi|x)$, where the expectation is taken wrt the distribution with density $f_{\Pi|x}$.

We recognize the posterior conditional distribution as a beta$(x + \alpha, n - x + \beta)$, so we have the Bayes estimator for squared-error loss and beta prior

$$\frac{\alpha + X}{\alpha + \beta + n}. \tag{4.45}$$

We should study this estimator from various perspectives.

linear combination of expectations

First, we note that it is a weighted average of the mean of the prior and the standard UMVUE. (We discuss the UMVUE for this problem in Examples 5.1 and 5.5.)

$$\left(\frac{\alpha+\beta}{\alpha+\beta+n}\right)\frac{\alpha}{\alpha+\beta}+\left(\frac{n}{\alpha+\beta+n}\right)\frac{X}{n}. \tag{4.46}$$

This is a useful insight, but we should not suppose that all Bayes estimators work that way.

unbiasedness

We see that the Bayes estimator cannot be *unbiased* if $\alpha \neq 0$ or $\beta \neq 0$ in the prior beta distribution (see Theorem 4.8). If $\alpha = 0$ and $\beta = 0$, the prior is improper because the integral of the prior density above does not converge. We can, however, set up the risk in the form of equation (4.39), and minimize it without ever determining a posterior distribution. The solution, X/n, which happens to be the UMVUE, is a generalized Bayes estimator. Because

$$\lim_{\alpha\to 0_+,\beta\to 0_+}\frac{\alpha+X}{\alpha+\beta+n}=\frac{X}{n}, \tag{4.47}$$

and for $\alpha > 0$ and $\beta > 0$, the prior is proper, we see that the UMVUE is a *limit of Bayes estimators*.

admissibility

By Theorem 3.10 we know that the biased Bayes estimator (4.45) is not admissible under a squared-error loss. The limiting Bayes estimator (4.47) is admissible under squared-error loss, and thus we see an application of Theorem 4.2.

minimaxity

Could the Bayes estimator with this prior and squared-error loss function be minimax? Work out the risk (Exercise 4.10), and determine values of α and β such that it is constant. This will be minimax. The solution (to make it independent of π) is

$$\alpha=\beta=\sqrt{n}/2. \tag{4.48}$$

Notice what this does: it tends to push the estimator toward 1/2, which has a maximum loss of 1/2, that is, the minimum maximum loss possible. Recall the randomized minimax estimator $\delta_{1/(n+1)}(X)$, equation (3.108) in Example 3.21.

Jeffreys's noninformative prior

The Jeffreys's noninformative prior in this case is proportional to $\sqrt{I(\pi)}$; see equation (4.41). Because the binomial is a member of the exponential family, we know $I(\pi) = 1/\mathrm{V}(T)$, where $\mathrm{E}(T) = \pi$. So $I(\pi) = n/\pi(1-\pi)$. Jeffreys's prior is therefore beta$(1/2, 1/2)$. The Bayes estimator corresponding to this noninformative prior is

$$\frac{X+\frac{1}{2}}{n+1}. \tag{4.49}$$

This is often used as an estimator of π in situations where $X > 0$ is rare. An estimator of π as 0 may not be very reasonable.

equivariance

For the group invariant problem in which $g(X) = n - X$ and $\bar{g}(\pi) = 1 - \pi$, we see that the loss function is invariant if $g^*(T) = 1 - T$. In this case, the Bayes estimator is *equivariant* if the prior is symmetric, that is, if $\alpha = \beta$.

empirical Bayes

We can make an empirical Bayes model from this example, as discussed in Section 4.2.6. We consider the observable random variable to be one of a set, X_k, each with conditional distribution binomial(n, π_k), where the π_k are all distributed independently as beta(α, β). An empirical Bayes procedure involves estimating α and β, and then proceeding as before. Although any (reasonable) estimates of α and β would work, we generally use the MLEs. We get those by forming the conditional likelihood of x given α and β, and then maximizing to get $\hat{\alpha}$ and $\hat{\beta}$. (We do this numerically because it cannot be done in closed form. We get the conditional likelihood of x given α and β by first forming the joint of x and the π_k's, and integrating out the π_k's.) The empirical Bayes estimator for π_k is

$$\frac{\hat{\alpha}+X_k}{\hat{\alpha}+\hat{\beta}+n}. \tag{4.50}$$

hierarchical Bayes

If we put prior distributions on α and β, say gamma distributions with different parameters, we could form a hierarchical Bayes model and use iterative conditional simulated sampling to compute the estimates. (This type of approach is called Markov chain Monte Carlo, or specifically in this case, Gibbs sampling. We discuss this approach in general in Section 4.7, and Gibbs sampling specifically beginning on page 669.) We would do this by working out the *full conditionals*.

The squared-error loss function is a very simple and common loss function. (In fact, the student must be very careful to remember that many simple properties of statistical methods depend on this special loss function.) We will consider the estimation problem of Example 4.6 with other loss functions in Example 4.11 and in Exercise 4.9. ∎

The prior distribution in Example 4.6 is a conjugate prior (when it exists; that is, when $\alpha > 0$ and $\beta > 0$), because the posterior is in the same parametric family. A conjugate prior and a squared-error loss function always yield Bayes estimators for $\mathrm{E}(X)$ that are linear in X, as we see in this specific case. Other priors may not be as easy to work with.

Example 4.7 Estimation of the Negative Binomial Parameter with Beta Prior and a Squared-Error Loss

In Example 3.12 on page 237 we discussed two different data-generating process, one of which led to the (conditional) binomial distribution of Example 4.6. The other, related data-generating process led to a (conditional) negative binomial distribution for a random variable N, that corresponds to the parameter n in Example 4.6 This negative binomial distribution has the same parameter π and another parameter that corresponds to the observed x in Example 4.6.

Given a fixed value x, we have a random variable N whose conditional distribution given π, has the probability function

$$p_{N|\pi}(n) = \binom{n-1}{x-1}\pi^x(1-\pi)^{n-x}\mathrm{I}_{x,x+1,\ldots}(n).$$

Again assuming a beta(α, β) prior, with known α and β, and going through the usual steps, we obtain the conditional of the parameter given the data as

$$p_{\Pi|n}(\pi) = \frac{\Gamma(x+\alpha)\Gamma(n-x+\beta)}{\Gamma(n+\alpha+\beta)}\pi^{x+\alpha-1}(1-\pi)^{n-x+\beta-1}\mathrm{I}_{(0,1)}(\pi),$$

which is a beta distribution. As for the binomial, the beta is a conjugate prior for the negative binomial. Now, we want to estimate π under a squared error loss. We know that the Bayesian estimate under a squared error loss is the posterior mean. In this case, because the distribution is a beta, and we easily work out the mean. Hence, we have have the Bayes estimator,

$$\hat{\pi} = \mathrm{E}_{\Pi|n}(\pi) = \frac{\alpha+x}{\alpha+\beta+N}. \tag{4.51}$$

Notice that this is the same estimator as expression (4.45) for the binomial parameter; thus, the estimators would conform to the likelihood principle.

As in Example 4.6 with the estimator of the binomial parameter, we could consider whether we can choose specific values of the hyperparameters so as to yield Bayes estimators with various properties (see Exercise 4.12). We see, for example, that with the given loss and any beta prior, it is not possible to obtain even a generalized estimator that is unbiased. ∎

Example 4.8 (Continuation of Example 4.5) Estimation of the Normal Mean and Variance with Inverted Chi-Squared and Conditional Normal Priors and a Squared-Error Loss

For estimating both μ and σ^2 in $\mathrm{N}(\mu, \sigma^2)$, as in Example 4.5, a conjugate prior family can be constructed by first defining a marginal prior on σ^2 and then a conditional prior on $\mu|\sigma^2$.

For the estimators, we minimize the expected loss with respect to the joint posterior distribution given in equation (4.32). For a squared-error loss, this yields the posterior means as the estimators. ∎

Another way this problem may be approached is by reparametrizing the normal, and in place of σ^2, using $1/(2\tau)$.

We now consider use of an improper prior when estimating the mean of a normal distribution with known variance.

Example 4.9 (Continuation of Example 4.4) Use of an Improper Prior and a Squared-Error Loss for Estimation of the Normal Mean When the Variance Is Known

Suppose we assume the observable random variable X has a $\mathrm{N}(\mu, \sigma^2)$ distribution in which σ^2 is known, and the parameter of interest, μ, is assumed to be a realization of an unobservable random variable $M \in \mathbb{R}$.

Let us use a prior distribution for M that is uniform over $\mathbb{R}$. This is obviously an improper prior, and the measure $\mathrm{d}F_M(\mu)$ is just the Lebesgue measure. Let us assume that we have n observations $x_1, \ldots, x_n$. Instead of going through the standard steps to get a posterior PDF, we go directly to the problem of determining the Bayes estimator by minimizing the risk in equation (4.39), if that minimum exists. We have

$$\begin{aligned} r(F_M, T) &= \int_{\mathrm{M}} \int_{\mathcal{X}} L(\mu, T(x)) \mathrm{d}F_{X|\mu}(x) \mathrm{d}F_M(\mu) \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (\mu - T(x))^2 \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^n} \mathrm{e}^{-\sum(x_i-\mu)^2/2\sigma^2} \mathrm{d}x \, \mathrm{d}\mu. \end{aligned}$$

The questions are whether we can reverse the integrations and whether the integral with respect to $\mathrm{d}\mu$ is finite. The two questions are the same, and we see that the answer is affirmative because for fixed x,

$$\int_{-\infty}^{\infty} (\mu - a)^2 \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^n} \mathrm{e}^{-\sum(x_i-\mu)^2/2\sigma^2} \, \mathrm{d}\mu < \infty.$$

We determine the estimator that minimizes the Bayes risk by differentiating the expression above wrt a, setting the result to 0, and solving for a. Using the "Pythagorean Theorem" of statistics, equation (0.0.99), in the exponential, we get the minimizer as $T(x) = \bar{x}$.

This generalized Bayes estimator is the optimal estimator under various other criteria: it is the MLE (Example 3.13), it is the MREE under a convex and even location-scale invariant loss function (Example 3.23), and it is UMVU (Example 5.6). ∎

4.3.4 Bayes Estimation with Other Loss Functions

For certain loss functions, even in relatively simple settings, a Bayes estimator may not exist; see Example 4.10. In some cases, however, we may choose a particular loss function so as to obtain an unbiased estimator (recall Theorem 4.8 concerning the squared-error loss). In other cases, we may seek a loss function that yields a constant risk. This gives us an admissible estimator.

Example 4.10 Nonexistence of a Bayes Estimator
Suppose that $X \sim \mathrm{N}(\theta, 1)$ given $\Theta = \theta$ and let the prior for Θ be $\mathrm{N}(0, 1)$. Consider the loss function

$$L(T, \theta) = \begin{cases} 0 \text{ if } T \geq \theta \\ 1 \text{ if } T < \theta. \end{cases}$$

(That is, we want to be sure not to underestimate θ.) Now, consider the constant estimator $T_n = n$ (where n is the sample size). The risk function is $R(T_n, \theta) = \mathrm{I}_{]-\infty,\theta[}(n)$. Hence, the average risk is $\Pr(\Theta > n)$ where $\Theta \sim \mathrm{N}(0, 1)$. Now, consider any Bayes estimator δ, and let Φ be the CDF for $\mathrm{N}(0, 1)$. We have

$$\begin{aligned} 0 &\leq \inf_{\delta} \int R(\delta, \theta) \mathrm{d}\Phi(\theta) \\ &\leq \inf_{n} \int R(T_n, \theta) \mathrm{d}\Phi(\theta) \\ &= \inf_{n} \Pr(\Theta > n) \\ &= 0. \end{aligned}$$

So, in order for any estimator δ to be a Bayes estimator, it must have an average risk of 0, which is not possible.

See Exercise 4.21 for further issues concerning this example. ∎

Example 4.11 (Continuation of Example 4.6) Estimation of the Binomial Parameter with Beta Prior and Other Loss Functions
We return to the problem in which we model the conditional distribution of an observable random variable X as a binomial(n, π) distribution, conditional on π, of course. Suppose we assume π comes from a beta(α, β) prior distribution; that is, we consider a random variable Π that has beta distribution. As in Example 4.6, we wish to estimate Π.

- Could we define a loss function so that the Bayes estimator is unbiased for a proper prior? Yes. Take

$$L(\pi, d) = \frac{(d - \pi)^2}{\pi(1 - \pi)}, \tag{4.52}$$

  and take a beta(1,1) (that is, uniform) prior. This yields the Bayes estimator

$$\frac{X}{n}. \tag{4.53}$$

- For any loss function other than the squared-error, will the Bayes estimator be minimax? Yes, the loss function (4.52) yields this property. The Bayes estimator X/n has constant risk (Exercise 4.22); therefore, it is minimax wrt that loss.

∎

4.3.5 Some Additional (Counter)Examples

Example 4.12 An Admissible Estimator that Is Not Bayes
∎

Example 4.13 A Bayes Estimator that Is Minimax but Not Admissible
If a Bayes estimator is unique under any loss function, then it is admissible under that loss (Theorem 4.3). Ferguson (1967) gave an example of a (nonunique) Bayes estimator that is not admissible, but has constant risk and so is minimax. ********** ∎

Example 4.14 A Limit of Unique Bayes Admissible Estimators that Is Not Admissible
∎

4.4 Probability Statements in Statistical Inference

The process of parametric point estimation, as discussed in Section 4.3, or of testing a simple hypothesis, as discussed in Section 4.5.4, is not consistent with the fundamental Bayesian description of the random nature of the parameter. Because of the widespread role of point estimation and simple hypothesis testing in science and in regulatory activities, however, Bayesian statistical procedures must be developed and made available. Tests of composite hypotheses and identification of Bayesian confidence sets are more consistent with the general Bayesian paradigm. (The standard terminology for a Bayesian analogue of a confidence set is *credible set*.)

In the classical (frequentist) approach to developing methods for hypothesis testing and for determining confidence sets, we assume a model P_θ for the state of nature and develop procedures by consideration of probabilities of the form $\Pr(T(X) \circ C(\theta)|\theta)$, where $T(X)$ is a statistic, $C(\theta)$ is some region determined by the true (unknown) value of θ, and $\circ$ is some relationship. The forms of $T(X)$ and $C(\theta)$ vary depending on the statistical procedure. The procedure may be a test, in which case we may have $T(X) = 1$ or 0, according to whether the hypothesis is rejected or not, or it may by a procedure to define a confidence set, in which case $T(X)$ is a set. For example, if θ is given to be in Θ_H, and the procedure $T(X)$ is an α-level test of H, then $\Pr(T(X) = 1|\theta \in \Theta_H) \leq \alpha$. In a procedure to define a confidence set, we may be able to say $\Pr(T(X) \ni \theta) = 1 - \alpha$.

These kinds of probability statements in the frequentist approach are somewhat awkward, and a person without training in statistics may find them particularly difficult to interpret. Instead of a statement of the form $\Pr(T(X)|\theta)$, many people would prefer a statement of the form $\Pr(\Theta \in \Theta_H|X = x)$.

In order to make such a statement, however, we first must think of the parameter as a random variable and then we must formulate a conditional

distribution for Θ, given $X = x$. In the usual Bayesian paradigm, we use a model that has several components: a marginal (prior) probability distribution for the *unobservable random variable* Θ; a conditional probability distribution for the *observable random variable* X, given $\Theta = \theta$; and other assumptions about the distributions. We denote the prior density of Θ as p_Θ, and the conditional density of X as $p_{X|\theta}$. The procedure is to determine the conditional (posterior) distribution of Θ, given $X = x$. Since we model our information about Θ as a probability distribution, it is natural and appropriate to speak of probabilities about Θ. This is the kind of approach Laplace took in analyzing the urn problem, as we described at the beginning of this chapter.

We can think of these differences in another way. If M is the model or hypothesis and D is the data, the difference is between

$$\Pr(D|M)$$

(a "frequentist" interpretation), and

$$\Pr(M|D)$$

(a "Bayesian" interpretation). People who support the latter interpretation will sometimes refer to the "prosecutor's fallacy" in which $\Pr(E|H)$ is confused with $\Pr(H|E)$, where E is some evidence and H is some hypothesis.

While in parametric point estimation, as discussed in Section 4.3, statements about probability may not be so meaningful, in tests of composite hypotheses and identification of credible sets, they are natural and appropriate. We discuss testing and determining credible sets in the next two sections.

4.5 Bayesian Testing

In statistical hypothesis testing, the basic problem is to decide whether or not to reject a statement about the distribution of a random variable. The statement must be expressible in terms of membership in a well-defined class. We usually formulate the testing problem as one of deciding between two statements:

$$H_0 : \theta \in \Theta_0$$

and

$$H_1 : \theta \in \Theta_1,$$

where $\Theta_0 \cap \Theta_1 = \emptyset$.

We do not treat H_0 and H_1 symmetrically; H_0 is the *hypothesis to be tested* and H_1 is the *alternative*. This distinction is important in developing a methodology of testing. We sometimes also refer to H_0 as the "null hypothesis" and to H_1 as the "alternative hypothesis".

In a Bayesian approach to this problem, we treat θ as a random variable, Θ and formulate the testing problem as beginning with prior probabilities

$$p_0 = \Pr(\Theta \in \Theta_0) \quad \text{and} \quad p_1 = \Pr(\Theta \in \Theta_1),$$

and then, given data x, determining posterior conditional probabilities

$$\hat{p}_0 = \Pr(\Theta \in \Theta_0) \quad \text{and} \quad \hat{p}_1 = \Pr(\Theta \in \Theta_1).$$

These latter probabilities can be identified with the posterior likelihoods, say L_0 and L_1.

In the Bayesian framework, we are interested in the probability that H_0 is true. The prior distribution provides an a priori probability, and the posterior distribution based on the data provides a posterior probability that H_0 is true. Clearly, we would choose to reject H_0 when the probability that it is true is small.

4.5.1 A First, Simple Example

Suppose we wish to test

$$H_0 : P = P_0 \qquad \text{versus} \qquad H_1 : P = P_1,$$

and suppose that known probabilities p_0 and $p_1 = 1 - p_0$ can be assigned to H_0 and H_1 prior to the experiment. We see

- The overall probability of an error resulting from the use of the test δ is

$$p_0 \mathrm{E}_0(\delta(X)) + p_1 \mathrm{E}_1(1 - \delta(X)).$$

- The Bayes test that minimizes this probability is given by

$$\delta(x) = \begin{cases} 1 \text{ when } \hat{p}_1(x) > k\hat{p}_0(x) \\ \\ 0 \text{ when } \hat{p}_1(x) < k\hat{p}_0(x), \end{cases}$$

  for $k = p_0/p_1$.
- The conditional probability of H_i given $X = x$, that is, the posterior probability of H_i, is

$$\frac{p_i \hat{p}_i(x)}{p_0\hat{p}_0(x) + p_1\hat{p}_1(x)}$$

  and the Bayes test therefore decides in favor of the hypothesis with the larger posterior probability.

Testing as an Estimation Problem

As an estimation problem, the testing problem is equivalent to estimating the indicator function $\mathrm{I}_{\Theta_0}(\theta)$. We use a statistic $S(X)$ as an estimator of $\mathrm{I}_{\Theta_0}(\theta)$. The estimand is in $\{0, 1\}$, and so $S(X)$ should be in $\{0, 1\}$, or at least in $[0, 1]$.

For a 0-1 loss function, the Bayes estimator of $\mathrm{I}_{\Theta_0}(\theta)$ is the function that minimizes the posterior risk, $\mathrm{E}_{\Theta|x}(L(\Theta, s))$. The risk is just the posterior probability, so the Bayesian solution using this loss is

$$S(x) = \begin{cases} 1 \text{ if } \Pr(\theta \in \Theta_0|x) > \Pr(\theta \notin \Theta_0|x) \\ 0 \text{ otherwise,} \end{cases}$$

where $\Pr(\cdot)$ is evaluated with respect to the posterior distribution $P_{\Theta|x}$.

4.5.2 Loss Functions

Due to the discrete nature of the decision regarding a test of an hypothesis, discrete loss functions are often more appropriate.

The 0-1-γ Loss Function

In a Bayesian approach to hypothesis testing using the test $\delta(X) \in \{0, 1\}$, we often formulate a loss function of the form

$$L(\theta, d) = \begin{cases} c_d \text{ for } \theta \in \Theta_0 \\ b_d \text{ for } \theta \in \Theta_1 \end{cases}$$

where $c_1 > c_0$ and $b_0 > b_1$, with $c_0 = b_1 = 0$, $b_0 = 1$, and $c_1 = \gamma > 0$. (This is a *0-1-γ loss function*; see page 261.)

A Bayesian action for hypothesis testing with a 0-1-γ loss function is fairly easy to determine. The posterior risk for choosing $\delta(X) = 1$, that is, for rejecting the hypothesis, is

$$c\Pr(\Theta \in \Theta_{H_0}|X = x),$$

and the posterior risk for choosing $\delta(X) = 0$ is

$$\Pr(\Theta \in \Theta_{H_1}|X = x),$$

hence the optimal decision is to choose $\delta(X) = 1$ if

$$c\Pr(\Theta \in \Theta_{H_0}|X = x) < \Pr(\Theta \in \Theta_{H_1}|X = x),$$

which is the same as

$$\Pr(\Theta \in \Theta_{H_0}|X = x) < \frac{1}{1+c}.$$

In other words, the Bayesian approach says to reject the hypothesis if its posterior probability is small. The Bayesian approach has a simpler interpretation than the frequentist approach. It also makes more sense for other loss functions.

The Weighted 0-1 or α_0-α_1 Loss Function

Another approach to account for all possibilities and to penalize errors differently when the null hypothesis is true or false is use a weighted 0-1 loss function such as a α_0-α_1 loss (see page 261). Using the estimator $S(X) = s \in \{0, 1\}$, as above, we define

$$L(\theta, s) = \begin{cases} 0 & \text{if } s = \mathrm{I}_{\Theta_0}(\theta) \\ \alpha_0 & \text{if } s = 0 \text{ and } \theta \in \Theta_0 \\ \alpha_1 & \text{if } s = 1 \text{ and } \theta \notin \Theta_0. \end{cases}$$

The 0-1-γ loss and the α_0-α_1 loss could be defined either in terms of the test rule δ or the estimator S; I chose to do one one way and the other another way just for illustration.

The Bayes estimator of $\mathrm{I}_{\Theta_0}(\theta)$ using this loss is

$$S(x) = \begin{cases} 1 \text{ if } \Pr(\theta \in \Theta_0 | x) > \frac{\alpha_1}{\alpha_0 + \alpha_1} \\ 0 \text{ otherwise,} \end{cases}$$

where again $\Pr(\cdot)$ is evaluated with respect to the posterior distribution. To see that this is the case, we write the posterior loss

$$\int_\Theta L(\theta, s) \mathrm{d}P_{\Theta|x} = a_0 \Pr(\theta \in \Theta_0 | x) \mathrm{I}_{\{0\}}(s) + a_1 \Pr(\theta \notin \Theta_0 | x) \mathrm{I}_{\{1\}}(s),$$

and then minimize it.

Under a α_0-α_1 loss, the null hypothesis H_0 is rejected whenever the posterior probability of H_0 is too small. The *acceptance level*, $\alpha_1/(\alpha_0 + \alpha_1)$, is determined by the specific values chosen in the loss function. The Bayes test, which is the Bayes estimator of $\mathrm{I}_{\Theta_0}(\theta)$, depends only on α_0/α_1. The larger α_0/α_1 is the smaller the posterior probability of H_0 that allows for it to be accepted. This is consistent with the interpretation that the larger α_0/α_1 is the more important a wrong decision under H_0 is relative to H_1.

Examples

Let us consider two familiar easy pieces using a α_0-α_1 loss.

Example 4.15 Binomial with Uniform Prior
First, let $X|\pi \sim$ binomial(π, n) and assume a prior on Π of U$(0, 1)$ (a special case of the conjugate beta prior from Example 4.2). Suppose $\Theta_0 = [0, 1/2]$.

The posterior probability that H_0 is true is

$$\frac{(n+1)!}{x!(n-x)!} \int_0^{1/2} \pi^x (1-\pi)^{n-x} \mathrm{d}\pi.$$

This is computed and then compared to the acceptance level. (Note that the integral is a sum of fractions.) ∎

Example 4.16 Normal with Known Variance and a Normal Prior on Mean
For another familiar example, consider $X|\mu \sim \mathrm{N}(\mu, \sigma^2)$, with σ^2 known, and μ a realization of a random variable from $\mathrm{N}(\mu_0, \sigma_0^2)$. We considered this problem in Example 4.4 on page 341. We recall that $M|x \sim \mathrm{N}(\mu_0(x), \omega^2)$, where

$$\mu_0(x) = \frac{\sigma^2\mu_0 + \sigma_0^2 x}{\sigma^2 + \sigma_0^2} \quad \text{and} \quad \omega^2 = \frac{\sigma^2\sigma_0^2}{\sigma^2 + \sigma_0^2}.$$

To test H_0, we compute the posterior probability of H_0. Suppose the null hypothesis is

$$H_0 : \mu \leq 0.$$

Then

$$\begin{aligned}\Pr(H_0|x) &= \Pr\left(\frac{\mu - \mu_0(x)}{\omega} \leq \frac{-\mu_0(x)}{\omega}\right) \\ &= \Phi(-\mu_0(x)/\omega).\end{aligned}$$

The decision depends on the $\alpha_1/(\alpha_0 + \alpha_1)$ quantile of $\mathrm{N}(0, 1)$. Let z_{α_0,α_1} be this quantile; that is, $\Phi(z_{\alpha_0,\alpha_1}) = \alpha_1/(\alpha_0 + \alpha_1)$. The H_0 is accepted if

$$-\mu_0(x) \geq z_{\alpha_0,\alpha_1}\omega.$$

Rewriting this, we see that the null hypothesis is rejected if

$$x > -\frac{\sigma^2}{\sigma_0^2}\mu_0 - \left(1 + \frac{\sigma^2}{\sigma_0^2}\right)\omega z_{\alpha_0,\alpha_1}.$$

∎

Notice a very interesting aspect of these tests. There is no predetermined acceptance level. The decision is based simply on the posterior probability that the null hypothesis is true.

A difficulty of the α_0-α_1 loss function, of course, is the choice of α_0 and α_1. Ideally, we would like to choose these based on some kind of utility considerations, but sometimes this takes considerable thought.

4.5.3 The Bayes Factor

Given a prior distribution P_Θ, let p_0 be the prior probability that H_0 is true, and p_1 be the prior probability that H_1 is true. The prior odds then is p_0/p_1. Similarly, let $\hat{p}_0$ be the posterior probability that H_0 is true given x, and $\hat{p}_1$ be the posterior probability that H_1 is true, yielding the posterior odds $\hat{p}_0/\hat{p}_1$. The posterior odds is the ratio of the posterior likelihoods, L_0/L_1.

The posterior probability of the event can be related to the relative odds. The posterior odds is

$$\frac{\hat{p}_0}{\hat{p}_1} = \frac{p_0}{p_1} \frac{f_{X|\theta_0}(x)}{\int f_{X|\theta}(x) \mathrm{d}F_\Theta}.$$

The term

$$\mathrm{BF}(x) = \frac{f_{X|\theta_0}(x)}{\int f_{X|\theta}(x) \mathrm{d}F_\Theta} \tag{4.54}$$

is called the *Bayes factor*. The Bayes factor obviously also depends on the prior $f_\Theta(\theta)$.

Rather than computing the posterior odds directly, we emphasize the Bayes factor, which for any stated prior odds yields the posterior odds. The Bayes factor is the posterior odds in favor of the hypothesis if $p_0 = 0.5$.

Note that, for the simple hypothesis versus a simple alternative, the Bayes factor simplifies to the likelihood ratio:

$$\frac{f_{X|\theta_0}(x)}{f_{X|\theta_1}(x)}.$$

One way of looking at this likelihood ratio is to use MLEs under the two hypotheses:

$$\frac{\sup_{\Theta_0} f_{X|\theta}(x)}{\sup_{\Theta_1} f_{X|\theta}(x)}.$$

This approach, however, assigns Dirac masses at the MLEs, $\hat{\theta}_0$ and $\hat{\theta}_1$.

The Bayes factor is more properly viewed as a Bayesian likelihood ratio,

$$\mathrm{BF}(x) = \frac{p_0 \int_{\Theta_0} f_{X|\theta}(x) \mathrm{d}\theta}{p_1 \int_{\Theta_1} f_{X|\theta}(x) \mathrm{d}\theta},$$

and, from a decision-theoretic point of view, it is entirely equivalent to the posterior probability of the null hypothesis. Under the α_0-α_1 loss function, H_0 is accepted when

$$\mathrm{BF}(x) > \frac{a_1}{a_0} / \frac{p_0}{p_1}$$

From this, we see that the Bayesian approach effectively gives an equal prior weight to the two hypotheses, $p_0 = p_1 = 1/2$ and then modifies the error penalties as $\tilde{a_i} = a_i p_i$, for $i = 0, 1$, or alternatively, incorporates the weighted error penalties directly into the prior probabilities:

$$\tilde{p_0} = \frac{a_0 p_0}{a_0 p_0 + a_1 p_1} \qquad \tilde{p_1} = \frac{a_1 p_1}{a_0 p_0 + a_1 p_1}.$$

The ratios such as likelihood ratios and relative odds that are used in testing carry the same information content if they are expressed as their reciprocals. These ratios can be thought of as evidence in favor of one hypothesis or model versus another hypothesis or model. The ratio provides a comparison of two alternatives, but there can be more than two alternatives under consideration. Instead of just H_0 and H_1 we may contemplate H_i and H_j, and

follow the same steps using p_i/p_j. The Bayes factor then depends on i and j, and of course whether we use the odds ratio p_i/p_j or p_j/p_i. We therefore sometimes write the Bayes factor as $\mathrm{BF}_{ij}(x)$ where the subscript ij indicates use of the ratio p_i/p_j. In this notation, the Bayes factor (4.54) would be written as $\mathrm{BF}_{01}(x)$.

Jeffreys (1961) suggested a subjective "scale" to judge the evidence of the data in favor of or against H_0. Kass and Raftery (1995) discussed Jeffreys's scale and other issues relating to the Bayes factor. They modified his original scale (by combining two categories), and suggested

- if $0 < \log_{10}(\mathrm{BF}_{10}) < 0.5$, the evidence against H_0 is "poor",
- if $0.5 \leq \log_{10}(\mathrm{BF}_{10}) < 1$, the evidence against H_0 is "substantial",
- if $1 \leq \log_{10}(\mathrm{BF}_{10}) < 2$, the evidence against H_0 is "strong", and
- if $2 \leq \log_{10}(\mathrm{BF}_{10})$, the evidence against H_0 is "decisive".

Note that the Bayes factor is the reciprocal of the one we first defined in equation (4.54). While this scale makes some sense, the separations are of course arbitrary, and the approach is not based on a decision theory foundation. Given such a foundation, however, we still have the subjectivity inherent in the choice of a_0 and a_1, or in the choice of a significance level.

Kass and Raftery (1995) also gave an interesting example illustrating the Bayesian approach to testing of the "hot hand" hypothesis in basketball. They formulate the null hypothesis (that players do not have a "hot hand") as the distribution of good shots by a given player, Y_i, out of n_i shots taken in game i as binomial(n_i, π), for games $i = 1, \ldots, g$; that is, the probability for a given player, the probability of making a shot is constant in all games (within some reasonable period). A general alternative is $H_1 : Y_i \sim \text{binomial}(n_i, \pi_i)$. We choose a flat $\mathrm{U}(0,1)$ conjugate prior for the H_0 model. For the H_1 model, we choose a conjugate prior beta(α, β) with $\alpha = \xi/\omega$ and $\beta = (1-\xi)/\omega$. Under this prior, the prior expectation $\mathrm{E}(\pi_i|\xi, \omega)$ has an expected value of ξ, which is distributed as $\mathrm{U}(0,1)$ for fixed ω. The Bayes factor is is very complicated, involving integrals that cannot be solved in closed form. Kass and Raftery use this to motivate and to compare various methods of evaluating the integrals that occur in Bayesian analysis. One simple method is Monte Carlo.

Often, however, the Bayes factor can be evaluated relatively easily for a given prior, and then it can be used to investigate the sensitivity of the results to the choice of the prior, by computing it for another prior.

From Jeffreys's Bayesian viewpoint, the purpose of hypothesis testing is to evaluate the evidence in favor of a particular scientific theory. Kass and Raftery make the following points in the use of the Bayes factor in the hypothesis testing problem:

- Bayes factors offer a straightforward way of evaluating evidence in favor of a null hypothesis.
- Bayes factors provide a way of incorporating external information into the evaluation of evidence about a hypothesis.

- Bayes factors are very general and do not require alternative models to be nested.
- Several techniques are available for computing Bayes factors, including asymptotic approximations that are easy to compute using the output from standard packages that maximize likelihoods.
- In "nonstandard" statistical models that do not satisfy common regularity conditions, it can be technically simpler to calculate Bayes factors than to derive non-Bayesian significance tests.
- The Schwarz criterion (or BIC) gives a rough approximation to the logarithm of the Bayes factor, which is easy to use and does not require evaluation of prior distributions. The BIC is

$$\text{BIC} = -2\log(L(\theta_m|x)) + k\log n,$$

  where θ_m is the value of the parameters that specify a given model, k is the number of unknown or free elements in θ_m, and n is the sample size. The relationship is

$$\frac{-\text{BIC}/2 - \log(\text{BF})}{\log(\text{BF})} \to 0,$$

  as $n \to \infty$.
- When we are interested in estimation or prediction, Bayes factors may be converted to weights to be attached to various models so that a composite estimate or prediction may be obtained that takes account of structural or model uncertainty.
- Algorithms have been proposed that allow model uncertainty to be taken into account when the class of models initially considered is very large.
- Bayes factors are useful for guiding an evolutionary model-building process.
- It is important, and feasible, to assess the sensitivity of conclusions to the prior distributions used.

  ***** stuff to add:
  pseudo-Bayes factors
  training sample
  arithmetic intrinsic Bayes factor
  geometric intrinsic Bayes factor
  median intrinsic Bayes factor

The Bayes Risk Set

A *risk set* can be useful in analyzing Bayesian procedures when the parameter space is finite. If

$$\Theta = \{\theta_1, \ldots, \theta_k\}, \tag{4.55}$$

the risk set for a procedure T is a set in $\mathbb{R}^k$:

$$\{(z_1, ..., z_k) : z_i = R(\theta_i, T)\}. \tag{4.56}$$

In the case of 0-1 loss, the risk set is a subset of the unit hypercube; specifically, for $\Theta = \{0, 1\}$, it is a subset of the unit square: $[0, 1] \times [0, 1]$.

4.5.4 Bayesian Tests of a Simple Hypothesis

Although the test of a simple hypothesis versus a simple alternative, as in the example Section 4.5.1, is easy to understand and helps to direct our thinking about the testing problem, it is somewhat limited in application. In a more common application, we may have a dense parameter space Θ, and hypotheses that specify different subsets of Θ. A common situation is the "one-sided" test for $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$. We can usually develop meaningful approaches to this problem, perhaps based on some boundary point of H_0. A "two-sided" test, in which, for example, the alternative specifies

$$\Theta_l = \{\theta : \theta < \theta_0\} \cup \Theta_u = \{\theta : \theta > \theta_0\}, \tag{4.57}$$

presents more problems for the development of reasonable procedures.

In a Bayesian approach, when the parameter space Θ is dense, but either hypothesis is simple, there is a particularly troubling situation. This is because of the Bayesian interpretation of the problem as one in which a probability is to be associated with a statement about a specific value of a continuous random variable.

Consider the problem in a Bayesian approach to deal with an hypothesis of the form $H_0 : \Theta = \theta_0$, that is $\Theta_0 = \{\theta_0\}$; versus the alternative $H_1 : \Theta \neq \theta_0$.

A reasonable prior for Θ with a continuous support would assign a probability of 0 to $\Theta = \theta_0$.

One way of getting around this problem may be to modify the hypothesis slightly so that the null is a small interval around θ_0. This may make sense, but it is not clear how to proceed.

Another approach is, as above, to assign a positive probability, say p_0, to the event $\Theta = \theta_0$. Although it may not appear how to choose p_0, just as it would not be clear how to choose an interval around θ_0, we can at least proceed to simplify the problem following this approach. We can write the joint density of X and Θ as

$$f_{X,\Theta}(x, \theta) = \begin{cases} p_0 f_{X|\theta_0}(x) & \text{if } \theta = \theta_0, \\ (1 - p_0) f_{X|\theta}(x) & \text{if } \theta \neq \theta_0. \end{cases} \tag{4.58}$$

There are a couple of ways of simplifying. Let us proceed by denoting the prior density of Θ over $\Theta - \theta_0$ as λ. We can write the marginal of the data (the observable X) as

$$f_X(x) = p_0 f_{X|\theta_0}(x) + (1 - p_0) \int f_{X|\theta}(x) \mathrm{d}\lambda(\theta). \tag{4.59}$$

We can then write the posterior density of Θ as

$$f_{\Theta|x}(\theta|x) = \begin{cases} p_1 & \text{if } \theta = \theta_0, \\ (1-p_1)\frac{f_{X|\theta}(x)}{f_X(x)} & \text{if } \theta \neq \theta_0, \end{cases} \tag{4.60}$$

where

$$p_1 = \frac{p_0 f_{X|\theta_0}(x)}{f_X(x)}. \tag{4.61}$$

This is the posterior probability of the event $\Theta = \theta_0$.

The Lindley-Jeffrey "Paradox"

In testing a simple null hypothesis against a composite alternative, an anomaly can occur in which a classical frequentist test can strongly reject the null, but a Bayesian test constructed with a mixed prior consisting of a point mass at the null and a diffuse continuous prior over the remainder of the parameter space.

Given a simple null hypothesis H_0, the result of an experiment x, and a prior distribution that favors H_0 weakly, a "paradox" occurs when the result x is significant by a frequentist test, indicating sufficient evidence to reject H_0 at a given level, but the posterior probability of H_0 given x is high, indicating strong evidence that H_0 is in fact true. This is called Lindley's paradox or the Lindley-Jeffrey paradox.

This can happen at the same time when the prior distribution is the sum of a sharp peak at H_0 with probability p and a broad distribution with the rest of the probability $1-p$. It is a result of the prior having a sharp feature at H_0 and no sharp features anywhere else.

Consider the testing problem in Example 4.16, except this time for a simple null hypothesis.

Example 4.17 Normal with Known Variance and a Normal Prior on Mean; Simple Null Hypothesis (Lindley, 1957)
Consider again $X|\mu \sim \mathrm{N}(\mu, \sigma^2)$, with σ^2 known. As before, to test H_0, we compute the posterior probability of H_0. Now, suppose the null hypothesis is

$$H_0 : \mu = 0.$$

In the case of the prior that supposed that μ a realization of a random variable from $\mathrm{N}(\mu_0, \sigma_0^2)$, which for a realization $X = x$ yielded $M|x \sim \mathrm{N}(\mu_0(x), \omega^2)$. For this, we get the posterior probability of H_0 to be 0.

Let us modify the prior so as to give a non-zero probability p_0 to the null hypothesis. As suggested above, we take a prior of the form

$$\tilde{f}_M(\theta|x) = \begin{cases} p_0 & \text{if } \mu = 0, \\ (1-p_0) f_M(\mu) & \text{if } \mu \neq 0, \end{cases} \tag{4.62}$$

where f_M is the PDF of a $\mathrm{N}(\mu_0, \sigma_0^2)$. Suppose, further, our prior beliefs about μ are not strong, so we choose σ_0^2 much greater than $\sigma^2 = 1$. Actually, it is

not important that the prior f_M be proportional to the normal. The overall likelihood of the alternative hypothesis is

$$L_1 = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp(-(x-\mu)^2/2)(1-p_0) f_M(\mu) \mathrm{d}\mu,$$

and the likelihood of the null hypothesis is

$$L_0 = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2).$$

The important fact to note is that L_1 can be quite small. This is because the prior gives a very small probability to a neighborhood of x, even a relatively large neighborhood in terms of $\sigma = 1$.

The posterior odds are

$$\frac{\Pr(\mu = 0 | X = x)}{\Pr(\mu \neq 0 | X = x)} = \frac{p_0}{1-p_0} \frac{L_0}{L_1}$$

which can be very large even if p_0 is very small. ∎

This rather unreasonable conclusion of a standard Bayesian analysis has been discussed in many articles and books; see, for example, Shafer (1982) or Johnson and Rossell (2010).

4.5.5 Least Favorable Prior Distributions

In testing composite hypotheses, we often ask what is the "worst case" within the hypothesis. In a sense, this is the attempt to reduce the composite hypothesis to a simple hypothesis. This is the idea behind a p-value. In a Bayesian testing problem, this corresponds to a bound on the posterior probability.

Again, consider the problem of testing $H_0 : \Theta = \theta_0$ versus the alternative $H_1 : \Theta \neq \theta_0$. ***

4.6 Bayesian Confidence Sets

4.6.1 Credible Sets

In a Bayesian setup, we define a random variable Θ that corresponds to the parameter of interest, and the usual steps in a Bayesian analysis allows us to compute $\Pr(\Theta \in \Theta_{H_0} | X = x)$. The problem in determining a confidence set is an inverse problem; that is, for a given α, we determine C_α such that $\Pr(\Theta \in C_\alpha | X = x) = 1 - \alpha$. Of course there may be many sets with this property. We need some additional condition(s).

In the frequentist approach, we add the property that the region be the smallest possible. "Smallest" means with respect to some simple measure such

as the usual Lebesgue measure; in the one-dimensional continuous case, we seek the shortest interval. In the Bayesian approach, we so something similar, except we use the posterior density as a measure.

The mechanics of determining credible sets begin with the standard Bayesian steps that yield the conditional distribution of the parameter given the observable random variable. If the density exists, we denote it as $f_{\Theta|x}$. At this point, we seek regions of θ in which $f_{\Theta|x}(\theta|x)$ is large. In general, the problem may be somewhat complicated, but in many situations of interest it is relatively straightforward. Just as in the frequentist approach, the identification of the region often depends on *pivotal* values, or pivotal functions. (Recall that a function $g(T, \theta)$ is said to be a pivotal function if its distribution does not depend on any unknown parameters.)

It is often straightforward to determine one with posterior probability content of $1 - \alpha$.

4.6.2 Highest Posterior Density Credible sets

If the posterior density is $f_{\Theta|x}(\theta|x)$, we determine a number c such that the set

$$C_\alpha(x) = \{\theta \ : \ f_{\Theta|x}(\theta) \geq c_\alpha\} \tag{4.63}$$

is such that $\Pr(\Theta \in C_\alpha|X = x) = 1 - \alpha$. Such a region is called a level $1 - \alpha$ *highest posterior density* or HPD credible set.

We may impose other conditions. For example, in a one-dimensional continuous parameter problem, we may require that one endpoint of the interval be infinite (that is, we may seek a one-sided confidence interval).

An HPD region can be disjoint if the posterior is multimodal.

If the posterior is symmetric, all HPD regions will be symmetric about x.

For a simple example, consider a $\mathrm{N}(0, 1)$ prior distribution on Θ and a $\mathrm{N}(\theta, 1)$ distribution on the observable. The posterior given $X = x$ is $\mathrm{N}(x, 1)$. All HPD regions will be symmetric about x. In the case of a symmetric density, the HPD is the same as the centered equal-tail credible set; that is, the one with equal probabilities outside of the credible set. In that case, it is straightforward to determine one with posterior probability content of $1 - \alpha$.

4.6.3 Decision-Theoretic Approach

We can also use a specified loss function to approach the problem of determining a confidence set.

We choose a region so as to minimize the expected posterior loss.

For example, to form a two-sided interval in a one-dimensional continuous parameter problem, a reasonable loss function may be

$$L(\theta, [c_1, c_2]) = \begin{cases} k_1(c_1 - \theta) & \text{if } \theta < c_1, \\ 0 & \text{if } c_1 \leq \theta \leq c_2, \\ k_2(\theta - c_2) & \text{if } \theta > c_2. \end{cases}$$

This loss function also leads to the interval between two quantiles of the posterior distribution.

It may not be HPD, and it may not be symmetric about some pivot quantity even if the posterior is symmetric.

4.6.4 Other Optimality Considerations

We may impose other conditions. For example, in a one-dimensional continuous parameter problem, we may require that one endpoint of the interval be infinite (that is, we may seek a one-sided confidence interval).

Example 4.18 Credible sets for the Binomial Parameter with a Beta Prior
Consider the problem of estimating π in a binomial(n, π) distribution with a beta(α, β) prior distribution, as in Example 4.6 on page 355.

Suppose we choose the hyperparameters in the beta prior as $\alpha = 3$ and $\beta = 5$. The prior, that is, the marginal distribution of Π, is as shown in Figure 4.1 and if n is 10 and we take one observation, $x = 2$ we have the conditional distribution of Π, as a beta with parameters $x + \alpha = 5$ and $n - x + \beta = 13$, as shown in Figure 4.2.

Now, given $x = 2$, and the original beta(3,5) prior, let's find an equal-tail 95% credible set. Here's some R code:

```
a<-3
b<-5
n<-10
x<-2
alpha<-0.05
lower<-qbeta(alpha/2,x+a,n-x+b)
upper<-qbeta(1-alpha/2,x+a,n-x+b)
pi<-seq(0,1,0.01)
plot(pi,dbeta(pi,x+a,n-x+b),type='l',
  main="95\% Credible set with x=2",
  ylab="Posterior",xlab=expression(pi))
lines(c(lower,lower),c(0,dbeta(lower,x+a,n-x+b)))
lines(c(upper,upper),c(0,dbeta(upper,x+a,n-x+b)))
lines(c(0,1),c(0,0))
```

We get the credible set shown in Figure 4.5. The probability in each tail is 0.025.

Because the posterior density is not symmetric, it is not an easy matter to get the HPD credible set.

The first question is whether the credible set is an interval. This depends on whether the posterior is unimodal. As we have already seen in Section 4.2, the posterior in this case is unimodal if $n > 0$, and so the credible set is indeed an interval.

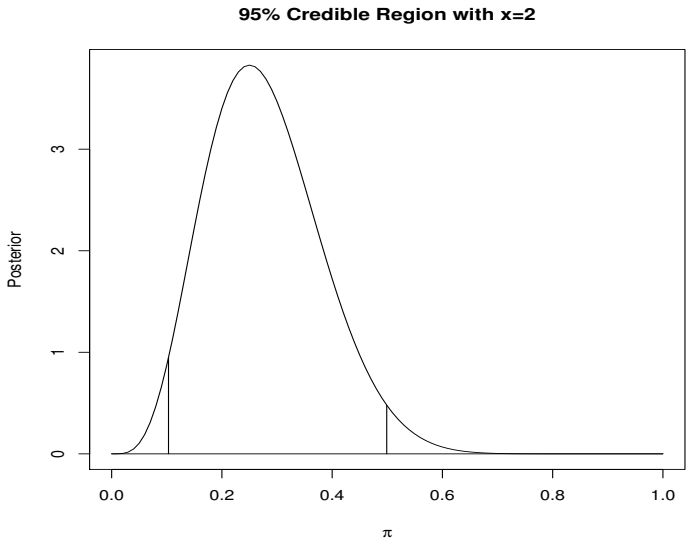


Figure 4.5. 95% Credible set after Observing $x = 2$

We can determine the region iteratively by starting with the equal-tail credible set. At each step in the iteration we have a candidate lower bound and upper bound. We determine which one has the higher value of the density, and then shift the interval in that direction. We continue this process, keeping the total probability constant at each step. Doing this we get the credible set shown in Figure 4.6. The probability in the lower tail is 0.014 and that in the upper tail is 0.036. The density is 0.642 at each endpoint; that is, in equation (4.63), $c_\alpha = 0.642$.

Here's the R code that yielded the HPD:

```
a<-3
b<-5
n<-10
x<-2
alpha<-0.05
#     start by determining the equal-tail CR, using the posterior
lower<-qbeta(alpha/2,x+a,n-x+b)
upper<-qbeta(1-alpha/2,x+a,n-x+b)
#     set a tolerance for convergence
tol <- 0.005 # to get the density values to agree to 3 decimal places
a10 <- 0
a20 <- 0
a1 <- alpha/2
```

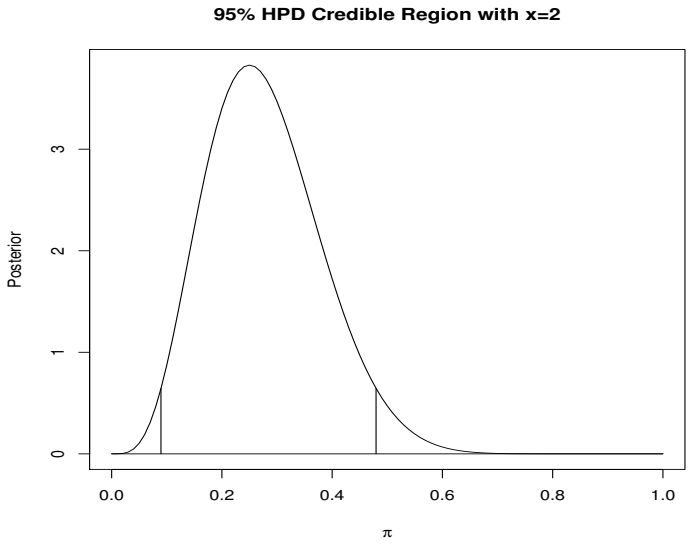


Figure 4.6. HPD 95% Credible set after Observing $x = 2$

```
a2 <- 1-alpha/2
adj <- a1
d <- 1
while (abs(d)>tol){
#      determine difference in the density at the two candidate points
d <- dbeta(lower,x+a,n-x+b)-dbeta(upper,x+a,n-x+b)
#      halve the adjustment in each iteration
adj <- adj/2
#      if density at lower boundary is higher, shift interval to the left
s <- 1
if(d>0) s <- -1
a1 <- a1 + s*adj
a2 <- a2 + s*adj
lower<-qbeta(a1,x+a,n-x+b)
upper<-qbeta(a2,x+a,n-x+b)
}
```

4.7 Computational Methods in Bayesian Inference;

Markov Chain Monte Carlo

Monte Carlo techniques often allow us to make statistical inferences when the statistical method involves intractable expressions. In applications in Bayesian inference, we can study the posterior distribution, which may be intractable or which may be known only proportionally, by studying random samples from that distribution.

In parametric Bayesian inference, the objective is to obtain the conditional posterior distribution of the parameter, given the observed data. This is Q_H in equation (4.2), and it is defined by the density in step 5 in the procedure outlined in Section 4.2.3. This density contains all of the information about the parameter of interest, although we may wish to use it for specific types of inference about the parameter, such as a point estimator or a credible set.

Understanding the Posterior Distribution

As with any probability distribution, a good way to understand the posterior distribution is to take a random sample from it. In the case of the posterior distribution, we cannot take a physical random sample. We can, however, simulate a random sample, using methods discussed in Section 0.0.7, beginning on page 663.

In single-parameter cases, random samples from the posterior distribution can often be generated using a direct acceptance/rejection method if the constant of proportionality is known. If the posterior density is known only proportionally, a Metropolis-Hastings method often can be used.

Often the posterior density is a fairly complicated function, especially in multi-parameter cases or in hierarchical models. In such cases, we may be able to express the conditional density of each parameter given all of the other parameters. In this case, it is fairly straightforward to use a Gibbs sampling method to generate samples from the multivariate distribution. Consider the relatively simple case in Example 4.5. The joint posterior PDF is given in equation (4.32). We can get a better picture of this distribution by simulating random observations from it. To do this we generate a realization σ^2 from the marginal posterior with PDF given in equation (4.33), and then with that value of σ^2, we generate a realization μ from the conditional posterior with PDF given in equation (4.34).

Example 4.19 illustrates this technique for a hierarchical model.

The simulated random samples from the posterior distribution gives us a picture of the density. It is often useful to make pair-wise scatter plots of the samples or estimated contour plots of the density based on the samples.

Simulated random samples can be used to approximate expectations of functions of the random parameters with respect to the posterior density (this is Monte Carlo quadrature), and they can also be used to identify other properties of the posterior distribution, such as its mode.

Computing the MAP

Computation of the MAP is essentially an optimization problem. In many cases, simulated annealing (see Section 0.4.3 on page 829) is a very effective method of determining the optimum point. The approach for optimizing the posterior probability density function is essentially the same as a Metropolis method for simulating random observations from the posterior distribution.

A Hierarchical Bayesian Model

Following custom, we use brackets to denote *densities*; $[X, Y]$, $[X|Y]$, and $[X]$ represent the joint, conditional, and marginal densities, respectively.

In a hierarchical Bayesian model, the joint distribution of the data and parameters is

$$[X|\theta_1] \times [\theta_1|\theta_2] \times [\theta_2|\theta_3] \times \cdots \times [\theta_{k-1}|\theta_k] \times [\theta_k]$$

The thing of interest is posterior density $[\theta_1|X]$.

The hierarchical structure implies

$$\begin{aligned}[\theta_1|X, \theta_{i,(i\neq 1)}] &= [\theta_1|X, \theta_2] \\ &= [\theta_i|\theta_{i-1}, \theta_{i+1}] \\ &= [\theta_k|\theta_{k-1}]\end{aligned}$$

Gibbs sampling can be used to estimate the marginal posterior densities.

Example 4.19 Gibbs Sampling Example from Gelfand and Smith, *JASA*

The paper by Gelfand and Smith (1990) was very important in popularizing the Gibbs method.

Consider an exchangeable Poisson model in which independent counts are observed over differing periods of time.

The data are $\{(s_i, t_i)\}$. Each yields a rate r_i.

Assume $[s_i|\lambda_i] = \mathrm{P}(\lambda_i t_i)$.

Assume a gamma prior distribution on the λ_i's with density

$$\frac{1}{\beta^\alpha \Gamma(\alpha)} \lambda_i^{\alpha-1} e^{-\lambda_i/\beta}$$

Further, assume β has an inverted gamma distribution with density

$$\frac{1}{\beta^{\gamma+1} \Gamma(\gamma)} \delta^\gamma e^{-\delta/\beta}$$

Beginning with $X = (s_1, s_2, \ldots, s_k)$, the conditional distribution of λ_i given X, β, and $\lambda_{j(j\neq i)}$ is merely the gamma with parameters $\alpha + s_j$ and

$\beta/(t_j+1)$, and the conditional distribution of β given X and the λ_i's is an inverted gamma with parameters $\gamma + k\alpha$ and $\sum \lambda_i + \delta$.

The various parameters (α, δ, γ) have interpretions that can be used to select reasonable values.

The Gibbs sampling method would estimate the marginal density of λ_i by generating $\lambda_i^{(1)}$ from the appropriate gamma distribution, i.e., with parameters $\alpha + s_i$ and $\beta^{(0)}/(t_i+1)$ for $i = 1, \ldots, k$, and then generating $\beta^{(1)}$ for the first iteration.

Continue this for k iterations.

Do it m times to have a density. ∎

Miscellaneous Results and Comments

Markov chain Monte Carlo has special applications when dealing with distributions that have densities known up to a constant of proportionality, that is densities specified as follows. Let h be a nonnegative integrable function that is not zero almost everywhere. Then h specifies a probability distribution, all we need to do to get the density f is normalize it.

$$f(x) = h(x)/c$$

where

$$c = \int h(x) d\mu(x)$$

The Hastings algorithm only uses h to simulate realizations from f, knowledge of the integral c is not required.

In Bayesian inference, h is the likelihood times the prior. This is always known, but the integral c is generally hard. MCMC permits easy simulations of realizations from the posterior (no knowledge of c necessary).

In most cases where there is complex dependence in the data, there is no simple probability model with c known, but it is easy to specify a model up to a constant of proportionality using an h. These are just very complicated exponential families.

Let t be a vector-valued statistic on the sample space and

$$h(x) = \exp(t(x)^{\mathrm{T}}\theta)$$

Then these specify a family of densities

$$f_\theta(x) = \exp(t(x)^{\mathrm{T}}\theta)/c(\theta).$$

In the expression

$$\exp(t(x)^{\mathrm{T}}\theta)/c(\theta),$$

$$c(\theta) = \int \exp(t(x)^{\mathrm{T}}\theta) d\mu(x),$$

but in MCMC it does not need to be known.

This is just an exponential family with canonical statistic $t(x)$ and canonical parameter θ.

Using Markov chain Monte Carlo we can simulate realizations from any distribution in the model, and using the simulations from any one distribution, we can calculate maximum likelihood estimates, bootstrap estimates of their sampling distribution and so forth.

There are also ways to get (randomized) significance tests with exact p-values using Markov chain Monte Carlo.

The output of the sampler is a Markov chain $X_1, X_2, \ldots$ whose equilibrium distribution is the distribution of interest, the one you want to sample from.

Averages with respect to that distribution are approximated by averages over the chain.

Notes and Further Reading

In this chapter we have presented Bayesian methods as an approach to the decision-theoretic principle of minimizing average risk. In this presentation we have glossed over the philosophic excitement that attended the evolution of the Bayesian approach to statistical inference.

In the early nineteenth century, Laplace developed a theory of "inverse probability", in which the frequency of observations are used to infer the probability that they arose from a particular data-generating process. Although inverse probability as a formal theory is not in current vogue, some of the underlying motivating ideas persist in inference based on likelihood and on "subjective probability". For more discussion of Laplace's work and the two examples at the beginning of this chapter, see Stigler (1986).

The idea that statistical inference can (and should) take into account not only strictly objective observations but also subjective and even personal evidence was first expounded in a clear mathematical theory by Savage in 1954 in the first edition of Savage (1972). Savage stated seven "postulates of a personalistic theory of decision" that lead to the existence of a subjective probability and a utility function. The essays in the volume edited by Kadane et al. (1999) address and expound on Savage's book. Kadane, Schervish, and Seidenfeld also consider the general cooperative Bayesian decision making. A satisfactory theory of group coherence in decisions may require the relaxation of one of Savage's postulates on the simple preferential ordering of decisions.

Good (1983) *** discuss

In a more applied context, Schlaifer (1959) incorporated a personalistic approach into statistical decision making. Many of the ideas in the Bayesian approach derive from those books and from the book by Jeffreys (1961).

An alternative approach to probabilistic reasoning is the Dempster-Shafer theory of belief functions (see Shafer (1976) and Yager and Liu (2008)).

In some cases, especially in hypothesis, the Bayesian approach is fundamentally different from the frequentist approach. The differences arise from the definition of the problem that is addressed. The articles by Casella and Berger (1987) (Roger) and (Jim) Berger and Sellke (1987) with accompanying discussion by several authors identify some of the differences in perspectives.

Berger (1985) and Robert (2001) provide extensive coverage of statistical inference from a Bayesian perspective. Both of these books compare the "frequentist" and Bayesian approaches and argue that the Bayesian paradigm is more solidly grounded.

Ghosh and Sen (1991) have considered Pitman closeness in the context of a posterior distribution, and defined *posterior Pitman closeness* in terms of probabilities evaluated with respect to the posterior distribution. Interestingly, the posterior Pitman closeness is transitive, while as we have seen on page 219, Pitman closeness does not have the transitive property.

Notation and Lingo

There are several instances in which the notation and terminology used in Bayesian statistics differ from the classical statistics that had evolved with a strong mathematical flavor.

I generally like to use uppercase letters to distinguish random variables from realizations of those random variables, which I generally represent by corresponding lowercase letters, but it is common in writing about a Bayesian analysis not to distinguish a random variable from its realization.

People who work with simple Bayes procedures began calling the distribution of the reciprocal of a chi-squared random variable an "inverse" chi-squared distribution. Because "inverse" is used in the names of distributions in a different way ("inverse Gaussian", for example), I prefer the term inverted chi-squared, or inverted gamma.

What is often called a "simple hypothesis" by most statisticians is often called a "sharp hypothesis in Bayesian analyses.

The Bayesian Religious Wars of the Mid-Twentieth Century

The analysis by Lindley and Phillips (1976) **********.

in Example 3.12

Hartley (1963)

A rather humorous historical survey of the antithetical Bayesian and frequentist approaches is given in McGrayne (2011).

Prior Distributions

Ghosh (2011) objective priors

Early versions of the maximum entropy principle were stated by Jaynes (1957a,b). Kass and Wasserman (1996) critique maximum entropy priors and other priors that are selected on the basis of being less informative.

Applications

Bayesian procedures have been somewhat slow to permeate the traditional areas of statistical applications, such as analysis of linear models, time series analysis and forecasting, and finite population sampling. This is not because the underlying theory has not been developed. Broemeling (1984) discusses Bayesian analysis of general linear models, and the articles in the book edited by Dey et al. (2000) provide an extensive coverage of Bayesian methods in generalized linear models. See Prado and West (2010) and West and Harrison (1997), for discussions of Bayesian methods in time series analysis. Bayesian methods for sampling from finite populations are discussed in Ghosh and Meeden (1998), and assessed further in Rao (2011).

Nonparametric Models

We have limited our discussion in this chapter to parametric models; that is, to situations in which the probability distributions of the observable random variables can be indexed by a real number of finite dimension. Nonparametric models can often be defined in terms of an index (or "parameter") of infinite dimension. A standard example in Bayesian analysis uses a Dirichlet process as a prior for an infinite discrete distribution (see Ferguson (1973), and Sethuraman (1994)).

Exercises

4.1. Formulate Laplace's urn problem at the beginning of this chapter in the modern Bayesian context; that is, identify the prior, the conditional of the observable data, the joint, the marginal, and the conditional posterior distributions.

4.2. Show that the family of distributions with PDF given in equation (4.23) is a conjugate family for an exponential family with PDF expressed in the form of equation (4.22).

4.3. Consider the exponential distribution with PDF

$$f_{X|\theta}(x|\theta) = \theta^{-1}\mathrm{e}^{x/\theta}\,\mathrm{I}_{\overline{\mathbb{R}}_+}(x).$$

a) Show that the inverted gamma distribution is a conjugate prior for this conditional distribution.

b) Given a random sample of size n from the exponential distribution and an inverted gamma with parameters α and β, determine the posterior conditional mean and variance.

4.4. Given the conditional PDF

$$f_{X|\gamma}(x) \propto (1 + (x - \gamma)^2)^{-1}.$$

a) Under the prior

$$f_\Gamma(\gamma) \propto \mathrm{e}^{-|\gamma-\mu|},$$

given a single observation, determine the MAP estimator of γ. Is this a meaningful estimator? Comment on why we might have expected such a useless estimator.

b) For the same distribution of the observables, consider the prior

$$f_\Gamma(\gamma) \propto \mathrm{e}^{-\alpha|\gamma-\mu|}.$$

For what values of $\alpha > 0$ will this prior yield a different estimator from that in the previous question?

c) Consider now an opposite kind of setup. Let the conditional density of the observable be

$$f_{X|\gamma}(x) \propto \mathrm{e}^{-|x-\gamma|},$$

and let the prior be

$$f_\Gamma(\gamma) \propto (1 + (\gamma - \mu)^2)^{-1}.$$

Determine the MAP estimator of γ. Comment on the difference in this estimator and that in the first part. Why might expect this situation?

4.5. Prove Theorem 4.5.

4.6. Prove Theorem 4.6.

4.7. Prove Theorem 4.7.

4.8. Consider the binomial(n, π) family of distributions in Example 4.6. Given a random sample $X_1, \ldots, X_n$ on the random variable X with conditional distribution in the binomial family, formulate the relevant PDF for obtaining $\hat{\alpha}$ and $\hat{\beta}$ in the empirical Bayes estimator of equation (4.50).

4.9. Consider again the binomial(n, π) family of distributions in Example 4.6. Given a random sample $X_1, \ldots, X_n$ on the random variable X with conditional distribution in the binomial family, determine the Bayes estimator of π under linex loss (equation (3.88) on page 262) with a beta(α, β) prior.

4.10. Consider again the binomial(n, π) family of distributions in Example 4.6. We wish to estimate π under squared-error loss with a beta(α, β) prior.

a) Determine the risk of the Bayes estimator (4.45), under squared-error loss.

b) Now consider the estimator

$$T^* = \frac{X}{n}\frac{n^{1/2}}{1 + n^{1/2}} + \frac{1}{2(1 + n^{1/2})},$$

in the form of equation (4.46). Determine a prior under which T^* is Bayes (one such prior is a beta distribution – which?), and show that T^* under squared-error loss has constant risk with respect to π.

4.11. Assume that in a batch of N items, M are defective. We are interested in the number of defective items in a random sample of n items from the batch of N items.
   a) Formulate this as a hypergeometric distribution.
   b) Now assume that $M \sim$ binomial(π, N). What is the Bayes estimator of the number of defective items in the random sample of size n using a squared-error loss?

4.12. For Example 4.7, consider each of the first five issues discussed in Example 4.6. Give the corresponding solutions for the negative binomial distribution, if the solutions are possible.

4.13. Consider the problem of estimating θ in the Poisson, assuming a random sample of size n. (The probability function, or density, is $f_{X|\theta}(x) = \theta^x e^{-\theta}/x!$ for nonnegative integers, and the parameter space is $\mathbb{R}_+$.)
   a) Determine the Bayes estimator of θ under squared-error loss and the prior $f_\Theta(\theta) = \theta_p \exp(-\theta_p \theta)$.
   b) Determine the Bayes estimator under linex loss and the prior $f_\Theta(\theta) = \theta_p \exp(-\theta_p \theta)$.
   c) Determine the Bayes estimator under zero-one loss and the prior $f_\Theta(\theta) = \theta_p \exp(-\theta_p \theta)$.
   d) In the previous questions, you should have noticed something about the prior. What is a more general prior that is a conjugate prior? Under that prior and the squared-error loss, what is the Bayes estimator? What property is shared by this estimator and the estimator in Exercise 4.13a)?
   e) Determine the Bayes estimator under squared-error loss and a uniform (improper) prior.
   f) Determine the Bayes estimator under zero-one loss and a uniform (improper) prior.
   g) Determine a minimax estimator under zero-one loss. Would you use this estimator? Why?
   h) Now restrict estimators of θ to $\delta_c(X) = cX$. Consider the loss function

$$L(\theta, \delta) = \left(\frac{\delta}{\theta} - 1\right)^2.$$

      i. Compute the risk $R(\delta_c, \theta)$. Determine whether δ_c is admissible if $c > 1$.
      ii. Compute the Bayes risk $r(f_\Theta, \delta_c)$ and determine the optimal value of c under f_Θ. (The prior f_Θ is the one used in Exercise 4.13a).)
      iii. Determine the optimal c for the minimax criterion applied to this class of estimators.
   i) As in Exercise 4.13a) with squared-error loss, consider the problem of estimating θ given the sample $X_1, \ldots, X_n$ and using the prior, $f_\Theta(\theta)$, where θ_p is empirically estimated from the data using a method-of-moments estimator.

4.14. Consider again the binomial(n, π) family of distributions in Example 4.6. Let $P_{\alpha,\beta}$ be the beta(α, β) distribution.
   a) Determine the gamma-minimax estimator of π under squared-error loss within the class of priors $\Gamma = \{P_{\alpha,\beta} : 0 < \alpha, \beta\}$.
   b) Determine the gamma-minimax estimator of π under squared-error loss within the class of priors $\Gamma = \{P_{\alpha,\beta} : 0 < \alpha, \beta \leq 1\}$.

4.15. Consider a generalization of the absolute-error loss function, $|\theta - d|$:

$$L(\theta, d) = \begin{cases} c(d-\theta) & \text{for } d \geq \theta \\ (1-c)(\theta - d) & \text{for } d < \theta \end{cases}$$

for $0 < c < 1$ (equation (3.87)). Given a random sample $X_1, \ldots, X_n$ on the random variable X, determine the Bayes estimator of $\theta = \mathrm{E}(X|\theta)$. (Assume whatever distributions are relevant.)

4.16. Let $X \sim \mathrm{U}(0, \theta)$ and the prior density of Θ be $\theta^{-2}\mathrm{I}_{[1,\infty)}(\theta)$. The posterior is therefore

$$f_{\Theta|x}(\theta|x) = \frac{2c^2}{\theta^3}\mathrm{I}_{[c,\infty)}(\theta),$$

where $c = \max(1, x)$.
   a) For squared-error loss, show that the Bayes estimator is the posterior mean. What is the posterior mean?
   b) Consider a reparametrization: $\tilde{\theta} = \theta^2$, and let $\tilde{\delta}$ be the Bayes estimator of $\tilde{\theta}$. The prior density now is

$$\frac{1}{2\tilde{\theta}^{3/2}}\mathrm{I}_{[1,\infty)}(\tilde{\theta}).$$

   In order to preserve the connection, take the loss function to be $L(\tilde{\theta}, \tilde{\delta}) = (\sqrt{\tilde{\delta}} - \sqrt{\tilde{\theta}})^2$. What is the posterior mean? What is the Bayes estimator of $\tilde{\theta}$?
   c) Compare the two estimators. Comment on the relevance of the loss functions and of the prior for the relationship between the two estimators.

4.17. Let X_1 depend on θ_1 and X_2 be independent of X_1 and depend on θ_2. Let θ_1 and θ_2 have independent prior distributions. Assume a squared-error loss. Let δ_1 and δ_2 be the Bayes estimators of θ_1 and θ_2 repectively.
   a) Show that $\delta_1 - \delta_2$ is the Bayes estimator of $\theta_1 - \theta_2$ given $X = (X_1, X_2)$ and the setup described.
   b) Now assume that $\theta_2 > 0$ (with probability 1), and let $\tilde{\delta_2}$ be the Bayes estimator of $1/\theta_2$ under the setup above. Show that $\delta_1\tilde{\delta_2}$ is the Bayes estimator of θ_1/θ_2 given $X = (X_1, X_2)$.

4.18. In the problem of estimating π given X from a binomial$(10, \pi)$ with beta(α, β) prior and squared-error loss, as in Example 4.6, sketch the risk functions, as in Figure 3.1 on page 277, for the unbiased estimator, the minimax estimator, and the estimator resulting from Jeffreys's noninformative prior.

4.19. Given an estimation problem with an integrable Lebesgue conditional PDF $f_{X|\theta}$ for the observables, an integrable Lebesgue prior PDF f_Θ, and with loss function $L(\theta, a) = w(\theta)(\theta - a)^2$, where $w(\theta)$ is a fixed weighting function. Determine the Bayes rule and comment on the role of $w(\theta)$ in this problem.

4.20. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma_0^2)$, with σ_0^2 known and μ unknown. Determine the generalized Bayes action for estimating μ under squared error loss and the noninformative prior of the Lebesgue measure on $]-\infty, \infty[$.

4.21. Refer to the problem described in Example 4.10.
   a) Show that every decision rule is inadmissible.
   b) We have implicitly assumed the action space to be $]-\infty, \infty[$. Show that if the action space is $[-\infty, \infty]$, then there is a Bayes rule, **and** that it is the only admissible rule.

4.22. Show that the unbiased Bayes estimator in Example 4.11 has constant risk wrt the loss function of equation (4.52).

4.23. As in Example 4.7, consider the problem of estimation of the negative binomial parameter with a beta prior, but instead of a squared-error loss, use the loss function of Example 4.11, given in equation (4.52). Determine the Bayes estimator. Do the estimators of the binomial parameter and the negative binomial parameter conform to the likelihood principle?

4.24. Consider the sample $(Y_1, x_1), \ldots, (Y_n, x_n)$ where the Y_i are iid as $\mathrm{N}(x_i^{\mathrm{T}}\beta, \sigma^2)$ for the fixed vectors x_i and for the unknown p-vector β. In the matrix representation $Y = X\beta + E$, assume that the $n \times p$ matrix X is of rank p. Let l be a given p-vector, and consider the problem of estimating $l^{\mathrm{T}}\beta$ under a squared-error loss.
   a) Assume $\sigma^2 = \sigma_0^2$, a known positive number. Using the prior distribution of β $\mathrm{N}_p(\beta_0, \Sigma)$, where β_0 is a known p-vector and Σ is a known positive definite matrix, determine the Bayes estimator of $l^{\mathrm{T}}\beta$.
   b) Now assume σ^2 is unknown. We will simplify the prior on β to be, conditional on σ^2, $\mathrm{N}_p(\beta_0, \sigma^2 V)$, where again β_0 is a known p-vector and V is a known positive definite matrix. Let the prior on σ^2 be the inverted gamma distribution with parameters α and β (see page 842). Determine the Bayes estimator of $l^{\mathrm{T}}\beta$.

4.25. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(\pi)$.
   a) Under the prior beta(α, β) and some given π_0, determine the Bayes factor and the Bayes test for

$$H_0 : \pi \leq \pi_0 \qquad \text{versus} \qquad H_1 : \pi > \pi_0.$$

   b) Now, consider testing

$$H_0 : \pi = \pi_0 \qquad \text{versus} \qquad H_1 : \pi \neq \pi_0.$$

      i. Make an appropriate modification to the beta prior.
      ii. Determine the Bayes factor and the Bayes test under your modified prior.

4.26. As in Exercise 4.13, consider the problem of making inferences about θ in the Poisson, assuming a random sample of size n under the prior $f_\Theta(\theta) = \theta_p \exp(-\theta_p \theta)$.

a) Let θ_0 be some given positive number. Determine the Bayes factor and the Bayes test for

$$H_0 : \theta \leq \theta_0 \qquad \text{versus} \qquad H_1 : \theta > \theta_0.$$

b) Now, consider testing

$$H_0 : \theta = \theta_0 \qquad \text{versus} \qquad H_1 : \theta \neq \theta_0.$$

i. Make an appropriate modification to the prior.

ii. Determine the Bayes factor and the Bayes test under your modified prior.

4.27. Let $X_1, \ldots, X_{n_1} \overset{\text{iid}}{\sim} \mathrm{N}(\mu_1, \sigma^2)$ and let $Y_1, \ldots, Y_{n_2} \overset{\text{iid}}{\sim} \mathrm{N}(\mu_2, \sigma^2)$.

a) Assume $\sigma^2 = \sigma_0^2$, a known positive number. As the prior distribution for $M = M_2 - M_1$ take $\mathrm{N}(\mu_p, \sigma_p^2)$, where mu_p and σ_p^2 are known constants. Determine the Bayes factor and the Bayes test for

$$H_0 : \mu_1 \leq \mu_2 \qquad \text{versus} \qquad H_1 : \mu_1 > \mu_2.$$

b) Now assume σ^2 is unknown. As the conditional prior distribution for $M = M_2 - M_1$ given σ^2, take $\mathrm{N}(\mu_p, \sigma^2/\kappa_p)$, where σ^2 is a realization of a random variable from a chi-squared distribution with parameters ν_p as degrees of freedom and σ_p as scale of σ. Determine the Bayes factor and the Bayes test for the test in the previous part.

4.28. Consider the problem of determining a credible set for a scalar parameter θ. Suppose that the conditional posterior has a Lebesgue PDF $f_{\Theta|x}(\theta)$ that is unimodal and not monotone. (It has a shape similar to that in Figure 4.6.)

a) Show that a $(1-\alpha)100\%$ HPD credible set is an interval and that the interval is unique.

b) Show that the $(1-\alpha)100\%$ HPD credible set has the shortest lenght of any interval $[a, b]$ satisfying

$$\int_a^b f_{\Theta|x}(\theta)\, \mathrm{d}\theta = 1 - \alpha.$$

5

Unbiased Point Estimation

In a decision-theoretic approach to statistical inference, we seek a method that minimizes the risk no matter what is the true state of nature. In a problem of point estimation, for example, we seek an estimator $T(X)$ which for a given loss function $L(g(\theta), T(X))$ yields a minimum of the risk, $\mathrm{E}_\theta(L(g(\theta), T(X)))$.

For some specific value of θ, say θ_1, one particular estimator, say T_1, may have the smallest expected loss, while for another value of θ, say θ_2, another estimator, say T_2, may a smaller expected loss.

What we would like is an estimator with least expected loss no matter what is the value of θ; that is, we would like an estimator with uniformly minimum risk. Because the risk depends on the value of θ, however, we see that we cannot devise such an estimator. The optimal estimator would somehow involve θ. We would prefer a procedure that does not depend on the unknown quantity we are trying to estimate, that is, we would like a procedure with *uniformly* good properties.

Since, in general, there is no procedure with uniformly minimum risk, we might consider restricting our procedures to some class of procedures that have some other desirable properties, for example, to procedures that are unbiased. As we will see in Section 5.1, this is often possible in the case of point estimation.

Unbiasedness also, of course, is desirable from an intuitive perspective. Although we may think that the concept does not have much practical importance because, after all, we are not going to repeat the procedure infinitely many times, we could raise the same issues of practical importance of minimum risk, which is also an expected value.

Unbiased Point Estimators

Our objective is to develop "good" estimators of statistical functions. The statistical function to be estimated is the estimand. It may be defined as a functional of the CDF, $\Upsilon(F)$, or, in a parametric setting, as a measurable

function of some underlying parameter, $g(\theta)$. In the following, I will generally represent the estimand as $g(\theta)$, but the concepts apply to more general estimands that may only be represented as some functional, $\Upsilon(F)$.

Although some "good" estimators are not unbiased, unbiasedness relates easily to fundamental concepts such as what does it mean to "estimate" a statistical function. Can any statistical function be estimated meaningfully? How many observations are required to yield a meaningful estimate?

An estimator $T(X)$ of a given estimand, $g(\theta)$, is unbiased with respect to θ if

$$\mathrm{E}_\theta(T(X)) = g(\theta) \; \forall \, \theta \in \Theta. \tag{5.1}$$

Thus we see that unbiasedness is a property of a statistic that relates to a parameter, but does not depend on the value of the parameter; hence, by definition, unbiasedness of a point estimator is a *uniform property*.

Unbiasedness depends on the distribution of the observable, which in turn depends on the data-generating process.

Example 5.1 Sampling in a Bernoulli distribution

In Example 3.12, we considered the problem of making inferences about π using data-generating processes governed by a family of Bernoulli distributions with parameter π. In one case, the approach was to take a random sample of size n, $X_1, \ldots, X_n$ which are iid as Bernoulli(π). This yielded a data-generating process in which $T = \sum X_i$ has a binomial distribution with parameters n and π. In another approach we took a sequential sample, $X_1, X_2, \ldots$, from the Bernoulli(π) until a fixed number t of 1's have occurred. In this data-generating process, the sample size N is random, and it is modeled by a negative binomial with parameters t and π. (Note that in a common formulation of a negative binomial distribution, the random variable is $N - t$ in the formulation we are using here. In the present formulation, $N \geq t$.)

In Examples 4.6 and 4.7 we see that the estimator of π under a squared error loss and a beta prior is the same for the two distributions that result from the two data-generating processes, and neither of them is unbiased.

In the first data-generating process, we see that an unbiased estimator of π is

$$W = \frac{T}{n}. \tag{5.2}$$

In the second data-generating process, we see that an unbiased estimator of π is

$$U = \begin{cases} \frac{t-1}{N-1} & \text{if } N > 1 \\ 1 & \text{otherwise} \end{cases} \tag{5.3}$$

(Exercise 5.1). The latter estimator is essentially the same as the former one, because by the definition of the data-generating process, the last observation does not count because its value is determined a priori. ∎

Estimability

A statistical function for which there is an unbiased estimator is said to be *U-estimable*. We often refer to such estimands simply as "estimable". There are estimands for which there is no unbiased estimator.

Example 5.2 an estimand that is not U-estimable
Consider the problem of estimating $1/\pi$ in binomial(n, π) for $\pi \in]0, 1[$. Suppose $T(X)$ is an unbiased estimator of $1/\pi$. Then

$$\sum_{x=0}^{n} T(x) \binom{n}{x} \pi^x (1-\pi)^{n-x} = 1/\pi.$$

If $1/\pi$ were U-estimable, the equation above would say that some polynomial in π is equal to $1/\pi$ for all $\pi \in]0, 1[$. That clearly cannot be; hence, $1/\pi$ is not U-estimable. Notice also as $\pi \to 0$, the left side tends to $T(0)$, which is finite, but the right side tends to ∞. ∎

Another related example, but one that corresponds to a more common parameter, is an estimator of the odds, $\pi/(1-\pi)$.

Example 5.3 another estimand that is not U-estimable
Consider the problem of estimating $\pi/(1-\pi)$ in binomial(n, π) for $\pi \in]0, 1[$. The possible realizations of the n Bernoulli trials are $(X_1, \ldots, X_n)$, where $X_i = 0$ or 1; hence, there are 2^n possibilities and any estimator T must take each realization into a number t_j, where j ranges from 1 to 2^n.

Now,

$$\mathrm{E}(T) = \sum_{j=1}^{2^n} t_j \pi^{n_j} (1-\pi)^{n-n_j},$$

where n_j is the number of ones in the j^{th} string of zeros and ones. If T is unbiased, then it must be the case that

$$\sum_{j=1}^{2^n} t_j \pi^{n_j} (1-\pi)^{n-n_j} = \frac{\pi}{1-\pi} \quad \forall \pi \in (0, 1).$$

But it is not possible that the polynomial in π on the left can equal $\pi/(1-\pi) \forall \pi \in (0, 1)$. ∎

Unbiasedness, while a uniform property, is not invariant to transformations. It is easy to see by simple examples that if $\mathrm{E}(T) = \theta$, in general, $\mathrm{E}(g(T)) \neq g(\theta)$.

Unbiasedness may lead to estimators that we would generally consider to be poor estimators, as the following example from Romano and Siegel (1986) shows.

Example 5.4 an unbiased estimator with poor properties
Consider the problem of using a sample of size 1 for estimating $g(\theta) = \mathrm{e}^{-3\theta}$ where θ is the parameter in a Poisson distribution. An unbiased is

$$T(X) = (-2)^X,$$

as you are asked to show in Exercise 5.2.

The estimator is ridiculous. It can be negative, even though $g(\theta) > 0$. It is increasing in the positive integers, even though $g(\theta)$ decreases over the positive integers. ∎

Degree of a Statistical Function

If a statistical function is estimable, we may ask how many observations are required to estimate it; that is, to estimate it unbiasedly. We refer to this number as the *degree* of the statistical function. Obviously, this depends on the distribution as well as the functional. A mean functional may not even be exist, for example, in a Cauchy distribution, but if the mean functional exists, it is estimable and its degree is 1. The variance functional in a normal distribution is estimable and its degree is 2 (see page 405).

5.1 Uniformly Minimum Variance Unbiased Point Estimation

An unbiased estimator may not be unique. If there are more than one unbiased estimator, we will seek one that has certain optimal properties.

5.1.1 Unbiased Estimators of Zero

Unbiased estimators of 0 play a useful role in UMVUE problems.

If $T(X)$ is unbiased for $g(\theta)$ then $T(X) - U(X)$ is also unbiased for $g(\theta)$ for any U such that $\mathrm{E}(U(X)) = 0$; in fact, **all unbiased estimators** of $g(\theta)$ belong to an equivalence class defined as

$$\{T(X) - U(X)\}, \tag{5.4}$$

where $\mathrm{E}_\theta(U(X)) = 0$.

In Theorem 5.2 and its corollary we will see ways that unbiased estimators of zero can be used to identify optimal unbiased estimators.

In some cases, there may be no such nontrivial $U(X)$ that yields a different unbiased estimator in (5.4). Consider for example, a single Bernoulli trial with probability of success π, yielding the random variable X, and consider $T(X) = X$ as an estimator of π. We immediately see that $T(X)$ is unbiased for π. Now, let S be an estimator of π, and let $S(0) = s_0$ and $S(1) = s_1$. For S to be unbiased, we must have

$$s_1\pi + s_0(1-\pi) = \pi,$$

but this means $(s_1 - s_0)\pi + s_0 = \pi$. This means $s_0 = 0$ and $s_1 = 1$; that is, $S(X) = T(X)$ for $X = 0$ or 1. In this case the unbiased point estimator is unique.

5.1.2 Optimal Unbiased Point Estimators

Restricting our attention to unbiased estimators, we return to the problem of selecting an estimator with uniform minimum risk (UMRU). We find that in general, no UMRUE exists for bounded loss functions. Such loss functions cannot be (strictly) convex. If, however, we consider only loss functions that are strictly convex, which means that they are unbounded, we may be able to find a UMRUE.

5.1.3 Unbiasedness and Squared-Error Loss; UMVUE

A squared-error loss function is particularly nice for an unbiased estimator that has a finite second moment, because in that case the expected loss is just the variance; that is, an unbiased estimator with minimum risk is an unbiased estimator with minimum variance.

Unbiasedness alone, of course, does not ensure that an estimator is good; the variance of the estimator may be quite large. Also, a biased estimator may in fact dominate a very good unbiased estimator; see Example 3.19 on page 272. In Theorem 3.10 on page 270, however, we saw that any bias in an admissible estimator under squared-error loss must have a negative correlation with the estimator.

The requirement of unbiasedness also protects us from "bad" estimators that have superior squared-error risk in some regions of the parameter space, such as in Example 3.1 on page 219. (The estimator in that example, of course, does not dominate the "good" estimator, as the shrunken estimator in Example 3.19 does.)

If the unbiased estimator has minimum variance among all unbiased estimators at each point in the parameter space, we say that such an estimator is a **uniformly** (for all values of θ) minimum variance unbiased estimator, that is, a UMVUE.

An unbiased estimator that has minimum variance among all unbiased estimators within a subspace of the parameter space is called a locally minimum variance unbiased estimator, or LMVUE.

UMVU is a special case of uniform minimum risk (UMRU), which generally only applies to convex loss functions.

Uniformity (the first "U") means the MVU property is independent of the estimand. "Unbiasedness" is itself a uniform property, because it is defined in terms of an expectation for any distribution in the given family.

UMVU is closely related to complete sufficiency, which means that it probably has nice properties (like being able to be identified easily) in exponential families. One of the most useful facts is the Lehmann-Scheffé theorem.

Theorem 5.1 (Lehmann-Scheffé Theorem)
Let T be a complete sufficient statistic for θ, and suppose T has finite second moment. If $g(\theta)$ is U-estimable, then there is a unique UMVUE of $g(\theta)$ of the form $h(T)$, where h is a Borel function.

The first part of this is just a corollary to the Rao-Blackwell theorem, Theorem 3.8. The uniqueness comes from the completeness, and of course, means unique a.e.

The Lehmann-Scheffé theorem may immediately identify a UMVUE.

Example 5.5 UMVUE of Bernoulli parameter
Consider the Bernoulli family of distributions with parameter π. Suppose we take a random sample $X_1, \ldots, X_n$. Now the Bernoulli (or in this case, the binomial(n, π)) is a complete one-parameter exponential family, and $T = \sum_{i=1}^{n} X_i$ is a complete sufficient statistic for π with expectation $n\pi$. By the Lehmann-Scheffé theorem, therefore, the unique UMVUE of π is

$$W = \sum_{i=1}^{n} X_i/n. \tag{5.5}$$

In Example 3.17, page 269, we showed that the variance of W achieves the CRLB; hence it must be UMVUE.

The random sample from a Bernoulli distribution is the same as a single binomial observation, and W is an unbiased estimator of π, as in Example 5.1. We also saw in that example that a constrained random sample from a Bernoulli distribution is the same as a single negative binomial observation N, and an unbiased estimator of π in that case is $(t-1)/(N-1)$, where t is the required number of 1's in the constrained random sample. This estimator is also UMVU for π (Exercise 5.1).

In the usual definition of this family, $\pi \in \Pi =]0, 1[$. Notice that if $\sum_{i=1}^{n} X_i = 0$ or if $\sum_{i=1}^{n} X_i = n$, $W \notin \Pi$. Hence, the UMVUE may not be valid in the sense of being a legitimate parameter for the probability distribution. ∎

Useful ways for checking that an estimator is UMVU are based on the following theorem and corollary.

Theorem 5.2
Let $\mathcal{P} = \{P_\theta\}$. Let T be unbiased for $g(\theta)$ and have finite second moment. Then T is a UMVUE for $g(\theta)$ iff $\mathrm{E}(TU) = 0\ \forall \theta \in \Theta$ and $\forall U \ni \mathrm{E}(U) = 0, \mathrm{E}(U^2) < \infty\ \forall \theta \in \Theta$.

Proof.
First consider "only if".

Let T be UMVUE for $g(\theta)$ and let U be such that $\mathrm{E}(U) = 0$ and $\mathrm{E}(U^2) < \infty$. Let c be any fixed constant, and let $T_c = T + cU$; then $\mathrm{E}(T) = g(\theta)$. Since T is UMVUE,

$$\mathrm{V}(T_c) \geq \mathrm{V}(T), \quad \forall \theta \in \Theta,$$

or

$$c^2\mathrm{V}(U) + 2c\mathrm{Cov}(T, U) \geq 0, \quad \forall \theta \in \Theta.$$

This implies $\mathrm{E}(TU) = 0\ \forall \theta \in \Theta$.

Now consider "if".
Assume $\mathrm{E}(T) = g(\theta)$ and $\mathrm{E}(T^2) < \infty$ and U is such that $\mathrm{E}(TU) = 0, \mathrm{E}(U) = 0, \mathrm{E}(U^2) < \infty\ \forall \theta \in \Theta$. Now let T_0 be an unbiased estimator of $g(\theta)$, that is, $\mathrm{E}(T_0) = g(\theta)$. Therefore, because $\mathrm{E}(TU) = 0$, $\mathrm{E}(T(T - T_0)) = 0\ \forall \theta \in \Theta$, and so $\mathrm{V}(T) = \mathrm{Cov}(T, T_0)$. Therefore, because $(\mathrm{Cov}(T, T_0))^2 \leq \mathrm{V}(T)\mathrm{V}(T_0)$, we have

$$\mathrm{V}(T) \leq \mathrm{V}(T_0)\ \forall \theta \in \Theta$$

implying that T is UMVUE. ∎

Corollary 5.2.1
Let $\widetilde{T}$ be a sufficient statistic for θ, and let $T = h(\widetilde{T})$ where h is a Borel function. Let r be any Borel function such that for $\widetilde{U} = r(\widetilde{T})$, $\mathrm{E}(\widetilde{U}) = 0$ and $\mathrm{E}(\widetilde{U}^2) < \infty\ \forall \theta \in \Theta$. Then T is a UMVUE for $g(\theta)$ iff $\mathrm{E}(T\widetilde{U}) = 0\ \forall \theta \in \Theta$.

Proof.
This follows from the theorem because if $\mathrm{E}(T\widetilde{U}) = 0\ \forall \theta \in \Theta$, then $\forall \theta \in \Theta$, $\mathrm{E}(TU) = 0$, $\mathrm{E}\left(\mathrm{E}(U|\widetilde{T})\right) = 0$, and $\mathrm{E}\left(\mathrm{E}(U|\widetilde{T})^2\right) < \infty$. This is the case because

$$\mathrm{E}(TU) = \mathrm{E}\left(\mathrm{E}\left(TU|\widetilde{T}\right)\right) = \mathrm{E}\left(\mathrm{E}\left(h\left(\widetilde{T}\right)U|\widetilde{T}\right)\right) = \mathrm{E}\left(h\left(\widetilde{T}\right)\mathrm{E}\left(U|\widetilde{T}\right)\right).$$

∎

How to find a UMVUE

We have seen how that is some cases, the Lehmann-Scheffé theorem may immediately identify a UMVUE.

In more complicated cases, we generally find an UMVUE by beginning with a "good" estimator and manipulating it to make it UMVUE. It might be unbiased to begin with, and we reduce its variance while keeping it unbiased. It might not be unbiased to begin with but it might have some other desirable property, and we manipulate it to be unbiased.

If we have a complete sufficient statistic T for θ, the Lehmann-Scheffé theorem leads to two methods. Another method uses unbiased estimators of zero and is based on the equivalence class of unbiased estimators (5.4).

1. Given the complete sufficient statistic T for θ, find a function of T that makes it unbiased; that is, find a UMVUE directly by finding $h(T)$ such that $\mathrm{E}_\theta(h(T)) = g(\theta)$.
2. Given the complete sufficient statistic T for θ and another statistic T_0 that is unbiased, condition the unbiased statistic on the complete sufficient statistic; that is, find a UMVUE as $h(T) = \mathrm{E}_\theta(T_0(X)|T)$. (This process is sometimes called "Rao-Blackwellization".)
3. Let T_0 be such that $\mathrm{E}_\theta(T_0) = g(\theta)$ and $\mathrm{E}_\theta(T_0^2) < \infty$. We find a UMVUE by finding U where $\mathrm{E}_\theta(U) = 0$ so as to minimize $\mathrm{E}((T_0 - U)^2)$. Useful estimators clearly must have finite second moment, otherwise, we cannot minimize a variance by combining the estimators. This method makes use of the equivalence class of unbiased estimators.

We will now consider examples of each of these methods.

Finding an UMVUE by Forming an Unbiased Function of a Complete Sufficient Statistic

Example 5.6 UMVUE of various parametric functions in a normal distribution

Let $X_1, X_2, \ldots, X_n$ be a random sample from a $\mathrm{N}(\mu, \sigma^2)$ distribution with unknown $\theta = (\mu, \sigma^2)$. (Notice that $n \geq 2$.) In Example 3.6, we have seen that $T = (\overline{X}, S^2)$ is sufficient and complete for θ.

For various $g(\theta)$ we will find the UMVUEs directly by finding $h(T)$ such that $\mathrm{E}_\theta(h(T)) = g(\theta)$:

- for $g(\theta) = \mu$:
$$h(T) = \overline{X} \tag{5.6}$$
- for $g(\theta) = \sigma^2$:
$$h(T) = S^2 \tag{5.7}$$
- for $g(\theta) = \mu^2$:
$$h(T) = \overline{X}^2 - S^2/n \tag{5.8}$$
- for $g(\theta) = \sigma^p$, with $p \geq 2$:
$$h(T) = \frac{(n-1)^{p/2}\Gamma((n-1)/2)}{2^{p/2}\Gamma((n-1+p)/2)} S^p \tag{5.9}$$
- for $g(\theta) = \mu/\sigma$ if $n \geq 3$:
$$h(T) = \frac{2^{1/2}\Gamma((n-1)/2)}{(n-1)^{1/2}\Gamma((n-2)/2)} \overline{X}/S. \tag{5.10}$$

We get the last two estimators by using the fact that $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$. ∎

Example 5.7 UMVUE of the variance in a Bernoulli distribution
Given random sample of size n from Bernoulli(π). We want to estimate $g(\pi) = \pi(1-\pi)$. We have a complete sufficient statistic, $T = \sum X_i$. The unbiasedness condition is

$$\sum_{t=0}^{n} \binom{n}{t} h(t)\pi^t(1-\pi)^{n-t} = \pi(1-\pi).$$

Rewriting this in terms of the odds $\rho = \pi/(1-\pi)$, we have, for all $\rho \in]0, \infty[$,

$$\begin{aligned}\sum_{t=0}^{n} \binom{n}{t} h(t)\rho^t &= \rho(1+\rho)^{n-2} \\ &= \sum_{t=1}^{n-1} \binom{n-2}{t-1} \rho^t.\end{aligned}$$

Now since for each t, the coefficient of ρ^t must be the same on both sides of the equation, we have the UMVUE of the Bernoulli variance to be

$$\frac{\sum x_i(n - \sum x_i)}{n(n-1)}. \tag{5.11}$$

Note that this is the same estimator as 5.7 for the variance in a normal distribution. ∎

Example 5.8 UMVUE of the upper limit in a uniform distribution
Consider the uniform distribution $\mathrm{U}(0, \theta)$. In Example 3.7 we saw that $X_{(n)}$ is complete sufficient for θ. An UMVUE for θ therefore is $(1 + 1/n)X_{(n)}$. ∎

Also see Example 3.1 in MS2.

Example 5.9 UMVUE in a two-parameter exponential distribution

Lebesgue PDF of the two-parameter exponential with parameter (α, θ) is

$$\theta^{-1}\mathrm{e}^{-(x-\alpha)/\theta}\mathrm{I}_{]\alpha,\infty[}(x)$$

Suppose we have observations $X_1, X_2, \ldots, X_n$. In Examples 1.11 and 1.18, we found the distributions of $X_{(1)}$ and $\sum X_i - nX_{(1)}$, and in Example 3.8 we showed that $T = (X_{(1)}, \sum X_i - nX_{(1)})$ is sufficient and complete for (α, θ). Hence, all we have to do is adjust them to be unbiased.

$$T_\alpha = X_{(1)} - \frac{1}{n(n-1)} \sum (X_i - X_{(1)}).$$

and

$$T_\theta = \frac{1}{n-1} \sum (X_i - X_{(1)}).$$

∎

Also see Example 3.2 in MS2.

UMVUE by Conditioning an Unbiased Estimator on a Sufficient Statistic

See Example 3.3 in MS2.

UMVUE by Minimizing the Variance within the Equivalence Class of Unbiased Estimators

******

5.1.4 Other Properties of UMVUEs

In addition to the obvious desirable properties of UMVUEs, we should point out that UMVUEs lack some other desirable properties. We can do this by citing examples.

First, as in Example 5.5, we see that the UMVUE may not be in the parameter space.

The next example shows that the UMVUE may not be a minimax estimation, even under the same loss function, that is, squared-error.

Example 5.10 UMVUE that is not minimax (continuation of Example 5.5)
Consider a random sample of size n from the Bernoulli family of distributions with parameter π. The UMVUE of π is $T = X/n$. Under the squared-error loss, the risk, that is, the variance in this case is $\pi(1-\pi)/n$. This is the smallest risk possible for an unbiased estimator, by inequality (3.39).

The maximum risk for T is easily seen to be $1/(4n)$ (when $\pi = 1/2$). Now, consider the estimator

$$T^* = \frac{X}{n}\frac{n^{1/2}}{1+n^{1/2}} + \frac{1}{2(1+n^{1/2})}.$$

This has risk

$$\begin{aligned}
R(T^*,\pi) &= \mathrm{E}_\pi((T^*-\pi)^2) \\
&= \mathrm{E}_\pi\left(\left(\frac{X}{n}\frac{n^{1/2}}{1+n^{1/2}} + \frac{\pi n^{1/2}}{2(1+n^{1/2})} - \frac{\pi n^{1/2}}{2(1+n^{1/2})} + \frac{1}{2(1+n^{1/2})} - \pi\right)^2\right) \\
&= \left(\frac{n^{1/2}}{1+n^{1/2}}\right)^2 \mathrm{E}_\pi\left(\left(\frac{X}{n}-\pi\right)^2\right) + \left(\frac{\pi n^{1/2}}{1+n^{1/2}} + \frac{1}{2(1+n^{1/2})} - \pi\right)^2 \\
&= \left(\frac{n^{1/2}}{1+n^{1/2}}\right)^2 \frac{\pi(1-\pi)}{n} + \left(\frac{1-2\pi}{2(1+n^{1/2})}\right)^2 \\
&= \frac{1}{4(1+n^{1/2})^2}. \qquad (5.12)
\end{aligned}$$

The risk of T^* is less than the maximum risk of T; therefore, T is not minimax.

Now we might ask is T^* minimax?

We first note that the risk (5.12) is constant, so T^* is minimax if it is admissible or if it is a Bayesian estimator (in either case with respect to the squared-error loss). We can see that T^* is Bayesian estimator (with a beta prior). (You are asked to prove this in Exercise 4.10 on page 383.) As we show in Chapter 4, a Bayes estimator with a constant risk is a minimax estimator; hence, δ^* is minimax. (This example is due to Lehmann.) ∎

Although we may initially be led to consideration of UMVU estimators by consideration of a squared-error loss, which leads to a mean squared-error risk, the UMVUE may not minimize the MSE. It was the fact that we could not minimize the MSE *uniformly* that led us to add on the requirement of unbiasedness. There may, however, be estimators that have a uniformly smaller MSE than the UMVUE. An example of this is in the estimation of the variance in a normal distribution. In Example 5.6 we have seen that the UMVUE of σ^2 in the normal distribution is S^2, while in Example 3.13 we have seen that the MLE of σ^2 is $(n-1)S^2/n$, and by equation (3.55) on page 243, we see that the MSE of the MLE is uniformly less than the MSE of the UMVUE.

There are other ways in which UMVUEs may not be very good as estimators; see, for example, Exercise 5.2. A further undesirable property of UMVUEs is that they are not invariant to transformation.

5.1.5 Lower Bounds on the Variance of Unbiased Estimators

The three Fisher information regularity conditions (see page 168) play a major role in UMVUE. In particular, these conditions allow us to develop a lower bound on the variance of any unbiased estimator.

The Information Inequality (CRLB) for Unbiased Estimators

What is the smallest variance an unbiased estimator can have? For an unbiased estimator T of $g(\theta)$ in a family of densities satisfying the regularity conditions and such that T has a finite second moment, the answer results from inequality (3.83) on page 256 for the scalar estimator T and estimand $g(\theta)$. (Note that θ itself may be a vector.) That is the information inequality or the Cramér-Rao lower bound (CRLB), and it results from the covariance inequality.

If $g(\theta)$ is a vector, then $\partial g(\theta)/\partial\theta$ is the Jacobian, and we have

$$\mathrm{V}(T(X)) \succeq \left(\frac{\partial}{\partial\theta} g(\theta)\right)^{\mathrm{T}} (I(\theta))^{-1} \frac{\partial}{\partial\theta} g(\theta), \tag{5.13}$$

where we assume the existence of all quantities in the expression.

Note the meaning of this relationship in the multiparameter case: it says that the matrix

$$\mathrm{V}(T(X)) - \left(\frac{\partial}{\partial\theta}g(\theta)\right)^{\mathrm{T}} (I(\theta))^{-1} \frac{\partial}{\partial\theta}g(\theta) \tag{5.14}$$

is nonnegative definite. (This includes the zero matrix; the zero matrix is nonnegative definite.)

Example 5.11 Fisher efficiency in a normal distribution
Consider a random sample $X_1, X_2, \ldots, X_n$ from the $\mathrm{N}(\mu, \sigma^2)$ distribution. In Example 3.9, we used the parametrization $\theta = (\mu, \sigma)$. Now we will use the parametrization $\theta = (\mu, \sigma^2)$. The joint log density is

$$\log p_{(\mu,\sigma)}(x) = c - \frac{n}{2}\log(\sigma^2) - \sum_i (x_i - \mu)^2/(2\sigma^2). \tag{5.15}$$

The information matrix is diagonal, so the inverse of the information matrix is particularly simple:

$$I(\theta)^{-1} = \begin{bmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{\sigma^4}{2(n-1)} \end{bmatrix}. \tag{5.16}$$

For the simple case of $g(\theta) = (\mu, \sigma^2)$, we have the unbiased estimator,

$$T(X) = \left(\overline{X},\ \sum_{i=1}^{n}(X_i - \overline{X})^2/(n-1)\right),$$

and

$$\mathrm{V}(T(X)) = \begin{bmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{\sigma^4}{2(n-1)} \end{bmatrix}, \tag{5.17}$$

which is the same as the inverse of the information matrix. The estimators are Fisher efficient. ∎

It is important to know in what situations an unbiased estimator can achieve the CRLB. Notice this would depend on both $p(X, \theta)$ and $g(\theta)$. Let us consider this question for the case of scalar θ and scalar function g. The necessary and sufficient condition that an estimator T of $g(\theta)$ attain the CRLB is that $(T - g(\theta))$ be proportional to $\partial \log(p(X, \theta))/\partial\theta$ a.e.; that is, for some a that does not depend on X,

$$\frac{\partial \log(p(X,\theta))}{\partial\theta} = a(\theta)(T - g(\theta)) \quad \text{a.e.} \tag{5.18}$$

This means that the CRLB can be obtained by an unbiased estimator only in the one-parameter exponential family.

For example, there are unbiased estimators of the mean in the normal, Poisson, and binomial families that attain the CRLB. There is no unbiased

estimator of θ that attains the CRLB in the family of distributions with Lebesgue densities proportional to $(1+(x-\theta)^2)^{-1}$ (this is the Cauchy family).

If the CRLB is attained for an estimator of $g(\theta)$, it cannot be attained for any other (independent) function of θ. For example, there is no unbiased estimator of μ^2 in the normal distribution that achieves the CRLB.

If the CRLB is not sharp, that is, if it cannot be attained, there may be other (larger) bounds, for example the Bhattacharyya bound. These sharper bounds are usually based on higher-order derivatives.

The following example is from Romano and Siegel (1986), who attribute it to Bickel and Doksum.

Example 5.12 UMVUE in Exponential Family That Does Not Attain the CRLB

Let X have a Poisson distribution with PDF

$$p(x) = \theta^y \mathrm{e}^{-\theta}/y!, \quad y = 0, 1, 2, \ldots,$$

and suppose we want to estimate $g(\theta) = \mathrm{e}^{-\theta}$.

For a sample of size 1, let

$$T(X) = \begin{cases} 1 \text{ if } X = 0 \\ 0 \text{ otherwise.} \end{cases}$$

We see that $T(X)$ has expectation $\mathrm{e}^{-\theta}$ and so is unbiased for the estimand. We know that X is sufficient, and we see that it is complete by considering a function g such that $\mathrm{E}(g(X)) = 0$ for all $\theta > 0$. For such a function, for all $\theta > 0$, we have

$$\mathrm{e}^{-\theta} \sum_{i=0}^{\infty} g(i) \frac{1}{i!} \theta^i = 0.$$

A power series that is identically zero in an interval must have all coefficients zero, and so $g(x) = 0$ a.e.; hence, X is complete.

Now, by the Lehmann-Scheffé theorem, $T = \mathrm{E}(T|X)$ is UMVUE for $\mathrm{e}^{-\theta}$, and since it has finite variance, $\mathrm{V}(T) = \mathrm{e}^{-\theta}(1-\mathrm{e}^{-\theta})$, it is the unique UMVUE.

We can work out the Fisher information to be

$$\begin{aligned} I\theta &= \mathrm{E}\left(\left(\frac{\partial \log(p(X;\theta))}{\partial \theta}\right)^2\right) \\ &= \mathrm{E}\left(\left(-1 + \frac{X}{\theta}\right)^2\right) \\ &= \frac{1}{\theta^2}\mathrm{E}(X^2) - \frac{2}{\theta}\mathrm{E}(X) + 1 \\ &= \frac{1}{\theta}. \end{aligned}$$

Hence, the CRLB for the variance of unbiased estimators of $g(\theta) = \mathrm{e}^{-\theta}$ is $\theta\mathrm{e}^{-2\theta}$. By expanding $\mathrm{e}^{-\theta}$ in a Taylor series, we see that $\mathrm{V}(T) = \mathrm{e}^{-\theta}(1-\mathrm{e}^{-\theta}) > \theta\mathrm{e}^{-2\theta}$; hence, the UMVUE does not attain the CRLB. ∎

The Bhattacharyya Lower Bound

We now consider a simple case in which θ is a scalar (and, hence the estimand $g(\theta)$ and the estimator $T(X)$ are scalars).

For the PDF $f(x;\theta)$ and the Borel scalar function $g(\theta)$ assume that each is differentiable r times, and write

$$f^{(r)} = \frac{\partial^r f(x;\theta)}{\partial^r}$$

and

$$g^{(r)} = \frac{\partial^r g(\theta)}{\partial^r}.$$

Let T be an unbiased estimator of $g(\theta)$.

Now, form the function

$$D_s = T - g(\theta) - \sum_{r=1}^{s} a_r f^{(r)}/f, \tag{5.19}$$

where the a_r are constants to be determined. Now, we have

$$\mathrm{E}(f^{(r)}/f) = 0 \tag{5.20}$$

as before, and since T be an unbiased estimator for $g(\theta)$, we have

$$\mathrm{E}(D_s) = 0.$$

The variance of D_s is therefore,

$$\mathrm{E}(D_s^2) = \int \left(T - g(\theta) - \sum_{r=1}^{s} a_r f^{(r)}/f \right)^2 f \mathrm{d}x. \tag{5.21}$$

We now seek to minimize this quantity in the a_r. To do so, for $p = 1, \ldots, s$, we differentiate and set equal to zero:

$$\int \left(T - g(\theta) - \sum_{r=1}^{s} a_r f^{(r)}/f \right) (f^{(p)}/f) f \mathrm{d}x = 0, \tag{5.22}$$

which yields

$$\int (T - g(\theta)) f^{(p)} \mathrm{d}x = \sum_{r=1}^{s} a_r \int \frac{f^{(r)}}{f} \frac{f^{(p)}}{f} f \mathrm{d}x. \tag{5.23}$$

Because of (5.20), the left-hand side of (5.23) is

$$\int T f^{(p)} \mathrm{d}x = g^{(p)}(\theta). \tag{5.24}$$

(Compare this with $\int Tf\mathrm{d}x = g(\theta)$.)

The right-hand side of (5.23) is

$$\sum_{r=1}^{s} a_r \mathrm{E}\left(\frac{f^{(r)}}{f}\frac{f^{(p)}}{f}\right).$$

Substituting back into (5.23) we have

$$g^{(p)}(\theta) = \sum_{r=1}^{s} a_r \mathrm{E}\left(\frac{f^{(r)}}{f}\frac{f^{(p)}}{f}\right), \tag{5.25}$$

for $p = 1, \ldots, s$. If the matrix of coefficients of the a_r is nonsingular, we can invert them to solve. For notational simplicity, let,

$$J_{rp} = \mathrm{E}\left(\frac{f^{(r)}}{f}\frac{f^{(p)}}{f}\right).$$

Then

$$a_r = \sum_{p=1}^{s} g^{(p)}(\theta) J_{rp}^{-1}.$$

Hence, at its minimum value

$$D_s = T - g(\theta) - \sum_{r=1}^{s}\sum_{p=1}^{s} g^{(p)}(\theta) J_{rp}^{-1} f^{(r)}/f. \tag{5.26}$$

and the variance of D_s from (5.21) is

$$\mathrm{E}(D_s^2) = \int \left(T - g(\theta) - \sum_{r=1}^{s}\sum_{p=1}^{s} g^{(p)}(\theta) J_{rp}^{-1} f^{(r)}/f\right)^2 f\mathrm{d}x. \tag{5.27}$$

We now use the fact that the derivative is zero, equation (5.22), to get

$$\mathrm{E}(D_s^2) = \int (T - g(\theta))^2 f\mathrm{d}x - \sum_{r=1}^{s}\sum_{p=1}^{s} g^{(p)}(\theta) J_{rp}^{-1} \int T f^{(r)}\mathrm{d}x, \tag{5.28}$$

which, because T is unbiased using equation (5.24), yields

$$\mathrm{E}(D_s^2) = \mathrm{V}(T) - \sum_{r=1}^{s}\sum_{p=1}^{s} g^{(p)}(\theta) J_{rp}^{-1} g^{(r)}(\theta).$$

Finally, because the left-hand side of this is nonnegative, we have the *Bhattacharyya bound* on the variance of T:

$$\mathrm{V}(T) \geq \sum_{r=1}^{s}\sum_{p=1}^{s} g^{(p)}(\theta) J_{rp}^{-1} g^{(r)}(\theta). \tag{5.29}$$

Notice that in the case of $s = 1$, this is the same as the CRLB.

5.2 U-Statistics

In estimation problems it is often fruitful to represent the estimand as some functional of the CDF, P. The mean, for example, if it exists is

$$M(P) = \int x \, \mathrm{d}P. \tag{5.30}$$

Given the exchangeable random variables $X_1, \ldots, X_n$ with CDF P, we can form a plug-in estimator of $M(P)$ by applying the functional to the ECDF.

In more complicated cases, the property of interest may be the quantile associated with π, that is, the unique value y_π defined by

$$\Xi_\pi(P) = \inf_y \{y : P(y) \geq \pi\}. \tag{5.31}$$

There is a basic difference in the functionals in equations (5.30) and (5.31). The first is an expected value, $\mathrm{E}(X_i)$ for each i. The second functional, however, cannot be written as an expectation. (Bickel and Lehmann (1969) showed this.)

5.2.1 Expectation Functionals and Kernels

In the following, we will consider the class of statistical functions that can be written as an expectation of a function h of some subsample, $X_{i_1}, \ldots, X_{i_m}$, where $i_1, \ldots, i_m$ are distinct elements of $\{1, \ldots, n\}$:

$$\begin{aligned} \theta &= \Theta(P) \\ &= \mathrm{E}(h(X_{i_1}, \ldots, X_{i_m})). \end{aligned} \tag{5.32}$$

Such Θs are called *expectation functionals*. The function h is called the *kernel* of the expectation functional. The number of arguments of the kernel is called the *order of the kernel*.

In the case of M in equation (5.30) above, h is the identity and the order m is 1.

Notice that we have unbiasedness of the kernel function for θ by the way we define the terms.

Expectation functionals that relate to parameter of interest are often easy to define. The simplest is just $\mathrm{E}(h(X_i))$. The utility of expectation functionals lies in the ease of working with them coupled with some useful general properties.

Note that without loss of generality we can assume that h is *symmetric* in its arguments because the X_is are exchangeable, and so even if h is not symmetric, any permutation $(i_1, \ldots, i_m)$ of the indexes has the same expectation, so we could form a function that is symmetric in the arguments and has the same expectation:

$$\bar{h}(X_1, \ldots, X_m) = \frac{1}{m!} \sum_{\text{all permutations}} h(X_{i_1}, \ldots, X_{i_m}).$$

Example 5.13 symmetric kernel
If $X_1 \overset{\mathrm{d}}{=} X_2 \overset{\mathrm{d}}{=} X$ and X_1, X_2, and X are exchangeable, we can write the variance of the random variable X as

$$\mathrm{V}(X) = \mathrm{E}(X_1^2) - \mathrm{E}(X_1)\mathrm{E}(X_2).$$

This may suggest the kernel

$$h(x_1, x_2) = x_1^2 - x_1 x_2, \tag{5.33}$$

which is not symmetric, that is, $h(x_1, x_2) \neq h(x_2, x_1)$. We can, however, form a kernel that is equivalent (in the sense of expected value) by a linear combination (with equal weights):

$$\begin{aligned} \bar{h}(x_1, x_2) &= \frac{1}{2}\left(h(x_1, x_2) + h(x_2, x_1)\right) \\ &= \frac{1}{2}(x_1 - x_2)^2, \end{aligned} \tag{5.34}$$

which is symmetric. ∎

Because of the symmetry, we will just need to consider h evaluated over the possible combinations of m items from the sample of size n. Furthermore, because the X_{i_j} are exchangeable, the properties of $h(X_{i_1}, \ldots, X_{i_m})$ are the same as the properties of $h(X_1, \ldots, X_m)$.

Degree of Expectation Functional

We might wonder what is the minimum number of arguments a kernel that is associated with a given expectation functional must have.

Example 5.14
Consider a single observation X from a $\mathrm{N}(\mu, \sigma^2)$ distribution with both μ and σ^2 unknown. Is there an unbiased estimator of σ^2 based on X? Suppose $T(X)$ is unbiased for σ^2. Now, suppose $\sigma^2 = \sigma_0^2$, some fixed value; that is, $\mathrm{E}(T(X)) = \sigma_0^2$. Because X is complete sufficient statistic for μ, $\mathrm{E}(T(X)) = \sigma_0^2$ for all μ implies $T(X) = \sigma_0^2$ a.e.; that is, $T(X)$ cannot be unbiased for σ^2.

We have seen that we do have an unbiased estimator of the variance from a sample of size 2, X_1 and X_2. It is the sample variance, which can be written as $\frac{1}{2}(X_1 - X_2)^2$, as suggested in Example 5.13. ∎

The analysis in the previous example leads us to the concept of the degree of an expectation functional or statistical function (see page 392). This is the minimum number of observations that can be combined in such a way that the expectation of the combination is the given functional. From the facts above, we see that the degree of the variance functional is 2.

5.2.2 Kernels and U-Statistics

Now consider the estimation of the expectation functional $\Theta(P)$ in equation (5.32), given a random sample $X_1, \ldots, X_n$, where $n \geq m$.

Clearly $h(X_1, \ldots, X_m)$ is an unbiased estimator of $\theta = \Theta(P)$, and so is $h(X_{i_1}, \ldots, X_{i_m})$ for any m-tuple, $1 \leq i_1 < \cdots < i_m \leq n$; hence, we have that

$$U = \frac{1}{\binom{n}{m}} \sum_{\text{all combinations}} h(X_{i_1}, \ldots, X_{i_m}) \tag{5.35}$$

is unbiased for θ.

A statistic of this form is called a *U-statistic*. The U-statistic is a function of all n items in the sample. The function h, which is called the *kernel of the U-statistic* is a function of m arguments. We also refer to the order of the kernel as the *order of the U-statistic*.

Examples

Example 5.15 r^{th} **raw moment:** $M'_r(P) = \mathrm{E}(X^r)$
In the simplest U-statistic for $r = 1$, the kernel is of order 1 and h is the identity, $h(x) = x$. This is just the sample mean. More generally, we have the r^{th} raw population moment by defining $h_r(x_i) = x_i^r$, yielding the first order U-statistic

$$U(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} X_i^r,$$

which is the r^{th} sample moment.

(The notation h_r will be used differently below.***) ∎

Example 5.16 r^{th} **power of the mean:** $(\mathrm{E}(X))^r$
Another simple U-statistic with expectation $(\mathrm{E}(X))^r$ where the r^{th} order kernel is $h(x_1, \ldots, x_r) = x_1 \cdots x_r$. The U-statistic

$$U(X_1, \ldots, X_n) = \frac{1}{\binom{n}{r}} \sum_{\text{all combinations}} X_{i_1} \cdots X_{i_r}$$

has expectation $(\mathrm{E}(X))^r$. ∎

Example 5.17 $\Pr(X \leq a)$**:** $\Theta(P) = \mathrm{E}\left(\mathrm{I}_{]\infty,a]}(X)\right) = P(a)$
Compare this with the quantile functional in equation (5.31), which cannot be expressed as an expectation functional. The quantile problem is related to an inverse problem in which the property of interest is the π; that is, given a value a, estimate $P(a)$. We can write an expectation functional and arrive at the U-statistic

$$\begin{aligned} U(X_1, \ldots, X_n) &= \frac{1}{\binom{n}{1}} \sum_{i=1}^{n} I_{]-\infty,a]}(X_i) \\ &= P_n(a), \end{aligned}$$

where P_n is the ECDF. ∎

Example 5.18 Gini's mean difference

$$\Theta(P) = \mathrm{E}\left(|X_1 - X_2|\right)$$

A familiar second order U-statistic is *Gini's mean difference*, in which $h(x_1, x_2) = |x_2 - x_1|$, for $n \geq 2$,

$$U = \frac{1}{\binom{n}{2}} \sum_{i<j} |X_j - X_i|. \tag{5.36}$$

∎

Example 5.19 covariance: $\Sigma(P_{YZ}) = \mathrm{Cov}(Y, Z)$
Let $X = (Y, Z)$. We form the second order kernel

$$h(x_1, x_2) = \frac{1}{2}(y_1 - y_2)(z_1 - z_2), \tag{5.37}$$

where $x_i = (y_i, z_i)$. We see that

$$\begin{aligned} \mathrm{E}(h(X_1, X_2)) &= \frac{1}{2}\left(\mathrm{E}(Y_1 Z_1) - \mathrm{E}(Y_1)\mathrm{E}(Z_2) + \mathrm{E}(Y_2 Z_2) - \mathrm{E}(Y_1)\mathrm{E}(Z_2)\right) \\ &= \mathrm{Cov}(U, Z). \end{aligned}$$

We form the U-statistic

$$U(X_1, \ldots, X_n) = \frac{1}{\binom{n}{2}} \sum_{i<j}^{n} h(X_i, X_j). \tag{5.38}$$

This U-statistic is the sample covariance $S^2_{y,z}$, that is,

$$U(x_1, \ldots, x_n) = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})(z_i - \bar{z}),$$

which is unbiased for the population covariance if it exists.

Notice that if $Y = Z$, the U-statistic

$$U(X_1, \ldots, X_n) = \frac{2}{n(n-1)} \sum_{i<j}^{n} h(X_i, X_j). \tag{5.39}$$

is the sample variance S^2, which is unbiased for the population variance if it exists. The kernel (5.37) is the same as (5.33), which we put in the symmetric form (5.34). ∎

Partitioning the Sample

Notice that while the kernel is a function of only m arguments, U is a function of all n random variables, $U(X_1, \ldots, X_n)$.

Some useful statistical techniques involve partitioning of a given sample. The jackknife (see Section 3.6.1 beginning on page 301) is based on a systematic partitioning of the sample, and the bootstrap (see Section 3.6.2 beginning on page 304) is based on a random resampling of the sample.

If we index the elements of a given sample of size n as $\{1, \ldots, n\}$, for given a integer r with $1 \leq r \leq n$, we may form the sample with indexes

$$\mathcal{S} = \{i_1, \ldots, i_r\} \subseteq \{1, \ldots, n\}.$$

Corresponding to a statistic T_n computed from the full sample, we often use the notation $T_{r,\mathcal{S}}$ to denote the corresponding statistic computed from the sample indexed by $\mathcal{S}$; that is,

$$T_{r,\mathcal{S}} = T_r(X_{i_1}, \ldots, X_{i_r}).$$

To analyze statistical properties of the U statistic, we need to know which elements of the sample occur in each term in the sum (5.35) over all combinations. Sometimes it is useful to order these combinations in a systematic way. A *lexicographic ordering* is often the best way to do this. In one lexicographic ordering, we write the labels as an m-tuple $(i_1, \ldots, i_m)$ and index the set of combinations such that $(i_1, \ldots, i_m)$ is less than $(j_1, \ldots, j_m)$, if $i_1 < j_1$ or else if for $r > 1$, $i_k = j_k$ for $k < r$ and $i_r < j_r$. This ordering makes it easy to identify a pattern of the terms in the sum (5.35) in which any particular X_i appears. The element X_1, for example, appears in the first $\binom{n-1}{m-1}$ terms, and the element X_2, appears in the first $\binom{n-2}{m-2}$ terms and in the $\binom{n-2}{m-3}$ terms following the first $\binom{n-1}{m-1}$ terms. Hence, X_1 and X_2 occur together in $\binom{n-2}{m-2}$ terms. These patterns become increasingly complicated of course.

It is instructive to note some simple results of the sum of $T_{r,\mathcal{S}}$, for various forms of T_n, over all combinations of a given sample, as in equation (5.35). We will denote such a summation as

$$\sum_{\mathrm{C}} T_{r,\mathcal{S}}.$$

Now, as an example, let

$$T_n = \sum_{i=1}^{n} X_i/n = \overline{X}.$$

Then

$$T_n^2 = \frac{1}{n^2}\left(\sum_{i=1}^{n} X_i^2 + \sum_{i \neq j} X_i X_j\right), \tag{5.40}$$

$$\sum_{\mathrm{C}} T_{r,\mathcal{S}} = \binom{n-1}{r-1}\frac{n}{r}T_n, \tag{5.41}$$

$$\begin{aligned}\sum_{\mathrm{C}} T_n T_{r,\mathcal{S}} &= T_n \sum_{\mathrm{C}} T_{r,\mathcal{S}} \\ &= \binom{n-1}{r-1}\frac{n}{r}T_n^2,\end{aligned} \tag{5.42}$$

and

$$\begin{aligned}\sum_{\mathrm{C}} T_{r,\mathcal{S}}^2 &= \binom{n-2}{r-2}\frac{1}{r^2}\left(\sum_{i=1}^{n} X_i^2 + \sum_{i\neq j} X_i X_j\right) \\ &\quad + \left(\binom{n-1}{r-1} - \binom{n-2}{r-2}\right)\frac{1}{r^2}\sum_{i=1}^{n} X_i^2 \\ &= \binom{n-2}{r-2}\frac{1}{r^2}\left(n^2 T_n^2 + \frac{n-r}{r-1}\sum_{i=1}^{n} X_i^2\right).\end{aligned} \tag{5.43}$$

U-Statistic as a Conditional Expectation of the Kernel

Notice that a U-statistic could be defined in terms of a conditional expectation of the kernel, given a sufficient statistic, say the order statistics. That is, if U is as given in equation (5.35), $X_1, \ldots, X_n$ is a random sample and $X_{(1)}, \ldots, X_{(n)}$ are the order statistics from the given distribution, and h is an m^{th} order kernel (with $m \leq n$), then

$$U = \mathrm{E}\left(h(X_1, \ldots, X_m) | X_{(1)}, \ldots, X_{(n)}\right). \tag{5.44}$$

Example 5.13 shows that the kernel is not unique; that is, the same U-statistic could be formed from different kernels.

Variations of the Order of the Kernel

We informally defined the order of the kernel as the "number of arguments" in the kernel, and by this we meant the number of sample items included in the kernel. Occasionally, the kernel will include some argument computed from the full sample; that is, an m^{th} order kernel involves more than m items from the sample; hence the precise meaning of "order" breaks down somewhat. An example of such a kernel is one that is a function of a single observation as well as of the sample mean, $h(x_i, \overline{x})$. Such a kernel obviously cannot be symmetric.

Example 5.20 variance
Writing the variance of the random variable X as

$$\mathrm{V}(X) = \mathrm{E}\left((X - \mathrm{E}(X))^2\right)$$

may suggest the kernel

$$h(x_i, \bar{x}) = (x_i - \overline{x})^2. \tag{5.45}$$

At first glance, we might think that the expected value of this kernel is σ^2. Because X_i is included in $\overline{X}$, however, we have

$$\begin{aligned}
\mathrm{E}\left(h(X_i, \overline{X})\right) &= \mathrm{E}\left(\left((n-1)X_i/n - \sum_{j\neq i} X_j/n\right)^2\right) \\
&= \mathrm{E}\left((n-1)^2 X_i^2/n^2 - 2(n-1)X_i \sum_{j\neq i} X_j/n^2 \right. \\
&\qquad \left. + \sum_{j\neq k\neq i} X_j X_k/n^2 + \sum_{j\neq i} X_j^2/n^2\right) \\
&= (n-1)^2\mu^2/n^2 + (n-1)^2\sigma^2/n^2 - 2(n-1)(n-1)\mu^2/n^2 \\
&\qquad +(n-1)(n-2)\mu^2/n^2 + (n-1)\mu^2/n^2 + (n-1)\sigma^2/n^2 \\
&= \frac{n-1}{n}\sigma^2,
\end{aligned}$$

and the U-statistic associated with this kernel of course also has expectation $\frac{n-1}{n}\sigma^2$. On more careful thought, we would expect the expected value of the kernel to be less than σ^2, because the expectation of $(X_i - \mu)^2$, which does not have $(n-1)X_i/n$ subtracted out, is σ^2.

This is not an example of a U-statistic that is "biased". A U-statistic is always (tautologically) unbiased for its expectation, if it exists. It we want a U-statistic for the variance, we have started with the wrong kernel!.

If instead of the kernel h above, we used the kernel

$$g(X_i, \overline{X}) = \frac{n}{n-1}(X_i - \overline{X})^2, \tag{5.46}$$

we would have an expectation functional of interest; that is, one such that $\mathrm{E}(g(X_1, \ldots, X_m))$ is something of interest, namely σ^2. ∎

Example 5.21 jackknife variance estimator
The jackknife variance estimator (3.166)

$$\frac{\sum_{j=1}^{N}(T_j^* - T)^2}{r(r-1)}.$$

is a U-statistic whose kernel is of order $n - d$, but the kernel also involves all n observations. ∎

Generalized U-Statistics

We can generalize U-statistics in an obvious way to independent random samples from more than one population. The sample sizes can be different. We do not even require that the number of elements used as arguments to the kernel be the same.

Example 5.22 two-sample Wilcoxon statistic
A common U-statistic involving two populations is the *two-sample Wilcoxon statistic*. For this, we assume that we have two samples $X_{11}, \ldots, X_{1n_1}$ and $X_{21}, \ldots, X_{2n_2}$. The kernel is $h(x_{1i}, x_{2j}) = \mathrm{I}_{]-\infty,0]}(x_{2j} - x_{1i})$. The two-sample Wilcoxon statistic is

$$U = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \mathrm{I}_{]-\infty,0]}(X_{2j} - X_{1i}). \tag{5.47}$$

This is an unbiased estimator of $\Pr(X_{11} \leq X_{21})$.

The more familiar form of this statistic is $n_1 n_2 U$, and in this form it is called the *Mann-Whitney statistic*.

The two sample Wilcoxon statistic or the Mann-Whitney statistic can be used to test that the distributions of two populations are the same versus the alternative that a realization from one distribution is typically smaller (or larger) than a realization from the other distribution. Although the two sample Wilcoxon statistic is sometimes used to test whether one population has a larger median than that of another population, if the distributions have quite different shapes, a typical value from the first population may tend to be smaller than a typical value from the second population. ∎

5.2.3 Properties of U-Statistics

U-statistics have a number of interesting properties. U-statistics are often useful in nonparametric inference because, among other reasons, they are asymptotically the same as the plug-in estimator that is based on the empirical CDF. Some of the important statistics used in modern computational statistical methods are U-statistics.

By conditioning on the order statistics, we can show that a UMVUE can be expressed as a U-statistic.

Theorem 5.3
Let $X_1, \ldots, X_n$ be a random sample from a distribution with parameter θ and with finite variance. Let $T = T(X_1, \ldots, X_n)$ be unbiased for θ. Then there is a U-statistic, U, that is also unbiased for θ, and

$$\mathrm{V}(U) \leq \mathrm{V}(T).$$

Proof.
We first define an n^{th} order expectation kernel to be associated with T; in fact, it is the function itself:

$$h(X_{i_1}, \ldots, X_{i_n}) = T(X_{i_1}, \ldots, X_{i_n}).$$

The associated U-statistic is

$$U = \frac{1}{n!} \sum_{\mathrm{C}} T(X_{i_1}, \ldots, X_{i_n})$$

Now, as in equation (5.44), we write

$$U = \mathrm{E}\left(T(X_{i_1}, \ldots, X_{i_n}) | X_{(1)}, \ldots, X_{(n)}\right).$$

Hence,

$$\begin{aligned} \mathrm{E}\left(U^2\right) &= \mathrm{E}\left(\left(\mathrm{E}\left(T|X_{(1)}, \ldots, X_{(n)}\right)\right)^2\right) \\ &\leq \mathrm{E}\left(\mathrm{E}\left(T^2|X_{(1)}, \ldots, X_{(n)}\right)\right) \\ &= \mathrm{E}\left(T^2\right). \end{aligned}$$

with equality if and only if $\mathrm{E}\left(T|X_{(1)}, \ldots, X_{(n)}\right)$ is degenerate and equals T with probability 1. ∎

We will assume $\mathrm{E}(h(X_1, \ldots, X_m)^2) < \infty$. We first introduce some additional notation for convenience.

(The notation h_r problem ***)

For $k = 1, \ldots, m$, let

$$\begin{aligned} h_k(x_1, \ldots, x_k) &= \mathrm{E}(h(X_1, \ldots, X_m)|X_1 = x_1, \ldots, X_k = x_k) \\ &= \mathrm{E}(h(x_1, \ldots, x_k, X_{k+1}, \ldots, X_m)). \end{aligned} \tag{5.48}$$

We have $h_m = h$ and

$$h_k(x_1, \ldots, x_k) = \mathrm{E}(h_{k+1}(x_1, \ldots, x_k, X_{k+1}, \ldots, X_m)). \tag{5.49}$$

Now, we define the centered versions of the h: for $k = 1, \ldots, m$,

$$\tilde{h}_k = h_k - \mathrm{E}(h(X_1, \ldots, X_m)), \tag{5.50}$$

and let

$$\tilde{h} = \tilde{h}_m$$

We see that the corresponding centered U-statistic is

$$U - \mathrm{E}(U) = \frac{1}{\binom{n}{m}} \sum_{\mathrm{C}} \tilde{h}(X_{i_1}, \ldots, X_{i_m}) \tag{5.51}$$

This notation is convenient in the demonstration that a sequence of adjusted kernels forms a martingale (see Serfling (1980), page 177).

It is also a simple matter to work out the variance of the corresponding U-statistic.

Theorem 5.4 (Hoeffding's Theorem)
Let U be a U-statistic with m^{th} order kernel h with $\mathrm{E}(h(X_1,\ldots,X_m)^2) < \infty$. Then

$$\mathrm{V}(U) = \frac{1}{\binom{n}{m}} \sum_{k=1}^{m} \binom{m}{k}\binom{n-m}{m-k}\zeta_k \tag{5.52}$$

where

$$\zeta_k = \mathrm{V}(h_k(X_1,\ldots,X_k)). \tag{5.53}$$

Proof.
MS2 p. 176. ∎

Projections of U-Statistics

One method of working out the asymptotic distribution of a U-statistic is by use of projections

We first relate Theorem 1.65 on page 118 to the U-statistic,

$$U_n = \frac{1}{\binom{n}{m}} \sum_{\text{all combinations}} h(X_{i_1},\ldots,X_{i_m}).$$

Let $\widetilde{U}_n$ be the projection of U_n onto $X_1,\ldots,X_n$. (See Section 1.5.3 beginning on page 115.) Recall, as in equation (5.48),

$$\begin{aligned} h_1(x_1) &= \mathrm{E}(h(X_1,X_2,\ldots,X_m)|X_1 = x_1) \\ &= \mathrm{E}(h(x_1,X_2,\ldots,X_m)). \end{aligned}$$

and

$$\tilde{h}_1 = h_1 - \mathrm{E}(h(X_1,\ldots,X_m)),$$

Then, starting with the definition of $\widetilde{U}_n$ as a projection, we have

$$\begin{aligned} \widetilde{U}_n &= \mathrm{E}(U_n) + \sum_{i=1}^{n} (\mathrm{E}(U_n|X_i) - \mathrm{E}(U_n)) \\ &= \mathrm{E}(U_n) + \frac{m}{n}\sum_{i=1}^{n} \tilde{h}_1(X_i). \end{aligned}$$

This yields

$$\mathrm{V}(\widetilde{U}_n) = \frac{m^2}{n}\zeta_1,$$

where, in the notation of equation (5.53), $\zeta_1 = \mathrm{V}(h_1(X_1))$.

Hence, by Hoeffding's theorem (actually a corollary of it), and Theorem 1.65, we have

$$\mathrm{E}((U_n - \widetilde{U}_n)^2) \in \mathrm{O}(n^{-2}).$$

If $\zeta_1 > 0$, this yields

$$\sqrt{n}(U_n - \mathrm{E}(U_n)) \xrightarrow{\mathrm{d}} \mathrm{N}(0, m^2\zeta_1).$$

(Theorem 3.5(i) in MS2.)

Computational Complexity

Evaluating a U-statistic can be computationally intensive, with the number of arithmetic operations of $O(n^m)$. As we discussed on page 303 for the delete-d jackknife, we may reduce the number of computations by using only some of the possible combinations. There are various ways that the combinations could be chosen, including, of course, just a random sampling. The U-statistic would be approximated by an average of the kernel evaluated only over the random sampling of the subsets.

5.3 Asymptotically Unbiased Estimation

A sequence of estimators that are unbiased for any finite sample size is unbiased in the limit and is asymptotically unbiased. There are, however, many situations when an unbiased estimator in a finite sample does not exist, or when we cannot form one easily, or when a biased estimator has better MSE for any finite sample than an unbiased estimator. A biased estimator that is asymptotically unbiased, and for which there is no dominating unbiased estimator, is often considered optimal.

Sometimes, by studying the nature of the bias, it may be possible to identify a correction, as in the following example that uses the jackknife (see Section 3.6.1 on page 301).

Example 5.23 Jackknife Bias Reduction
Suppose that we can represent the bias of T as a power series in n^{-1}; that is,

$$\begin{aligned} \text{Bias}(T) &= \text{E}(T) - \theta \\ &= \sum_{q=1}^{\infty} \frac{a_q}{n^q}, \end{aligned} \tag{5.54}$$

where the a_q do not involve n. If all $a_q = 0$, the estimator is unbiased. If $a_1 \neq 0$, the order of the bias is n^{-1}. (Such an estimator is sometimes called "second-order accurate". "First-order" accuracy implies a bias of order $n^{-1/2}$.)

Using the power series representation for the bias of T, we see that the bias of the jackknife estimator is

$$\begin{aligned}
\text{Bias}(\text{J}(T)) &= \text{E}(\text{J}(T)) - \theta \\
&= n(\text{E}(T) - \theta) - \frac{n-1}{n}\sum_{j=1}^{n}\text{E}(T_{(-j)} - \theta) \\
&= n\sum_{q=1}^{\infty}\frac{a_q}{n^q} - (n-1)\left(\sum_{q=1}^{\infty}\frac{a_q}{(n-1)^q}\right) \\
&= a_2\left(\frac{1}{n} - \frac{1}{n-1}\right) + a_3\left(\frac{1}{n^2} - \frac{1}{(n-1)^2}\right) + \ldots \\
&= -a_2\left(\frac{1}{n(n-1)}\right) + a_3\left(\frac{1}{n^2} - \frac{1}{(n-1)^2}\right) + \ldots; \qquad (5.55)
\end{aligned}$$

that is, the bias of the jackknife estimator, $\text{Bias}(\text{J}(T))$, is at most of order n^{-2}. If $a_q = 0$ for $q = 2, \ldots$, the jackknife estimator is unbiased.

This reduction in the bias is a major reason for using the jackknife. Any explicit analysis of the bias reduction, however, depends on a representation of the bias in a power series in n^{-1} with constant coefficients. This may not be possible, of course.

From

$$\text{E}(\text{J}(T)) - \theta \;=\; \text{E}(T) - \theta + (n-1)\left(\text{E}(T) - \frac{1}{n}\sum_{j=1}^{n}\text{E}(T_{(-j)})\right),$$

we have the jackknife estimator of the bias in T,

$$B_{\text{J}} = (n-1)\left(\overline{T}_{(\bullet)} - T\right), \qquad (5.56)$$

and the jackknife bias-corrected estimator of θ,

$$T_{\text{J}} = nT - (n-1)\overline{T}_{(\bullet)}. \qquad (5.57)$$

∎

Example 5.24 Higher-Order Bias Corrections
Suppose that we pursue the bias correction to higher orders by using a second application of the jackknife. The pseudovalues are

$$T_j^{**} = n\text{J}(T) - (n-1)\text{J}(T_{(-j)}). \qquad (5.58)$$

Assuming the same series representations for the bias as before, a second-order jackknife estimator,

$$\text{J}^2(T) = \frac{n^2\text{J}(T) - (n-1)^2\sum_{j=1}^{n}\text{J}(T)_{(-j)}/n}{n^2 - (n-1)^2}, \qquad (5.59)$$

is unbiased to order $\text{O}(n^{-3})$.

There are two major differences between this estimator and the first-order jackknifed estimator. For the first-order jackknife, $\mathrm{J}(T)$ differs from T by a quantity of order n^{-1}; hence, if T has variance of order n^{-1} (as we usually hope), the variance of $\mathrm{J}(T)$ is asymptotically the same as that of T. In other words, the bias reduction carries no penalty in increased variance. This is not the case for higher-order bias correction of $\mathrm{J}^2(T)$.

The other difference is that in the bias expansion,

$$\mathrm{E}(T) - \theta = \sum_{q=1}^{\infty} a_q/n^q,$$

if $a_q = 0$ for $q \geq 2$, then the first-order jackknifed estimator is unbiased. For the second-order jackknifed estimator, even if $a_q = 0$ for $q \geq 3$, the estimator may not be unbiased. Its bias is

$$\mathrm{Bias}(\mathrm{J}^2(T)) = \frac{a_2}{(n-1)(n-2)(2n-1)}; \tag{5.60}$$

that is, it is still of order n^{-3}. ∎

We will consider four general kinds of estimators that may be of this type: estimators based on the method of moments, functions of unbiased estimators, V-statistics, and quantile estimators. Some of these estimators arise as plug-in statistics in the ECDF, such as those based on the method of moments, and others from a general plug-in rule, in which individual estimators are used in different parts of the formula for the estimand, such as ratio estimators.

We would like for such biased estimators to have either limiting bias or asymptotic bias of zero.

5.3.1 Method of Moments Estimators

If the estimand is written as a functional of the CDF, $\theta = \Theta(P)$, an estimator formed by applying Θ to the ECDF, $\widehat{\theta} = \Theta(P_n)$ is call a plug-in estimator.

If Θ is an expectation functional of the form $\int x^r \mathrm{d}P(x)$, that is, if Θ is a raw moment, then the plug-in estimator $\Theta(P_n)$ is unbiased for Θ. Central moments are more often of interest interest. A plug-in estimator of a central moment, just as the central moment itself, can be written as a function of the corresponding raw moment and the first moment. Such estimators are called method of moments estimators.

An example of an estimator based on the method of moments is $\widetilde{S}^2 = (n-1)S^2/n$ as an estimator of the population variance, σ^2. This is the second central moment of the sample, just as σ^2 is the second central moment of the population. We have seen that, in certain conditions, the MSE of $\widetilde{S}^2$ is less than that of S^2, and while it is biased, its limiting and asymptotic bias is zero and is of order $1/n$.

Although the second central sample moment is biased, the raw sample moments are unbiased for the corresponding raw population moments, if they exist.

5.3.2 Ratio Estimators

Ratio estimators, that is, estimators composed of the ratio of two separate estimators, arise often in sampling applications. Another situation is when an estimator is based on a linear combination of observations with different variances. If we have some way of estimating the variances so we can form a weighted linear combination, the resulting estimator will be biased, but its MSE may be better than the unweighted estimator. Also, it is often the case that the biased estimator is asymptotically normal and unbiased.

5.3.3 V-Statistics

As we have seen, a U-statistic is an unbiased estimator of an expectation functional; specifically, if $\Theta(P) = \mathrm{E}(h(X_1, \ldots, X_m))$ the U-statistic with kernel h is unbiased for $\Theta(P)$. Applying the functional Θ to the ECDF P_n, we have

$$\begin{aligned} \Theta(P_n) &= \frac{1}{n^m} \sum_{i_1=1}^{n} \cdots \sum_{i_m=1}^{n} h(X_{i_1}, \ldots, X_{i_m}) \\ &= V \quad \text{(say)}, \end{aligned} \tag{5.61}$$

which we call the *V-statistic* associated with the kernel h, or equivalently associated with the U-statistic with kernel h. Recalling that $\Theta(P_n)$ in general is not unbiased for $\Theta(P)$, we do not expect a V-statistic to be unbiased in general. However, in view of the asymptotic properties of P_n, we might expect V-statistics to have good asymptotic properties.

A simple example is the variance, for which the U-statistic in equation (5.39) is unbiased. The V-statistic with the same kernel is

$$\begin{aligned} V &= \frac{1}{2n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} (X_i - X_j)^2 \\ &= \frac{1}{2n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} (X_i^2 + X_j^2 - 2X_i X_j) \\ &= \frac{n-1}{n} S^2, \end{aligned} \tag{5.62}$$

where S^2 is the sample variance. This V-statistic is the same as the plug-in estimator of the population variance, and as with the plug-in estimator, no particular underlying distribution is assumed. It is also the same as the MLE estimator given an assumed underlying normal distribution. The V-statistic is biased for the population variance; but as we have seen, it has a smaller MSE than the unbiased U-statistic.

The development of V-statistics can be based on the idea of applying the same functional to the ECDF F_n as the functional that defines the estimand when applied to the CDF F, and which is the basis for the U-statistics. Since

the ECDF assigns probability $1/n$ to each point of the values $X_1, \ldots, X_n$, any m independent variables with CDF F_n take on each of the possible m-tuples $(X_{i_1}, \ldots, X_{i_m})$ with probability $1/n^m$. The plug-in estimator, call it V, of θ is therefore

$$V = \frac{1}{n^m} \sum_{i_1=1}^{n} \cdots \sum_{i_m=1}^{n} h(X_{i_1}, \ldots, X_{i_m}).$$

Notice for $m = 1$, V is a U-statistic; but consider $m = 2$, as above. We have

$$U = \frac{1}{n(n-1)} \sum_{i} \sum_{j \neq i} h(X_i, X_j),$$

however

$$\begin{aligned} V &= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} h(X_i, X_j) \\ &= \frac{1}{n^2} \sum_{i} \sum_{j \neq i} h(X_i, X_j) + \frac{1}{n^2} \sum_{i=1}^{n} h(X_i, X_i) \end{aligned}$$

While, as we have seen U is unbiased for θ, we see that V is biased:

$$\begin{aligned} \mathrm{E}(V) &= \frac{n-1}{n}\theta + \frac{1}{n}\mathrm{E}(h(X_1, X_1)) \\ &= \theta + \frac{1}{n}\left(\mathrm{E}(h(X_1, X_1)) - \theta\right). \end{aligned}$$

An example of a V-statistic with $m = 2$ uses $h(x_1, x_2) = (x_1 - x_2)^2/2$, as in equation (5.62), and results in $(n-1)S^2/n$ as an estimator of σ^2, which is of course asymptotically unbiased.

Theorem 3.16 in MS2 shows that under certain general conditions, V-statistics have limiting normal distributions and are asymptotically unbiased.

5.3.4 Estimation of Quantiles

plug-in
- finite sample properties – Harrel-Davis estimator
- asymptotic normality (ch 1)

5.4 Asymptotic Efficiency

Often a statistical procedure does not have some desirable property for any finite sample size, but the procedure does have that property asymptotically. The asymptotic properties that are of most interest are those defined in terms

of a sequence that has a limiting standard normal distribution, $\mathrm{N}(0,1)$, or more generally, $\mathrm{N}_k(0, I_k)$. A standard normal distribution of a statistic is desirable because in that case, it is easy to associate statements of probabilities with values of the statistic. It is also desirable because it is often easy to work out the distribution of functions of a statistic that has a normal distribution.

It is important to remember the difference in an asymptotic property and a limiting property. An *asymptotic distribution* is the same as a *limiting distribution*, but other asymptotic properties are defined, somewhat arbitrarily, in terms of a limiting distribution of some function of the sequence of statistics and of a finite divergent or convergent sequence, a_n. This seems to mean that a particular asymptotic property, such as, say, the asymptotic variance, depends on what function of the sequence of statistics that we choose. Although there may be some degree of arbitrariness in "an" asymptotic expectation, there is a certain uniqueness, as expressed in Proposition 2.3 in MS2.

5.4.1 Asymptotic Relative Efficiency

We assume a family of distributions $\mathcal{P}$, a sequence of estimators $\{T_n\}$ of $g(\theta)$, and a sequence of constants $\{a_n\}$ with $\lim_{n\to\infty} a_n = \infty$ or with $\lim_{n\to\infty} a_n = a > 0$, and such that $a_n T_n(X) \xrightarrow{\mathrm{d}} T$ and $0 < \mathrm{E}(T) < \infty$. We define the asymptotic mean squared error of $\{T_n\}$ for estimating $g(\theta)$ wrt $\mathcal{P}$ as an asymptotic expectation of $(T_n - g(\theta))^2$; that is, $\mathrm{E}((T - g(\theta))^2)/a_n$, which we denote as $\mathrm{AMSE}(T_n, g(\theta), \mathcal{P})$.

For comparing two estimators, we may use the *asymptotic relative efficiency*, which for the estimators S_n and T_n of $g(\theta)$ wrt $\mathcal{P}$ is

$$\mathrm{ARE}(S_n, T_n, \mathcal{P}) = \mathrm{AMSE}(S_n, g(\theta), \mathcal{P})/\mathrm{AMSE}(T_n, g(\theta), \mathcal{P}).$$

5.4.2 Asymptotically Efficient Estimators

Relative efficiency is a useful concept for comparing two estimators, whether or not they are unbiased. When we restrict our attention to unbiased estimators we use the phrase *Fisher efficient* to refer to an estimator that attains its Cramér-Rao lower bound (Definition 3.8). Again, notice the slight difference in "efficiency" and "efficient"; while one meaning of "efficiency" is a relative term that is not restricted to unbiased estimators (or other unbiased procedures, as we will see later), "efficient" is absolute. "Efficient" only applies to unbiased estimators, and an estimator either is or is not efficient. The state of being efficient, of course is called "efficiency". This is another meaning of the term. The phrase "Fisher efficiency" helps to emphasis this difference.

We consider the problem of estimating the k-vector θ based on a random sample $X_1, \ldots, X_n$. We denote the sequence of estimators as $\{\hat{\theta}_n\}$. Suppose

$$(V_n(\theta))^{-\frac{1}{2}} \left(\hat{\theta}_n - \theta\right) \xrightarrow{\mathrm{d}} \mathrm{N}_k(0, I_k),$$

where, for each n, $V_n(\theta)$ is a $k \times k$ positive definite matrix. From the definition of asymptotic expectation of $\left(\hat{\theta}_n - \theta\right)^2$, $V_n(\theta)$ is the asymptotic variance-covariance matrix of $\hat{\theta}_n$. Note that this matrix may depend on θ. We should note that for any fixed n, $V_n(\theta)$ is not necessarily the variance-covariance matrix of $\hat{\theta}_n$; that is, it is possible that $V_n(\theta) \neq \mathrm{V}(\hat{\theta}_n)$.

Just as we have defined Fisher efficiency for an unbiased estimator of fixed size, we define a sequence to be *asymptotically Fisher efficient* if the sequence is asymptotically unbiased, the Fisher information matrix $I_n(\theta)$ exists and is positive definite for each n, and $V_n(\theta) = (I_n(\theta))^{-1}$ for each n. The definition of asymptotically (Fisher) efficiency is often limited even further so as to apply only to estimators that are asymptotically normal. (MS2 uses the restricted definition.)

Being asymptotically efficient does not mean for any fixed n that $\hat{\theta}_n$ is efficient. First of all, for fixed n, $\hat{\theta}_n$ may not even be unbiased; even if it is unbiased, however, it may not be efficient.

As we have emphasized many times, asymptotic properties are different from limiting properties. As a striking example of this, consider a very simple example from Romano and Siegel (1986).

Example 5.25 Asymptotic and Limiting Properties
Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{N}_1(\mu, 1)$, and consider a randomized estimator $\hat{\mu}_n$ of μ defined by

$$\hat{\mu}_n = \begin{cases} \overline{X}_n & \text{with probability } 1 - \frac{1}{n} \\ n^2 & \text{with probability } \frac{1}{n}. \end{cases}$$

It is clear that $n^{1/2}(\hat{\mu}_n - \mu) \xrightarrow{\text{d}} \mathrm{N}(0, 1)$, and furthermore, the Fisher information for μ is $n^{-1/2}$. The estimator $\hat{\mu}_n$ is therefore asymptotically Fisher efficient. The bias of $\hat{\mu}_n$, however, is

$$\mathrm{E}(\hat{\mu}_n - \mu) = \mu\left(1 - \frac{1}{n}\right) + n - \mu = n - \mu/n,$$

which tends to infinity, and the variance is

$$\begin{aligned} \mathrm{V}(\hat{\mu}_n) &= \mathrm{E}(\hat{\mu}^2) - (\mathrm{E}(\hat{\mu}))^2 \\ &= \left(1 - \frac{1}{n}\right)\frac{1}{n} + \left(\frac{1}{n}\right)n^4 - \left(\mu\left(1 - \frac{1}{n}\right) + n\right)^2 \\ &= n^3 + \mathrm{O}(n^2), \end{aligned}$$

which also tends to infinity. Hence, we have an asymptotically Fisher efficient estimator whose limiting bias and limiting variance are both infinite. ∎

The example can be generalized to any estimator T_n of $g(\theta)$ such that $\mathrm{V}(T_n) = 1/n$ and $n^{1/2}(T_n - g(\theta)) \xrightarrow{\text{d}} \mathrm{N}(0, 1)$. From T_n form the estimator

$$\widetilde{T}_n = \begin{cases} T_n \text{ with probability } 1 - \frac{1}{n} \\ n^2 \text{ with probability } \frac{1}{n}. \end{cases}$$

The estimator $\widetilde{T}_n$ is also asymptotically Fisher efficient but has infinite limiting bias and infinite limiting variance.

Asymptotic Efficiency and Consistency

Although asymptotic efficiency implies that the estimator is asymptotically unbiased, even if the limiting variance is zero, asymptotic efficiency does not imply consistency. The counterexample above shows this.

Likewise, of course, consistency does not imply asymptotic efficiency. There are many reasons. First, asymptotic efficiency is usually only defined in the case of asymptotic normality (of course, it is unlikely that a consistent estimator would not be asymptotically normal). More importantly, the fact that both the bias and the variance go to zero as required by consistency, is not very strong. There are many ways both of these can go to zero without requiring asymptotic unbiasedness or that the asymptotic variance satisfy the asymptotic version of the information inequality.

The Asymptotic Variance-Covariance Matrix

In the problem of estimating the k-vector θ based on a random sample $X_1, \ldots, X_n$ with the sequence of estimators as $\{\hat{\theta}_n\}$, if

$$(V_n(\theta))^{-\frac{1}{2}} \left(\hat{\theta}_n - \theta\right) \xrightarrow{\mathrm{d}} \mathrm{N}_k(0, I_k),$$

where, for each n, $V_n(\theta)$ is a $k \times k$ positive definite matrix, then $V_n(\theta)$ is the asymptotic variance-covariance matrix of $\hat{\theta}_n$. As we have noted, for any fixed n, $V_n(\theta)$ is not necessarily the variance-covariance matrix of $\hat{\theta}_n$.

If $V_n(\theta) = \mathrm{V}(\hat{\theta}_n)$, then under the information inequality regularity conditions that yield the CRLB, we know that

$$V_n(\theta) \succeq (I_n(\theta))^{-1},$$

where $I_n(\theta)$ is the Fisher information matrix.

Superefficiency

Although if $V_n(\theta) \neq \mathrm{V}(\hat{\theta}_n)$, the CRLB says nothing about the relationship between $V_n(\theta)$ and $(I_n(\theta))^{-1}$, we might expect that $V_n(\theta) \succeq (I_n(\theta))^{-1}$. That this is not necessarily the case is shown by a simple example given by Joseph Hodges in a lecture in 1951, published in Le Cam (1953) (see also Romano and Siegel (1986)).

Example 5.26 Hodges' Superefficient Estimator
Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{N}_1(\mu, 1)$, and consider an estimator $\hat{\mu}_n$ of μ defined by

$$\hat{\mu}_n = \begin{cases} \overline{X}_n & \text{if } |\overline{X}_n| \geq n^{-1/4} \\ t\overline{X}_n & \text{otherwise,} \end{cases}$$

for some fixed t with $|t| < 1$.

We have

$$\sqrt{n}(T_n - g(\theta)) \xrightarrow{\text{d}} \mathrm{N}(0, v(\theta)),$$

where $v(\theta) = 1$ if $\theta \neq 0$ and $v(\theta) = t^2$ if $\theta = 0$. (This takes a little working out; consider the two parts.)

Notice that $I(\theta) = 1$ and $g'(\theta) = 1$, hence, at $\theta = 0$, with $|t| < 1$, we have

$$v(\theta) < \frac{(g'(\theta))^2}{I(\theta)}.$$

The estimator $\hat{\mu}_n$ is sometimes called "Hodges' superefficient estimator". ∎

What gives Example 5.26 its kick is the dependence of the asymptotic distribution of $\hat{\mu}_n$ on μ. If $\mu \neq 0$, $\hat{\mu}_n$ has the same asymptotic distribution as $\overline{X}_n$, and obeys the CRLB, both in its variance for finite n (even though it is biased) and in its asymptotic variance. However, if $\mu = 0$, $\hat{\mu}_n$ is still asymptotically unbiased, but the asymptotic variance of $\hat{\mu}_n$ is t^2/n, which is smaller than the inverse of asymptotic Fisher information, $1/n$.

A point in the parameter space at which this anomaly occurs is called a *point of superefficiency*. Le Cam has shown that under certain regularity conditions (that are slightly more stringent than the information inequality regularity conditions, see page 169) the number of points of superefficiency is countable. I list all of these regularity conditions in the statement of the following theorem, which is due to Le Cam (1953).

Theorem 5.5
Let $X_1, \ldots, X_n$ be iid **********

Proof.

∎

Superefficiency is not important in applications (that is, where n is finite) any decrease in mean squared error at a point of superefficiency is accompanied by an increase in mean squared error at nearby points (and, of course, if we knew the parameter was a point of superefficiency, we would probably not be estimating it.

5.5 Applications

Many methods of statistical inference rely on samples of identically distributed random variables. Two major areas of application of the methods are in analysis of linear models and sampling of finite populations.

5.5.1 Estimation in Linear Models

In a simple variation on the requirement of identical distributions, we assume a model with two components, one "systematic" and one random, and the distributions of the observable random variables depend on the systematic component.

Systematic and Random Components

The most common form of linear model is one in which a random variable Y is the sum of a systematic component that determines its expected value and random component that is the value of an underlying unobservable random variable that has an expected value of 0. The systematic component may be a function of some additional variables x and parameters θ. If we represent the underlying unobservable random with expectation 0, as ϵ, we have

$$Y = f(x, \theta) + \epsilon. \tag{5.63}$$

In this setup the mean of the random variable Y is determined by the parameter θ and the values of the x variables, which are *covariates* (also called *regressors*, *carriers*, or *independent variables*). We generally treat the covariates as fixed variables, that is, whether or not we could also model the covariates as random variables, in the simplest cases, we will use their observed values without regard to their origin.

Regression Models

The model above is a regression model. In the simplest variation, the observable random variables are independent, and have distributions in the same location family: $\mathcal{P} = \{P_{f(x,\theta),P_\epsilon}\}$. The family $\mathcal{P}_\epsilon$ of distributions P_ϵ of the random component may be a parametric family, such as $\mathrm{N}(0, \sigma^2)$, or it may be a nonparametric family. Whatever other assumptions on P_ϵ, we assume $\mathrm{E}(\epsilon) = 0$.

Linear Models

Often we assume that the systematic component is a linear combination of the covariates. This setup is called a *linear model*, and is usually written in the form

$$Y = x^{\mathrm{T}}\beta + E, \tag{5.64}$$

where Y is the observable random variable, x is an observable p-vector of covariates, β is an unknown and unobservable p-vector of parameters, and E is an unobservable random variable with $\mathrm{E}(E) = 0$ and $\mathrm{V}(E) = \sigma^2 I$. The parameter space for β is $B \subseteq \mathbb{R}^p$.

An item of a random sample from this model may be denoted

$$Y_i = x_i^{\mathrm{T}}\beta + E_i, \tag{5.65}$$

and a random sample be written in the vector-matrix form

$$Y = X\beta + E, \tag{5.66}$$

where Y and E are n-vectors, X is an $n \times p$ matrix whose rows are the x_i^{T}, and β is the p-vector above. A sample of realizations may be written in the vector-matrix form

$$y = X\beta + \epsilon. \tag{5.67}$$

where y and ϵ are n-vectors. This is the most commonly used notation.

Inference in a Linear Model

For estimation in a linear model, rather than formulating a decision problem and seeking a minimum risk estimator, we usually begin with a different approach. Estimation in a linear model is most commonly developed based on two simple heuristics: least squares and unbiasedness.

The degree of β is p, meaning that the minimum number of observations required for unbiased estimation of β is p. Inferences about characteristics of the distribution of ϵ require additional observations, however, and so we assume $n > p$ in the following.

In statistical inference, we can think of β either as an unobservable random variable or as an unknown constant. If we think of it as an unknown constant and we want to determine a value of it that optimizes some objective function (such as a likelihood or a sum of squares), then we first must substitute a variable for the constant. Although we often skip over this step, it is important conceptually.

Least Squares Solutions of Overdetermined Linear Systems

Having substituted the variable b is in place of the unknown model parameter β, we have an overdetermined linear system

$$y \approx Xb, \tag{5.68}$$

where y and X are given, b is unknown, and $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times p}$, and $b \in \mathbb{R}^p$. Solving for b in this system is a common problem in linear algebra. It is

one aspect of the statistical problem of fitting the model (5.66), in which we assume that y is a realization of a random variable Y with $\mathrm{E}(Y) = X\beta$, but for the time being we will just consider the algebraic issues in solving, or "fitting", the overdetermined system.

Fitting an overdetermined system $y \approx Xb$ involves a choice of a criterion for the goodness of the approximation. A common choice is the squared error; that is, a solution is a vector b that minimizes $\|y - Xb\|_2$. This follows the approach to statistical inference discussed in Section 3.2.3. The solution to the linear algebra problem (5.68) is often called an "estimator" even though there is no underlying probability distribution.

We define a *least squares* estimator (LSE) of b or of β in equation (5.66) as

$$b^* = \arg\min_{b \in B} \|y - Xb\|^2, \tag{5.69}$$

where $\|c\| = \|c\|_2 = \sqrt{c^{\mathrm{T}}c} = \sqrt{\sum_{i=1}^{p} c_i^2}$ for the p-vector c.

A least squares estimator of β may or may not be unique. Whether or not b^* is unique,

$$\|y - Xb^*\|^2 \tag{5.70}$$

is unique. This is because the objective function is convex and bounded below.

The least squares estimator is obtained by direct minimization of

$$\begin{aligned} s(b) &= \|y - Xb\|^2 \\ &= y^{\mathrm{T}}y - 2b^{\mathrm{T}}X^{\mathrm{T}}y + b^{\mathrm{T}}X^{\mathrm{T}}Xb. \end{aligned}$$

First of all, we note that $s(b)$ is differentiable, and

$$\frac{\partial^2}{\partial b^2} s(b) = X^{\mathrm{T}}X$$

is nonnegative definitive. We therefore know that at the minimum, we have the estimating equation

$$\partial s(b)/\partial b = 0. \tag{5.71}$$

The estimating equation leads to the *normal equations*:

$$X^{\mathrm{T}}Xb = X^{\mathrm{T}}y. \tag{5.72}$$

The coefficient matrix in these equations has a special form; it is a *Gramian* matrix. We may use b^* to denote any solution to the normal equations formed from the linear system $y = Xb$, that is

$$b^* = (X^{\mathrm{T}}X)^{-}X^{\mathrm{T}}y. \tag{5.73}$$

Notice that if X is not of full rank, b^* is not unique.

A unique solution to these equations is

$$\widehat{\beta} = (X^{\mathrm{T}}X)^{+}X^{\mathrm{T}}y; \tag{5.74}$$

that is, the solution arising from the Moore-Penrose inverse (see page 784).

LSE in a Probability Model

The mechanical aspects of least squares fitting do not rely on any probability distributions.

An LSE of β yields LSEs of other quantities. In general, for an estimand θ that can be expressed as

$$\theta = \mathrm{E}g(Y, \widehat{\beta}), \tag{5.75}$$

we call $\widehat{\theta} = g(y, \widehat{\beta})$ the LSE of θ. Notice that this definition preserves unbiasedness if the relationships are linear.

If the quantities in the equations correspond to n observations that follow the model (5.64), then we form an LSE of $l^{\mathrm{T}}\beta$, for given $l \in \mathbb{R}^p$, as

$$l^{\mathrm{T}}\widehat{\beta}. \tag{5.76}$$

While this quantity may not be unique, the quantity

$$\|Y - X\widehat{\beta}\|^2/(n-p) \tag{5.77}$$

is unique; it is the LSE of $\mathrm{V}(\epsilon) = \sigma^2$; and furthermore, it is unbiased for σ^2 (exercise).

Linear U-Estimability

One of the most important questions for statistical inference involves estimating or testing some linear combination of the elements of the parameter β; for example, we may wish to estimate $\beta_1 - \beta_2$ or to test the hypothesis that $\beta_1 - \beta_2 = c_1$ for some constant c_1. In general, we will consider the linear combination $l^{\mathrm{T}}\beta$. Whether or not it makes sense to estimate such a linear combination depends on whether there is a function of the observable random variable Y such that

$$g(\mathrm{E}(Y)) = l^{\mathrm{T}}\beta. \tag{5.78}$$

We generally restrict our attention to linear functions of $\mathrm{E}(Y)$ and formally define a linear combination $l^{\mathrm{T}}\beta$ to be (linearly) *U-estimable* if and only if there exists a vector t such that

$$t^{\mathrm{T}}\mathrm{E}(Y) = l^{\mathrm{T}}\beta \tag{5.79}$$

for any β.

It is clear that if X is of full column rank, then $l^{\mathrm{T}}\beta$ is linearly estimable for any l. More generally, it is easy to see that $l^{\mathrm{T}}\beta$ is linearly estimable for any $l \in \mathrm{span}(X^{\mathrm{T}})$. (The t vector in equation (5.79) is just the normalized coefficients expressing l in terms of the columns of X.)

Estimability depends only on the simplest distributional assumption about the model; that is, that $\mathrm{E}(\epsilon) = 0$.

Theorem 5.6
Let $Y = X\beta + \epsilon$ where $\mathrm{E}(\epsilon) = 0$. Let $l^{\mathrm{T}}\beta$ be a linearly estimable function and let $\widehat{\beta} = (X^{\mathrm{T}}X)^{+}X^{\mathrm{T}}Y$. Then $l^{\mathrm{T}}\widehat{\beta}$ is unbiased for $l^{\mathrm{T}}\beta$.

Proof.
Because $l \in \text{span}(X^{\mathrm{T}}) = \text{span}(X^{\mathrm{T}}X)$, we can write

$$l = X^{\mathrm{T}}X\tilde{t}, \tag{5.80}$$

for some vector $\tilde{t}$ Now, we have

$$\begin{aligned} \mathrm{E}(l^{\mathrm{T}}\widehat{\beta}) &= \mathrm{E}(l^{\mathrm{T}}(X^{\mathrm{T}}X)^{+}X^{\mathrm{T}}Y) \\ &= \tilde{t}^{\mathrm{T}}X^{\mathrm{T}}X(X^{\mathrm{T}}X)^{+}X^{\mathrm{T}}X\beta \\ &= \tilde{t}^{\mathrm{T}}X^{\mathrm{T}}X\beta \\ &= l^{\mathrm{T}}\beta. \end{aligned} \tag{5.81}$$

∎

Although we have been taking $\widehat{\beta}$ to be $(X^{\mathrm{T}}X)^{+}X^{\mathrm{T}}Y$, the equations above follow for other least squares fits, $b^* = (X^{\mathrm{T}}X)^{-}X^{\mathrm{T}}Y$, for any generalized inverse. In fact, the estimator of $l^{\mathrm{T}}\beta$ is invariant to the choice of the generalized inverse.

Theorem 5.7
Let $Y = X\beta + \epsilon$ where $\mathrm{E}(\epsilon) = 0$. Let $l^{\mathrm{T}}\beta$ be a linearly estimable function, let $\widehat{\beta} = (X^{\mathrm{T}}X)^{+}X^{\mathrm{T}}Y$ and let $b^* = (X^{\mathrm{T}}X)^{-}X^{\mathrm{T}}Y$. Then $l^{\mathrm{T}}b^* = l^{\mathrm{T}}\widehat{\beta}$.

Proof.
If $b^* = (X^{\mathrm{T}}X)^{-}X^{\mathrm{T}}Y$, we have $X^{\mathrm{T}}Xb^* = X^{\mathrm{T}}Y$, and so

$$l^{\mathrm{T}}\widehat{\beta} - l^{\mathrm{T}}b^* = \tilde{t}^{\mathrm{T}}X^{\mathrm{T}}X(\widehat{\beta} - b^*) = \tilde{t}^{\mathrm{T}}(X^{\mathrm{T}}Y - X^{\mathrm{T}}Y) = 0.$$

∎

Gauss-Markov Theorem

The Gauss-Markov theorem provides a restricted optimality property for estimators of estimable functions of β under the condition that $\mathrm{E}(\epsilon) = 0$ and $\mathrm{V}(\epsilon) = \sigma^2 I$; that is, in addition to the assumption of zero expectation, which we have used above, we also assume that the elements of ϵ have constant variance and that their covariances are zero. Note that we do not assume independence or normality.

The Gauss-Markov theorem states that $l^{\mathrm{T}}\widehat{\beta}$ is the unique *best linear unbiased estimator* (BLUE) of the estimable function $l^{\mathrm{T}}\beta$. (Recall that Theorem 5.7 tells us that the inner product is invariant to the choice of the generalized inverse; that is, $l^{\mathrm{T}}b^* = l^{\mathrm{T}}\widehat{\beta}$, where b^* and $\widehat{\beta}$ are given in equations (5.73) and (5.74) respectively.) "Linear" estimator in this context means a linear combination of X; that is, an estimator in the form $a^{\mathrm{T}}X$. It is clear that $l^{\mathrm{T}}\widehat{\beta}$ is linear, and we have already seen that it is unbiased for $l^{\mathrm{T}}\beta$. "Best" in this context means that its variance is no greater than any other estimator that fits the requirements.

Theorem 5.8 (Gauss-Markov theorem)
Let $Y = X\beta + \epsilon$ where $\mathrm{E}(\epsilon) = 0$ and $\mathrm{V}(\epsilon) = \sigma^2 I$, and assume $l^{\mathrm{T}}\beta$ is linearly estimable. Let $\widehat{\beta} = (X^{\mathrm{T}}X)^{+}X^{\mathrm{T}}Y$. Then $l^{\mathrm{T}}\widehat{\beta}$ is the a.s. unique BLUE of $l^{\mathrm{T}}\beta$.

Proof.
Let $a^{\mathrm{T}}Y$ be any unbiased estimator of $l^{\mathrm{T}}\beta$, and write $l = X^{\mathrm{T}}Y\tilde{t}$ as in equation (5.80) above. Because $a^{\mathrm{T}}Y$ is unbiased for any β, as we saw above, it must be the case that $a^{\mathrm{T}}X = l^{\mathrm{T}}$. Recalling that $X^{\mathrm{T}}X\widehat{\beta} = X^{\mathrm{T}}Y$, we have

$$\begin{aligned}
\mathrm{V}(a^{\mathrm{T}}Y) &= \mathrm{V}(a^{\mathrm{T}}Y - l^{\mathrm{T}}\widehat{\beta} + l^{\mathrm{T}}\widehat{\beta}) \\
&= \mathrm{V}(a^{\mathrm{T}}Y - \tilde{t}^{\mathrm{T}}X^{\mathrm{T}}Y + l^{\mathrm{T}}\widehat{\beta}) \\
&= \mathrm{V}(a^{\mathrm{T}}Y - \tilde{t}^{\mathrm{T}}X^{\mathrm{T}}Y) + \mathrm{V}(l^{\mathrm{T}}\widehat{\beta}) + 2\mathrm{Cov}(a^{\mathrm{T}}Y - \tilde{t}^{\mathrm{T}}X^{\mathrm{T}}Y,\ \tilde{t}^{\mathrm{T}}X^{\mathrm{T}}Y).
\end{aligned}$$

Now, under the assumptions on the variance-covariance matrix of ϵ, which is also the (conditional, given X) variance-covariance matrix of Y, we have

$$\begin{aligned}
\mathrm{Cov}(a^{\mathrm{T}}Y - \tilde{t}^{\mathrm{T}}X^{\mathrm{T}}Y,\ l^{\mathrm{T}}\widehat{\beta}) &= (a^{\mathrm{T}} - \tilde{t}^{\mathrm{T}}X^{\mathrm{T}})\sigma^2 IX\tilde{t} \\
&= (a^{\mathrm{T}}X - \tilde{t}^{\mathrm{T}}X^{\mathrm{T}}X)\sigma^2 I\tilde{t} \\
&= (l^{\mathrm{T}} - l^{\mathrm{T}})\sigma^2 I\tilde{t} \\
&= 0;
\end{aligned}$$

that is,

$$\mathrm{V}(a^{\mathrm{T}}Y) = \mathrm{V}(a^{\mathrm{T}}Y - \tilde{t}^{\mathrm{T}}X^{\mathrm{T}}Y) + \mathrm{V}(l^{\mathrm{T}}\widehat{\beta}).$$

This implies that

$$\mathrm{V}(a^{\mathrm{T}}Y) \geq \mathrm{V}(l^{\mathrm{T}}\widehat{\beta});$$

that is, $l^{\mathrm{T}}\widehat{\beta}$ has minimum variance among the linear unbiased estimators of $l^{\mathrm{T}}\beta$.

To see that it is unique, we consider the case in which $\mathrm{V}(a^{\mathrm{T}}Y) = \mathrm{V}(l^{\mathrm{T}}\widehat{\beta})$; that is, $\mathrm{V}(a^{\mathrm{T}}Y - \tilde{t}^{\mathrm{T}}X^{\mathrm{T}}Y) = 0$. For this variance to equal 0, it must be the case that $a^{\mathrm{T}} - \tilde{t}^{\mathrm{T}}X^{\mathrm{T}} = 0$ or $a^{\mathrm{T}}Y = \tilde{t}^{\mathrm{T}}X^{\mathrm{T}}Y = l^{\mathrm{T}}\widehat{\beta}$ a.s.; that is, $l^{\mathrm{T}}\widehat{\beta}$ is the a.s. unique linear unbiased estimator that achieves the minimum variance. ∎

If we assume further that $\epsilon \sim \mathrm{N}_n(0, \sigma^2 I)$, we see that $l^{\mathrm{T}}\widehat{\beta}$ is the uniformly minimum variance unbiased estimator (UMVUE) for $l^{\mathrm{T}}\beta$. This is because $(X^{\mathrm{T}}Y,\ (Y - X\widehat{\beta})^{\mathrm{T}}(Y - X\widehat{\beta}))$ is complete and sufficient for (β, σ^2). This line of reasoning also implies that $(Y - X\widehat{\beta})^{\mathrm{T}}(Y - X\widehat{\beta})/(n-r)$, where $r = \mathrm{rank}(X)$, is UMVUE for σ^2.

****** biased estimator with smaller MSE

Example 5.27 Inadmissibility of the LSE in the Linear Model
inadmissible under squared-error loss regularization; see page 252 ∎

Optimal Properties of the Moore-Penrose Inverse

The solution corresponding to the Moore-Penrose inverse is unique because that generalized inverse is unique. That solution is interesting for another reason.

Theorem 5.9
Let b^* be any solution to the normal equations (5.72), that is,

$$b^* = (X^{\mathrm{T}}X)^- X^{\mathrm{T}}Y,$$

and let

$$\widehat{\beta} = (X^{\mathrm{T}}X)^+ X^{\mathrm{T}}Y$$

then

$$\|\widehat{\beta}\|_2 \leq \|b^*\|_2.$$

Proof.
To see that this solution has minimum norm, first factor Z, as

$$X = QRU^{\mathrm{T}},$$

and form the Moore-Penrose inverse as

$$X^+ = U \begin{bmatrix} R_1^{-1} & 0 \\ 0 & 0 \end{bmatrix} Q^{\mathrm{T}}.$$

Now let

$$\widehat{\beta} = X^+Y.$$

This is a least squares solution (that is, we have chosen a specific least squares solution).

Now, let

$$Q^{\mathrm{T}}Y = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix},$$

where c_1 has exactly r elements and c_2 has $n-r$ elements, and let

$$U^{\mathrm{T}}b = \begin{pmatrix} t_1 \\ t_2 \end{pmatrix},$$

where b is the variable in the norm $\|Y - Xb\|_2$ that we seek to minimize, and where t_1 has r elements.

Because multiplication by an orthogonal matrix does not change the norm, we have

$$\begin{aligned} \|Y - Xb\|_2 &= \|Q^{\mathrm{T}}(Y - XUU^{\mathrm{T}}b)\|_2 \\ &= \left\| \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} - \begin{bmatrix} R_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{pmatrix} t_1 \\ t_2 \end{pmatrix} \right\|_2 \\ &= \left\| \begin{pmatrix} c_1 - R_1t_1 \\ c_2 \end{pmatrix} \right\|_2. \end{aligned}$$

The residual norm is minimized for $t_1 = R_1^{-1}c_1$ and t_2 arbitrary. However, if $t_2 = 0$, then $\|t\|_2$ is also minimized. Because $U^{\mathrm{T}}b = t$ and U is orthogonal, $\|b\|_2 = \|t\|_2 = \|t_1\|_2 + \|t_2\|_2$, and so with $t_2 = 0$, that is, with $b = \widehat{\beta}$, $\|\widehat{\beta}\|_2$ is the minimum among the norms of all least squares solutions, $\|b^*\|_2$. ∎

Quadratic Forms

Quadratic forms in nonnegative definite or positive definite matrices arise often in statistical applications, especially in the analysis of linear models. The analysis often involves the decomposition of a quadratic form in the positive definite matrix A, $y^{\mathrm{T}}Ay$, into a sum, $y^{\mathrm{T}}A_1y + y^{\mathrm{T}}A_2y$, where $A_1 + A_2 = A$ and A_1 and A_2 are nonnegative definite matrices.

Cochran's Theorems

There are various facts that are sometimes called *Cochran's theorem*. The simplest one concerns k symmetric idempotent $n \times n$ matrices, $A_1, \ldots, A_k$ that sum to the identity matrix.

Theorem 5.10 (Cochran's theorem I)
Let $A_1, \ldots, A_k$ be symmetric idempotent $n \times n$ matrices such that

$$I_n = A_1 + \cdots + A_k.$$

Then

$$A_iA_j = 0 \text{ for all } i \neq j.$$

Proof.
For an arbitrary j, for some matrix V, we have

$$V^{\mathrm{T}}A_jV = \mathrm{diag}(I_r, 0),$$

where $r = \mathrm{rank}(A_j)$. Now

$$\begin{aligned} I_n &= V^{\mathrm{T}}I_nV \\ &= \sum_{i=1}^{k} V^{\mathrm{T}}A_iV \\ &= \mathrm{diag}(I_r, 0) + \sum_{i \neq j} V^{\mathrm{T}}A_iV, \end{aligned}$$

which implies

$$\sum_{i \neq j} V^{\mathrm{T}}A_iV = \mathrm{diag}(0, I_{n-r}).$$

Now for each i, $V^{\mathrm{T}}A_iV$ is idempotent, and because the diagonal elements of a symmetric idempotent matrix are all nonnegative, and hence the equation

implies implies that for each $i \neq j$, the first r diagonal elements are 0. Furthermore, since these diagonal elements are 0, all elements in the first r rows and columns are 0. We have, therefore, for each $i \neq j$,

$$V^{\mathrm{T}} A_i V = \mathrm{diag}(0, B_i)$$

for some $(n-r) \times (n-r)$ symmetric idempotent matrix B_i. Now, for any $i \neq j$, consider $A_i A_j$ and form $V^{\mathrm{T}} A_i A_j V$. We have

$$\begin{aligned} V^{\mathrm{T}} A_i A_j V &= (V^{\mathrm{T}} A_i V)(V^{\mathrm{T}} A_j V) \\ &= \mathrm{diag}(0, B_i)\mathrm{diag}(I_r, 0) \\ &= 0. \end{aligned}$$

Because V is nonsingular, this implies the desired conclusion; that is, that $A_i A_j = 0$ for any $i \neq j$. ∎

We can now extend this result to an idempotent matrix in place of I; that is, for an idempotent matrix A with $A = A_1 + \cdots + A_k$.

Theorem 5.11 (Cochran's theorem II)
Let $A_1, \ldots, A_k$ be $n \times n$ symmetric matrices and let

$$A = A_1 + \cdots + A_k.$$

Then any two of the following conditions imply the third one:

(a). A is idempotent.
(b). A_i is idempotent for $i = 1, \ldots, k$.
(c). $A_i A_j = 0$ for all $i \neq j$.

(The theorem also applies to nonsymmetric matrices if condition (c) is augmented with the requirement that $\mathrm{rank}(A_i^2) = \mathrm{rank}(A_i)$ for all i. We will restrict our attention to symmetric matrices, however, because in most applications of these results, the matrices are symmetric.)

Proof.
First, if we assume properties (a) and (b), we can show that property (c) follows for the special case $A = I$.

Now, let us assume properties (b) and (c) and show that property (a) holds. With properties (b) and (c), we have

$$\begin{aligned} AA &= (A_1 + \cdots + A_k)(A_1 + \cdots + A_k) \\ &= \sum_{i=1}^{k} A_i A_i + \sum_{i \neq j} \sum_{j=1}^{k} A_i A_j \\ &= \sum_{i=1}^{k} A_i \\ &= A. \end{aligned}$$

Hence, we have property (a); that is, A is idempotent.

Finally, let us assume properties (a) and (c). Property (b) follows immediately from

$$A_i^2 = A_i A_i = A_i A = A_i A A = A_i^2 A = A_i^3$$

and the fact that $A^{p+1} = A^p \Longrightarrow A$ is idempotent. ∎

Theorem 5.12 (Cochran's theorem IIa)
Any two of the properties (a) through (c) also imply a fourth property:

(d). $\text{rank}(A) = \text{rank}(A_1) + \cdots + \text{rank}(A_k)$.

Proof.
We first note that any two of properties (a) through (c) imply the third one, so we will just use properties (a) and (b). Property (a) gives

$$\text{rank}(A) = \text{tr}(A) = \text{tr}(A_1 + \cdots + A_k) = \text{tr}(A_1) + \cdots + \text{tr}(A_k),$$

and property (b) states that the latter expression is $\text{rank}(A_1)+\cdots+\text{rank}(A_k)$, thus yielding property (d). ∎

There is also a partial converse: properties (a) and (d) imply the other properties.

One of the most important special cases of Cochran's theorem is when $A = I$ in the sum:

$$I_n = A_1 + \cdots + A_k.$$

The identity matrix is idempotent, so if $\text{rank}(A_1) + \cdots + \text{rank}(A_k) = n$, all the properties above hold. (See Gentle (2007), pages 283–285.)

In applications of linear models, a quadratic form involving Y is often partitioned into a sum of quadratic forms. The most important statistical application of Cochran's theorem is for the distribution of quadratic forms of normally distributed random vectors.

Theorem 5.13 (Cochran's theorem III)
Assume that Y is distributed as $\text{N}_d(\mu, I_d)$, and for $i = 1, \ldots k$, let A_i be a $d \times d$ symmetric matrix with rank r_i such that $\sum_i A_i = I_d$. This yields a partition of the total sum of squares $Y^{\text{T}}Y$ into k components:

$$Y^{\text{T}}Y = Y^{\text{T}}A_1 Y + \cdots + Y^{\text{T}}A_k Y.$$

Then the $Y^{\text{T}}A_i Y$ have independent noncentral chi-squared distributions $\chi^2_{r_i}(\delta_i)$ with $\delta_i = \mu^{\text{T}}A_i\mu$ if and only if $\sum_i r_i = d$.

Proof.
This follows from the results above and the multivariate normal distribution. (See Gentle (2007), pages 324–325.) ∎

The "Sum of Squares" Quadratic Form

In statistical analysis, we often compare the variability within various subsamples with the overall variability of the full sample. This is the basic idea in the common method called analysis of variance (AOV). The variability within any sample is usually measured by the sum of squares of the elements in the sample from their overall mean, $\sum(y_i - \bar{y})^2$.

This sum of squares can be expressed as a quadratic form in an idempotent matrix. We can develop this matrix by use of the expressions for recursive computation of the variance. The basic matrix is the *Helmert matrix* (see Gentle (2007), page 308):

$$H_n = \begin{bmatrix} 1/\sqrt{n} & 1/\sqrt{n} & 1/\sqrt{n} & \cdots & 1/\sqrt{n} \\ 1/\sqrt{2} & -1/\sqrt{2} & 0 & \cdots & 0 \\ 1/\sqrt{6} & 1/\sqrt{6} & -2/\sqrt{6} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \cdots & -\frac{(n-1)}{\sqrt{n(n-1)}} \end{bmatrix} \tag{5.82}$$

Note that the Helmert matrix is orthogonal:

$$H_n^{\mathrm{T}} H_n = H_n H_n^{\mathrm{T}} = I_n.$$

The $(n-1)\times n$ matrix below the first row of the Helmert matrix is of particular interest. Let

$$H_n = \begin{bmatrix} 1/\sqrt{n}\, 1_n^{\mathrm{T}} \\ \cdots\cdots\cdots \\ K_{n-1} \end{bmatrix}. \tag{5.83}$$

First note that the two partitions are orthogonal to each other:

$$1/\sqrt{n}\, 1_n^{\mathrm{T}} K_{n-1} = 0. \tag{5.84}$$

(This also follows from the orthogonality of H_n of course.)

Now let

$$A = K_{n-1}^{\mathrm{T}} K_{n-1}, \tag{5.85}$$

that is,

$$A = \begin{bmatrix} \frac{n-1}{n} & \frac{-1}{n} & \cdots & \frac{-1}{n} \\ \frac{-1}{n} & \frac{n-1}{n} & \cdots & \frac{-1}{n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{-1}{n} & \frac{-1}{n} & \cdots & \frac{n-1}{n} \end{bmatrix} \tag{5.86}$$

Note that, for a sample of size n, A is the matrix of the quadratic form that yields $\sum(x_i - \bar{x})^2$:

$$y^{\mathrm{T}} A y = \sum(y_i - \bar{y})^2. \tag{5.87}$$

We can form similar matrices for subsamples so as to decompose a sum of squares.

Example 5.28 one-way fixed-effects AOV model
Consider the linear model

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad i = 1, \ldots, m; \quad j = 1, \ldots, n, \tag{5.88}$$

where we assume that $\mathrm{E}(\epsilon_{ij}) = 0$ and $\mathrm{V}(\epsilon_{ij}) = \sigma^2$ for all i, j, and $\mathrm{Cov}(\epsilon_{ij}, \epsilon_{i'j'}) = 0$ if $i \neq i'$ or $j \neq j'$. This can be expressed in the form of the linear model (5.66), $Y = X\beta + E$, where $\beta = (\mu, \alpha_1, \ldots, \alpha_m)$ and

$$X = \begin{bmatrix} 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \cdots & 1 \end{bmatrix} \tag{5.89}$$

Letting

$$\overline{Y}_i = \sum_{j=1}^{n} Y_{ij}/n \tag{5.90}$$

and

$$\overline{Y} = \sum_{i=1}^{m} \overline{Y}_i/m, \tag{5.91}$$

we may form two sums of squares

$$\mathrm{SSA} = n \sum_{i=1}^{m} (\overline{Y}_i - \overline{Y})^2 \tag{5.92}$$

and

$$\mathrm{SSE} = \sum_{i=1}^{m} \sum_{j=1}^{n} (Y_{ij} - \overline{Y}_i)^2, \tag{5.93}$$

which have the property that

$$\sum_{i=1}^{m} \sum_{j=1}^{n} (Y_{ij} - \overline{Y})^2 = \mathrm{SSA} + \mathrm{SSE}. \tag{5.94}$$

Both SSA and SSE can be expressed as quadratic forms in matrices similar to $K_{n-1}^{\mathrm{T}} K_{n-1}$, where K_{n-1} is given in equation (5.83). This is what you are asked to do in Exercise 5.7. ∎

The question of interest in a model such is this is whether the α_i are different from one another; that is, whether or not it is meaningful to group the Y_{ij} based on the i.

Example 5.29 estimating the effects in a one-way fixed-effects AOV model (continuation of Example 5.28)
The individual α_i are not U-estimable. We can see this because in this case, $l = (0, \ldots, 1, \ldots)$ and so l is not in the row space of X in (5.89). (This argument follows from the condition in equation (5.79).) We see that $l = (1, \ldots, 1, \ldots)$ and so l is not in the row space of X and so $\mu + \alpha_i$ is estimable, and its UMVUE is $\overline{Y}_i$. Also, $\alpha_i - \alpha_j$ for $i \neq j$ is estimable because it corresponds to an l with first element 0, and all other elements 0 except for two, one of which is 1 the other is -1. Such vectors are called *contrasts*.

For any linear combination of $\beta = (\mu, \alpha_1, \ldots, \alpha_m)$ that is estimable, say $l^{\mathrm{T}}\beta$, we see that the a.s. unique UMVUE is $l^{\mathrm{T}}\widehat{\beta}$, where $\widehat{\beta} = (X^{\mathrm{T}}X)^{+}X^{\mathrm{T}}Y$ (equation (5.74)).

Although the form of the AOV model (5.88) is the one that is commonly used, we see that a closely related model could be formed by restricting this model so that $\sum_i \alpha_i = 0$. This related model is $Y_{ij} = \theta_i + \epsilon_{ij}$. The θ_i in this restricted model are U-estimable. ∎

Notice that so far we have not assumed any specific family of distributions for the AOV model. We have unique UMVUEs. To answer the question posed above of whether the α_i are actually different from one another, however, we need a basis for a statistical test. We might attempt some kind of nonparametric test based on rankings, but in the next example, we will make the common assumption that the random components have a normal distribution. Note that the previous assumption of 0 covariances gives independence if we assume normality. Cochran's theorem tells us what the distributions are.

Example 5.30 distribution of the sums of squares in a one-way fixed-effects AOV model (continuation of Example 5.28)
If we assume that $\epsilon_{ij} \sim \mathrm{N}(0, 1)$, we know the distributions of functions of SSA and SSE, and on that basis we can assess the significance of the α_i. We have

$$\frac{1}{\sigma^2}\left(\mathrm{SSA} - \frac{n}{m-1}\sum_{i=1}^{m}\left(\alpha_i - \sum_{i=1}^{m}\alpha_i/m\right)^2\right) \sim \chi^2_{m-1} \tag{5.95}$$

and

$$\frac{1}{\sigma^2}\mathrm{SSE} \sim \chi^2_{m(n-1)}. \tag{5.96}$$

(Exercise 5.8.) ∎

The UMVUE of σ^2 is $\mathrm{SSE}/(m(n-1))$. Note that the UMVUE of σ^2 is the same as the general result given in equation (5.77). (Exercise 5.9.) The UMVUE is consistent in n for m fixed, and is consistent in m for n fixed.

You are to show this in Exercise 5.10. Compare this with the MLE of σ^2 in Example 6.27 in Chapter 6.

The model in equation (5.88) is called the one-way AOV model. If the α_i in this model are assumed to be constants, it is called a "fixed-effects model". A fixed-effects model is also sometimes called "model I". Now let's consider a variant of this called a "random-effects model" or "model II", because the α_i in this model are assumed to be iid random variables.

Example 5.31 UMVUEs of the variances in the one-way random-effects AOV model
Consider the linear model

$$Y_{ij} = \mu + \delta_i + \epsilon_{ij}, \quad i = 1, \ldots, m; \quad j = 1, \ldots, n, \tag{5.97}$$

where the δ_i are identically distributed with $\mathrm{E}(\delta_i) = 0$, $\mathrm{V}(\delta_i) = \sigma_\delta^2$, and $\mathrm{Cov}(\delta_i, \delta_{\tilde{i}}) = 0$ for $i \neq \tilde{i}$, and the ϵ_{ij} are independent of the δ_i and are identically distributed with with $\mathrm{E}(\epsilon_{ij}) = 0$, $\mathrm{V}(\epsilon_{ij}) = \sigma_\epsilon^2$, and $\mathrm{Cov}(\epsilon_{ij}, \epsilon_{\tilde{i}\tilde{j}}) = 0$ for either $i \neq \tilde{i}$ or $j \neq \tilde{j}$.

An important difference in the random-effects model and the fixed-effects model is that in the random-effects model, we do not have independence of the observables. We have

$$\mathrm{Cov}(Y_{ij}, Y_{\tilde{i}\tilde{j}}) = \begin{cases} \sigma_\delta^2 + \sigma_\epsilon^2 & \text{for } i = \tilde{i},\ j = \tilde{j}, \\ \sigma_\delta^2 & \text{for } i = \tilde{i},\ j \neq \tilde{j}, \\ 0 & \text{for } i \neq \tilde{i}. \end{cases} \tag{5.98}$$

A model such as this may be appropriate when there are a large number of possible treatments and m of them are chosen randomly and applied to experimental units whose responses Y_{ij} are observed. While in the fixed-effects model (5.88), we are interested in whether $\alpha_1 = \cdots = \alpha_m = 0$, in the random-effects model, we are interested in whether $\sigma_\delta^2 = 0$, which would result in a similar practical decision about the treatments.

In the model (5.97) the variance of each Y_{ij} is $\sigma_\delta^2 + \sigma_\epsilon^2$, and our interest in using the model is to make inference on the relative sizes of the components of the variance σ_δ^2 and σ_ϵ^2. The model is sometimes called a "variance components model".

Let us suppose now that $\delta_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma_\delta^2)$, where $\sigma_\delta^2 \geq 0$, and $\epsilon_{ij} \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma_\epsilon^2)$, where as usual $\sigma^2 > 0$. This will allow us to determine exact sampling distributions of the relevant statistics.

We transform the model using Helmert matrices H_m and H_n as in equation (5.82).

Let

$$Y = \begin{bmatrix} Y_{11} & \cdots & Y_{1n} \\ \vdots & & \vdots \\ Y_{m1} & \cdots & Y_{mn} \end{bmatrix}; \quad \delta = \begin{bmatrix} \delta_1 \\ \vdots \\ \delta_m \end{bmatrix}; \quad \text{and } \epsilon = \begin{bmatrix} \epsilon_{11} & \cdots & \epsilon_{1n} \\ \vdots & & \vdots \\ \epsilon_{m1} & \cdots & \epsilon_{mn} \end{bmatrix}.$$

We now write the original model as

$$Y = \delta 1_n^{\mathrm{T}} + \epsilon.$$

Now, for the transformations. Let

$$Z = H_m X H_n^{\mathrm{T}},$$

$$\widetilde{\delta} = H_m \delta,$$

and

$$\widetilde{\epsilon} = H_m \epsilon H_n^{\mathrm{T}}.$$

We first of all note that the transformations are all nonsingular and

$$Z = H 1_n^{\mathrm{T}} + \widetilde{\epsilon}.$$

Next, we see because of the orthonormality of the Helmert matrices that the distributions of $\widetilde{\delta}$ and $\widetilde{\epsilon}$ are the same as those of δ and ϵ and they are still independent. Furthermore, the Z_{ij} are independent, and we have

$$Z_{i1} \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma_a^2 + \sigma^2), \quad \text{for } i = 1, \ldots, m$$

and

$$Z_{ij} \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2), \quad \text{for } i = 1, \ldots, m; j = 2, \ldots, n0.$$

To continue with the analysis, we follow the same steps as in Example 5.28, and get the same decomposition of the "adjusted total sum of squares" as in equation (5.94):

$$\sum_{i=1}^{m} \sum_{j=1}^{n} (Z_{ij} - \overline{Z})^2 = \mathrm{SSA} + \mathrm{SSE}. \tag{5.99}$$

Again, we get chi-squared distributions, but the distribution involving SSA is not the same as in expression (5.95) for the fixed-effects model.

Forming

$$\mathrm{MSA} = \mathrm{SSA}/(m-1)$$

and

$$\mathrm{MSE} = \mathrm{SSE}/(m(n-1)),$$

we see that

$$\mathrm{E}(\mathrm{MSA}) = n\sigma_\delta^2 + \sigma_\epsilon^2$$

and

$$\mathrm{E}(\mathrm{MSE}) = \sigma_\epsilon^2.$$

Unbiased estimators of σ_δ^2 and σ_ϵ^2 are therefore

$$s_\delta^2 = (\mathrm{MSA} - \mathrm{MSE})/n \tag{5.100}$$

and

$$s_\epsilon^2 = \text{MSE}, \tag{5.101}$$

and we can also see that these are UMVUEs.

Now we note something that might at first glance be surprising: s_δ^2 in equation (5.100) may be negative. This occurs if $(m-1)\text{MSA}/m < \text{MSE}$. This will be the case if the variation among Y_{ij} for a fixed i is relatively large compared to the variation among $\overline{Y}_i$ (or similarly, if the variation among Z_{ij} for a fixed i is relatively large compared to the variation among $\overline{Z}_i$). ∎

Compare this with the MLEs in Example 6.29 in Chapter 6

Predictions in the Linear Model

Given a vector x_0, use of $\widehat{\beta}$ in equation (5.64), with E set to $\text{E}(E)$, we have the predicted value of Y given x_0:

$$\begin{aligned}\widehat{Y}_0 &= \widehat{\beta}^{\mathrm{T}} x_0 \\ &= ((X^{\mathrm{T}}X)^+ X^{\mathrm{T}} y)^{\mathrm{T}} x_0.\end{aligned} \tag{5.102}$$

If $x_0 \in \text{span}(X)$, then from Theorem 5.7, $(b^*)^{\mathrm{T}} x_0 = \widehat{\beta}^{\mathrm{T}} x_0$, so in this case the predicted value of Y is invariant to choice of the generalized inverse.

In the model (5.66) corresponding to a set of n observations on the model (5.64), we have predicted values of the response Y at all rows within X:

$$\begin{aligned}\widehat{Y} &= X\widehat{\beta} \\ &= X(X^{\mathrm{T}}X)^+ X^{\mathrm{T}} Y.\end{aligned} \tag{5.103}$$

From equation (3.42), we see that this has the minimum MSE of any function of X.

The idempotent projection matrix $X(X^{\mathrm{T}}X)^+X^{\mathrm{T}}$ is called the "hat matrix" because given Y, it provides $\widehat{Y}$. (See page 795 for properties of projection matrices.)

We see from Definition 1.46 page 116 that $\widehat{Y}$ is the projection of Y onto the column space of X. (This is a slightly different meaning of the word "projection", but obviously the meanings are related.) From Theorem 1.64 we see that the "residual vector" $Y - \widehat{Y}$ is orthogonal to the columns of X; that is, $\text{Cov}(Y - \widehat{Y}, x) = 0$ for any column x of X, and since $\widehat{Y}$ is a linear combination of the columns of X, $\text{Cov}(Y - \widehat{Y}, \widehat{Y}) = 0$. If we assume a normal distribution for ϵ, then 0 covariance implies independence.

5.5.2 Estimation in Survey Samples of Finite Populations

A substantial proportion of all applications of statistics deal with sample surveys in finite populations. Some aspects of this kind of application distinguish

it from other areas of applied statistics. Särndal et al. (1997) provide a general coverage of the theory and methods. Valliant et al. (2000) provide a different perspective on some of the particular issues of inference in finite populations.

Finite Populations

We think of a finite population as being a finite set $\mathcal{P} = \{(1, y_1), \ldots, (N, y_N)\}$. Our interest will be in making inferences about the population using a sample $\mathcal{S} = \{(L_1, X_1), \ldots, (L_n, X_n)\}$. We will also refer to $X = \{X_1, \ldots, X_n\}$ as the "sample". In discussions of sampling it is common to use n to denote the size of the sample and N to denote the size of the population. Another common notation used in sampling is Y to denote the population total, $Y = \sum_{i=1}^{N} y_i$. Estimation of the total is one of the most basic objectives in sampling applications.

The parameter that characterizes the population is $\theta = (y_1, \ldots, y_N)$. The parameter space, Θ, is the subspace of $\mathbb{R}^N$ containing all possible values of the y_i.

There are two approaches to the analysis of the problem. In one, which is the more common and which we will follow, $\mathcal{P}$ is essentially the sample space. In another approach $\mathcal{P}$ or θ is thought of as some random sample from a sample space or parameter space, called a "superpopulation".

The sample is completely determined by the set $\mathcal{L}_S = \{i_1, \ldots, i_n\}$ of indexes of $\mathcal{P}$ that correspond to elements in X. For analysis of sampling methods, we define an indicator

$$I_i = \begin{cases} 1 \text{ if } i \in \mathcal{L}_S \\ 0 \text{ othersise.} \end{cases}$$

"Sampling" can be thought of as selecting the elements of $\mathcal{L}_S$, that is, the labels of the population elements.

Probability-based inferences about $\mathcal{P}$ are determined by the method of selection of $\mathcal{S}$. This determines the probability of getting any particular $\mathcal{S}$, which we will denote by $p(\mathcal{S})$. If $p(\mathcal{S})$ is constant for all $\mathcal{S}$, we call the selected sample a *simple random sample*.

A sample may be collected *without replacement* or *with replacement*. (The meanings of these are just what the words mean. In sampling without replacement, the elements of $\mathcal{S}$ are distinct.) Sampling with replacement is generally easier to analyze, because it is the same as taking a random sample from a discrete uniform distribution. Sampling without replacement is more common and it is what we will assume throughout.

There are many variations on the method of collecting a sample. Both a general knowledge of the population and some consideration of the mechanical aspects of collecting the sample may lead to the use of *stratified sampling*, *cluster sampling*, *multi-stage sampling*, *systematic sampling*, or other variations.

Estimation

We are interested in "good" estimators, specifically UMVUEs, of estimable functions of θ. An interesting estimable function of θ is $Y = \sum_{i=1}^{N} \theta_i$.

One of the most important results is the following theorem.

Theorem 5.14
(i) if $p(\mathcal{S}) > 0$ for all $\mathcal{S}$, then the set of order statistics $X_{(1)}, \ldots, X_{(n)}$ is complete for all $\theta \in \Theta$.
and
(ii) if $p(\mathcal{S})$ is constant for all $\mathcal{S}$, then the order statistics $X_{(1)}, \ldots, X_{(n)}$ are sufficient for all $\theta \in \Theta$.

This theorem is somewhat similar to Corollary 3.1.1, which applied to the family of distributions dominated by Lebesgue measure. The sufficiency is generally straightforward, and we expect it to hold in any iid case.

The completeness is a little more complicated, and the proof of Theorem 3.13 in MS2 is worth looking at. The set of order statistics may be complete in some family, such as *the* family of distributions dominated by Lebesgue measure, but may not be complete in some subfamily, such as the family of normal distributions with mean 0.

After we have (i) and (ii), we have
(iii): For any estimable function of θ, its unique UMVUE is the unbiased estimator $T(X_1, \ldots, X_n)$ that is symmetric in its arguments. (The symmetry makes the connection to the order statistics.)

Example 5.32 UMVUE of population total using simple random sample
Consider estimation of $Y = g(\theta) = \sum_{i=1}^{N} y_i$ from the simple random sample $X_1, \ldots, X_n$. We first note that

$$\begin{aligned} \widehat{Y} &= \frac{N}{n} \sum_{i \in \mathcal{L}_S} y_i \\ &= \frac{N}{n} \sum_{i=1}^{N} I_i y_i \end{aligned}$$

is unbiased for Y:

$$\begin{aligned} \mathrm{E}(\widehat{Y}) &= \frac{N}{n} \sum_{i=1}^{N} y_i \mathrm{E}(I_i) \\ &= \sum_{i=1}^{N} y_i. \end{aligned}$$

From Theorem 5.14, we can see easily that $\widehat{Y} = N\overline{y}$ is the UMVUE of Y.

Now we consider the variance of $\widehat{Y}$. First, note that

$$\mathrm{V}(I_i) = \frac{n}{N}\left(1 - \frac{n}{N}\right)$$

(it's Bernoulli), and for $i \neq j$,

$$\begin{aligned}\mathrm{Cov}(I_i, I_j) &= \mathrm{E}(I_i I_j) - \mathrm{E}(I_i)\mathrm{E}(I_j) \\ &= \frac{n(n-1)}{N(N-1)} - \frac{n^2}{N^2}.\end{aligned}$$

Hence,

$$\begin{aligned}\mathrm{V}(\widehat{Y}) &= \frac{N^2}{n^2}\mathrm{V}\left(\sum_{i=1}^{N} I_i y_i\right) \\ &= \frac{N^2}{n^2}\left(\sum_{i=1}^{N} y_i^2 \mathrm{V}(I_i) + 2\sum_{1\leq i\leq j\leq N} y_i y_j \mathrm{Cov}(I_i, I_j)\right) \\ &= \frac{N}{n}\left(1 - \frac{n}{N}\right)\left(\sum_{i=1}^{N} y_i^2 - \frac{2}{N-1}\sum_{1\leq i\leq j\leq N} y_i y_j\right) \\ &= \frac{N^2}{n}\left(1 - \frac{n}{N}\right)\frac{1}{N-1}\sum_{i=1}^{N}\left(y_i - \frac{Y}{N}\right)^2. \qquad (5.104)\end{aligned}$$

We see that the variance of $\widehat{Y}$ is composed of three factors, an expansion factor N^2/n, a finite population correction factor $(1-n/N)$, and the variance of a selection from a finite population,

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(y_i - \frac{Y}{N}\right)^2. \qquad (5.105)$$

∎

The sample variance S^2 is unbiased for σ^2, and so from this we have immediately the UMVUE of $\mathrm{V}(\widehat{Y})$ (Exercise 5.11).

Horvitz-Thompson Estimation

The properties of any statistic derived from a sample $X_1, \ldots, X_n$ depend on the *sampling design*; that is, on how the items in the sample were selected. The two main properties of the design are the probability that a specific population item, say y_i, is selected, and the probability that two specific population items, say y_i and y_j are both selected. Probabilities of combinations of larger sets may also be of interest, but we can work out simple expectations and variances just based on these two kinds of probabilities.

Let π_i be the probability that y_i is included in the sample, and let π_{ij} be the probability that both y_i and y_j are included.

If $\pi_i > 0$ for all i, the *Horvitz-Thompson estimator* of the population total is

$$\widehat{Y}_{\mathrm{HT}} = \sum_{i \in \mathcal{L}_S} \frac{y_i}{\pi_i}. \tag{5.106}$$

It is easy to see that $\widehat{Y}_{\mathrm{HT}}$ is unbiased for Y:

$$\begin{aligned} \mathrm{E}\left(\widehat{Y}_{\mathrm{HT}}\right) &= \mathrm{E}\left(\sum_{i \in \mathcal{L}_S} \frac{y_i}{\pi_i}\right) \\ &= \sum_{i=1}^{N} \left(\frac{y_i}{\pi_i}\pi_i\right) \\ &= \sum_{i=1}^{N} y_i. \end{aligned}$$

The variance of the Horvitz-Thompson estimator depends on the π_{ij} as well as the π_i:

$$\mathrm{V}\left(\widehat{Y}_{\mathrm{HT}}\right) = \sum_{i=1}^{N} \sum_{j=i+1}^{N} (\pi_i \pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 \tag{5.107}$$

(Exercise 5.12). Expressions for other sampling estimators are often shown in a similar manner.

An important approximation for working out variances of more complicated sampling estimators is linearization, especially when the estimator involves a ratio.

Notes and Further Reading

Most of the material in this chapter is covered in MS2 Chapter 3 and Section 4.5, and in TPE2 Chapter 2.

Unbiasedness

The property of unbiasedness for point estimators was given a solid but preliminary treatment by Halmos (1946).

Unbiasedness has a heuristic appeal, although people will sometimes question its relevance by pointing out that it provides no guarantee of the goodness of an estimator in a single set of data. That argument, however, could apply to most measures of the quality of an estimator. Similar types of arguments could bring into question any consideration of asymptotic properties of statistical procedures.

Unbiasedness is particularly useful when the loss is squared-error, because in that case unbiasedness may lead to uniformly minimum risk estimators. For absolute-error loss functions, a corresponding approach would be to require median unbiasedness.

Fisher Efficient Estimators and Exponential Families

Fisher efficient estimators occur only in exponential families, and there is always one in an exponential family. This fact had been know for some time, but the first rigorous proof was given by Wijsman (1973).

U-Statistics

The fundamental paper by Hoeffding (1948) considered the asymptotic normality of certain unbiased point estimators and introduced the class of estimators that he named U-statistics. Serfling (1980) provides an extensive discussion of U-statistics, as well as V-statistics. The statement and proof Theorem 5.3 and the use of the conditional kernels h_k as in equation (5.48) follow Serfling. Kowalski and Tu (2008) consider several applications of U-statistics in a variety of settings.

Exercises

5.1. Show that the estimator (5.3) in Example 5.1 is the UMVUE of π. (Note that there are three things to show: $(t-1)/(N-1)$ is unbiased, it has minimum variance among all unbiased estimators, and it is unique — "the" implies uniqueness.)

5.2. Consider the problem of using a sample of size 1 for estimating $g(\theta) = \mathrm{e}^{-3\theta}$ where θ is the parameter in a Poisson distribution.
   a) Show that $T(X) = (-2)^X$ is unbiased for $g(\theta)$.
   b) Show that $T(X) = (-2)^X$ is a UMVUE $g(\theta)$.
   c) What is wrong with this estimator?

5.3. Show that the estimators (5.11) and (5.7) are the same.

5.4. Show that the $h(T)$s in Example 5.6 are unbiased for the $g(\theta)$s given.

5.5. Define an alternative kernel for U-statistic that is unbiased for the covariance in Example 5.19; that is, instead of the kernel in equation (5.37), give a kernel similar to that in equation (5.46). Show that the resulting U-statistic is unbiased for the covariance.

5.6. In the setup of model (5.64), show that the LSE $\|Y - X\widehat{\beta}\|^2/(n-p)$ is unbiased for σ^2.

5.7. Let $X_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad i = 1, \ldots, m, \quad j = 1, \ldots, n$, where α_i's and ϵ_{ij}'s are independent random variables, $\alpha_i \sim \mathrm{N}(0, \sigma_\alpha^2)$, $\epsilon_{ij} \sim \mathrm{N}(0, \sigma_\epsilon^2)$, and μ, σ_α^2, and σ_ϵ^2 are unknown parameters. Let

$$\overline{X}_i = \sum_{j=1}^{n} X_{ij}/n,$$

$$\overline{X} = \sum_{i=1}^{m} \overline{X}_i/m,$$

$$\text{MSA} = n\sum_{i=1}^{m}(\overline{X}_i - \overline{X})^2/(m-1),$$

and

$$\text{MSE} = \sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij} - \overline{X}_i)^2/(m(n-1)).$$

Express MSA and MSE as quadratic forms using parts of Helmert matrices and use Chochran's theorem to show that they are independent.

5.8. Show that the quantities in expressions (5.95) and (5.96) have the chi-squared distributions claimed.

5.9. Show that the UMVUE of σ^2, $\text{SSE}/(m(n-1))$, given in Example 5.28 is the same as the UMVUE of σ^2 for the general linear model given in equation (5.77).
Hint: Write the model given in equation (5.88) in the form of the general linear model in equation (5.67).

5.10. Suppose $X_{ij} \overset{\text{iid}}{\sim} \text{N}(\mu_i, \sigma^2)$ for $i = 1, \ldots, m$ and $j = 1, \ldots, n$. (Compare the one-way AOV model of Examples 5.28, 5.29, and 5.30.)
a) Determine the UMVUE $T_{mn}(X)$ of σ^2.
b) Show that $T_{mn}(X)$ is consistent in mean squared error for σ^2 as $m \to \infty$ and n remains fixed.
c) Show that $T_{mn}(X)$ is consistent in mean squared error for σ^2 as $n \to \infty$ and m remains fixed.

5.11. Show that the sample variance S^2 is the UMVUE of σ^2 in equation (5.105) of Example 5.32. Hence, determine the UMVUE of $\text{V}(\widehat{Y})$.

5.12. Show that the variance of the Horvitz-Thompson estimator is as shown in equation (5.107), for given π_i and π_{ij}. This is tedious, but it requires very little other than "advanced arithmetic" and simple properties of variances of sums.

6

Statistical Inference Based on Likelihood

The concepts of probability theory can be applied to statistical analyses in a very straightforward manner: we assume that observed events are governed by some *data-generating process* that depends on a probability distribution P, and our observations of those events can be used to make inferences about the probability distribution. The various ways that we use the observations to make those inferences constitute the main body of statistical theory. One of the general approaches that I outlined in Section 3.2 involves the use of a likelihood function. We considered this approach briefly in Section 3.2.1. In this chapter, we will explore the use of likelihood in statistical inference more fully. In this chapter, the emphasis will be on estimation, and in Chapter 7, we will consider use of likelihood in testing statistical hypotheses.

Although methods based on the likelihood may not have the logical appeal of methods based on a decision-theoretic approach, they do have an intuitive appeal. More importantly, estimators and tests based on this approach have a number of desirable mathematical properties, especially asymptotic properties. Methods based on maximizing the likelihood are grounded on the likelihood principle.

We begin with some general definitions and notation for methods based on the likelihood principle, and then look at specific applications.

6.1 The Likelihood Function and Its Use in Statistical Inference

Definition 6.1 (likelihood function)
Given a sample $x_1, \ldots, x_n$ from distributions with probability densities $p_i(x)$ with respect to a common σ-finite measure, the *likelihood function* is defined as

$$L_n(p_i\,;\,x) = c\prod_{i=1}^{n} p_i(x_i), \tag{6.1}$$

where $c \in \mathbb{R}_+$ is any constant independent of the p_i. ∎

It is common to speak of $L_n(p_i\,;\,X)$ with $c = 1$ as "the" likelihood function, and in the following, we will not write the c.

Methods based on the likelihood function are often chosen because of their asymptotic properties, and so it is common to use the n subscript as in equation (6.1); in the following, however, we will usually find it convenient to drop the n.

As we generally do in discussing methods of statistical inference, in some cases, we will view the sample $x_1, \ldots, x_n$ as a set of constants. In cases when we want to consider the probabilistic or statistical properties of the statistical methods, we will view the observations as a vector of random variables.

In equation (6.1), the domain of the likelihood function is some class of distributions specified by their probability densities, $\mathcal{P} = \{p_i(x)\}$, where all PDFs are dominated by a common σ-finite measure. In applications, often the PDFs are of a common parametric form, so equivalently, we can think of the domain of the likelihood function as being a parameter space, say Θ. In that case, the family of densities can be written as $\mathcal{P} = \{p_\theta(x)\}$ where $\theta \in \Theta$, the known parameter space. It is usually more convenient to write $p_\theta(x)$ as $p(x\,;\,\theta)$, and we often write the likelihood function (6.1) as

$$L(\theta\,;\,x) = \prod_{i=1}^{n} p(x_i\,;\,\theta). \tag{6.2}$$

Although in equation (6.2), we have written $L(\theta\,;\,x)$, the expression $L(p_\theta\,;\,x)$ may be more appropriate because it reminds us of an essential ingredient in the likelihood, namely a PDF.

What Likelihood Is Not

The differences in a likelihood and a PDF are illustrated clearly in Example 1.5 on page 20. A likelihood is neither a probability nor a probability density. Notice, for example, that while the definite integrals over $\mathbb{R}_+$ of both PDFs in in Example 1.5 are 1, the definite integrals over $\mathbb{R}_+$ of the likelihood (1.21) in Example 1.5 are not the same, as we can easily see from the plots on the right side of Figure 1.2.

It is not appropriate to refer to the "likelihood of an observation". We use the term "likelihood" in the sense of the likelihood of a model or the likelihood of a distribution *given observations*.

The Log-Likelihood Function

The *log-likelihood function*,

$$l_L(\theta\,;\,x) = \log L(\theta\,;\,x), \tag{6.3}$$

is a sum rather than a product. We often denote the log-likelihood without the "L" subscript. The notation for the likelihood and the log-likelihood varies with authors. My own choice of an uppercase "L" for the likelihood and a lowercase "l" for the log-likelihood is long-standing, and not based on any notational optimality consideration. Because of the variation in the notation for the log-likelihood, I will often use the "l_L" notation because this expression is suggestive of the meaning.

We will often work with either the likelihood or the log-likelihood as if there is only one observation.

Likelihood Principle

According to the *likelihood principle* in statistical inference all of the information that the data provide concerning the relative merits of two hypotheses is contained in the likelihood ratio of those hypotheses and the data; that is, if for x and y,

$$\frac{L(\theta\,;\,x)}{L(\theta\,;\,y)} = c(x,y) \quad \forall\theta, \tag{6.4}$$

where $c(x,y)$ is constant for given x and y, then any inference about θ based on x should be in agreement with any inference about θ based on y.

Although at first glance, we may think that the likelihood principle is so obviously the right way to make decisions, Example 6.1 may cause us to think more critically about this principle.

The likelihood principle asserts that for making inferences about a probability distribution, the overall data-generating process need not be considered; only the observed data are relevant.

Example 6.1 The likelihood principle in sampling from a Bernoulli distribution

In Example 3.12 we considered the problem of making inferences on the parameter π in a family of Bernoulli distributions.

One approach was to take a random sample of size n, $X_1, \ldots, X_n$ from the Bernoulli(π), and then use $T = \sum X_i$, which has a binomial distribution with parameters n and π.

Another approach was to take a sequential sample, $X_1, X_2, \ldots$, until a fixed number t of 1's have occurred. The size of the sample N is random and the random variable N has a negative binomial distribution with parameters t and π.

Now, suppose we take the first approach with $n = n_0$ and we observe $T = t_0$; and then we take the second approach with $t = t_0$ and we observe $N = n_0$. Using the PDFs in equations 3.43 and 3.44 we get the likelihoods

$$L_{\mathrm{B}}(\pi) = \binom{n_0}{t_0}\pi^{t_0}(1-\pi)^{n_0-t_0} \tag{6.5}$$

and

$$L_{\mathrm{NB}}(\pi) = \binom{n_0 - 1}{t_0 - 1} \pi^{t_0} (1-\pi)^9. \tag{6.6}$$

Because $L_{\mathrm{B}}(\pi)/L_{\mathrm{NB}}(\pi)$ does not involve π, the maxima of the likelihoods will occur at the same point. A maximum likelihood estimator of π based on a binomial observation of t_0 out of n_0 is the same as a maximum likelihood estimator of π based on a negative binomial observation of n_0 for t_0 1's because the maximum of the likelihood occurs at the same place, t_0/n_0. The estimators conform to the likelihood principle. Recall that the UMVU estimators are different. (Example 5.1 and follow-up in Example 5.5 and Exercise 5.1.) ∎

Further comments on Example 6.1

We see that the likelihood principle allows the likelihood function to be defined as any member of an equivalence class $\{cL : c \in \mathbb{R}_+\}$, as in the definition (3.45).

The likelihood principle, however, is stronger than just the requirement that the estimator be invariant. It says that because $L_{\mathrm{B}}(\pi)/L_{\mathrm{NB}}(\pi)$ does not involve π, any decision about π based on a binomial observation of 3 out of 12 should be the same as any decision about π based on a negative binomial observation of 12 for 3 1's. Because the variance of $\hat{\pi}$ *does* depend on whether a binomial distribution or a negative binomial distribution is assumed, the fact that the estimators are the same does not imply that the inference follows the likelihood principle. See Example 6.9. ∎

We will revisit this example again in Example 7.12 on page 539, where we wish to test a statistical hypothesis concerning π. We get different conclusions in a significance test.

6.2 Maximum Likelihood Parametric Estimation

Let us assume a parametric model; that is, a family of densities $\mathcal{P} = \{p(x\,;\,\theta)\}$ where $\theta \in \Theta$, a known parameter space.

For a sample $X_1, \ldots, X_n$ from a distribution with probability density $p(x\,;\,\theta)$, we write the likelihood function as a function of a variable in place of the parameter:

$$L(t\,;\,x) = \prod_{i=1}^{n} p(x_i\,;\,t). \tag{6.7}$$

Note the reversal in roles of variables and parameters. While I really like to write the likelihood as a function of a variable of something other than the parameter, which I think of as fixed, I usually write it like everyone else; that is, I write

$$L(\theta\,;\,x) = \prod_{i=1}^{n} p(x_i\,;\,\theta).$$

In the likelihood function the data, that is, the realizations of the variables in the density function, are considered as fixed, and the parameters are considered as variables of the optimization problem,

$$\max_{\theta} L(\theta\,;\,x). \tag{6.8}$$

For given x, the relative values of $L(\theta\,;\,x)$ are important. For given x_1 and x_2, the relative values of $L(\theta\,;\,x_1)$ and $L(\theta\,;\,x_2)$ are not relevant. Notice in Example 1.5, while $L(\theta\,;\,5) \leq L(\theta\,;\,1)$ for all θ, $\max L(\theta\,;\,5)$ occurs at $\theta = 5$, and $\max L(\theta\,;\,1)$ occurs at $\theta = 1$. Notice also in Example 6.1, while $L_{\mathrm{B}}(\pi)$ in equation (6.5) is uniformly less than $L_{\mathrm{NB}}(\pi)$ in equation (6.6), they both achieve their maximum at the same point, $\pi = 1/4$.

Closure of the Parameter Space

It is important to specify the domain of the likelihood function. If Θ is the domain of L in equation (6.7), we want to maximize L for $t \in \Theta$; that is, maximum likelihood often involves a constrained optimization problem.

There may be difficulties with this maximization problem (6.8), however, because of open sets. The first kind of problem is because the parameter space may be open. We address that problem in our definition the optimal estimator below. See Example 6.4. The second kind of open set may be the region over which the likelihood function is positive. This problem may arise because the support of the distribution is open and is dependent on the parameter to be estimated. We address that problem by adding a zero-probability set to the support (see Example 6.5 below).

For certain properties of statistics that are derived from a likelihood approach, it is necessary to consider the parameter space Θ to be closed (see, for example, Wald (1949)). Often in a given probability model, such as the exponential or the binomial, we do not assume Θ to be closed. If Θ is not a closed set, however, the maximum in (6.8) may not exist, so we consider the closure of Θ, $\overline{\Theta}$. (If Θ is closed $\overline{\Theta}$ is the same set, so we can always just consider $\overline{\Theta}$.)

6.2.1 Definition and Examples

Definition 6.2 (maximum likelihood estimate; estimator)
Let $L(\theta\,;\,x)$ be the likelihood of $\theta \in \Theta$ for the observations x from a distribution with PDF with respect to a σ-finite measure ν. A *maximum likelihood estimate*, or MLE, of θ, written $\widehat{\theta}$, is defined as

$$\widehat{\theta} = \arg\max_{\theta \in \overline{\Theta}} L(\theta\,;\,x), \tag{6.9}$$

if it exists. There may be more than solution; any one is an MLE. If x is viewed as a random variable, then $\widehat{\theta}$ is called a *maximum likelihood estimator* of θ. ∎

While I like to use the "hat" notation to mean an MLE, I also sometimes use it to mean any estimate or estimator.

The estimate (or estimator) $\widehat{\theta}$ is a Borel function of the observations or of the random variables.

We use "MLE" to denote either a maximum likelihood estimate or estimator, or to denote the method of maximum likelihood estimation. The proper meaning can be determined from the context. If the term MLE is used in a statement about a maximum likelihood estimate or estimator, the statement can be assumed to apply to both the estimate and the estimator.

If $\widehat{\theta}$ in (6.9) exists, we also have

$$\widehat{\theta} = \arg\max_{\theta \in \overline{\Theta}} l_L(\theta\,;\,x), \tag{6.10}$$

that is, the MLE can be identified either from the likelihood function or from the log-likelihood.

The Likelihood Equations

Notice that finding an MLE means to solve a constrained optimization problem. In simple cases, the constraints may not be active. In even simpler cases, the likelihood is differentiable, and the MLE occurs at a stationary point in the interior of the constraint space. In these happy cases, the MLE can be identified by differentiation.

If the likelihood function or the log-likelihood function is differentiable within Θ°, we call

$$\nabla L(\theta\,;\,x) = 0 \tag{6.11}$$

or

$$\nabla l_L(\theta\,;\,x) = 0 \tag{6.12}$$

the *likelihood equations*.

If $\theta_r \in \Theta^\circ$ is a root of the likelihood equations and if the Hessian $\mathrm{H}_L(\theta_r)$ evaluated at θ_r is negative definite, then $\theta_r \in \Theta^\circ$ is a local optimizer of L (and of l_L). (See Theorem 0.0.13.)

If the maximum occurs within Θ°, then every MLE is a root of the likelihood equations. There may be other roots within Θ°, of course. Any such root of the likelihood equation, called an RLE, may be of interest.

Example 6.2 MLE in the exponential family (continuation of Example 1.5)
In the exponential family of Example 1.5, with a sample $x_1, \ldots, x_n$, the likelihood in equation (1.21) becomes

$$L(\theta\,;\,x) = \theta^{-n} \mathrm{e}^{-\sum_{i=1}^n x_i/\theta} \mathrm{I}_{\mathbb{R}_+}(\theta),$$

whose derivative wrt θ is

$$\left(-n\theta^{-n-1}\mathrm{e}^{-\sum_{i=1}^{n} x_i/\theta} + \theta^{-n-2}\sum_{i=1}^{n} x_i \mathrm{e}^{-\sum_{i=1}^{n} x_i/\theta}\right) \mathrm{I}_{\mathbb{R}_+}(\theta).$$

Equating this to zero, we obtain

$$\widehat{\theta} = \sum_{i=1}^{n} x_i/n$$

as a stationary point. Checking the second derivative, we find it is negative at $\widehat{\theta}$, and so we conclude that $\widehat{\theta}$ is indeed the MLE of θ, and it is the only maximizer. Also, from the plot on the right side of Figure 1.2, we have visual confirmation. Of course, Figure 1.2 is for a sample of size one.

We can easily see that for a sample of size n this graph would be similar, but it would have a sharper peak; see Figure 6.1.

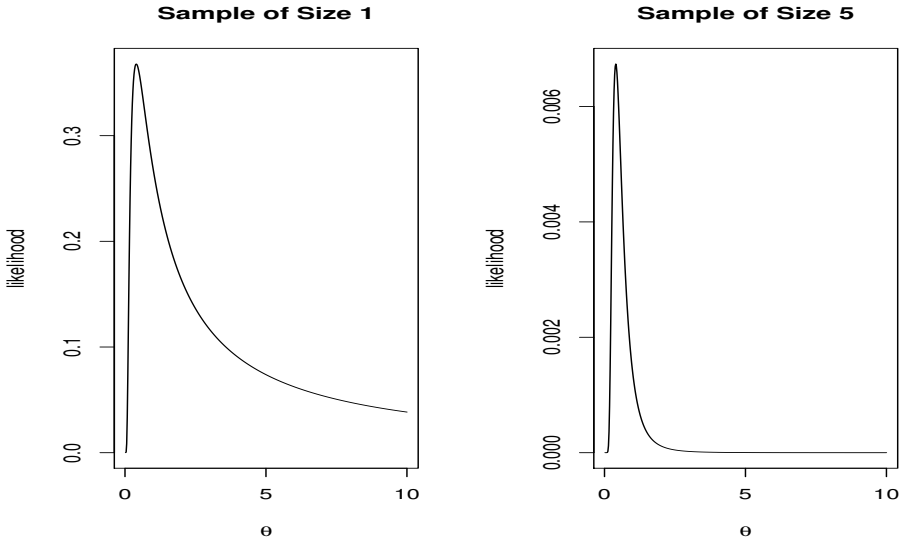


Figure 6.1. Likelihood for Different Sample Sizes

The fact that the likelihood has a sharper peak is in agreement with our expectation that the estimate should be "better" if we have a larger sample. ∎

Example 6.3 MLE in the exponential family with right censoring
In use of the exponential family for modeling "lifetimes", say of lightbulbs, it is often the case that the experiment is terminated before all of the random variables are realized; that is, we may have a potential sample $x_1, \ldots, x_n$,

but actually we only have values for the $x_i < t_c$, where t_c is some fixed and known value. It might be called the "censoring time". This setup is yields *censored data*, in particular, it is *right censored data*, because the larger values are censored. Suppose that $t_1, \ldots, t_r$ observations are obtained, leaving $n - r$ unobserved values of the potential sample. In this setup, the time t_c is fixed, and so r is a random variable. We could also contemplate an experimental setup in which r is chosen in advance, and so the censoring time t_c is a random variable. (These two data-generating processes are similar to the two experiments we described for Bernoulli data in Example 3.12, and to which we have alluded in other examples.) The first method is called "Type I censoring" (upper bound on the observation fixed) and the other method is called "Type II censoring" (fixed number of observed values to be taken).

Censoring is different from a situation in which the distribution is truncated, as in Exercise 2.14 on page 203.

For right censored data with n, r, and t_c as described above from any distribution with PDF $f(x;\theta)$ and CDF $F(x;\theta)$, the likelihood function is

$$L(\theta; x) = \prod_{i=1}^{r} f(t_i;\theta)(1 - F(t_c;\theta))^{n-r}.$$

We may note in passing that the likelihood is the same for type I and type II censoring, just as we saw it to be in the binomial and negative binomial distributions arising from Bernoulli data in Example 3.12.

Now, for the case where the distribution is exponential with parameter θ, we have the likelihood function

$$L(\theta; x)\frac{1}{\theta^r}\mathrm{e}^{\sum_{i=1}^{r} t_i/\theta}\mathrm{e}^{(n-r)t_c/\theta}.$$

The maximum, which we can find by differentiation, occurs at

$$\widehat{\theta} = T/r,$$

where $T = \sum_{i=1}^{r} t_i + (n-r)t_c$ is called the "total time on test". ∎

MLE in $\partial\Theta$

If Θ is open, and if the maximizer in equation (6.9) is in $\partial\Theta$, the distribution defined by the MLE may be degenerate, as can be the case in the following example.

Example 6.4 MLE of Bernoulli parameter
Consider the Bernoulli family of distributions with parameter π. In the usual definition of this family, $\pi \in \Pi =]0, 1[$. Suppose we take a random sample $X_1, \ldots, X_n$. The log-likelihood is

$$l_L(\pi\,;\, x) = \sum_{i=1}^{n} x_i \log(\pi) + \left(n - \sum_{i=1}^{n} x_i\right) \log(1-\pi). \tag{6.13}$$

This is a concave differentiable function, so we can get the maximum by differentiating and setting the result to zero. We obtain

$$\widehat{\pi} = \sum_{i=1}^{n} x_i/n. \tag{6.14}$$

If $\sum_{i=1}^{n} x_i = 0$ or if $\sum_{i=1}^{n} x_i = n$, $\widehat{\pi} \notin \Pi$, but $\widehat{\pi} \in \overline{\Pi}$ so $\widehat{\pi}$ is the MLE of π.

Note that in this case, the MLE corresponds to the Bayes estimator with loss function (4.52) and uniform prior (see page 360) and to the UMVUE (see page 394). ∎

Further comments on Example 6.4

In Example 6.1 we considered the problem of making inferences on the parameter π in a family of Bernoulli distributions, and considered two different approaches. One approach was to take a random sample of size n, $X_1, \ldots, X_n$ from the Bernoulli(π), and then use $T = \sum X_i$, which has a binomial distribution with parameters n and π. Another approach was to take a sequential sample, $X_1, X_2, \ldots$, until a fixed number t of 1's have occurred. The likelihood principle tells us that if the data are the same, we should reach the same conclusions. In Example 6.1 we wrote the likelihood functions based on these two different approaches. One was the same as in equation (6.13) and so the MLE under that setup would be that given in equation (6.14). After canceling constants, the other log-likelihood in Example 6.1 was also

$$\sum_{i=1}^{n} x_i \log(\pi) + \left(n - \sum_{i=1}^{n} x_i\right) \log(1-\pi),$$

so a that sampling scheme yields the same MLE, if n and $\sum_{i=1}^{n} x_i$ are the same.

Of course making inferences about a parameter involves more than just obtaining a good estimate of it. We will consider the problem again in Examples 6.9 and 7.12. ∎

Allowing an MLE to be in $\overline{\Theta} - \Theta$ is preferable to saying that an MLE does not exist. It does, however, ignore the question of continuity of $L(\theta\,;x)$ over $\overline{\Theta}$, and it allows an estimated PDF that is degenerate.

We have encountered this situation before in the case of UMVUEs; see Example 5.5.

While the open parameter space in Example 6.4 would lead to a problem with existence of the MLE if its definition was as a maximum over the parameter space instead of its closure, an open support can likewise lead to a problem. Consider a distribution with Lebesgue PDF

$$p_X(x) = h(x,\theta)\mathrm{I}_{S(\theta)}(x) \tag{6.15}$$

where $S(\theta)$ is open. In this case, the likelihood has the form

$$L(\theta\,;\,x) = h(x,\theta)\mathrm{I}_{R(x)}(\theta), \tag{6.16}$$

where $R(x)$ is open. It is quite possible that $\sup L(\theta\,;\,x)$ will occur on $\overline{R}(x) - R(x)$.

Example 6.5 MLE in $\mathrm{U}(0,\theta)$; closed support
Consider $X_1,\ldots,X_n \overset{\text{iid}}{\sim} \mathrm{U}(0,\theta)$, with $\theta \in \Theta = \mathbb{R}_+$. The PDF is

$$p_X(x) = \frac{1}{\theta}\mathrm{I}_{[0,\theta]}(x). \tag{6.17}$$

The likelihood is

$$L(\theta\,;\,x) = \frac{1}{\theta}\mathrm{I}_{[x_{(n)},\infty[}(\theta). \tag{6.18}$$

The MLE is easily seen to be $\widehat{\theta} = x_{(n)}$. In Example 5.8, we saw that the UMVUE of θ is $(1+1/n)x_{(n)}$.

Suppose we take the support to be the open interval $]0,\theta[$. (Despite Definition 1.12, such a support is often assumed.) The likelihood function then is

$$L(\theta\,;\,x) = \frac{1}{\theta}\mathrm{I}_{]x_{(n)},\infty[}(\theta).$$

This is discontinuous and it does not have a maximum, as we see in Figure 6.2.

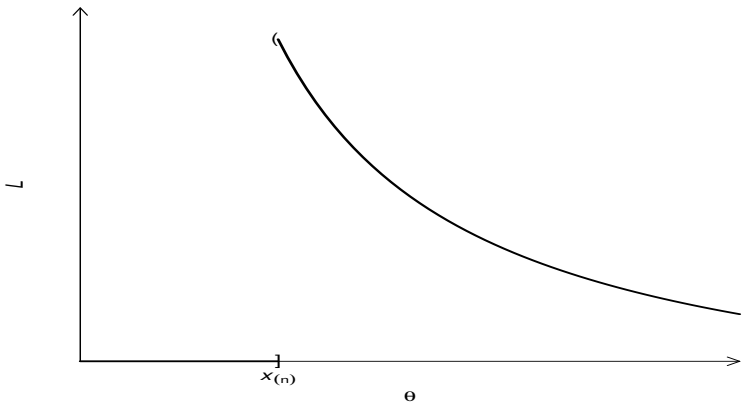


Figure 6.2. Discontinuous Likelihood with No Maximum

In this case the maximum of the likelihood does not exist, but the supremum of the likelihood occurs at $x_{(n)}$ and it is finite. We would like to call $x_{(n)}$ the MLE of θ.

We can reasonably do this by modifying the definition of the family of distributions by adding a zero-probability set to the support. We redefine the family in equation (6.15) to have the Lebesgue PDF

$$p_X(x) = \frac{1}{\theta} \mathrm{I}_{[0,\theta]}(x). \tag{6.19}$$

Now, the open interval $]x_{(n)}, \infty[$ where the likelihood was positive before becomes a half-closed interval $[x_{(n)}, \infty[$, and the maximum of the likelihood occurs at $x_{(n)}$.

This is one reason why we define the support to be closed.

This approach is cleaner than solving the logical problem by defining the MLE in terms of the sup rather than the max. A definition in terms of the sup may not address problems that could arise due to various types of discontinuity of $L(\theta\,;\,x)$ at the boundary of $S(\theta)$. ∎

MLE of More than One Parameter

It is usually more difficult to determine the MLE of more than one parameter. The likelihood equation in that case is a system of equations. Also, of course, the likelihood equation, whether a single equation or a system, may not be easy to solve, as the following example shows.

Example 6.6 MLE of the parameters in a gamma distribution
Consider the gamma family of distributions with parameters α and β. Given a random sample $x_1, \ldots, x_n$, the log-likelihood of α and β is

$$l_L(\alpha, \beta\,;\,x) = -n\alpha \log(\beta) - n \log(\Gamma(\alpha))) + (\alpha - 1) \sum \log(x_i) - \frac{1}{\beta} \sum x_i. \tag{6.20}$$

This yields the likelihood equations

$$-n \log(\beta) - n \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} + \sum \log(x_i) = 0 \tag{6.21}$$

and

$$-\frac{n\alpha}{\beta} + \frac{1}{\beta^2} \sum x_i = 0. \tag{6.22}$$

Checking the Hessian (at any point in the domain), we see that a root of the likelihood equations is a local minimizer.

At the solution we have

$$\widehat{\beta} = \sum x_i / (n\widehat{\alpha}) \tag{6.23}$$

and

$$\log(\widehat{\alpha}) - \frac{\Gamma'(\widehat{\alpha})}{\Gamma(\widehat{\alpha})} + \sum \log(x_i)/n - \log(\sum x_i/n) = 0. \tag{6.24}$$

There is no closed form solution. A numerical method must be used; see Example 6.14. ∎

Sometimes in multiple-parameter models, the parameters are functionally independent and the optimization can be performed on different parts of the separable likelihood function. This is the case for a normal distribution, as we see in Example 6.25.

Example 6.7 MLE in the exponential family with range dependency

Consider the two-parameter exponential family, that is, a shifted version of the exponential family of distributions. This family is the subject of Example 5.9 on page 397. The Lebesgue PDF is

$$\theta^{-1}\mathrm{e}^{-(x-\alpha)/\theta}\mathrm{I}_{]\alpha,\infty[}(x)$$

Suppose we have observations $X_1, X_2, \ldots, X_n$. The likelihood function is

$$L(\alpha, \theta; X) = \theta^{-n} \exp\left(-\sum (X_i - \alpha)/\theta\right) \mathrm{I}_{]0,X_{(1)}]}(\alpha)\mathrm{I}_{]0,\infty[}(\theta).$$

This is 0 when $\alpha > X_{(1)}$, but it is increasing in α on $]0, X_{(1)}]$ independently of θ.

Hence, the MLE of α is $X_{(1)}$.

Now, we substitute this back into $L(\alpha, \theta; X)$ and maximize wrt θ, that is, we solve

$$\max_{\theta}\left(\theta^{-n} \exp\left(-\sum (X_i - X_{(1)})/\theta\right)\right).$$

We do this by forming and solving the likelihood equation, noting that it yields a maximum within the parameter space. We get

$$\widehat{\theta} = \frac{1}{n}\sum (X_i - X_{(1)}).$$

In Example 5.9, we found the UMVUEs:

$$T_\alpha = X_{(1)} - \frac{1}{n(n-1)}\sum (X_i - X_{(1)})$$

and

$$T_\theta = \frac{1}{n-1}\sum (X_i - X_{(1)}).$$

(Recall that we find a complete sufficient statistic and then manipulate it to be unbiased.) Notice the similarity of these to the MLEs, which are biased. ∎

6.2.2 Finite Sample Properties of MLEs

Any approach to estimation may occasionally yield very poor or meaningless estimators. In addition to the possibly negative UMVUEs for variance components in Example 5.31, we have seen in Exercise 5.2 that a UMVUE of $g(\theta) = \mathrm{e}^{-3\theta}$ in a Poisson distribution is not a very good estimator. While in some cases the MLE is more reasonable (see Exercise 6.1), in other cases the MLE may be very poor.

As we have mentioned, MLEs have a nice intuitive property. In Section 6.3 we will see that they also often have good asymptotic properties.

We now consider some other properties; some useful and some less desirable.

Relation to Sufficient Statistics

Theorem 6.1
If there is a sufficient statistic and an MLE exists, then an MLE is a function of the sufficient statistic.

Proof.
This follows directly from the factorization theorem. ∎

Relation to Efficient Statistics

Given the three Fisher information regularity conditions (see page 168) we have defined "Fisher efficient estimators" as unbiased estimators that achieve the lower bound on their variance.

Theorem 6.2

Assume the FI regularity conditions for a family of distributions $\{P_\theta\}$ with the additional Le Cam-type requirement that the Fisher information matrix $I(\theta)$ is positive definite for all θ. Let $T(X)$ be a Fisher efficient estimator of θ. Then $T(X)$ is an MLE of θ.

Proof.
Let $p_\theta(x)$ be the PDF. We have

$$\frac{\partial}{\partial\theta}\log(p_\theta(x)) = I(\theta)(T(x) - \theta)$$

for any θ and x. Clearly, for $\theta = T(x)$, this equation is 0 (hence, $T(X)$ is an RLE). Because $I(\theta)$, which is the negative of the Hessian of the likelihood, is positive definite for all θ, the likelihood is convex in θ and $T(x)$ maximizes the likelihood. ∎

Notice that without the additional requirement of a positive definite information matrix, Theorem 6.2 would yield only the conclusion that $T(X)$ is an RLE.

Equivariance of MLEs

If $\widehat{\theta}$ is a good estimator of θ, it would seem to be reasonable that $g(\widehat{\theta})$ is a good estimator of $g(\theta)$, where g is a Borel function. "Good", of course, is relative to some criterion. In a decision-theoretic approach, we seek L-invariance; that is, invariance of the loss function (see page 266). Even if the loss function is invariant, other properties may not be preserved. If the criterion is UMVU, then the estimator in general will not have this equivariance property; that is, if $\widehat{\theta}$ is a UMVUE of θ, then $g(\widehat{\theta})$ may not be a UMVUE of $g(\theta)$. (It is not even unbiased in general.)

We now consider the problem of determining the MLE of $g(\theta)$ when we have an MLE $\widehat{\theta}$ of θ. Following the definition of an MLE, the MLE of $g(\theta)$ should be the maximizer of the likelihood function of $g(\theta)$. If the function g is not one-to-one, the likelihood function of $g(\theta)$ may not be well-defined. We therefore introduce the induced likelihood.

Definition 6.3 (induced likelihood)
Let $\{p_\theta : \theta \in \Theta\}$ with $\Theta \subseteq \mathbb{R}^d$ be a family of PDFs wrt a common σ-finite measure, and let $L(\theta)$ be the likelihood associated with this family, given observations. Now let g be a Borel function from Θ to $\Lambda \subseteq \mathbb{R}^{d_1}$ where $1 \leq d_1 \leq d$. Then

$$\widetilde{L}(\lambda) = \sup_{\{\theta\,:\,\theta\in\Theta \text{ and } g(\theta)=\lambda\}} L(\theta) \tag{6.25}$$

is called the *induced likelihood function* for the transformed parameter. ∎

The induced likelihood provides an appropriate MLE for $g(\theta)$ in the sense of the following theorem.

Theorem 6.3
Suppose $\{p_\theta : \theta \in \Theta\}$ with $\Theta \subseteq \mathbb{R}^d$ is a family of PDFs wrt a common σ-finite measure with associated likelihood $L(\theta)$. Let $\widehat{\theta}$ be an MLE of θ. Now let g be a Borel function from Θ to $\Lambda \subseteq \mathbb{R}^{d_1}$ where $1 \leq d_1 \leq d$ and let $\widetilde{L}(\lambda)$ be the resulting induced likelihood. Then $g(\widehat{\theta})$ maximizes $\widetilde{L}(\lambda)$.

Proof.
Follows directly from definitions, but it is an exercise to fill in the details. ∎

Usually when we consider reparametrizations, as in Section 2.6, with one-to-one functions. This provides a clean approach to the question of the MLE of $g(\theta)$ without having to introduce an induced likelihood.

Given the distribution P_θ for the random variable X, suppose we seek an MLE of $\tilde{g}(\theta)$. If $\tilde{g}$ is not one-to-one, then $\tilde{g}(\theta)$ does not provide enough information to define the distribution $P_{\tilde{g}(\theta)}$ for X. Therefore, we cannot define the likelihood for $\tilde{g}(\theta)$.

If $\tilde{g}(\theta)$ is one-to-one, let $g(\theta) = \tilde{g}(\theta)$, otherwise, define

$$g(\theta) = (\tilde{g}(\theta), h(\theta))$$

in such a way that $g(\theta)$ is one-to-one. The function h is not unique, but g^{-1} is unique; the likelihood is well-defined; $g(\widehat{\theta})$ is an MLE of $g(\theta)$; and so $\tilde{g}(\widehat{\theta})$ is an MLE of $\tilde{g}(\theta)$. Compare this with the results of Theorem 6.3 above.

Example 6.8 MLE of the variance in a Bernoulli distribution
Consider the Bernoulli family of distributions with parameter π. The variance of a Bernoulli distribution is $g(\pi) = \pi(1-\pi)$. Given a random sample $x_1, \ldots, x_n$, the MLE of π is

$$\widehat{\pi} = \sum_{i=1}^{n} x_i/n,$$

as we saw in Example 6.4, hence the MLE of the variance is

$$\frac{1}{n}\sum x_i \left(1 - \sum x_i/n\right). \tag{6.26}$$

Note that this estimator is biased and that it is the same estimator as that of the variance in a normal distribution from Example 3.13:

$$\sum (x_i - \bar{x})^2/n.$$

■

As we saw in Example 5.7, the UMVUE of the variance in a Bernoulli distribution is, as in equation (5.11),

$$\frac{1}{n-1}\sum x_i \left(1 - \sum x_i/n\right).$$

The difference in the MLE and the UMVUE of the variance in the Bernoulli distribution is the same as the difference in the estimators of the variance in the normal distribution that we encountered in Example 3.13 and Example 5.6. How do the MSEs of the estimators of the variance in a Bernoulli distribution compare? (Exercise 6.6.)

Whenever the variance of a distribution can be expressed as a function of other parameters $g(\theta)$, as in the case of the Bernoulli distribution, the estimator of the variance is $g(\widehat{\theta})$, where $\widehat{\theta}$ is an MLE of θ. The MLE of the variance of the gamma distribution, for example, is $\widehat{\alpha}\widehat{\beta}^2$, where $\widehat{\alpha}$ and $\widehat{\beta}$ are the MLEs in Example 6.6. The plug-in estimator of the variance of the gamma distribution, given the sample, $X_1, X_2 \ldots, X_n$, as always, is

$$\frac{1}{n}\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2.$$

Example 6.9 The likelihood principle in sampling from a Bernoulli distribution

In Example 6.1 we considered the problem of making inferences on the parameter π in a family of Bernoulli distributions by either taking a random sample of size n and using $T = \sum X_i$, which has a binomial distribution, or by taking a sample, $X_1, X_2, \ldots$, until a fixed number t of 1's have occurred and observing the size of the sample N, which has a negative binomial distribution. Given $T = t$ or $N = n$, either way, we get the MLE

$$\hat{\pi} = t/n.$$

To make inferences on π using $\hat{\pi}$ we need the variance $\mathrm{V}(\hat{\pi})$. Under the binomial distribution, we need the variance of T/n, which is $\pi(1-\pi)/n$, whose MLE as in Example 6.8 is $\hat{\pi}(1-\hat{\pi})/n$. Under the negative binomial distribution, we need the variance of t/N. The variance of N is $t(1-\pi)/\pi^2$ and its MLE is the same with $\hat{\pi}$ in place of π. The variance of t/N cannot be expressed in closed form. (See Stephan (1945).)

Although we have seen in equations (6.5) and (6.6) that the ratio of the likelihoods does not involve π and the MLEs based on the two data-generating processes conform to the likelihood principle, the variances of the MLEs are different. ∎

Other Properties of MLEs

Some properties of MLEs are not always desirable.

First of all, we note that an MLE may be biased. The most familiar example of this is the MLE of the variance, as seen in Examples 6.8 and 3.13. Another example is the MLE of the location parameter in the uniform distribution in Example 6.5.

Although the MLE approach is usually an intuitively logical one, it is not based on a formal decision theory, so it is not surprising that MLEs may not possess certain desirable properties that are formulated from that perspective.

An example of a likelihood function that is not very useful without some modification is in nonparametric probability density estimation. Suppose we assume that a sample comes from a distribution with continuous PDF $p(x)$. The likelihood is $\prod_{i=1}^{n} p(x_i)$. Even under the assumption of continuity, there is no solution. We will discuss this problem in Chapter 8.

C. R. Rao cites another example in which the likelihood function is not very meaningful.

Example 6.10 a meaningless MLE
Consider an urn containing N balls labeled $1, \ldots, N$ and also labeled with distinct real numbers $\theta_1, \ldots, \theta_N$ (with N known). For a sample without replacement of size $n < N$ where we observe $(x_i, y_i) = (\text{label}, \theta_{\text{label}})$, what is the likelihood function? It is either 0, if the label and θ_{label} for at least one observation is inconsistent, or $\binom{N}{n}^{-1}$, otherwise; and, of course, we don't know!

This likelihood function is not informative, and could not be used, for example, for estimating $\theta = \theta_1 + \cdots + \theta_N$. (There is a pretty good estimator of θ; it is $N(\sum y_i)/n$.) ∎

There are other interesting examples in which MLEs do not have desirable (or expected) properties.

- An MLE may be discontinuous in the data. This is obviously the case for a discrete distribution, but it can also occur in a contaminated continuous distribution as, for example, in the case of ϵ-mixture distribution family with CDF
$$P_{x_c,\epsilon}(x) = (1-\epsilon)P(x) + \epsilon \mathrm{I}_{[x_c,\infty[}(x), \tag{6.27}$$
where $0 \leq \epsilon \leq 1$.
- An MLE may not be a function of a sufficient statistic (if the MLE is not unique).
- An MLE may not satisfy the likelihood equation as, for example, when the likelihood function is not differentiable at its maximum, as in Example 6.5.
- The likelihood equation may have a unique root, yet no MLE exists. While there are examples in which the roots of the likelihood equations occur at minima of the likelihood, this situation does not arise in any realistic distribution (that I am aware of). Romano and Siegel (1986) construct a location family of distributions with support on
$$\mathbb{R} - \{x_1 + \theta, x_2 + \theta : x_1 < x_2\},$$
where x_1 and x_2 are known but θ is unknown, with a Lebesgue density $p(x)$ that rises as $x \nearrow x_1$ to a singularity at x_1 and rises as $x \swarrow x_2$ to a singularity at x_2 and that is continuous and strictly convex over $]x_1, x_2[$ and singular at both x_1 and x_2. With a single observation, the likelihood equation has a root at the minimum of the convex portion of the density between $x_1 + \theta$ and $x_2 + \theta$, but the likelihood increases without bound at both $x_1 + \theta$ and $x_2 + \theta$.
- An MLE may differ from an MME; in particular an MLE of the population mean may not be the sample mean.

Note that Theorem 6.1 hints at two other issues: nonuniqueness of an MLE and existence of an MLE. We now consider these.

Nonuniqueness

There are many cases in which the MLEs are not unique (and I'm not just referring to RLEs). The following examples illustrate this.

Example 6.11 likelihood in a Cauchy family
Consider the Cauchy distribution with location parameter θ. The likelihood equation is

$$\sum_{i=1}^{n} \frac{2(x_i - \theta)}{1 + (x_i - \theta)^2} = 0.$$

This may have multiple roots (depending on the sample), and so the one yielding the maximum would be the MLE. Depending on the sample, however, multiple roots can yield the same value of the likelihood function. ∎

Another example in which the MLE is not unique is $\mathrm{U}(\theta - 1/2,\ \theta + 1/2)$.

Example 6.12 likelihood in a uniform family with fixed range
Given the sample $x_1, \ldots, x_n$, the likelihood function for $\mathrm{U}(\theta - 1/2,\ \theta + 1/2)$ is

$$\mathrm{I}_{[x_{(n)} - 1/2,\ x_{(1)} + 1/2]}(\theta).$$

It is maximized at any value between $x_{(n)} - 1/2$ and $x_{(1)} + 1/2$. ∎

Nonexistence and Other Properties

We have already mentioned situations in which the likelihood approach does not seem to be the logical way, and have seen that sometimes in nonparametric problems, the MLE does not exist. This often happens when there are more "things to estimate" than there are observations. This can also happen in parametric problems. It may happen that the maximum does not exist because the likelihood is unbounded from above. In this case the argmax does not exist, and the maximum likelihood estimate does not exist.

Example 6.13 nonexistence of MLE
Consider the normal family of distributions with parameters μ and σ^2. Suppose we have one observation x. The log-likelihood is

$$l_L(\mu, \sigma^2\,;\, x) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2},$$

which is unbounded when $\mu = x$ and σ^2 approaches zero. It is therefore clear that no MLE of σ^2 exists. Strictly speaking, we could also say that no MLE of μ exists either; however, for any fixed value of σ^2 in the (open) parameter space, $\mu = x$ maximizes the likelihood, so it is reasonable to call x the MLE of μ.

Recall from Example 5.14 that the degree of the variance functional is 2. ∎

In this case, some people prefer to say that the *likelihood function does not exist*; that is, they suggest that the definition of a likelihood function include boundedness.

6.2.3 The Score Function and the Likelihood Equations

In several of the preceding examples, we found the MLEs by differentiating the likelihood and equating the derivative to zero. In many cases, of course, we cannot find an MLE by just differentiating the likelihood; Example 6.5 is such a case. We will discuss methods of finding an MLE in Section 6.2.4 beginning on page 465.

In the following we will generally consider only the log-likelihood, and we will assume that it is differentiable within Θ°.

The derivative of the log-likelihood is the score function $s_n(\theta\,;\,x)$ (equation (3.57) on page 244). The score function is important in computations for determining an MLE, as we see in Section 6.2.4, but it is also important in studying properties of roots of the likelihood equation, especially asymptotic properties, as we see in Section 6.3.

The score function is an estimating function and leads to the likelihood equation $\nabla l_L(\theta\,;\,x) = 0$ or

$$s_n(\theta\,;\,x) = 0, \tag{6.28}$$

which is an estimating equation, similar to the estimating equation (5.71) for least squares estimators. Generalizations of these equations are called "generalized estimating equations", or GEEs; see Section 3.2.5.

Any root of the likelihood equations, which is called an RLE, may be an MLE. A theorem from functional analysis, usually proved in the context of numerical optimization, states that if θ_* is an RLE and $\mathrm{H}_{l_L}(\theta_*\,;\,x)$ is negative definite, then there is a *local maximum* at θ_*. This may allow us to determine that an RLE is an MLE. There are, of course, other ways of determining whether an RLE is an MLE. In MLE, the determination that an RLE is actually an MLE is an important step in the process.

The Log-Likelihood Function and the Score Function in Regular Families

In the regular case satisfying the three Fisher information regularity conditions (see page 168), the likelihood function and consequently the log-likelihood are twice differentiable within Θ°, and the operations of differentiation and integration can be interchanged. In this case, the score estimating function is unbiased (see Definition 3.7):

$$\begin{aligned}
\mathrm{E}_\theta(s_n(\theta\,;\,X)) &= \int_{\mathcal{X}} \frac{\partial}{\partial\theta} l_L(\theta\,;\,x) p(x;\theta)\mathrm{d}x \\
&= \int_{\mathcal{X}} \frac{\partial}{\partial\theta} p(x;\theta)\mathrm{d}x \\
&= \frac{\partial}{\partial\theta} \int_{\mathcal{X}} p(x;\theta)\mathrm{d}x \\
&= 0.
\end{aligned} \tag{6.29}$$

The derivatives of the log-likelihood function relate directly to useful concepts in statistical inference. If it exists, the derivative of the log-likelihood is the relative rate of change, with respect to the parameter placeholder θ, of the probability density function at a fixed observation. If θ is a scalar, some positive function of the derivative, such as its square or its absolute value, is obviously a measure of the effect of change in the parameter, or of change in the estimate of the parameter. More generally, an outer product of the derivative with itself is a useful measure of the changes in the components of the parameter:

$$\nabla l_L\big(\theta^{(k)}\,;\,x\big)\,\Big(\nabla l_L\big(\theta^{(k)}\,;\,x\big)\Big)^{\mathrm{T}}.$$

Notice that the average of this quantity with respect to the probability density of the random variable X,

$$I(\theta_1\,;\,X) = \mathrm{E}_{\theta_1}\left(\nabla l_L\big(\theta^{(k)}\,;\,X\big)\,\Big(\nabla l_L\big(\theta^{(k)}\,;\,X\big)\Big)^{\mathrm{T}}\right), \tag{6.30}$$

is the *information matrix* for an observation on Y about the parameter θ.

If θ is a scalar, the square of the first derivative is the negative of the second derivative,

$$\left(\frac{\partial}{\partial\theta} l_L(\theta\,;\,x)\right)^2 \;=\; -\frac{\partial^2}{\partial\theta^2} l_L(\theta\,;\,x),$$

or, in general,

$$\nabla l_L\big(\theta^{(k)}\,;\,x\big)\,\Big(\nabla l_L\big(\theta^{(k)}\,;\,x\big)\Big)^{\mathrm{T}} \;=\; -\,\mathrm{H}_{l_L}\big(\theta^{(k)}\,;\,x\big). \tag{6.31}$$

MLEs in Exponential Families

If X has a distribution in the exponential class and we write its density in the natural or canonical form, the likelihood has the form

$$L(\eta\,;\,x) = \exp(\eta^{\mathrm{T}}T(x) - \zeta(\eta))h(x). \tag{6.32}$$

The log-likelihood equation is particularly simple:

$$T(x) - \frac{\partial\zeta(\eta)}{\partial\eta} = 0. \tag{6.33}$$

Newton's method for solving the likelihood equation is

$$\eta^{(k)} = \eta^{(k-1)} - \left(\frac{\partial^2\zeta(\eta)}{\partial\eta(\partial\eta)^{\mathrm{T}}}\Big|_{\eta=\eta^{(k-1)}}\right)^{-1}\left(T(x) - \frac{\partial\zeta(\eta)}{\partial\eta}\Big|_{\eta=\eta^{(k-1)}}\right)$$

Note that the second term includes the Fisher information matrix for η. (The expectation is constant.) (Note that the FI matrix is not for a *distribution*; it is for a *parametrization* of a distribution.)

We have

$$\mathrm{V}(T(X)) = \frac{\partial^2 \zeta(\eta)}{\partial\eta(\partial\eta)^{\mathrm{T}}} \,|_{\eta=\eta} \,.$$

Note that the variance is evaluated at the true η (even though in an expression such as $\partial\eta$ it must be a variable).

If we have a full-rank member of the exponential class then V is positive definite, and hence there is a unique maximum.

If we write

$$\mu(\eta) = \frac{\partial\zeta(\eta)}{\partial\eta},$$

in the full-rank case, μ^{-1} exists and so we have the solution to the likelihood equation:

$$\widehat{\eta} = \mu^{-1}(T(x)). \tag{6.34}$$

So maximum likelihood estimation is very nice for the exponential class.

6.2.4 Finding an MLE

Notice that the problem of obtaining an MLE is a *constrained optimization problem*; that is, an objective function is to be optimized subject to the constraints that the solution be within the closure of the parameter space.

In some cases the MLE occurs at a stationary point, which can be identified by differentiation. That is not always the case, however. A standard example in which the MLE does not occur at a stationary point is a distribution in which the range depends on the parameter, and the simplest such distribution is the uniform $\mathrm{U}(0,\theta)$, which was the subject of Example 6.5.

In this section, we will discuss some standard methods of maximizing a likelihood function and also some methods that are useful in more complicated situations.

Computations

If the log-likelihood is twice differentiable and if the range does not depend on the parameter, Equation (6.31) is interesting because the second derivative, or an approximation of it, is used in a Newton-like method to solve the maximization problem (6.10). Newton's equation

$$\mathrm{H}_{l_L}(\theta^{(k-1)}\,;\,x)\,d^{(k)} = \nabla l_L(\theta^{(k-1)}\,;\,x) \tag{6.35}$$

is used to determine the step direction in the k^{th} iteration. A quasi-Newton method uses a matrix $\widetilde{H}_{l_L}(\theta^{(k-1)})$ in place of the Hessian $\mathrm{H}_{l_L}(\theta^{(k-1)})$. (See notes on optimization in Appendix 0.4.)

In terms of the score function, and taking the step length to be 1, equation (6.35) gives the iteration

$$\theta^{(k)} = \theta^{(k-1)} - \left(\nabla s_n(\theta^{(k-1)}\,;\,x)\right)^{-1} s_n(\theta^{(k-1)}\,;\,x). \tag{6.36}$$

Fisher Scoring on the Log-Likelihood

In "Fisher scoring", the Hessian in Newton's method is replaced by its expected value. The iterates then are

$$\widehat{\theta}_{k+1} = \widehat{\theta}_k - H_l^{-1}(\widehat{\theta}_k|x)\nabla l(\widehat{\theta}_k|x).$$

Example 6.14 Computing the MLE of the parameters in a gamma distribution (continuation of Example 6.6)
The likelihood equations for the gamma(α, β) distribution in Example 6.6 led to the two equations

$$\widehat{\beta} = \bar{x}/\widehat{\alpha}$$

and

$$\log(\widehat{\alpha}) - \frac{\Gamma'(\widehat{\alpha})}{\Gamma(\widehat{\alpha})} + \sum \log(x_i)/n - \log(\bar{x}) = 0. \tag{6.37}$$

The two unknowns in these equations are separable, so we merely need to solve (6.37) in one unknown. The Hessian and gradient in equation (6.35) or (6.36) are scalars.

The function

$$\Psi(\alpha) = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} \tag{6.38}$$

is called the psi function or the digamma function (see Olver et al. (2010)). The R software package has a function for evaluation of the digamma function. If Newton's method (see Appendix 0.4) is to be used to solve equation (6.37), we also need $\Psi'(\alpha)$. This is called the trigamma function, and it is also available in R.

To see how we may compute this, let us generate some artificial data and solve the likelihood equations using the iterations in equation (6.36).

```
#      Generate artificial data
alpha <- 2
beta  <- 5
n <- 10
x <- rgamma(n,alpha,scale=beta)

#      Define functions to solve likelihood equation
sna <- function(meanlog,logmean,a0){
 log(a0)-digamma(a0)+meanlog-logmean
 }
```

```
snaprime <- function(meanlog,logmean,a0){
 1/a0 -trigamma(a0)
 }

#      Initialize data for algorithm
n <- 10
meanlog <- sum(log(x))/n
logmean <- log(mean(x))

#      Initialize starting value, set tolerance, loop
tol <- 10e-7
ak <- 3; akp1 <- ak+3*tol
iter <- 100
i <- 0
while (abs(akp1-ak)>tol&i<iter){
   i <- i+1
   ak <- max(tol,akp1)
   akp1 <- ak -
        sna(meanlog,logmean,ak)/snaprime(meanlog,logmean,ak)
}
bk <- mean(x)/ak
ak
bk
i
```

Depending on the sample, this code will converge in 5 to 10 iterations. ∎

One-Step MLE

Perhaps surprisingly, one iteration of equation (6.36) often gives a good approximation using any "reasonable" starting value. The result of one iteration is called a "one-step MLE".

Example 6.15 Computing the one-step MLEs of the parameters in a gamma distribution (continuation of Example 6.14)

```
#      Generate artificial data and initialize for algorithm
x <- rgamma(n,alpha,scale=beta)
n <- 10
meanlog <- sum(log(x))/n
logmean <- log(mean(x))

#      Initialize starting value, set tolerance, loop
tol <- 10e-7
ak <- 3; akp1 <- ak+3*tol
```

```
iter <- 100
i <- 0
while (abs(akp1-ak)>tol&i<iter){
   i <- i+1
   ak <- max(tol,akp1)
   akp1 <- ak -
        sna(meanlog,logmean,ak)/snaprime(meanlog,logmean,ak)
}
bk <- mean(x)/ak
ak
bk
ak <- 3; akp1 <- ak+3*tol; ak1 <- akp1
akp1 <- ak1 -
        sna(meanlog,logmean,ak1)/snaprime(meanlog,logmean,ak1)
ak1 <- akp1
bk1 <- mean(x)/ak1
ak1
bk1
```

Here are some results from several runs on artificial data. The one-step MLE is generally close to the MLE.

converged	3.017 4.001 4.297 1.687 2.703 2.499 4.955
one-step	3.017 3.746 3.892 0.584 2.668 2.393 4.161

Nondifferentiable Likelihood Functions

The definition of MLEs does not depend on the existence of a likelihood equation. The likelihood function may not be differentiable with respect to the parameter, as in the following example in which the parameter space is countable.

Example 6.16 MLE in the hypergeometric distribution
Consider a common problem in quality assurance. A batch of N items contains an unknown number M of defective items. We take a random sample of n items from the batch, and observing that the sample contains x defective items, we wish to estimate M. The likelihood is

$$L(M,\ N, n;\ x) = \binom{M}{x}\binom{N-M}{n-x}\Big/\binom{N}{n}, \tag{6.39}$$

over the appropriate ranges of N, M, n, and x, which are all nonnegative integers. This is a single-parameter distribution, because we assume N and n are known due to the setup of the problem. The parameter space is

$$\mathrm{M} = \{0, 1, \ldots, N\},$$

with $M = 0$ yielding a degenerate distribution.

The likelihood is not differentiable in M (it is not even continuous in M), so there is no likelihood equation. The function does have a maximum, however, and so an MLE exits.

Even if a function is not differentiable, we can seek a maximum by identifying a point of change from increasing to decreasing. We approximate a derivative:

$$\frac{L(M)}{L(M-1)} =$$

This is larger than 1, that is, the function is increasing, so long as $M < (N+1)x/n$ and greater than 1 otherwise. Hence, the MLE is

$$\widehat{M} = \lceil (N+1)x/n \rceil.$$

Note that this is biased. The UMVUE of M is Nx/n. ∎

EM Methods

Expectation Maximization methods are iterative methods for finding an MLE. Although EM methods do not rely on missing data, they can be explained most easily in terms of a random sample that consists of two components, one observed and one unobserved or missing. EM methods can also be used for other applications.

Missing Data

A simple example of missing data occurs in life-testing, when, for example, a number of electrical units are switched on and the time when each fails is recorded.

In such an experiment, it is usually necessary to curtail the recordings prior to the failure of all units.

The failure times of the units still working are unobserved, but the number of censored observations and the time of the censoring obviously provide information about the distribution of the failure times.

Mixtures

Another common example that motivates the EM algorithm is a finite mixture model.

Each observation comes from an unknown one of an assumed set of distributions. The missing data is the distribution indicator.

The parameters of the distributions are to be estimated. As a side benefit, the class membership indicator is estimated.

Applications of EM Methods

The missing data can be missing observations on the same random variable that yields the observed sample, as in the case of the censoring example; or the missing data can be from a different random variable that is related somehow to the random variable observed.

Many common applications of EM methods involve missing-data problems, but this is not necessary.

Often, an EM method can be constructed based on an artificial "missing" random variable to supplement the observable data.

Example 6.17 MLE in a multinomial model

One of the simplest examples of the EM method was given by Dempster et al. (1977).

Consider the multinomial distribution with four outcomes, that is, the multinomial with probability function,

$$p(x_1, x_2, x_3, x_4) = \frac{n!}{x_1!x_2!x_3!x_4!}\pi_1^{x_1}\pi_2^{x_2}\pi_3^{x_3}\pi_4^{x_4},$$

with $n = x_1+x_2+x_3+x_4$ and $1 = \pi_1+\pi_2+\pi_3+\pi_4$. Suppose the probabilities are related by a single parameter, θ, with $0 \leq \theta \leq 1$:

$$\begin{aligned}
\pi_1 &= \frac{1}{2} + \frac{1}{4}\theta \\
\pi_2 &= \frac{1}{4} - \frac{1}{4}\theta \\
\pi_3 &= \frac{1}{4} - \frac{1}{4}\theta \\
\pi_4 &= \frac{1}{4}\theta.
\end{aligned}$$

Given an observation (x_1, x_2, x_3, x_4), the log-likelihood function is

$$l(\theta) = x_1 \log(2+\theta) + (x_2 + x_3)\log(1-\theta) + x_4 \log(\theta) + c$$

and

$$\mathrm{d}l(\theta)/\mathrm{d}\theta = \frac{x_1}{2+\theta} - \frac{x_2+x_3}{1-\theta} + \frac{x_4}{\theta}.$$

The objective is to estimate θ.

Dempster, Laird, and Rubin used $n = 197$ and $x = (125, 18, 20, 34)$. (For this simple problem, the MLE of θ can be determined by solving a simple polynomial equation, but let's proceed with an EM formulation.)

To use the EM algorithm on this problem, we can think of a multinomial with five classes, which is formed from the original multinomial by splitting

the first class into two with associated probabilities $1/2$ and $\theta/4$. The original variable x_1 is now the sum of u_1 and u_2. Under this reformulation, we now have a maximum likelihood estimate of θ by considering $u_2 + x_4$ (or $x_2 + x_3$) to be a realization of a binomial with $n = u_2 + x_4 + x_2 + x_3$ and $\pi = \theta$ (or $1 - \theta$). However, we do not know u_2 (or u_1). Proceeding as if we had a five-outcome multinomial observation with two missing elements, we have the log-likelihood for the complete data,

$$l_c(\theta) = (u_2 + x_4)\log(\theta) + (x_2 + x_3)\log(1 - \theta),$$

and the maximum likelihood estimate for θ is

$$\frac{u_2 + x_4}{u_2 + x_2 + x_3 + x_4}.$$

The E-step of the iterative EM algorithm fills in the missing or unobservable value with its expected value given a current value of the parameter, $\theta^{(k)}$, and the observed data. Because $l_c(\theta)$ is linear in the data, we have

$$\mathrm{E}\left(l_c(\theta)\right) = \mathrm{E}(u_2 + x_4)\log(\theta) + \mathrm{E}(x_2 + x_3)\log(1 - \theta).$$

Under this setup, with $\theta = \theta^{(k)}$,

$$\begin{aligned}\mathrm{E}_{\theta^{(k)}}(u_2) &= \frac{1}{4}x_1\theta^{(k)} \Big/ \left(\frac{1}{2} + \frac{1}{4}x_1\theta^{(k)}\right) \\ &= u_2^{(k)}.\end{aligned}$$

We now maximize $\mathrm{E}_{\theta^{(k)}}\left(l_c(\theta)\right)$. This maximum occurs at

$$\theta^{(k+1)} = (u_2^{(k)} + x_4)/(u_2^{(k)} + x_2 + x_3 + x_4).$$

The following R statements execute a single iteration, after `tk` has been initialized to some value between 0 and 1.

```
u2kp1 <- x[1]*tk/(2+tk)
tk <- (u2kp1 + x[4])/(sum(x)-x[1]+u2kp1)
```

Within just a few iterations, `tk` settles to approximately 0.62682. ∎

Example 6.18 MLE in a variation of the life-testing experiment
Consider an experiment described by Flury and Zoppè (2000). It is assumed that the lifetime of light bulbs follows the exponential distribution with mean θ. To estimate θ, n light bulbs were tested until they all failed. Their failure times were recorded as $x_1, \ldots, x_n$. In a separate experiment, m bulbs were tested, but the individual failure times were not recorded. Only the number of bulbs, r, that had failed at time t was recorded. This is a slightly different setup as in Example 6.3.

The missing data are the failure times of the bulbs in the second experiment, $u_1, \ldots, u_m$. We have

$$l_c(\theta\,;\,x,u) = -n(\log\theta + \bar{x}/\theta) - \sum_{i=1}^{m}(\log\theta + u_i/\theta).$$

The expected value for a bulb still burning is

$$t + \theta$$

and the expected value of one that has burned out is

$$\theta - \frac{t\mathrm{e}^{-t/\theta^{(k)}}}{1 - \mathrm{e}^{-t/\theta^{(k)}}}.$$

Therefore, using a provisional value $\theta^{(k)}$, and the fact that r out of m bulbs have burned out, we have $\mathrm{E}_{U|x,\theta^{(k)}}(l_c)$ as

$$\begin{aligned} q^{(k)}(x,\theta) = &-(n+m)\log(\theta) \\ &-\frac{1}{\theta}\left(n\bar{x} + (m-r)(t+\theta^{(k)}) + r(\theta^{(k)} - th^{(k)})\right), \end{aligned}$$

where $h^{(k)}$ is given by

$$h^{(k)} = \frac{\mathrm{e}^{-t/\theta^{(k)}}}{1 - \mathrm{e}^{-t/\theta^{(k)}}}.$$

The k^{th} M step determines the maximum with respect to the variable θ, which, given $\theta^{(k)}$, occurs at

$$\theta^{(k+1)} = \frac{1}{n+m}\left(n\bar{x} + (m-r)(t+\theta^{(k)}) + r(\theta^{(k)} - th^{(k)})\right). \tag{6.40}$$

Starting with a positive number $\theta^{(0)}$, equation (6.40) is iterated until convergence. The expectation $q^{(k)}$ does not need to be updated explicitly.

To see how this works, let's generate some artificial data and try it out. Some R code to implement this is:

```
# Generate data from the exponential with theta=2,
# and with the second experiment truncated at t=3.
# Note that R uses a form of the exponential in
# which the parameter is a multiplier; i.e., the R
# parameter is 1/theta.
# Set the seed, so computations are reproducible.
set.seed(4)
n <- 100
m <- 500
theta <- 2
t <- 3
x <- rexp(n,1/theta)
r<-min(which(sort(rexp(m,1/theta))>=3))-1
```

Some R code to implement the EM algorithm:

```
# We begin with  theta=1.
# (Note theta.k is set to theta.kp1 at
#  the beginning of the loop.)
theta.k<-.01
theta.kp1<-1
# Do some preliminary computations.
n.xbar<-sum(x)
# Then loop and test for convergence
   theta.k <- theta.kp1
   theta.kp1 <- (n.xbar +
                 (m-r)*(t+theta.k) +
                 r*(theta.k-
                     t*exp(-t/theta.k)/(1-exp(-t/theta.k))
                   )
                )/(n+m)
```

The value of θ stabilizes to less than 0.1% change at 1.912 in 6 iterations. ∎

This example is interesting because if we assume that the distribution of the light bulbs is uniform, $\mathrm{U}(0,\theta)$ (such bulbs are called "heavybulbs"!), the EM algorithm cannot be applied.

Maximum likelihood methods must be used with some care whenever the range of the distribution depends on the parameter.

In this case, however, there is another problem. It is in computing $q^{(k)}(x,\theta)$, which does not exist for $\theta < \theta^{(k-1)}$.

Example 6.19 MLE in a normal mixtures model

A two-component normal mixture model can be defined by two normal distributions, $\mathrm{N}(\mu_1,\sigma_1^2)$ and $\mathrm{N}(\mu_2,\sigma_2^2)$, and the probability that the random variable (the observable) arises from the first distribution is w.

The parameter in this model is the vector $\theta = (w,\mu_1,\sigma_1^2,\mu_2,\sigma_2^2)$. (Note that w and the σs have the obvious constraints.)

The pdf of the mixture is

$$p(y;\theta) = wp_1(y;\mu_1,\sigma_1^2) + (1-w)p_2(y;\mu_2,\sigma_2^2),$$

where $p_j(y;\mu_j,\sigma_j^2)$ is the normal pdf with parameters μ_j and σ_j^2. (I am just writing them this way for convenience; p_1 and p_2 are actually the same parametrized function of course.)

In the standard formulation with $C = (X,U)$, X represents the observed data, and the unobserved U represents class membership.

Let $U = 1$ if the observation is from the first distribution and $U = 0$ if the observation is from the second distribution.

The unconditional $\mathrm{E}(U)$ is the probability that an observation comes from the first distribution, which of course is w.

Suppose we have n observations on X, $x_1, \ldots, x_n$.

Given a provisional value of θ, we can compute the conditional expected value $\mathrm{E}(U|x)$ for any realization of X. It is merely

$$\mathrm{E}(U|x, \theta^{(k)}) = \frac{w^{(k)} p_1(x; \mu_1^{(k)}, \sigma_1^{2(k)})}{p(x; w^{(k)}, \mu_1^{(k)}, \sigma_1^{2(k)}, \mu_2^{(k)}, \sigma_2^{2(k)})}$$

The M step is just the familiar MLE of the parameters:

$$w^{(k+1)} = \frac{1}{n} \sum \mathrm{E}(U|x_i, \theta^{(k)})$$

$$\mu_1^{(k+1)} = \frac{1}{n w^{(k+1)}} \sum q^{(k)}(x_i, \theta^{(k)}) x_i$$

$$\sigma_1^{2^{(k+1)}} = \frac{1}{n w^{(k+1)}} \sum q^{(k)}(x_i, \theta^{(k)})(x_i - \mu_1^{(k+1)})^2$$

$$\mu_2^{(k+1)} = \frac{1}{n(1 - w^{(k+1)})} \sum q^{(k)}(x_i, \theta^{(k)}) x_i$$

$$\sigma_2^{2^{(k+1)}} = \frac{1}{n(1 - w^{(k+1)})} \sum q^{(k)}(x_i, \theta^{(k)})(x_i - \mu_2^{(k+1)})^2$$

**** variations **** relate to AOV

∎

****** rewrite all this:

In maximum likelihood estimation, the objective function is the likelihood, $L_X(\theta; x)$ or the log-likelihood, $l_X(\theta; x)$. (Recall that a likelihood depends on a known distributional form for the data; that is why we use the notation $L_X(\theta; x)$ and $l_X(\theta; x)$, where "X" represents the random variable of the distribution.)

The variable for the optimization is θ; thus in an iterative algorithm, we find $\theta^{(1)}, \theta^{(2)}, \ldots$.

One type of alternating method is based on conditional optimization and a conditional bounding function alternates between updating $\theta^{(k)}$ using maximum likelihood and conditional expected values. This method is called the *EM method* because the alternating steps involve an expectation and a maximization.

Given $\theta^{(k-1)}$ we seek a function $q_k(x, \theta)$ that has a known relationship with $l_X(\theta;\ x)$, and then we determine $\theta^{(k)}$ to maximize $q_k(x, \theta)$ (subject to any constraints on acceptable values of θ).

The minorizing function $q_k(x, \theta)$ is formed as a conditional expectation of a joint likelihood. In addition to the data we have observed, call it X, we assume we have some unobserved data U.

Thus, we have "complete" data $C = (X, U)$ given the actual observed data X, and the other component, U, of C that is not observed.

Let $L_C(\theta; c)$ be the likelihood of the complete data, and let $L_X(\theta; x)$ be the likelihood of the observed data, with similar notation for the log-likelihoods. We refer to $L_C(\theta; c)$ as the "complete likelihood".

There are thus two likelihoods, one based on the complete (but unknown) sample, and one based only on the observed sample.

We wish to estimate the parameter θ, which figures in the distribution of both components of C.

The conditional likelihood of C given X is

$$L_{C|X}(\theta; c|x) = L_C(\theta;\ x, u)/L_X(\theta;\ x),$$

or

$$l_{C|X}(\theta; c|x) = l_C(\theta;\ x, u) - l_X(\theta;\ x).$$

Note that the conditional of C given X is the same as the conditional of U given X, and we may write it either way, either $C|X$ or $U|X$.

Because we do not have all the observations, $L_{C|X}(\theta; c|x)$ and $L_C(\theta;\ c)$ have

- unknown variables (the unobserved U)
- the usual unknown parameter.

Hence, we cannot follow the usual approach of maximizing the likelihood with given data.

We concentrate on the unobserved or missing data first.

We use a provisional value of $\theta^{(k-1)}$ to approximate the complete likelihood based on the expected value of U given $X = x$.

The expected value of the likelihood, which will generally be a function of both θ and $\theta^{(k-1)}$, minorizes the objective function of interest, $L_X(\theta;\ x)$, as we will see.

We then maximize this minorizing function with respect to θ to get $\theta^{(k)}$.

Let $L_C(\theta\ ;\ x, u)$ and $l_C(\theta\ ;\ x, u)$ denote, respectively, the likelihood and the log-likelihood for the complete sample. The objective function, that is, the likelihood for the observed X, is

$$L_X(\theta\ ;\ x) = \int L_C(\theta\ ;\ x, u)\, \mathrm{d}u,$$

and $l_X(\theta\ ;\ x) = \log L_X(\theta\ ;\ x)$.

After representing the function of interest, $L_X(\theta\ ;\ x)$, as an integral, the problem is to determine this function; that is, to average over U. (This is what the integral does, but we do not know what to integrate.) The average over U is the expected value with respect to the marginal distribution of U.

This is a standard problem in statistics: we estimate an expectation using observed data.

In this case, however, even the values that we average to estimate the expectation depends on θ, so we use a provisional value of θ.

We begin with a provisional value of θ, call it $\theta^{(0)}$.

Given any provisional value $\theta^{(k-1)}$, we will compute a provisional value $\theta^{(k)}$ that increases (or at least does not decrease) the conditional expected value of the complete likelihood.

The EM approach to maximizing $L_X(\theta\,;\,x)$ has two alternating steps. The steps are iterated until convergence.

E step : compute $q_k(x,\theta) = \mathrm{E}_{U|x,\theta^{(k-1)}}\big(l_C(\theta;\,x,U)\big)$.
M step : determine $\theta^{(k)}$ to maximize $q_k(x,\theta)$, or at least to increase it (subject to any constraints on acceptable values of θ).

Convergence of the EM Method

Is $l_X(\theta^{(k)};\,x) \geq l_X(\theta^{(k-1)};\,x)$?
(If it is, of course, then $L_X(\theta^{(k)};\,x) \geq L_X(\theta^{(k-1)};\,x)$, because the log is monotone increasing.)

The sequence $\theta^{(1)}, \theta^{(2)}, \ldots$ converges to a local maximum of the observed-data likelihood $L(\theta\,;\,x)$ under fairly general conditions. (It can be very slow to converge, however.)

Why EM Works

The real issue is whether the EM sequence

$$\{\theta^{(k)}\} \to \arg\max_\theta l_X(\theta;\,x)$$
$$(= \arg\max_\theta L_X(\theta;\,x)).$$

If $l_X(\cdot)$ is bounded (and it better be!), this is essentially equivalent to to asking if

$$l_X(\theta^{(k)};\,x) \geq l_X(\theta^{(k-1)};\,x).$$

(So long as in a sufficient number of steps the inequality is strict.)

Using an equation from before, we first write

$$l_X(\theta;\,X) = l_C(\theta;\,(X,U)) - l_{U|X}(\theta; U|X),$$

and then take the conditional expectation of functions of U given x and under the assumption that θ has the provisional value $\theta^{(k-1)}$:

$$l_X(\theta;\,X) = \mathrm{E}_{U|x,\theta^{(k-1)}}\big(l_C(\theta;\,(x,U))\big) - \mathrm{E}_{U|x,\theta^{(k-1)}}\big(l_{U|X}(\theta; U|x)\big)$$

$$= q_k(x,\theta) - h_k(x,\theta),$$

where

$$h_k(x,\theta) = \mathrm{E}_{U|x,\theta^{(k-1)}}\left(l_{U|X}(\theta;U|x)\right).$$

Now, consider

$$l_X(\theta^{(k)};\ X) - l_X(\theta^{(k-1)};\ X).$$

This has two parts:

$$q_k(x,\theta^{(k)}) - q_k(x,\theta^{(k-1)})$$

and

$$-\left(h_k(x,\theta^{(k)}) - h_k(x,\theta^{(k-1)})\right).$$

The first part is nonnegative from the M part of the k^{th} step.

What about the second part? We will show that it is nonnegative also (or without the minus sign it is nonpositive).

For the other part, for given $\theta^{(k-1)}$ and any θ, ignoring the minus sign, ...

$$\begin{aligned}
&h_k(x,\theta) - h_k(x,\theta^{(k-1)}) \\
&= \mathrm{E}_{U|x,\theta^{(k-1)}}\left(l_{U|X}(\theta;U|x)\right) - \mathrm{E}_{U|x,\theta^{(k-1)}}\left(l_{U|X}(\theta^{(k-1)};U|x)\right) \\
&= \mathrm{E}_{U|x,\theta^{(k-1)}}\left(\log\left(L_{U|x}(\theta;U|x)/L_{U|X}(\theta^{(k-1)};U|x)\right)\right) \\
&\le \log\left(\mathrm{E}_{U|x,\theta^{(k-1)}}\left(L_{U|X}(\theta;U|x)/L_{U|X}(\theta^{(k-1)};U|x)\right)\right) \\
&\qquad \text{(by Jensen's inequality)} \\
&= \log\int_{\mathcal{D}(U)} \frac{L_{U|X}(\theta;U|x)}{L_{U|X}(\theta^{(k-1)};U|x)}\, L_{U|X}(\theta^{(k-1)};U|x)\,\mathrm{d}u \\
&= \log\int_{\mathcal{D}(U)} L_{U|X}(\theta;U|x)\,\mathrm{d}u \\
&= 0.
\end{aligned}$$

So the second term is also nonnegative, and hence,

$$l_X(\theta^{(k)};\ x) \ge l_X(\theta^{(k-1)};\ x).$$

A Minorizing Function in EM Algorithms

With $l_X(\theta;\ x) = q_k(x,\theta) - h_k(x,\theta)$, and $h_k(x,\theta) \le h_k(x,\theta^{(k-1)})$ from the previous pages, we have

$$l_X(\theta^{(k-1)};\ x) - q_k(x,\theta^{(k-1)}) \le l_X(\theta;\ x) - q_k(x,\theta);$$

and so

$$q_k(x, \theta) + c(x, \theta^{(k-1)}) \leq l_X(\theta;\ x),$$

where $c(x, \theta^{(k-1)})$ is constant with respect to θ.

Therefore for given $\theta^{(k-1)}$ and any x,

$$g(\theta) = l_X(\theta^{(k-1)};\ X) - q_k(x, \theta^{(k-1)})$$

is a minorizing function for $l_X(\theta;\ x)$.

Alternative Ways of Performing the Computations

There are two kinds of computations that must be performed in each iteration:

- E step : compute $q_k(x, \theta) = \mathrm{E}_{U|x,\theta^{(k-1)}}\big(l_c(\theta;\ x, U)\big)$.
- M step : determine $\theta^{(k)}$ to maximize $q_k(x, \theta)$, subject to any constraints on acceptable values of θ.

There are obviously various ways to perform each of these computations.

A number of papers since 1977 have suggested specific methods for the computations.

For each specification of a method for doing the computations or each little modification, a new name is given, just as if it were a new idea:
GEM, ECM, ECME, AECM, GAECM, PXEM, MCEM, AEM, EM1, SEM

E Step

There are various ways the expectation step can be carried out.

In the happy case of a "nice" distribution, the expectation can be computed in closed form. Otherwise, computing the expectation is a numerical quadrature problem. There are various procedures for quadrature, including Monte Carlo.

Some people have called an EM method that uses Monte Carlo to evaluate the expectation an MCEM method. (If a Newton-Cotes method is used, however, we do not call it an NCEM method!) The additional Monte Carlo computations add a lot to the overall time required for convergence of the EM method.

An additional problem in using Monte Carlo in the expectation step may be that the distribution of C is difficult to simulate. The convergence criterion for optimization methods that involve Monte Carlo generally should be tighter than for deterministic methods.

M Step

For the maximization step, there are even more choices.

The first thing to note, as we mentioned earlier for alternating algorithms generally, is that rather than maximizing q_k, we can just require that the overall sequence increase.

Dempster et al. (1977) suggested requiring only an increase in the expected value; that is, take $\theta^{(k)}$ so that

$$q_k(u, \theta^{(k)}) \geq q_{k-1}(u, \theta^{(k-1)}).$$

They called this a generalized EM algorithm, or GEM. (Even in the paper that introduced the "EM" acronym, another acronym was suggested for a variation.) If a one-step Newton method is used to do this, some people have called this a EM1 method.

Meng and Rubin (1993) describe a GEM algorithm in which the M-step is an alternating conditional maximization; that is, if $\theta = (\theta_1, \theta_2)$, first $\theta_1^{(k)}$ is determined to maximize q subject to the constraint $\theta_2 = \theta_2^{(k-1)}$; then $\theta_2^{(k)}$ is determined to maximize q_k subject to the constraint $\theta_1 = \theta_1^{(k)}$. This sometimes simplifies the maximization problem so that it can be done in closed form. They call this an expectation conditional maximization method, ECM.

Alternate Ways of Terminating the Computations

In any iterative algorithm, we must have some way of deciding to terminate the computations. (The generally-accepted definition of "algorithm" requires that it terminate. In any event, of course, we want the computations to cease at some point.)

One way of deciding to terminate the computations is based on convergence; if the computations have converged we quit. In addition, we also have some criterion by which we decide to quit anyway.

In an iterative optimization algorithm, there are two obvious ways of deciding when convergence has occurred. One is when the decision variables (the estimates in MLE) are no longer changing appreciably, and the other is when the value of the objective function (the likelihood) is no longer changing appreciably.

Convergence

It is easy to think of cases in which the objective function converges, but the decision variables do not. All that is required is that the objective function is flat over a region at its maximum. In statistical terms, the corresponds to unidentifiability.

The Variance of Estimators Defined by the EM Method

As is usual for estimators defined as solutions to optimization problems, we may have some difficulty in determining the statistical properties of the estimators.

Louis (1982) suggested a method of estimating the variance-covariance matrix of the estimator by use of the gradient and Hessian of the complete-data log-likelihood, $l_{L_c}(\theta\ ;\ u, v)$. Kim and Taylor (1995) also described ways of estimating the variance-covariance matrix using computations that are part of the EM steps.

It is interesting to note that under certain assumptions on the distribution, the iteratively reweighted least squares method can be formulated as an EM method (see Dempster et al. (1980)).

Missing Data

Although EM methods do not rely on missing data, they can be explained most easily in terms of a random sample that consists of two components, one observed and one unobserved or missing.

A simple example of missing data occurs in life-testing, when, for example, a number of electrical units are switched on and the time when each fails is recorded.

In such an experiment, it is usually necessary to curtail the recordings prior to the failure of all units.

The failure times of the units still working are unobserved, but the number of censored observations and the time of the censoring obviously provide information about the distribution of the failure times.

Mixtures

Another common example that motivates the EM algorithm is a finite mixture model.

Each observation comes from an unknown one of an assumed set of distributions. The missing data is the distribution indicator.

The parameters of the distributions are to be estimated. As a side benefit, the class membership indicator is estimated.

Applications of EM Methods

The missing data can be missing observations on the same random variable that yields the observed sample, as in the case of the censoring example; or the missing data can be from a different random variable that is related somehow to the random variable observed.

Many common applications of EM methods involve missing-data problems, but this is not necessary.

Often, an EM method can be constructed based on an artificial "missing" random variable to supplement the observable data.

6.3 Asymptotic Properties of MLEs, RLEs, and GEE Estimators

The argmax of the likelihood function, that is, the MLE of the argument of the likelihood function, is obviously an important statistic.

In many cases, a likelihood equation exists, and often in those cases, the MLE is a root of the likelihood equation. In some cases there are roots of the likelihood equation (RLEs) that may or may not be an MLE.

6.3.1 Asymptotic Distributions of MLEs and RLEs

We recall that asymptotic expectations are defined as expectations in asymptotic distributions (rather than as limits of expectations). The first step in studying asymptotic properties is to determine the asymptotic distribution.

Example 6.20 asymptotic distribution of the MLE of the variance in a Bernoulli family
In Example 6.8 we determined the MLE of the variance $g(\pi) = \pi(1-\pi)$ in the Bernoulli family of distributions with parameter π. The MLE of $g(\pi)$ is $T_n = \overline{X}(1-\overline{X})$.

From Example 1.25 on page 94, we get its asymptotic distributions as

$$\sqrt{n}(g(\pi) - T_n) \to \mathrm{N}(0, \pi(1-\pi)(1-2\pi)^2),$$

if $\pi \neq 1/2$, and if $\pi = 1/2$,

$$4n(g(\pi) - T_n) \xrightarrow{\mathrm{d}} \chi^2_1.$$

∎

6.3.2 Asymptotic Efficiency of MLEs and RLEs

One of the most important properties of roots of the likelihood equation, given the Le Cam regularity conditions (see page 169), is asymptotic efficiency. The regularity conditions are the same as those for Le Cam's theorem on the countability of superefficient estimators (see page 421). ********** fix

For distributions that satisfy the Le Cam regularity conditions (these conditions are essentially the FI regularity conditions plus a condition on the FI matrix), there is a nice sequence of the likelihood equation (6.12) that is formed from the sequence of score functions,

$$s_n(\theta) = \nabla l_{L_n}(\theta\,;\,x). \tag{6.41}$$

Theorem 6.4
Assume the Le Cam regularity conditions for a family of distributions $\{P_\theta\}$*, and let* $s_n(\theta)$ *be the score function for a sample of size* n*. There is a sequence of estimators* $\widehat{\theta}_n$ *such that*

$$\Pr(s_n(\widehat{\theta}_n) = 0) \to 1,$$

and

$$\widehat{\theta}_n \xrightarrow{\text{p}} \theta.$$

Proof.
∎

For a sequence in a Le Cam regular family that satisfies the conclusion of Theorem 6.4, there is an even more remarkable fact. We have seen in Theorem 6.2 that, with the FI regularity conditions, *if* there is an efficient estimator, then that estimator is an MLE.

Theorem 6.5
Assume the Le Cam regularity conditions for a family of distributions $\{P_\theta\}$*, and let* $s_n(\theta)$ *be the score function for a sample of size* n*. Any consistent sequence of RLEs, that is, any consistent sequence* $\widehat{\theta}_n$ *that satisfies*

$$s_n(\widehat{\theta}_n) = 0,$$

is asymptotically efficient.

Proof.
∎

Notice the differences in Theorems 6.2 and 6.5. Theorem 6.2 for finite sample efficiency requires only the FI regularity conditions for RLEs (or with the additional requirement of a positive definite information matrix for an MLE), but is predicated on the existence of an efficient estimator. As is often the case in asymptotic efficiency, Theorem 6.5 requires the Le Cam regularity conditions but it gives a stronger result: consistency yields asymptotic efficiency.

It is important to be clear on what these theorems say. They apply to RLEs, which may be MLEs of the parameter, which as a variable is the variable of differentiation, say θ, in the score function. If $\widehat{\theta}$ is the MLE of θ, then by definition, $g(\widehat{\theta})$ is the MLE of $g(\theta)$. If $\widehat{\theta}$ is asymptotically efficient for estimating θ, that does not mean that $g(\widehat{\theta})$ is asymptotically efficient for estimating $g(\theta)$.

Example 6.21 an MLE that is not asymptotically efficient
In Example 6.8 we determined that the MLE of the variance $g(\pi) = \pi(1-\pi)$ in the Bernoulli family of distributions with parameter π is $T_n = \overline{X}(1-\overline{X})$, and in Example 6.20 we determined its asymptotic distribution.

When $\pi \neq 1/2$, we have

$$\sqrt{n}(g(\pi) - T_n) \to \mathrm{N}(0, \pi(1-\pi)(1-2\pi)^2),$$

and when $\pi = 1/2$, we have

$$4n(g(\pi) - T_n) \xrightarrow{\mathrm{d}} \chi^2_1.$$

Hence, T_n is asymptotically biased. ∎

Asymptotic Relative Efficiency

Remember that the ARE is the ratio of two asymptotic expectations — not the asymptotic expectation of a ratio, and certainly not the limit of a ratio; although of course sometimes these three things are the same.

Example 6.22 ARE(MLE, UNVUE) in the exponential family with range dependency

In Examples 5.9 and 6.7 we considered the two parameter exponential distribution with Lebesgue PDF

$$\theta^{-1}\mathrm{e}^{-(x-\alpha)/\theta}\mathrm{I}_{]\alpha,\infty[}(x).$$

In Example 5.9, we found the UMVUEs:

$$T_{\alpha_n} = X_{(1)} - \frac{1}{n(n-1)}\sum(X_i - X_{(1)})$$

and

$$T_{\theta_n} = \frac{1}{n-1}\sum(X_i - X_{(1)}).$$

In Example 6.7, we found the MLEs:

$$\widehat{\alpha}_n = X_{(1)}$$

and

$$\widehat{\theta}_n = \frac{1}{n}\sum(X_i - X_{(1)}).$$

The distributions for $\widehat{\theta}$ and T_θ are relatively easy. We worked out the distribution of T_θ in Example 1.18, and $\widehat{\theta}$ is just a scalar multiple of T_θ. Because of the relationship between $\widehat{\theta}$ and T_θ, however, we do not even need the asymptotic distributions.

In Example 1.11 we found that the distribution of $\widehat{\alpha}$ is a two-parameter exponential distribution with parameters α and θ/n; hence,

$$n(X_{(1)} - \alpha) \xrightarrow{\mathrm{d}} \text{ exponential}(0, \theta).$$

Now let us consider the ARE of the MLE to the UMVUE for these two parameters.

- ARE(MLE,UMVUE) for θ.
  This is an easy case, because the estimators always differ by the ratio $n/(n-1)$. We do not even need the asymptotic distributions. The ARE is 1.
- ARE(MLE,UMVUE) for α.
  We have found the asymptotic distributions of $U = \widehat{\alpha} - \alpha$ and $V = T_\alpha - \alpha$, so we just work out the asymptotic expectations of U^2 and V^2. We get $\mathrm{E}(V^2) = \theta$ and $\mathrm{E}(U^2) = \theta + \theta^2$. Therefore, the ARE is $\theta/(\theta + \theta^2)$.

∎

6.3.3 Inconsistent MLEs

In previous sections, we have seen that sometimes MLEs do not have some statistical properties that we usually expect of good estimators.

The discussion in this section has focused on MLEs (or RLEs) that are consistent. It is not necessarily the case that MLEs are consistent, however. The following example is from Romano and Siegel (1986).

Example 6.23 rational, irrational estimand
Let $X_1, \ldots, X_n$ be a sample from $\mathrm{N}(\theta, 1)$. Define the estimand $g(\theta)$ as

$$g(\theta) = \begin{cases} -\theta \text{ if } \theta \text{ is irrational} \\ \theta \text{ if } \theta \text{ is rational.} \end{cases}$$

Because $\overline{X}_n$ is the MLE of θ, $g(\overline{X}_n)$ is the MLE of $g(\theta)$. Now $\overline{X}_n \sim \mathrm{N}(\theta, 1/n)$ and so is almost surely irrational; hence, $g(\overline{X}_n) = -\overline{X}_n$ a.s. Now, by the SLLN, we have $g(\overline{X}_n) = -\theta$ a.s. Hence, if θ is a rational number $\neq 0$, then

$$g(\overline{X}_n) \overset{\text{a.s.}}{\to} -\theta \neq \theta = g(\theta).$$

∎

While that example may seem somewhat contrived, consider an example due to Ferguson.

Example 6.24 mixtures
Let $X_1, \ldots, X_n$ be a sample from from the distribution with PDF wrt Lebesgue measure

$$p_X(x;\theta) = (1-\theta)p_{\mathrm{T}}(x;\theta,\delta(\theta)) + \theta p_{\mathrm{U}}(x),$$

where $\theta \in [0,1]$, $\delta(\theta)$ is a continuous decreasing function of θ with $\delta(0) = 1$ and $0 < \delta(\theta) \leq 1 - \theta$ for $0 < \theta < 1$, and

$$p_{\mathrm{T}}(x;\theta,\delta(\theta)) = \frac{1}{\delta(\theta)}\left(1 - \frac{|x-\theta|}{\delta(\theta)}\right) \mathrm{I}_{[\theta-\delta(\theta),\theta+\delta(\theta)]}(x)$$

and

$$p_{\mathrm{U}}(x) = \frac{1}{2}\mathrm{I}_{[-1,1]}(x).$$

The distribution is a mixture of a triangular distribution centered on θ and the $\mathrm{U}(-1,1)$ distribution.

Note that the densities are continuous in θ for any x and is defined on $[0,1]$ and therefore an MLE exists.

Let $\widehat{\theta}_n = \widehat{\theta}(X_1,\ldots,X_n)$ denote any MLE of θ. Now, if $\theta < 1$, then

$$p_X(x;\theta) \le (1-\theta)/\delta(\theta) + \theta/2 < 1/\delta(\theta) + \frac{1}{2},$$

and so for any $\alpha < 1$

$$\max_{0\le\theta\le\alpha} \frac{l_n(\theta)}{n} \le \log\left(\frac{1}{\delta(\theta)} + \frac{1}{2}\right) < \infty.$$

Now, if we could choose $\delta(\theta)$ so that

$$\max_{0\le\theta\le 1} \frac{l_n(\theta)}{n} \xrightarrow{\text{a.s.}} \infty,$$

then $\widehat{\theta}_n$ will eventually be greater than α for any $\alpha < 1$, and so the MLE is not consistent.

So, can we choose such a $\delta(\theta)$?

Let

$$M_n = \max(X_1,\ldots,X_n),$$

hence $M_n \xrightarrow{\text{a.s.}} \infty$, and

$$\begin{aligned}\max_{0\le\theta\le 1} \frac{l_n(\theta)}{n} &\ge \frac{l_n(M_n)}{n} \\ &\ge \frac{n-1}{n}\log\left(\frac{M_n}{2}\right) + \frac{1}{n}\log\left(\frac{1-M_n}{\delta(M_n)}\right),\end{aligned}$$

and so

$$\liminf_n \max_{0\le\theta\le 1} \frac{l_n(\theta)}{n} \ge \log\left(\frac{1}{2}\right) + \liminf_n \log\left(\frac{1-M_n}{\delta(M_n)}\right) \text{ a.s.}$$

So we need to choose $\delta(\theta)$ so that the last limit is infinite a.s. Now, $\forall\theta M_n \xrightarrow{\text{a.s.}} \infty$, and the slowest rate is for $\theta = 1$, because that distribution has the smallest mass in a sufficiently small neighborhood of 1. Therefore, all we need to do is choose $\delta(\theta) \to 0$ as $\theta \to 1$ fast enough so that the limit is infinite a.s. when $\theta = 0$.

So now for $0 < \epsilon < 1$,

$$\begin{aligned}\sum_n \Pr_{\theta=0}(n^{1/4}(1-M_n) > \epsilon) &= \sum_n \Pr_{\theta=0}(M_n < 1 - \epsilon n^{-1/4}) \\ &= \sum_n \left(1 - \epsilon^2 \frac{n^{-1/4}}{2}\right)^n \\ &\leq \sum_n \exp\left(-\epsilon^2 \frac{n^{-1/4}}{2}\right) \\ &< \infty.\end{aligned}$$

Hence, by the Borel-Cantelli lemma, $n^{1/4}(1-M_n) \to 0$ a.s. Finally, choosing

$$\delta(\theta) = (1-\theta)\exp\left(-(1-\theta)^{-4}+1\right),$$

we have a function that satisfies the requirements above (it is continuous decreasing with $\delta(0) = 1$ and $0 < \delta(\theta) \leq 1-\theta$ for $0 < \theta < 1$) and it is such that

$$\begin{aligned}\frac{1}{n}\log\left(\frac{1-M_n}{\delta(M_n)}\right) &= \frac{1}{n(1-M_n)^4} - \frac{1}{n} \\ &\xrightarrow{\text{a.s.}} \infty.\end{aligned}$$

This says that *any* MLE of θ must tend to 1 a.s., and so cannot be consistent. ∎

In addition to these examples, we recall the Neyman-Scott problem in Example 6.27, where the ordinary MLE of the variance is not consistent, but we were able to reformulate the problem so as to obtain an MLE of the variance that is consistent.

6.3.4 Properties of GEE Estimators

Consistency of GEE Estimators

The roots of a generalized estimating equation

$$s_n(\gamma) = 0$$

often have good asymptotic properties.

If the GEE is chosen so that

$$\mathrm{E}_\theta(s_n(\theta)) = 0,$$

or else so that the asymptotic expectation of $\{x_n\}$ is zero **********

The class of estimators arising from the generalized estimating equations (3.77) and (3.79), under very general assumptions have an asymptotic normal distribution. This is Theorem 5.13 in MS2.

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{\mathrm{d}} \mathrm{N}(0, \sigma_F^2),$$

where $\{\widehat{\theta}_n\}$ is a sequence of GEE estimators and

$$\sigma_F^2 = \int (\psi(x,\theta))^2 \mathrm{d}F(x)/(\psi'(x,\theta))^2.$$

6.4 Application: MLEs in Generalized Linear Models

6.4.1 MLEs in Linear Models

In the case of underlying normal probability distribution, estimation of the mean based on least squares is the same as MLE. Consider a linear model (5.66) as discussed in Section 5.5.1.

Example 6.25 MLE in a linear model
Let

$$Y = X\beta + E, \tag{6.42}$$

where Y and E are n-vectors with $\mathrm{E}(E) = 0$ and $\mathrm{V}(E) = \sigma^2 I_n$, X is an $n \times p$ matrix whose rows are the x_i^{T}, and β is the p-vector parameter. In Section 5.5.1 we studied a least squares estimator of β; that is,

$$b^* = \arg\min_{b \in B} \|Y - Xb\|^2 \tag{6.43}$$

$$= (X^{\mathrm{T}}X)^- X^{\mathrm{T}}Y. \tag{6.44}$$

Even if X is not of full rank, in which case the least squares estimator is not unique, we found that the least squares estimator has certain optimal properties for estimable functions.

Of course at this point, we could not use MLE — we do not have a distribution. We could define a least squares estimator without an assumption on the distribution of Y or E, but for an MLE we need an assumption on the distribution.

After we considered the least-squares estimator without a specific distribution, next in Section 5.5.1, we considered the additional assumption in the model that

$$E \sim \mathrm{N}_n(0, \sigma^2 I_n),$$

or

$$Y \sim \mathrm{N}_n(X\beta, \sigma^2 I_n).$$

In that case, we found that the least squares estimator yielded the unique UMVUE for any estimable function of β and for σ^2. Again, let us assume

$$Y \sim \mathrm{N}_n(X\beta, \sigma^2 I_n),$$

yielding, for an observed y, the log-likelihood

$$l_L(\beta, \sigma^2\ y, X) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y - X\beta)^{\mathrm{T}}(y - X\beta). \tag{6.45}$$

Maximizing this function with respect to β is the same as minimizing the expression in equation (6.43), and so an MLE of β is the same as a least squares estimator of β.

Estimation of σ^2, however, is different. In the case of least squares estimation on page 426, with no specific assumptions about the distribution of E in the model (6.42), we had no basis for forming an objective function of squares to minimize. With an assumption of normality, however, instead of explicitly forming a least squares problem for estimating σ^2, using a least squares estimator of β, b^*, we merely used the distribution of $(y - Xb^*)^{\mathrm{T}}(y - Xb^*)$ to form a UMVUE of σ^2,

$$s^2 = (Y - Xb^*)^{\mathrm{T}}(Y - Xb^*)/(n - r), \tag{6.46}$$

where $r = \mathrm{rank}(X)$.

In the case of maximum likelihood, we directly determine the value of σ^2 that maximizes the expression in equation (6.45). This is an easy optimization problem. The solution is

$$\widehat{\sigma}^2 = (y - X\widehat{\beta})^{\mathrm{T}}(y - X\widehat{\beta})/n \tag{6.47}$$

where $\widehat{\beta} = b^*$ is an MLE of β. Compare the MLE of σ^2 with the least squares estimator, and note that the MLE is biased. Recall that we have encountered these two estimators in the simpler cases of Example 3.13 (MLE) and 5.6 (UMVUE). See also equation (3.55). ∎

In Examples 5.28, 5.29 and 5.30 (starting on page 434), we considered UMVUE in a special case of the linear model called the fixed-effects one-way AOV model. We now consider MLE in this model.

Example 6.26 MLE in the one-way fixed-effects AOV model
We consider the model

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad i = 1, \ldots, m; \quad j = 1, \ldots, n. \tag{6.48}$$

For least squares or UMVUE we do not need to assume any particular distribution; all we need assume is that $\mathrm{E}(\epsilon_{ij}) = 0$ and $\mathrm{V}(\epsilon_{ij}) = \sigma^2$ for all i, j, and $\mathrm{Cov}(\epsilon_{ij}, \epsilon_{i'j'}) = 0$ if $i \neq i'$ or $j \neq j'$. For MLE, however, we need to assume a distribution, and so we will assume each ϵ_{ij} has a normal distribution with the additional assumptions about expected values.

Proceeding to write the likelihood under the normal assumption, we see that an MLE is $\widehat{\beta} = (X^{\mathrm{T}}X)^{-}X^{\mathrm{T}}Y$ for any generalized inverse of $(X^{\mathrm{T}}X)$, which is the same as the least squares estimator obtained in equation (5.73). ∎

Example 6.27 ML estimation of the variance in the one-way fixed-effects AOV model
In Example 5.30, we assumed a normal distribution for the residuals, and obtained the distribution of the sum of squares

$$\text{SSE} = \sum_{i=1}^{m}\sum_{j=1}^{n}(Y_{ij} - \overline{Y}_i)^2,$$

and from that we obtained the UMVUE of σ^2 as $\text{SSE}/m(n-1)$.

From maximization of the likelihood, we obtain the MLE of σ^2 as

$$\widehat{\sigma^2} = \frac{1}{nm}\sum_{i=1}^{m}\sum_{j=1}^{n}(Y_{ij} - \overline{Y}_i)^2 \tag{6.49}$$

(exercise).

While the MLE of σ^2 is consistent in mean squared error as $n \to \infty$ and m remains fixed, it is not consistent as $m \to \infty$ and n remains fixed (Exercise 6.3). ∎

There are interesting ways of getting around the lack of consistency of the variance estimator in Example 6.27. In the next example, we will illustrate an approach that is a simple use of a more general method called REML, for "residual maximum likelihood" (also called "restricted maximum likelihood").

Example 6.28 REML estimation of the variance in the one-way fixed-effects AOV model
In the preceding examples suppose there are only two observations per group; that is, the model is

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad i = 1, \ldots, m; \quad j = 1, 2,$$

with all of the other assumptions made above.

The MLE of σ^2 in equation (6.49) can be written as

$$\widehat{\sigma^2} = \frac{1}{4m}\sum_{i=1}^{m}\sum_{j=1}^{2}(Y_{i1} - Y_{i2})^2. \tag{6.50}$$

We see that the limiting expectation of $\widehat{\sigma^2}$ as $m \to \infty$ is $\sigma/2$; that is, the estimator is not consistent. (This particular setup is called the "Neyman-Scott problem". In a fixed sample, of course, the estimator is biased, and there is no reason to expect any change unless n instead of m were to increase.)

We see that the problem is caused by the unknown means, and as m increases the number of unknown parameters increases linearly in m. We can, however, reformulate the problem so as to focus on σ^2. For $i = 1, \ldots, m$,

let $Z_i = Y_{i1} - Y_{i2}$. Now, using Z_i, the likelihood is based on the $\mathrm{N}(0, 2\sigma^2)$ distribution. Under the likelihood for this setup, which we call REML, we get the maximum likelihood estimate

$$\widehat{\sigma^2}_{\mathrm{REML}} = \frac{1}{2m}\sum_{i=1}^{m}\sum_{j=1}^{2}(Z_i)^2. \tag{6.51}$$

which is consistent in m for σ^2. ∎

In the next example, we will consider a random-effects model.

Example 6.29 MLE in the one-way random-effects AOV model
Consider the linear model in Example 5.31 on page 436,

$$Y_{ij} = \mu + \delta_i + \epsilon_{ij}, \quad i = 1, \ldots, m; \quad j = 1, \ldots, n, \tag{6.52}$$

where the δ_i are identically distributed with $\mathrm{E}(\delta_i) = 0$, $\mathrm{V}(\delta_i) = \sigma_\delta^2$, and $\mathrm{Cov}(\delta_i, \delta_{\tilde{i}}) = 0$ for $i \neq \tilde{i}$, and the ϵ_{ij} are independent of the δ_i and are identically distributed with with $\mathrm{E}(\epsilon_{ij}) = 0$, $\mathrm{V}(\epsilon_{ij}) = \sigma_\epsilon^2$, and $\mathrm{Cov}(\epsilon_{ij}, \epsilon_{\tilde{i}\tilde{j}}) = 0$ for either $i \neq \tilde{i}$ or $j \neq \tilde{j}$.

In order to use a likelihood approach, of course, we need to make assumptions about the distributions of the random variables. Let us suppose now that $\delta_i \overset{\mathrm{iid}}{\sim} \mathrm{N}(0, \sigma_\delta^2)$, where $\sigma_\delta^2 \geq 0$, and $\epsilon_{ij} \overset{\mathrm{iid}}{\sim} \mathrm{N}(0, \sigma_\epsilon^2)$, where as usual $\sigma^2 > 0$.

Our interest in using the model is to make inference on the relative sizes of the components of the variance σ_δ^2 and σ_ϵ^2.

In Example 5.31, we obtained the UMVUEs of σ_δ^2 and σ_ϵ^2, and noted that the unbiased estimator of σ_δ^2 may be negative.

Now we consider the MLE of σ_δ^2 and σ_ϵ^2. In the case of the model (6.52) with the assumption of normality and independence, using the PDF obtained in Example exa:onewayAOVmodel2, it is relatively easy to write the log-likelihood,

$$\begin{aligned} l_L(\mu, \sigma_\delta^2, \sigma_\epsilon^2 \,;\, y) = -\tfrac{1}{2}\big(&mn\log(2\pi) + m(n-1)\log(\sigma_\epsilon^2) + m\log(\sigma_\delta^2) \\ &+ \mathrm{SSE}/\sigma_\epsilon^2 + \mathrm{SSA}/(\sigma_\epsilon^2 + n\sigma_\delta^2) + mn(\bar{y} - \mu)^2/(\sigma_\epsilon^2 + n\sigma_\delta^2)\big)\,. \end{aligned} \tag{6.53}$$

The MLEs must be in the closure of the parameter space, which for σ_δ^2 and σ_ϵ^2 is $\overline{\mathbb{R}}_+$. From this we have the MLEs

- if $(m-1)\mathrm{MSA}/m \geq \mathrm{MSE}$

$$\widehat{\sigma}_\delta^2 = \frac{1}{n}\left(\frac{m-1}{m}\mathrm{MSA} - \mathrm{MSE}\right) \tag{6.54}$$

$$\widehat{\sigma}_\epsilon^2 = \mathrm{MSE} \tag{6.55}$$

- if $(m-1)\mathrm{MSA}/m < \mathrm{MSE}$

$$\widehat{\sigma}_\delta^2 = 0 \tag{6.56}$$

$$\widehat{\sigma}_\epsilon^2 = \frac{m-1}{m}\mathrm{MSE}. \tag{6.57}$$

Again we should note that we are dealing with a special model in Example 6.29; it is "balanced"; that is, for each i, there is a constant number of j's. If, instead, we had $j = 1, \ldots, n_i$, we would not be able to write out the log-likelihood so easily, and the MLEs would be very difficult to determine.

6.4.2 MLEs in Generalized Linear Models

Regression models such as (3.5)

$$Y = f(X\,;\theta) + E$$

are very useful in statistical applications. In this form, we assume independent observations $(Y_1, x_1), \ldots, (Y_n, x_n)$ and Y to represent an n-vector, X to represent an $n \times p$ matrix whose rows are the x_i^{T}, and E to represent an unobservable n-vector of random variables$\sim P_\tau$, with unknown P_τ, but with $\mathrm{E}(E) = 0$ and $\mathrm{V}(E) = \sigma^2 I$.

The expression "$f(\cdot)$" represents a systematic effect related to the values of "X", and "E" represents a random effect, an unexplained effect, or simply a "residual" that is added to the systematic effect.

A model in which the parameters are additively separable and with an additive random effect is sometimes called an additive model:

$$Y = f(X)\theta + \epsilon.$$

A simple version of this is called a linear (additive) model:

$$Y = X\beta + \epsilon, \tag{6.58}$$

where β is a p-vector of parameters. We have considered specific instances of this model in Examples 6.25 and 5.31.

Either form of the additive model can be generalized with a "link function" to be a *generalized additive model*.

In the following, we will concentrate on the linear model, $Y = X\beta + \epsilon$, and we will discuss the link function and the generalization of the linear model, which is called a generalized linear model (GLM or GLIM).

Let us assume that the distribution of the residual has a first moment and that it is known. In that case, we can take its mean to be 0, otherwise, we can incorporate it into $X\beta$. (If the first moment does not exist, we can work with the median.) Hence, assuming the mean of the residual exists, the model can be written as

$$\mathrm{E}(Y) = X\beta,$$

that is, the expected value of Y is the systematic effect in the model. More generally, we can think of the model as being a location family with PDF

$$p_\epsilon(\epsilon) = p_\epsilon(y - X\beta), \tag{6.59}$$

wrt a given σ-finite measure.

In the linear model (6.58), if $\epsilon \sim \mathrm{N}(0, \sigma^2)$, as we usually assume, we can easily identify η_i, $T(y_i)$, and $\zeta(\eta_i)$ in equation (6.32), and of course, $h(y_i) \equiv 1$. This is a location-scale family.

Generalized Linear Models

A model as in equation (6.58) has limitations. Suppose, for example, that we are interested in modeling a response that is binary, for example, two states of a medical patient, "diseased" or "disease-free". As usual, we set up a random variable to map the sample space to $\mathbb{R}$:

$$Y : \{\text{disease-free,diseased}\} \mapsto \{0, 1\}.$$

The linear model $X = Z\beta + \epsilon$ does not make sense. It is continuous and unbounded.

A more useful model may address $\Pr(X = 0)$.

To make this more concrete, consider the situation in which several groups of subjects are each administered a given dose of a drug, and the number responding in each group is recorded. The data consist of the counts y_i responding in the i^{th} group, which received a level x_i of the drug.

A basic model is

$$\begin{aligned} \mathrm{P}(Y_i = 0|x_i) &= 1 - \pi_i \\ \mathrm{P}(Y_i = 1|x_i) &= \pi_i \end{aligned} \tag{6.60}$$

The question is how does π depend on x?

A linear dependence, $\pi = \beta_0 + \beta_1 x$ does not fit well in this kind of situation – unless we impose restrictions, π would not be between 0 and 1.

We can try a transformation to $[0, 1]$.

Suppose we impose an invertible function on

$$\eta = \beta_0 + \beta_1 x$$

that will map it into $[0, 1]$:

$$\pi = h(\eta), \tag{6.61}$$

or

$$g(\pi) = \eta. \tag{6.62}$$

We call this a *link function*.

A common model following this setup is

$$\pi_x = \Phi(\beta_0 + \beta_1 x), \tag{6.63}$$

where Φ is the normal cumulative distribution function, and β_0 and β_1 are unknown parameters to be estimated. This is called a *probit model*. The link function in this case is Φ^{-1}.

The related *logit model*, in which the log odds ratio $\log(\pi/(1-\pi))$ is of interest, has as link function

$$\eta = \log\left(\frac{\pi}{1-\pi}\right). \tag{6.64}$$

Other possibilities are the complementary log-log function

$$\eta = \log(-\log(1-\pi)), \tag{6.65}$$

and the log-log function,

$$\eta = -\log(-\log(\pi)). \tag{6.66}$$

Link Functions

The link function relates the systematic component to the mean of the random variable.

In the case of the linear model, let η_i be the systematic component for a given value of the independent variable,

$$\eta_i = \beta_0 + \beta_1 x_{1i} + \cdots \beta_p x_{pi},$$

and let $\mu_i = \mathrm{E}(Y)$, as before. Let g be the link function:

$$\eta_i = g(\mu_i).$$

In this case, the link function is linear in a set of parameters, β_j, and it is usually more natural to think in terms of these parameters rather than θ,

$$g\left(\frac{\mathrm{d}}{\mathrm{d}\theta}b(\theta_i)\right) = g(\mu_i) = \eta_i = x_i^{\mathrm{T}}\beta.$$

The generalized linear model can now be thought of as consisting of three parts:

1. the systematic component
2. the random component
3. the link between the systematic and random components.

In the context of generalized linear models, a standard linear model has a systematic component of

$$\beta_0 + \beta_1 x_{1i} + \cdots \beta_m x_{mi},$$

a random component that is an identical and independent normal distribution for each observation, and a link function that is the identity.

Fitting Generalized Linear Models

Our initial objective is to fit the model, that is, to determine estimates of the β_j.

The model parameters are usually determined either by a maximum likelihood method or by minimizing some function of the residuals. One approach is to use the link function and do a least squares fit of η using the residuals $y_i - \mu_i$. It is better, however, to maximize the likelihood or, alternatively, the log-likelihood,

$$l(\theta, \phi|y) = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi).$$

The most common method of optimizing this function is "Fisher scoring", which is a method like Newton's method, except that some quantities are replaced by their expected values.

In the generalized linear model, where the likelihood is linked to the parameters that are really of interest, this still must be cast in terms that will yield values for $\widehat{\beta}$.

Analysis of Deviance

Our approach to modeling involves using the observations (including the realizations of the random variables) as fixed values and treating the parameters as variables (not random variables, however). The original model was then encapsulated into a likelihood function, $L(\theta|y)$, and the principle of fitting the model was maximization of the likelihood with respect to the parameters. The log likelihood, $l(\theta|x)$, is usually used.

In model fitting an important issue is how well does the model fit the data? How do we measure the fit? Maybe use residuals. (Remember, some methods of model fitting work this way; they minimize some function of the residuals.) We compare different models by means of the measure of the fit based on the residuals. We make inference about parameters based on changes in the measure of fit.

Using the likelihood approach, we make inference about parameters based on changes in the likelihood. Likelihood ratio tests are based on this principle.

A convenient way of comparing models or making inference about the parameters is with the *deviance function*, which is a likelihood ratio:

$$D(y|\widehat{\theta}) = 2\left(l(\theta_{\max}|y) - l(\widehat{\theta}|y)\right),$$

where $\widehat{\theta}$ is the fit of a potential model.

For generalized linear models the analysis of deviance plays a role similar to that of the analysis of sums of squares (analysis of "variance") in linear models.

Under appropriate assumptions, when θ_1 is a subvector of θ_2, the difference in deviances of two models, $D(y|\widehat{\theta_2}) - D(y|\widehat{\theta_1})$ has an asymptotic chi-squared distribution with degrees of freedom equal to the difference in the number of parameters.

********** repeat below For models with a binary response variable, we need a different measure of residuals. Because we are measuring the model fit in terms of the deviance, D, we may think of the observations as each contributing a quantity d_i, such that $\sum d_i = D$. (Exactly what that value is depends on the form of the systematic component and the link function that are in the likelihood.) The quantity

$$r_i = \text{sign}(y_i - \widehat{\mu}_i)\sqrt{d_i}$$

increases in $(y_i - \widehat{\mu}_i)$ and $\sum r_i^2 = D$. We call r_i the *deviance residual*.

For the logit model,

$$r_i = \text{sign}(y_i - \widehat{\mu}_i)\sqrt{-2\left(y_i \log(\widehat{\pi}_i) + (1 - y_i)\log(1 - \widehat{\pi}_i)\right)}$$

Generalized Additive Models

The mechanical process of dealing with generalized additive models parallels that of dealing with generalized linear models. There are some very important differences, however. The most important is probably that the distribution of the deviances is not worked out.

The meaning of degrees of freedom is also somewhat different.

So, first, we work out an analogous concept for degrees of freedom.

The response variable is Bernoulli (or binomial). We model the log odds ratios as

$$\begin{aligned} \log\left(\frac{\pi_i}{1-\pi_i}\right) &= \eta_i \\ &= \beta_0 + \beta_1 x_{1i} + \cdots + \beta_6 x_{6i} \\ &= x_i^{\mathrm{T}}\beta. \end{aligned}$$

For a binomial with number m_i, we write the log-likelihood,

$$l(\pi|y) = \sum_{i=1}^{n} \left(y_i \log(\pi_i/(1-\pi_i)) + m_i \log(1-\pi_i)\right),$$

where a constant involving m_i and y_i has been omitted. Substituting, we have,

$$l(\beta|y) = \sum_{i=1}^{n} y_i x_i^{\mathrm{T}}\beta - \sum_{i=1}^{n} m_i \log\left(1 + \exp(x_i^{\mathrm{T}}\beta)\right).$$

The log likelihood depends on y only through $X^{\mathrm{T}}y$.

$$\frac{\partial l}{\partial \pi_i} = \frac{y_i - m_i\pi_i}{\pi_i(1-\pi_i)}$$

Using the chain rule, we have

$$\begin{aligned}\frac{\partial l}{\partial \beta_j} &= \sum_{i=1}^{n} \frac{y_i - m_i\pi_i}{\pi_i(1-\pi_i)} \frac{\partial \pi_i}{\partial \beta_j} \\ &= \sum_{i=1}^{n} \frac{y_i - m_i\pi_i}{\pi_i(1-\pi_i)} \frac{d\pi_i}{d\eta_i} x_{ij}\end{aligned}$$

The Fisher information is

$$\begin{aligned}-\mathrm{E}\left(\frac{\partial^2 l}{\partial \beta_j \partial \beta_k}\right) &= \sum_{i=1}^{n} \frac{m_i}{\pi_i(1-\pi_i)} \frac{\partial \pi_i}{\partial \beta_j} \frac{\partial \pi_i}{\partial \beta_k} \\ &= \sum_{i=1}^{n} \frac{m_i (d\pi_i/d\eta_i)^2}{\pi_i(1-\pi_i)} x_{ij} x_{ik} \\ &= (X^{\mathrm{T}} W X)_{jk},\end{aligned}$$

where W is a diagonal matrix of weights,

$$\frac{m_i(\mathrm{d}\pi_i/\mathrm{d}\eta_i)^2}{\pi_i(1-\pi_i)}$$

Notice

$$\frac{\mathrm{d}\pi_i}{\mathrm{d}\eta_i} = \pi_i(1-\pi_i),$$

so we have the simple expression,

$$\frac{\partial l}{\partial \beta} = X^{\mathrm{T}}(y - m\pi)$$

in matrix notation, and for the weights we have,

$$m_i\pi_i(1-\pi_i)$$

Use Newton's method,

$$\widehat{\beta}^{(k+1)} = \widehat{\beta}^{(k)} - H_l^{-1}(\widehat{\beta}^{(k)}) \nabla l(\widehat{\beta}^{(k)}),$$

in which H_l is replaced by

$$\mathrm{E}\left(-\frac{\partial^2 l}{\partial \beta \partial \beta^{\mathrm{T}}}.\right)$$

Using $\widehat{\beta}^{(k)}$, we form $\widehat{\pi}^{(k)}$ and $\widehat{\eta}^{(k)}$, and then, an adjusted $y^{(k)}$,

$$y_i^{(k)} = \widehat{\eta}^{(k)} + \frac{(y - m_i\widehat{\pi}_i^{(k)})}{m_i} \frac{\mathrm{d}\eta_i}{\mathrm{d}\pi_i}$$

This leads to

$$\widehat{\beta}^{(k+1)} = (X^{\mathrm{T}} W^{(k)} X)^{-1} X^{\mathrm{T}} W^{(k)} y^{(k)},$$

and it suggests an iteratively reweighted least squares (IRLS) algorithm.

Residuals

For models with a binary response variable, we need a different measure of residuals. Because we are measuring the model fit in terms of the deviance, D, we may think of the observations as each contributing a quantity d_i, such that $\sum d_i = D$. (Exactly what that value is depends on the form of the systematic component and the link function that are in the likelihood.) The quantity

$$r_i^D = \text{sign}(y_i - \widehat{\mu}_i)\sqrt{d_i}$$

increases in $(y_i - \widehat{\mu}_i)$ and $\sum (r_i^D)^2 = D$. We call r_i^D the *deviance residual*.

For the logit model,

$$r_i^D = \text{sign}(y_i - \widehat{\mu}_i)\sqrt{-2\left(y_i \log(\widehat{\pi}_i) + (1 - y_i)\log(1 - \widehat{\pi}_i)\right)}.$$

Another kind of residual is called the "working" residual. It is

$$r_i^W = (y_i - \widehat{\mu}_i)\frac{\partial \widehat{\eta}_i}{\partial \widehat{\mu}_i},$$

where the derivatives are evaluated at the final iteration of the scoring algorithm.

In the logistic regression model, these working residuals are

$$\frac{y_i - \widehat{\pi}_i}{\widehat{\pi}_i(1 - \widehat{\pi}_i)}$$

Residuals can be standardized by taking into account their different standard deviations that result from the influence.

This is the same kind of concept as influence in linear models. Here, however, we have

$$\widehat{\beta}^{(k+1)} = (X^{\mathrm{T}} W^{(k)} X)^{-1} X^{\mathrm{T}} W^{(k)} y^{(k)},$$

where the weights are

$$m_i \widehat{\pi}_i^{(k)}(1 - \widehat{\pi}_i^{(k)}).$$

One measure is the diagonal of the hat matrix:

$$W^{\frac{1}{2}} X (X^{\mathrm{T}} W X)^{-1} X^{\mathrm{T}} W^{\frac{1}{2}}$$

In the case of generalized linear models, the hat matrix is only the prediction transformation matrix for the linear, systematic component.

Data consisting of counts, for example, the number of certain events within a fixed period of time, give rise naturally to a Poisson model. The relationship between the mean and the covariates is often assumed to be multiplicative, giving rise to a log-linear model,

$$\log(\mu) = \eta = x^{\mathrm{T}}\beta.$$

Another possibility for count data is that the covariates have an additive effect and the direct relation

$$\mu = x^{\mathrm{T}}\beta$$

can be used.

Notice that the mean of the binomial and the Poisson distributions determine the variance.

In practice the variance of discrete response data, such as binomial or Poisson data, is observed to exceed the nominal variance that would be determined by the mean.

This phenomenon is referred to as "over-dispersion". There may be logical explanations for over-dispersion, such as additional heterogeneity over and above what is accounted for by the covariates, or some more complicated variance structure arising from correlations among the responses.

6.5 Variations on the Likelihood

There are situations in which a likelihood equation either cannot be written or else it is not solvable. This may happen because of too many parameters, for example. In such cases an approximate likelihood equation may be more appropriate. In other cases, there may be a nuisance parameter that complicates the computation of the MLE for the parameter of interest. In both kind of these situations, we use approximate likelihood methods.

6.5.1 Quasi-likelihood Methods

Another way we deal with nuisance parameters in maximum likelihood estimation is by making some simplifying approximations. One type of simplification is to reduce the dimensionality of the nuisance parameters by assuming some relationship among them. This yields a "quasi-likelihood" function. This may allow us to solve what otherwise might be a very difficult problem. In some cases it may not affect the MLE for the parameters of interest. A common application in which quasi-likelihood methods are useful is in estimation of parameters in a generalized linear model.

Quasi-likelihood Methods in Generalized Linear Models

Over-dispersion in the generalized linear model can often be accounted for by the nuisance parameter ϕ in the likelihood. For example, we modify the simple binomial model so the variance is

$$\mathrm{V}(y_i|x_i) = \phi\frac{\pi_i(1-\pi_i)}{n_i}.$$

Notice the multiplier ϕ is constant, while π depends on the covariates and n depends on the group size. This of course leads to a more complicated likelihood function, but it may not be necessary to use the actual likelihood.

Quasi-likelihood and need not correspond to any particular distribution; rather quasi can be used to combine any available link and variance function.

Wedderburn (1974) introduced a quasi-likelihood function to allow

$$\mathrm{E}(y|x) = \mu = h(x^{\mathrm{T}}\beta)$$

and

$$\mathrm{V}(y|x) = \sigma^2(\mu) = \phi v(\mu),$$

where ϕ is the (nuisance) dispersion parameter in the likelihood and $v(\mu)$ is a variance function that is entirely separate from the likelihood.

Quasi-likelihood methods require only specification of a relationship between the mean and variance of the response.

In a multiparameter case, $\theta = (\theta_1, \theta_2)$, we may be interested in only some of the parameters, or in some function of the parameters, perhaps a transformation into a lower-dimensional space. There are various ways of approaching this.

6.5.2 Nonparametric and Semiparametric Models

Empirical Likelihood

Profile Likelihood

If $\theta = (\theta_1, \theta_2)$ and our interest is only in θ_1, the simplest way of handling this is just to consider θ_2 to be fixed, perhaps at several different values, one at a time. If θ_2 is fixed, the likelihood $L(\theta_1\,;\,\theta_2, x)$ is called a *profile likelihood* or *concentrated likelihood* of θ_1 for given θ_2 and x.

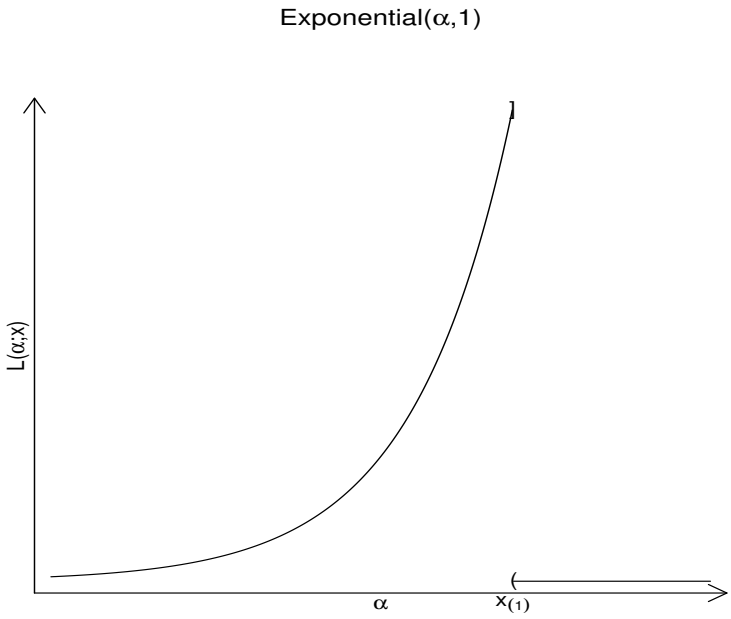
*** build up Example 6.7 ... 2-D, then profile ******** the derivative is not useful in finding the MLE is in a parametric-support family. For example, assume $X_1, \ldots, X_n \overset{\text{iid}}{\sim}$exponential$(\alpha, 1)$. The likelihood is

$$L(\alpha\,;\,x) = \mathrm{e}^{-\sum(x_i-\alpha)}\mathrm{I}_{]-\infty, x_{(1)}]}(\alpha).$$

Setting the derivative to 0 is not a useful way to find a stationary point. (Note that the derivative of the indicator function is the Dirac delta function.) In fact, the max does not occur at a stationary point. The MLE of α is $x_{(1)}$.

********** relate this to Examples 1.5 and 6.5

****** make 2-D plot



In some cases, it turns out that the estimation of a subset of the parameters does not depend on the value of some other subset. A good method of estimation of β in a linear model $X = Z\beta + \epsilon$ where the residuals ϵ have a common variance σ^2 and zero correlation can be performed equally well no matter what the value of σ^2 is. (The Gauss-Markov theorem tells us that the least-squares method yields a good estimator.) If the residuals are independently distributed as normals with a common variance, we can formulate the problem as a problem in maximum likelihood estimation. The MLE for β (which just happens to be the same as the LSE) can be thought of in terms of a profile likelihood, because a particular value of σ^2 could be chosen a priori. (This is of course not necessary because the maximum or the likelihood with respect to β occurs at the same point regardless of the value of σ^2.)

Conditional Likelihood

When there is a nuisance parameter for which we have a sufficient statistic, a simple approach is to use the PDF conditional on the sufficient statistic to form the likelihood function for the parameter of interest. After doing this, the MLE procedure continues as in the usual case. If the PDFs can be factored so that one factor includes θ_2 and some function of the sample, $S(x)$, and the other factor, given $S(x)$, is free of θ_2, then this factorization can be carried

into the likelihood. Such a likelihood is called a *conditional likelihood* of θ_1 given $S(x)$.

Conditional likelihood methods often arise in applications in which the parameters of two different distributions are to be compared; that is, when only their relative values are of interest. Suppose $\mu = (\mu_1, \mu_2)$ and let $\theta_1 = \mu_1/\mu_2$. Although our interest is in θ_1, we may not be able to write the likelihood as a function of θ_1. If, however, we can find θ_2 for which we have a sufficient statistic, $T_2(X)$, and we can factor the likelihood using the factorization theorem so that the factor corresponding to conditional distribution of X given $T_2(X)$ does not depend on θ_2. This factor, as a function of θ_1, is the conditional likelihood function.

Sometimes a profile likelihood can be thought of as a particularly simple conditional likelihood. The linear model estimation problem referred to above could be formulated as a conditional likelihood. The actual form of the likelihood would be more complicated, but the solution is equivalent to the solution in which we think of the likelihood as a profile likelihood.

Conditional Likelihood for the Exponential Class

If X has a distribution in the exponential class with $\theta = (\eta_1, \eta_2)$, and its likelihood can be written in the form

$$L(\theta\,;\,x) = \exp(\eta_1^{\mathrm{T}} T_1(x) + \eta_2^{\mathrm{T}} T_2(x) - \zeta(\eta_1, \eta_2))h(x),$$

or, in the log-likelihood form,

$$l_L(\theta\,;\,x) = \eta_1^{\mathrm{T}} T_1(x) + \eta_2^{\mathrm{T}} T_2(x) - \zeta(\eta_1, \eta_2) + c(x),$$

we can easily write the conditional log-likelihood:

$$l_L(\eta_1\,;\,x\,;\,T_2) = \eta_1^{\mathrm{T}} T_1(x) + \tilde{\zeta}(\eta_1, T_2) + c(x).$$

Notice that this decomposition can be achieved iff η_1 is a linear function of θ.

If our interest is only in η_1, we only determine the argument that maximizes the function

$$\eta_1^{\mathrm{T}} T_1(x) + \tilde{\zeta}(\eta_1, T_2),$$

which is does not depend on η_2.

****

Partial Likelihood

The idea of partial likelihood is somewhat similar to conditional likelihood. The most common area of application is in *semiparametric models*. These are models of the form

$$f(x;\theta) = g(x;\theta)h(x), \tag{6.67}$$

where x is observable, θ is unknown and unobservable, g is a function of known form, but f and h are of unknown form. The estimation problem has two components: the estimation of parameter θ and the nonparametric estimation of the function h.

In the setup of equation (6.67) when $f(x;\theta)$ is the PDF of the observable, we form a *partial likelihood* function based on $g(x;\theta)$. This partial likelihood is an likelihood in the sense that it is a constant multiple (wrt θ) of the full likelihood function. The parameter θ can be estimated using the ordinary method for MLE.

The most common example of this kind of problem in statistical inference is estimation of the proportional hazards model. Rather than discuss partial likelihood further here, I will postpone consideration of this semiparametric problem to Section 8.4 beginning on page 576.

Notes and Further Reading

Most of the material in this chapter is covered in MS2 Section 4.4, Section 4.5,and Section 5.4, and in TPE2 Chapter 6.

Likelihood and Probability

Although it is natural to think of the distribution that yields the largest lieklihood as the "most probable" distribution that gave rise to an observed sample, it is important not to think of the likelihood function, even if it could be properly normalized, as a probability density. In the likelihood approach to statistical inference, there is no posterior conditional probability distribution as there is in a Bayesian approach. The book by Edwards (1992) provides a good discussion of the fundamental concept of likelihood.

EM Methods

EM methods were first discussed systematically by Dempster et al. (1977). A general reference for EM methods is Ng et al. (2012).

Computations

The R function `fitdistr` in the `MASS` library computes the MLEs for a number of common distributions.

Multiple RLEs

There are interesting open questions associated with determining if an RLE yields a global maximum. See, for example, Biernacki (2005).

Maximum Likelihood in Linear Models

Variance Components

The problem of estimation of variance components in linear models received considerable attention during the later heyday of the development of statistical methods for analysis of variance. The MLEs for the between-variance, σ^2_δ, and the residual variance, σ^2_ϵ, in the balanced one-way random-effects model (equations (6.54) through (6.57)) were given by Herbach (1959). Thompson Jr. (1962) suggested a restricted maximum likelihood approach in which the estimator is required to be equivariant. This method has come to be the most commonly used method of variance components estimation. This method is a more general form of REML, which we used in a special case in Example 6.27.

As we mentioned, there are great differences in methods of estimation of variance components depending on whether the data are balanced, as in Example 5.31, or unbalanced. Various problems of variance component estimation in unbalanced one-way models were discussed by Harville (1969). Searle et al. (1992) provide a thorough coverage of the various ways of estimating variance components and the underlying theory.

Unbiasedness and Consistency

While many MLEs are biased, most of the ones encountered in common situations are at least consistent in mean squared error. Neyman and Scott (1948) give an example, which is a simplified version of an example due to Wald, of an MLEs that is not consistent. The problem is the standard one-way ANOVA model with two observations per class. The asymptotics are in the number of classes, and hence, of course in the number of observations. The model is $X_{ij} \sim \mathrm{N}(\mu_j, \sigma^2)$ with $i = 1, 2$ and $j = 1, 2, \ldots$. The asymptotic (and constant) expectation of the MLE of σ^2 is $\sigma^2/2$. This example certainly shows that MLEs may behave very poorly, but its special property should be recognized. The dimension of the parameter space is growing at the same rate as the number of observations.

Quasilikelihood

The idea of a quasilikelihood began with the work of Wedderburn (1974) on generalized linear models. This work merged with the earlier work of Durbin (1960) and Godambe (1960) on estimating functions. Heyde (1997) covers the important topics in the area.

Empirical Likelihood

The initial studies of empirical likelihood were in the application of likelihood ratio methods in nonparametric inference in the 1970s. Owen (2001) provides and introduction and summary.

Exercises

6.1. Consider the problem in Exercise 5.2 of using a sample of size 1 for estimating $g(\theta) = \mathrm{e}^{-3\theta}$ where θ is the parameter in a Poisson distribution. What is the MLE of $g(\theta)$?

6.2. Show that the MLE of σ^2 in the one-way fixed-effects model is as given in equation (6.49) on page 489.

6.3. Suppose $X_{ij} \overset{\text{iid}}{\sim} \mathrm{N}(\mu_i, \sigma^2)$ for $i = 1, \ldots, m$ and $j = 1, \ldots, n$. (Compare Exercise 5.10 on page 444.)
   a) Show that the MLE of σ^2 is not consistent in mean squared error as $m \to \infty$ and n remains fixed.
   b) Show that the MLE of σ^2 is consistent in mean squared error as $n \to \infty$ and m remains fixed.

6.4. a) Show that the log-likelihood given in equation (6.53) is correct.
   b) Show that the MLEs given in equations (6.54) through (6.57) maximize the likelihood over $(\bar{\mathbb{R}}_+)^2$.

6.5. Fill in the details for the proof of Theorem 6.3.

6.6. Given a Bernoulli distribution with parameter π. We wish to estimate the variance $g(\pi) = \pi(1-\pi)$. Compare the MSE of the UMVUE in equation (5.11) with the MLE in equation (6.26).

6.7. Determine the MLE of μ for the distribution with CDF given in equation (6.27) if $P(x)$ is the CDF of the distribution $\mathrm{N}(\mu, \sigma^2)$. At what point is it discontinuous in the data?

6.8. Computations for variations on Example 6.19. Use a computer program, maybe R to generate some artificial data to use to experiment with the EM method in some variations of the normal mixtures model. Take $\theta = (0.7, 0, 1, 1, 2)$. The following R code will generate 300 observations from such a model.

```
# Generate data from normal mixture.
# Note that R uses sigma, rather than sigma^2 in rnorm.
#  Set the seed, so computations are reproducible.
set.seed(4)
n <- 300
w <- 0.7
mu1 <- 0
sigma21 <- 1
mu2 <- 5
sigma22 <- 2
x <- ifelse(runif(n)<w,
rnorm(n,mu1,sqrt(sigma21)),rnorm(n,mu2,sqrt(sigma22)))
```

   a) Assume that μ_1, σ_1^2, μ_2, and σ_2^2 are all known and use EM to estimate $\theta_1 = w$.

b) Assume that σ_1^2 and σ_2^2 are known and use EM to estimate θ_1, θ_2, and θ_4.

c) Assume that all are unknown and use EM to estimate θ.

6.9. Consider another variation on the normal mixture in Example 6.19. Assume that w is known and $\sigma_1^2 = \sigma_2^2 = \sigma^2$, but μ_1, μ_2, and σ^2 are unknown. If $w = 1/2$, this setup is similar to the two-class one-way fixed effects AOV model in Example 6.26. What are the differences? Compare the estimators in the two setups.

7

Statistical Hypotheses and Confidence Sets

In a frequentist approach to statistical hypothesis testing, the basic problem is to decide whether or not to reject a statement about the distribution of a random variable. The statement must be expressible in terms of membership in a well-defined class. The hypothesis can therefore be expressed by the statement that the distribution of the random variable X is in the class $\mathcal{P}_H = \{P_\theta : \theta \in \Theta_H\}$. An hypothesis of this form is called a statistical hypothesis.

The basic paradigm of statistical hypothesis testing was described in Section 3.5.1, beginning on page 290. We first review some of those ideas in Section 7.1, and then in Section 7.2 we consider the issue of optimality of tests. We first consider the Neyman-Pearson Fundamental Lemma, which identifies the optimal procedure for testing one simple hypothesis versus another simple hypothesis. Then we discuss tests that are *uniformly* optimal in Section 7.2.2. As we saw in the point estimation problem, it is often not possible to develop a procedure that is uniformly optimal, so just as with the estimation problem, we can impose restrictions, such as unbiasedness or invariance, or we can define uniformity in terms of some global risk. Because hypothesis testing is essentially a binary decision problem, a minimax criterion usually is not relevant, but use of global averaging may be appropriate. (This is done in the Bayesian approaches described in Section 4.5, and we will not pursue it further in this chapter.)

If we impose restrictions on certain properties of the acceptable tests, we then proceed to find uniformly most powerful tests under those restrictions. We discuss unbiasedness of tests in Section 7.2.3, and we discuss uniformly most powerful unbiased tests in Section 7.2.4. In Section 7.3, we discuss general methods for constructing tests based on asymptotic distributions. Next we consider additional topics in testing statistical hypotheses, such as nonparametric tests, multiple tests, and sequential tests.

Confidence sets are closely related to hypothesis testing. In general, rejection of an hypothesis is equivalent to the hypothesis corresponding to a set of parameters or of distributions outside of a confidence set constructed at

a level of confidence that corresponds to the level of significance of the test. The basic ideas of confidence sets were discussed in Section 3.5.2, beginning on page 296. The related concept of credible sets was described in Section 4.6.1, beginning on page 372. Beginning in Section 7.1 of the present chapter, we discuss confidence sets in somewhat more detail.

The Decisions in Hypothesis Testing

It is in hypothesis testing more than in any other type of statistical inference that the conflict among various fundamental philosophies come into sharpest focus.

Neyman-Pearson; two
Fisher significance test; one
one, where the other is "all others"
evidence as measured by likelihood

7.1 Statistical Hypotheses

A problem in statistical hypothesis testing is set in the context of a given broad family of distributions, $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. As in other problems in statistical inference, the objective is to decide whether the given observations arose from some subset of distributions $\mathcal{P}_H \subseteq \mathcal{P}$.

The statistical hypothesis is a statement of the form "the family of distributions is $\mathcal{P}_H$", where $\mathcal{P}_H \subseteq \mathcal{P}$, or perhaps "$\theta \in \Theta_H$", where $\Theta_H \subseteq \Theta$.

The full statement consists of two pieces, one part an assumption, "assume the distribution of X is in the class", and the other part the hypothesis, "$\theta \in \Theta_H$, where $\Theta_H \subseteq \Theta$." Given the assumptions, and the definition of Θ_H, we often denote the hypothesis as H, and write it as

$$H : \theta \in \Theta_H. \tag{7.1}$$

Two Hypotheses

While, in general, to reject the hypothesis H would mean to decide that $\theta \notin \Theta_H$, it is generally more convenient to formulate the testing problem as one of deciding between two statements:

$$H_0 : \theta \in \Theta_0 \tag{7.2}$$

and

$$H_1 : \theta \in \Theta_1, \tag{7.3}$$

where $\Theta_0 \cap \Theta_1 = \emptyset$. These two hypotheses could also be expressed as "the family of distributions is $\mathcal{P}_0$" and "the family of distributions is $\mathcal{P}_1$", respectively, with the obvious meanings of $\mathcal{P}_0$ and $\mathcal{P}_1$.

We do not treat H_0 and H_1 symmetrically; H_0 is the *hypothesis* (or "null hypothesis") to be tested and H_1 is the *alternative*. This distinction is important in developing a methodology of testing.

Tests of Hypotheses

To test the hypotheses means to choose one hypothesis or the other; that is, to make a decision, d, where d is a real number that indicates the hypothesis accepted. As usual in statistical inference, we have a sample X from the relevant family of distributions and a statistic $T(X)$ on which we base our decision.

A *nonrandomized test procedure* is a rule $\delta(X)$ that assigns two decisions to two disjoint subsets, C_0 and C_1, of the range of $T(X)$ that we call the test statistic. We equate those two decisions with the real numbers 0 and 1, so $\delta(X)$ is a real-valued function,

$$\delta(x) = \begin{cases} 0 \text{ for } T(x) \in C_0 \\ 1 \text{ for } T(x) \in C_1. \end{cases} \tag{7.4}$$

Note for $i = 0, 1$,

$$\Pr(\delta(X) = i) = \Pr(T(X) \in C_i). \tag{7.5}$$

We call C_1 the *critical region*, and generally denote it by just C.

A test $\delta(X)$ is associated with a critical region C. We may use the term "critical region" either to denote a set of values of a statistic $T(X)$ or just of the sample X itself.

If $\delta(X)$ takes the value 0, the decision is not to reject; if $\delta(X)$ takes the value 1, the decision is to reject. If the range of $\delta(X)$ is $\{0, 1\}$, the test is a nonrandomized test.

Although occasionally it may be useful to choose the range of $\delta(X)$ as some other set of real numbers, such as $\{d_0, d_1\}$ or even a set with cardinality greater than 2, we generally define the decision rule so that $\delta(X) \in [0, 1]$. If the range is taken to be the closed interval $[0, 1]$, we can interpret a value of $\delta(X)$ as the probability that the null hypothesis is rejected.

If it is not the case that $\delta(X)$ equals 0 or 1 a.s., we call the test a *randomized test*.

Errors in Decisions Made in Testing

There are four possibilities in a test of an hypothesis: the hypothesis may be true, and the test may or may not reject it, or the hypothesis may be false, and the test may or may not reject it. The result of a statistical hypothesis test can be incorrect in two distinct ways: it can reject a true hypothesis or it can fail to reject a false hypothesis. We call rejecting a true hypothesis a "type I error", and failing to reject a false hypothesis a "type II error".

Our standard approach in hypothesis testing is to control the level of the probability of a type I error under the assumptions, and to try to find a test subject to that level that has a small probability of a type II error.

We call the maximum allowable probability of a type I error the "significance level", and usually denote it by α.

We call the probability of rejecting the null hypothesis the *power of the test*, and will denote it by β. If the alternate hypothesis is the true state of nature, the power is one minus the probability of a type II error.

It is clear that we can easily decrease the probability of one type of error (if its probability is positive) at the cost of increasing the probability of the other.

In a common approach to hypothesis testing under the given assumptions on X (and using the notation above), we choose $\alpha \in]0,1[$ and require that $\delta(X)$ be such that

$$\Pr(\delta(X) = 1 \mid \theta \in \Theta_0) \leq \alpha. \tag{7.6}$$

and, subject to this, find $\delta(X)$ so as to minimize

$$\Pr(\delta(X) = 0 \mid \theta \in \Theta_1). \tag{7.7}$$

Optimality of a test T is defined in terms of this constrained optimization problem.

Notice that the restriction on the type I error applies $\forall \theta \in \Theta_0$. We call

$$\sup_{\theta \in \Theta_0} \Pr(\delta(X) = 1 \mid \theta) \tag{7.8}$$

the *size of the test*. If the size of the test is less than the significance level, then the test can be modified, possibly by use of an auxiliary random mechanism.

In common applications, $\Theta_0 \cup \Theta_1$ forms a convex region in $\mathbb{R}^k$, and Θ_0 contains the set of common closure points of Θ_0 and Θ_1 and $\Pr(\delta(X) = 1 \mid \theta)$ is a continuous function of θ; hence the sup in equation (7.8) is generally a max. (The set of common closure points, that is, the boundary between Θ_0 and Θ_1, will have a prominent role in identifying optimal tests.)

If the size is less than the level of significance, the test is said to be *conservative*, and in that case, we often refer to α as the "nominal size".

Example 7.1 Testing in the exponential family

Suppose we have observations $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{exponential}(\theta)$. The Lebesgue PDF is

$$p_\theta(x) = \theta^{-1} \mathrm{e}^{-x/\theta} \mathrm{I}_{]0,\infty[}(x),$$

with $\theta \in]0,\infty[$. Suppose now we wish to test

$$H_0 : \theta \leq \theta_0 \qquad \text{versus} \qquad H_1 : \theta > \theta_0.$$

We know that $\overline{X}$ is sufficient for θ. If H_1 is true, $\overline{X}$ is likely to be larger than if H_0 is true, so a reasonable test may be to reject H_0 if $T(X) = \overline{X} > c_\alpha$, where c_α is some fixed positive constant; that is,

$$\delta(X) = \mathrm{I}_{]c_\alpha,\infty[}(T(X)).$$

We choose c_α so as to control the probability of a type I error. We call $T(X)$ the test statistic.

Knowing the distribution of $\overline{X}$ to be gamma$(n, \theta/n)$, we can now work out

$$\Pr(\delta(X) = 1 \mid \theta) = \Pr(T(X) > c_\alpha \mid \theta),$$

which, for $\theta < \theta_0$ is the probability of a type I error. We set up the testing procedure so as to limit the probability of this type of error to be no more than α.

For $\theta \geq \theta_0$

$$1 - \Pr(\delta(X) = 1 \mid \theta)$$

is the probability of a type II error.

Over the full range of θ, we identify the power of the test as

$$\beta(\theta) = \Pr(\delta(X) = 1 \mid \theta).$$

These probabilities for $n = 1$, as a function of θ, are shown in Figure 7.1.

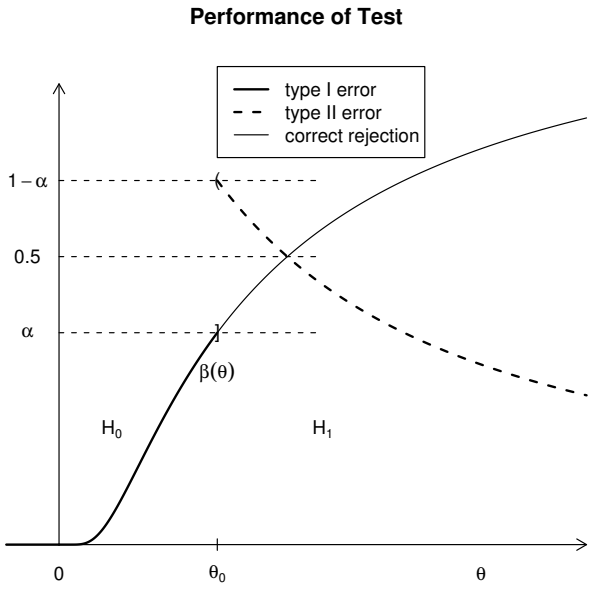


Figure 7.1. Probabilities of Type I and Type II Errors

Now, for a given significance level α, we choose c_α so that

$$\Pr(T(X) > c_\alpha \mid \theta \leq \theta_0) \leq \alpha.$$

This is satisfied for c_α if $\Pr(Y > c_\alpha) = \alpha$, where Y is a random variable with the gamma$(n, \theta_0/n)$ distribution. ∎

p-Values

Note that there is a difference in *choosing* the test procedure, and in *using* the test. To use the test, the question of the choice of α comes back. Does it make sense to choose α first, and then proceed to apply the test just to end up with a decision d_0 or d_1? It is not likely that this rigid approach would be very useful for most objectives. In statistical data analysis our objectives are usually broader than just deciding which of two hypotheses appears to be true based on some arbitrary standard for "truth". On the other hand, if we have a well-developed procedure for testing the two hypotheses, the decision rule in this procedure could be very useful in data analysis.

One common approach is to use the *functional form* of the rule, but not to pre-define the critical region. Then, *given the same setup* of null hypothesis and alternative, to collect data $X = x$, and to determine the smallest value $\hat{\alpha}(x)$ at which the null hypothesis would be rejected. The value $\hat{\alpha}(x)$ is called the *p-value* of x associated with the hypotheses. The p-value indicates the strength of the evidence of the data against the null hypothesis. If the alternative hypothesis is "everything else", a test based a p-value is a significance test.

Although use of p-values represents a fundamentally different approach to hypothesis testing than an approach based on a pre-selected significance level, the p-value does correspond to the "smallest" significance under which the null hypothesis would be rejected. Because of practical considerations, computer software packages implementing statistical hypothesis testing procedures report p-values instead of "reject" or "do not reject".

Example 7.2 Testing in the exponential family; p-value
Consider again the problem in Example 7.1, where we had observations $X_1, \ldots, X_n \overset{\text{iid}}{\sim}$exponential$(\theta)$, and wished to test

$$H_0 : \theta \leq \theta_0 \qquad \text{versus} \qquad H_1 : \theta > \theta_0.$$

Our test was based on $T(X) = \overline{X} > c$, where c was some fixed positive constant chosen so that $\Pr(Y > c) = \alpha$, where Y is a random variable distributed as gamma$(n, \theta_0/n)$.

Suppose instead of choosing c, we merely compute $\Pr(Y > \bar{x})$, where $\bar{x}$ is the mean of the set of observations. This is the p-value for the null hypothesis and the given data.

If the p-value is less than a prechosen significance level α, then the null hypothesis is rejected. ∎

***** tests

Power of a Statistical Test

We call the probability of rejecting H_0 the *power* of the test, and denote it by β, or for the particular test $\delta(X)$, by β_δ. The power is defined over the full set of distributions in the union of the hypotheses. For hypotheses concerning the parameter θ, as in Example 7.1, the power can be represented as a curve $\beta(\theta)$, as shown in Figure 7.1. We see that the *power function* of the test, for any given $\theta \in \Theta$ as

$$\beta_\delta(\theta) = \mathrm{E}_\theta(\delta(X)). \tag{7.9}$$

The power in the case that H_1 is true is 1 minus the probability of a type II error. Thus, minimizing the error in equation (7.7) is equivalent to maximizing the power within Θ_1.

The probability of a type II error is generally a function of the true distribution of the sample P_θ, and hence so is the power, which we may emphasize by the notation $\beta_\delta(P_\theta)$ or $\beta_\delta(\theta)$. In much of the following, we will assume that $\theta \in \Theta \subseteq \mathbb{R}^k$; that is, the statistical inference is "parametric". This setup is primarily one of convenience, because most concepts carry over to more general nonparametric situations. There are some cases, however, when they do not, as for example, when we speak of continuity wrt θ. We now can focus on the test under either hypothesis (that is, under either subset of the family of distributions) in a unified fashion.

Because the power is generally a function of θ, what does maximizing the power mean? That is, maximize it *for what values of* θ? Ideally, we would like a procedure that yields the maximum for all values of θ; that is, one that is most powerful for all values of θ. We call such a procedure a *uniformly most powerful* or UMP test. For a given problem, finding such procedures, or establishing that they do not exist, will be one of our primary objectives.

In some cases, $\beta_\delta(\theta)$ may be a continuous function of θ. Such cases may allow us to apply analytic methods for identifying most powerful tests within a class of tests satisfying certain desirable restrictions. (We do this on page 525.)

Randomized Tests

We defined a randomized test (page 293) as one whose range is not a.s. $\{0, 1\}$. Because in this definition, a randomized test does not yield a "yes/no" decision about the hypothesis being tested, a *test with a random component* is more useful.

Given a randomized test $\delta(X)$ that maps $\mathcal{X}$ onto $\{0, 1\} \cup D_R$, we can construct a test with a random component using the rule that if if $\delta(X) \in D_R$, then the experiment R is performed with $\delta_R(X)$ chosen so that the overall probability of a type I error is the desired level. The experiment R is independent of the random variable about whose distribution the hypothesis applies to. As a practical matter a $\mathrm{U}(0, 1)$ random variable can be used to define the random experiment. The random variable itself is often simulated on a computer.

A test with a random component may be useful for establishing properties of tests or as counterexamples to some statement about a given test. (We often use randomized estimators in this way; see Example 5.25.)

Another use of tests with random components is in problems with countable sample spaces when a critical region within the sample space cannot be constructed so that the test has a specified size.

While randomized estimators rarely have application in practice, randomized test procedures can actually be used to increase the power of a conservative test. Use of a randomized test in this way would not make much sense in real-world data analysis, but if there are regulatory conditions to satisfy, it might needed to achieve an exact size.

7.2 Optimal Tests

Testing statistical hypotheses involves making a decision whether or not to reject a null hypothesis. If the decision is not to reject we may possibly make a secondary decision as to whether or not to continue collecting data, as we discuss in Section 7.6. For the moment, we will ignore the sequential testing problem and address the more basic question of optimality in testing. We first need a measure or criterion.

A general approach to defining optimality is to define a loss function that increases in the "badness" of the statistical decision, and to formulate the *risk* as the expected value of that loss function within the context of the family of probability models being considered. Optimal procedures are those that minimize the risk. The decision-theoretic approach formalizes these concepts.

Decision-Theoretic Approach

The decision space in a testing problem is usually $\{0, 1\}$, which corresponds respectively to not rejecting and rejecting the hypothesis. (We may also allow for another alternative corresponding to "making no decision".) As in the decision-theoretic setup, we seek to minimize the *risk*:

$$R(P, \delta) = \mathrm{E}\big(L(P, \delta(X))\big). \tag{7.10}$$

In the case of the 0-1 loss function and the four possibilities, the risk is just the probability of either type of error.

We want a test procedure that minimizes the risk, but rather than taking into account the total expected loss in the risk (7.10), we generally prefer to restrict the probability of a type I error as in inequality (7.6) and then, subject to that, minimize the probability of a type II error as in equation (7.7), which is equivalent to maximizing the power under the alternative hypothesis. This approach is minimizes the risk subject to a restriction that the contribution to the risk from one type of loss is no greater than a specified amount.

The issue of a uniformly most powerful test is similar to the issue of a uniformly minimum risk test subject to a restriction.

An Optimal Test in a Simple Situation

First, consider the problem of picking the optimal critical region C in a problem of testing the hypothesis that a discrete random variable has the probability mass function $p_0(x)$ versus the alternative that it has the probability mass function $p_1(x)$. We will develop an optimal test for any given significance level based on one observation.

For $x \ni p_0(x) > 0$, let

$$r(x) = \frac{p_1(x)}{p_0(x)}, \tag{7.11}$$

and label the values of x for which r is defined so that

$$r(x_{r_1}) \geq r(x_{r_2}) \geq \cdots .$$

Let N be the set of x for which $p_0(x) = 0$ and $p_1(x) > 0$. Assume that there exists a j such that

$$\sum_{i=1}^{j} p_0(x_{r_i}) = \alpha.$$

If S is the set of x for which we reject the test, we see that the significance level is

$$\sum_{x \in S} p_0(x).$$

and the power over the region of the alternative hypothesis is

$$\sum_{x \in S} p_1(x).$$

Then it is clear that if $C = \{x_{r_1}, \ldots, x_{r_j}\} \cup N$, then $\sum_{x \in S} p_1(x)$ is maximized over all sets C subject to the restriction on the size of the test.

If there does not exist a j such that $\sum_{i=1}^{j} p_0(x_{r_i}) = \alpha$, the rule is to put $x_{r_1}, \ldots, x_{r_j}$ in C so long as

$$\sum_{i=1}^{j} p_0(x_{r_i}) = \alpha^* < \alpha.$$

We then define a randomized auxiliary test R

$$\begin{aligned} \Pr(R = d_1) &= \delta_R(x_{r_{j+1}}) \\ &= (\alpha - \alpha^*)/p_0(x_{r_{j+1}}) \end{aligned}$$

It is clear in this way that $\sum_{x \in S} p_1(x)$ is maximized subject to the restriction on the size of the test.

Example 7.3 Testing between two discrete distributions
Consider two distributions with support on a subset of $\{0, 1, 2, 3, 4, 5\}$. Let $p_0(x)$ and $p_1(x)$ be the probability mass functions. Based on one observation, we want to test $H_0 : p_0(x)$ is the mass function versus $H_1 : p_1(x)$ is the mass function.

Suppose the distributions are as shown in Table 7.1, where we also show the values of r and the labels on x determined by r.

Table 7.1. Two Probability Distributions

x	0	1	2	3	4	5
p_0	.05	.10	.15	0	.50	.20
p_1	.15	.40	.30	.05	.05	.05
r	3	4	2	-	1/10	2/5
label	2	1	3	-	5	4

Thus, for example, we see $x_{r_1} = 1$ and $x_{r_2} = 0$. Also, $N = \{3\}$.

For given α, we choose C such that

$$\sum_{x \in C} p_0(x) \leq \alpha$$

and so as to maximize

$$\sum_{x \in C} p_1(x).$$

We find the optimal C by first ordering $r(x_{i_1}) \geq r(x_{i_2}) \geq \cdots$ and then satisfying $\sum_{x \in C} p_0(x) \leq \alpha$. The ordered possibilities for C in this example are

$$\{1\} \cup \{3\}, \quad \{1, 0\} \cup \{3\}, \quad \{1, 0, 2\} \cup \{3\}, \quad \cdots .$$

Notice that including N in the critical region does not cost us anything (in terms of the type I error that we are controlling).

Now, for any given significance level, we can determine the optimum test based on one observation.

- Suppose $\alpha = .10$. Then the optimal critical region is $C = \{1, 3\}$, and the power for the null hypothesis is $\beta_\delta(p_1) = .45$.
- Suppose $\alpha = .15$. Then the optimal critical region is $C = \{0, 1, 3\}$, and the power for the null hypothesis is $\beta_\delta(p_1) = .60$.
- Suppose $\alpha = .05$. We cannot put 1 in C, with probability 1, but if we put 1 in C with probability 0.5, the α level is satisfied, and the power for the null hypothesis is $\beta_\delta(p_1) = .25$.
- Suppose $\alpha = .20$. We choose $C = \{0, 1, 3\}$ with probability $2/3$ and $C = \{0, 1, 2, 3\}$ with probability $1/3$. The α level is satisfied, and the power for the null hypothesis is $\beta_\delta(p_1) = .75$.

All of these tests are most powerful based on a single observation for the given values of α. ∎

We can extend this idea to tests based on two observations. We see immediately that the ordered critical regions are

$$C_1 = \{1,3\} \times \{1,3\}, \quad C_1 \cup \{1,3\} \times \{0,3\}, \quad \cdots .$$

Extending this direct enumeration would be tedious, but, at this point we have grasped the implication: the ratio of the likelihoods is the basis for the most powerful test. This is the Neyman-Pearson Fundamental Lemma.

7.2.1 The Neyman-Pearson Fundamental Lemma

Example 7.3 illustrates the way we can approach the problem of testing any simple hypothesis against another simple hypothesis so long as we have a PDF. Notice the pivotal role played by ratio r in equation (7.11). This is a ratio of likelihoods.

Thinking of the hypotheses in terms of a parameter θ that indexes these two PDFs by θ_0 and θ_1, for a sample $X = x$, we have the likelihoods associated with the two hypotheses as $L(\theta_0; x)$ and $L(\theta_1; x)$. We may be able to define an α-level critical region for nonrandomized tests in terms of the ratio of these likelihoods: Let us assume that a positive number k exists such that there is a subset of the sample space C with complement with respect to the sample space C^c, such that

$$\begin{aligned} \frac{L(\theta_1; x)}{L(\theta_0; x)} &\geq k \quad \forall x \in C \\ \frac{L(\theta_1; x)}{L(\theta_0; x)} &\leq k \quad \forall x \in C^c \end{aligned} \tag{7.12}$$

and

$$\alpha = \Pr(X \in C \mid H_0).$$

(Notice that such a k and C may not exist.) For testing H_0 that the distribution of X is P_0 versus the alternative H_1 that the distribution of X is P_1, we can see that C is the best critical region of size α for testing H_0 versus H_1; that is, if A is any critical region of size α, then

$$\int_C L(\theta_1) - \int_A L(\theta_1) \geq 0 \tag{7.13}$$

(exercise).

The critical region defined in equation (7.12) illustrates the basic concepts, but it leaves some open questions. Following these ideas, the Neyman-Pearson Fundamental Lemma precisely identifies the most powerful test for a simple hypothesis versus another simple hypothesis and furthermore shows that the test is a.s. unique.

Theorem 7.1 (Neyman-Pearson Fundamental Lemma)
Let P_0 and P_1 be distributions with PDFs that are defined with respect to a common σ-finite measure. For given data X, let $L(P_0; X)$ and $L(P_1; X)$ be the respective likelihood functions.

To test $H_0 : P_0$ versus $H_1 : P_1$ at the level $\alpha \in]0, 1[$,

(i) there exists a test δ such that

$$\mathrm{E}_{P_0}(\delta(X)) = \alpha \tag{7.14}$$

and

$$\delta(X) = \begin{cases} 1 & L(P_1; X) > cL(P_0; X) \\ \gamma & L(P_1; X) = cL(P_0; X) \\ 0 & L(\theta_1; X) < cL(P_0; X) \end{cases}; \tag{7.15}$$

(ii) δ is most powerful test
(iii) if $\tilde{\delta}$ is a test that is as powerful as δ, then $\tilde{\delta}(X) = \delta(X)$ a.e. μ.

Proof. ∎

Example 7.4 Testing hypotheses about the parameter in a Bernoulli distribution
Suppose we assume a Bernoulli(π) distribution for the independently observed random variables $X_1, \ldots, X_n$, and we wish to test $H_0 : \pi = 1/4$ versus $H_1 : \pi = 3/4$ at the level $\alpha = 0.05$. If X is the number of 1s; that is, if $X = \sum_{i=1}^{n} X_i$, then X has a binomial distribution under either H_0 or H_1 and so the likelihoods in equation (7.15) are based on binomial PDFs. The optimal test $\delta(X)$, following the Neyman-Pearson setup, is based on the relationship of the ratio $L(P_1, x)/L(P_0, x)$ to a constant c as in equation (7.15) such that $\mathrm{E}_{P_0}(\delta(X)) = 0.05$.

Now, suppose $n = 30$ and the number of 1s observed is x. The ratio of the likelihoods is

$$\frac{L(P_1, x)}{L(P_0, x)} = 3^{2x-30}, \tag{7.16}$$

and larger values of x yield a value of 1 for $\delta(X)$. Since x can take on only the values $0, 1, \ldots, 30$, the ratio can take on only 31 different values. If the value of c is chosen so that the test rejects H_0 if $x \geq 12$, then

$$\mathrm{E}_{P_0}(\delta(X)) = \Pr(X \geq 12) = 0.05065828 = \alpha_-, \tag{7.17}$$

say, while if c is chosen so that the test rejects H_0 if $x \geq 13$, then

$$\mathrm{E}_{P_0}(\delta(X)) = 0.02159364 = \alpha_+. \tag{7.18}$$

From equation (7.16), we see that $c = 3^{-4}$ and from equations (7.17) and (7.18) we see that $\gamma = (\alpha - \alpha_-)/(\alpha_+ - \alpha_-) = 0.9773512$ in equation (7.15). ∎

Use of Sufficient Statistics

It is a useful fact that if there is a sufficient statistic $S(X)$ for θ, and $\tilde{\delta}(X)$ is an α-level test for an hypothesis specifying values of θ, then there exists an α-level test for the same hypothesis, $\delta(S)$ that depends only on $S(X)$, and which has power at least as great as that of $\tilde{\delta}(X)$. We see this by factoring the likelihoods.

Nuisance Parameters and Similar Regions

It is often the case that there are additional parameters not specified by the hypotheses being tested. In this situation we have $\theta = (\theta_s, \theta_u)$, and the hypothesis may be of the form $H_0 : \theta_s = \theta_{s0}$ or more generally, for the sample space,

$$H_0 : \Theta = \Theta_0,$$

where Θ_0 does not restrict some of the parameters. The hypothesis specifies the family of distributions as $\mathcal{P}_0 = \{P_\theta ; \theta \in \Theta_0\}$.

The problem is that the performance of the test, that is, $\mathrm{E}(\delta(X))$ may depend on the value of θ_u, even though we are not interested in θ_u. There is nothing we can do about this over the full parameter space, but since we think it is important to control the size of the test, we require, for given α,

$$\mathrm{E}_{H_0}(\delta(X)) = \alpha.$$

(Strictly speaking, we may only require that this expectation be bounded above by α.) Hence, we seek a procedure $\delta(X)$ such that $\mathrm{E}(\delta(X)) = \alpha$ over the subspace $\theta = (\theta_{s0}, \theta_u)$.

Is this possible? It certainly is if $\alpha = 1$; that is, if the rejection region is the entire sample space. Are there regions similar to the sample space in this regard? Maybe.

If a critical region R is such that $\Pr_{H_0}(X \in R) = \alpha$ for all values of θ, the region is called an α-level *similar region* with respect to $\theta = (\theta_{s0}, \theta_u)$, or with respect to H_0, or with respect to $\mathcal{P}_0$.

A test $\delta(X)$ such that $\mathrm{E}_\theta(\delta(X)) = \alpha$ for all $\theta \in H_0$ is called an α-level similar test with respect to $\theta = (\theta_{s0}, \theta_u)$.

Now, suppose S is a sufficient statistic for the family $\mathcal{P}_0 = \{P_\theta ; \theta \in \Theta_0\}$. Let $\delta(X)$ be a test that satisfies

$$\mathrm{E}(\delta(X)|S) = \alpha \quad \text{a.e. } \mathcal{P}_0.$$

In this case, we have

$$\begin{aligned} \mathrm{E}_{H_0}(\delta(X)) &= \mathrm{E}_{H_0}(\mathrm{E}(\delta(X)|S)) \\ &= \alpha; \end{aligned}$$

hence, the test is similar wrt $\mathcal{P}_0$. This condition on the critical region is called *Neyman structure*.

The concepts of similarity and Neyman structure are relevant for unbiased tests, which we will consider in Section 7.2.3.

Now suppose that U is boundedly complete sufficient for θ_u. If

$$\mathrm{E}(\mathrm{E}_{H_0}(\delta(X)|U)) = \alpha,$$

then the test has Neyman structure. While the power may still depend on θ_u, this fact may allow us to determine optimal tests of given size without regard to the nuisance parameters.

7.2.2 Uniformly Most Powerful Tests

The probability of a type I error is limited to α or less. We seek a procedure that yields the minimum probability of a type II error, given that bound on the probability of a type I error. This would be a "most powerful" test. Ideally, the test would be most powerful for all values of $\theta \in \Theta_1$. We call such a procedure a *uniformly most powerful* or UMP α-level test. For a given problem, finding such tests, or establishing that they do not exist, will be one of our primary objectives. The Neyman-Pearson Lemma gives us a way of determining whether a UMP test exists, and if so how to find one. The main issue is the likelihood ratio as a function of the parameter in the region specified by a composite H_1. If the likelihood ratio is monotone, then we have a UMP based on the ratio.

Generalizing the Optimal Test to Hypotheses of Intervals: UMP Tests

Although it applies to a simple alternative (and hence "uniform" properties do not make much sense), the Neyman-Pearson Lemma gives us a way of determining whether a *uniformly most powerful* (UMP) test exists, and if so how to find one. We are often interested in testing hypotheses in which either or both of Θ_0 and Θ_1 are convex regions of $\mathbb{R}$ (or $\mathbb{R}^k$).

We must look at the likelihood ratio as a function both of a scalar parameter θ and of a scalar function of x. The question is whether, for given θ_0 and any $\theta_1 > \theta_0$ (or equivalently any $\theta_1 < \theta_0$), the likelihood is monotone in some scalar function of x; that is, whether the family of distributions of interest is parameterized by a scalar in such a way that it has a *monotone likelihood ratio* (see page 167 and Exercise 2.5). In that case, it is clear that we can extend the test in (7.15) to be uniformly most powerful for testing $H_0 : \theta = \theta_0$ against an alternative $H_1 : \theta > \theta_0$ (or $\theta_1 < \theta_0$).

Example 7.5 Testing hypotheses about the parameter in a Bernoulli distribution (continuation of Example 7.4)

Syppose we assume a Bernoulli(π) distribution, and we wish to test $H_0 : \pi = 1/4$ versus $H_1 : \pi > 1/4$ at the level $\alpha = 0.05$. The alternative hypothesis is certainly more reasonable than the one in Example 7.4.

********** modify Example 7.4

****** also mention randomization *** don't do it

In this case, we do not have the problem of conflicting evidence mentioned on page 541. ∎

The exponential class of distributions is important because UMP tests are easy to find for families of distributions in that class. Discrete distributions are especially simple, but there is nothing special about them. Example 7.1 developed a test for $H_0 : \theta \leq \theta_0$ versus the alternative $H_1 : \theta > \theta_0$ in a one-parameter exponential distribution, that is clearly UMP, as we can see by using the formulation (7.15) in a pointwise fashion. (The one-parameter exponential distribution, with density over the positive reals $\theta^{-1}\mathrm{e}^{-x/\theta}$ is a member of the exponential class. Recall that the two-parameter exponential distribution used is not a member of the exponential family.)

Let us first identify some classes of hypotheses.

- simple versus simple

$$H_0 : \theta = \theta_0 \qquad \text{versus} \qquad H_1 : \theta = \theta_1. \tag{7.19}$$

- one-sided

$$H_0 : \theta \leq \theta_0 \qquad \text{versus} \qquad H_1 : \theta > \theta_0. \tag{7.20}$$

- two-sided; null on extremes

$$H_0 : \theta \leq \theta_1 \text{ or } \theta \geq \theta_2 \qquad \text{versus} \qquad H_1 : \theta_1 < \theta < \theta_2. \tag{7.21}$$

- two-sided; null in center

$$H_0 : \theta_1 \leq \theta \leq \theta_2 \qquad \text{versus} \qquad H_1 : \theta < \theta_1 \text{ or } \theta > \theta_2. \tag{7.22}$$

We can examine tests of these hypotheses in the context of the exponential family of Example 7.1.

Example 7.6 Testing in the exponential family

Suppose we have observations $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{exponential}(\theta)$. In Example 7.1, we developed a "reasonable" test of the hypothesis $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$. We will now use the test of equation (7.15) for these hypotheses, as well as the other hypotheses listed above.

We will consider a test at the α level for each type of hypothesis. We will develop a test based on the statistic $T(X) = \overline{X}$, which is sufficient for θ.

For the given observations, the likelihood is $L(\theta; X) = \theta^{-n}\mathrm{e}^{-\sum X_i/\theta}\mathrm{I}_{]0,\infty[}(\theta)$.

- First, we wish to test

$$H_0 : \theta = \theta_0 \qquad \text{versus} \qquad H_1 : \theta = \theta_1.$$

We must find a c_α such that

$$\begin{aligned}\Pr_{\theta_0}\left(\theta_1^{-n}\mathrm{e}^{-\sum X_i/\theta_1} > c_\alpha\theta_0^{-n}\mathrm{e}^{-\sum X_i/\theta_0}\right)\\ +\lambda\Pr_{\theta_0}\left(\theta_1^{-n}\mathrm{e}^{-\sum X_i/\theta_1} = c_\alpha\theta_0^{-n}\mathrm{e}^{-\sum X_i/\theta_0}\right) = \alpha.\end{aligned}$$

Because the second probability is 0, we have

$$\Pr_{\theta_0}\left(\theta_1^{-n}\mathrm{e}^{-n\overline{X}/\theta_1} > c_\alpha\theta_0^{-n}\mathrm{e}^{-n\overline{X}/\theta_0}\right) = \alpha;$$

that is,

$$\Pr_{\theta_0}\left(n\log(\theta_1) + n\overline{X}/\theta_1 < -\log(c_\alpha) + n\log(\theta_0) + n\overline{X}/\theta_0\right) = \alpha.$$

- We wish to test

$$H_0 : \theta \le \theta_0 \qquad \text{versus} \qquad H_1 : \theta > \theta_0.$$

These are the hypotheses of Example 7.1. A reasonable test may be to reject H_0 if $T(X) = \overline{X} > c_\alpha$, where c_α is some fixed positive constant; that is,

$$\delta(X) = \mathrm{I}_{]c_\alpha,\infty[}(T(X)).$$

We choose c_α so as to control the probability of a type I error.
Knowing the distribution of $\overline{X}$ to be gamma$(n, \theta/n)$, we can now work out

$$\Pr(\delta(X) = 1 \mid \theta) = \Pr(T(X) > c_\alpha \mid \theta),$$

which, for $\theta < \theta_0$ is the probability of a type I error. We set up the testing procedure so as to limit the probability of this type of error to be no more than α.
For $\theta \ge \theta_0$

$$1 - \Pr(\delta(X) = 1 \mid \theta)$$

is the probability of a type II error.
Over the full range of θ, we identify the power of the test as

$$\beta(\theta) = \Pr(\delta(X) = 1 \mid \theta).$$

- We wish to test

$$H_0 : \theta \le \theta_1 \text{ or } \theta \ge \theta_2 \qquad \text{versus} \qquad H_1 : \theta_1 < \theta < \theta_2.$$

This we can do in the same manner as above.

- We wish to test

$$H_0 : \theta_1 \le \theta \le \theta_2 \qquad \text{versus} \qquad H_1 : \theta < \theta_1 \text{ or } \theta > \theta_2.$$

We cannot do this as above; in fact, there is no UMP test. We will explore our options in the next section.

∎

Nonexistence of UMP Tests

One of the most interesting cases in which a UMP test cannot exist is when the alternative hypothesis is two-sided as in hypothesis (7.22). Hypothesis (7.22) is essentially equivalent to the pair of hypotheses with a simple null:

$$H_0 : \theta = \theta_0 \qquad \text{versus} \qquad H_1 : \theta \neq \theta_0.$$

If $\Theta = \mathbb{R}^d$, it is easy to see that in most cases of practical interest no UMP test can exist for these hypotheses, and you should reason through this statement to see that it is true.

So what can we do?

This is similar to the problem in point estimation when we realized we could not have an estimator that would uniformly minimize the risk. In that case, we added a requirement of unbiasedness or invariance, or else we added some global property of the risk, such as minimum averaged risk or minimum maximum risk. We might introduce similar criteria for the testing problem.

First, let's consider a desirable property of tests that we will call *unbiasedness*.

7.2.3 Unbiasedness of Tests

Recall that there are a couple of standard definitions of unbiasedness.

- If a random variable X has a distribution with parameter θ, for a point estimator $T(X)$ of an estimand $g(\theta)$ to be unbiased means that

  $$\mathrm{E}_\theta(T(X)) = g(\theta).$$

  Although no loss function is specified in this meaning of unbiasedness, we know that such an estimator minimizes the risk based on a squared-error loss function. (This last statement is not iff. Under squared-error loss the conditions of minimum risk and unbiasedness defined in this way are equivalent if g is continuous and not constant over any open subset of the parameter space and if $\mathrm{E}_\theta(T(X))$ is a continuous function of θ.)
- Another definition of unbiasedness is given with direct reference to a loss function. This is sometimes called L-unbiasedness. The estimator (or more generally, the procedure) $T(X)$ is said to be L-unbiased under the loss function L, if for all θ and $\tilde{\theta}$,

  $$\mathrm{E}_\theta(L(\theta, T(X))) \leq \mathrm{E}_\theta(L(\tilde{\theta}, T(X))).$$

  Notice the subtle differences in this property and the property of an estimator that may result from an approach in which we seek a minimum-risk estimator; that is, an approach in which we seek to solve the minimization problem,

$$\min_T \mathrm{E}_\theta(L(\theta, T(X)))$$

for all θ. This latter problem does not have a solution. (Recall the approach is to add other restrictions on $T(X)$.)
L-unbiasedness under a squared-error also leads to the previous definition of unbiasedness.

Unbiasedness in hypothesis testing is the property that the test is more likely to reject the null hypothesis at any point in the parameter space specified by the alternative hypothesis than it is at any point in the parameter space specified by the null hypothesis.

Definition 7.1 (unbiased test)
The α-level test δ with power function $\beta_\delta(\theta) = \mathrm{E}_\theta(\delta(X))$ for the hypothesis $H_0 : \theta \in \Theta_{H_0}$ versus $H_1 : \theta \in \Theta_{H_1}$ is said to be *unbiased* if

$$\beta_\delta(\theta) \leq \alpha \quad \forall\, \theta \in \Theta_{H_0}$$

and

$$\beta_\delta(\theta) \geq \alpha \quad \forall\, \theta \in \Theta_{H_1}.$$

∎

Notice that this unbiasedness depends not only on the hypotheses, but also on the significance level.

This definition of unbiasedness for a test is L-unbiasedness if the loss function is 0-1.

In many cases of interest, the power function $\beta_\delta(\theta)$ is a continuous function of θ. In such cases, we may be particularly interested in the power on any common boundary point of Θ_{H_0} and Θ_{H_1}, that is,

$$B = \partial\Theta_{H_0} \cap \partial\Theta_{H_1}.$$

The condition of unbiasedness of Definition 7.1 implies that $\beta_\delta(\theta) = \alpha$ for any $\theta \in B$. We recognize this condition in terms of the similar regions that we have previously defined, and we immediately have

Theorem 7.2 *An unbiased test with continuous power function is similar on the boundary.*

7.2.4 UMP Unbiased (UMPU) Tests

We will be interested in UMP tests that are unbiased; that is, in UMPU tests.

We first note that if an α-level UMP test exists, it is unbiased, because its power is at least as great as the power of the constant test (for all x), $\delta(x) = \alpha$. Hence, any UMP test is automatically UMPU.

Unbiasedness becomes relevant when no UMP exists, such as when the alternative hypothesis is two-sided:

$$H_0 : \theta = \theta_0 \qquad \text{versus} \qquad H_1 : \theta \neq \theta_0.$$

Hence, we may restrict our attention to tests with the desirable property of unbiasedness.

In the following we consider the hypothesis $H_0 : \theta \in \Theta_{H_0}$ versus $H_1 : \theta \in \Theta_{H_1}$, and we seek a test that is UMP within the restricted class of unbiased tests.

We will also restrict our attention to hypotheses in which

$$B = \partial\Theta_{H_0} \cap \partial\Theta_{H_1} \neq \emptyset, \tag{7.23}$$

and to tests with power functions that are continuous in θ.

Theorem 7.3 *Let $\delta(X)$ be an α-level test of hypotheses satisfying (7.23) that is similar on B and that has continuous power function in $\theta \in \Theta_{H_0} \cup \Theta_{H_1}$. If $\delta_*(X)$ is uniformly most powerful among such tests, then $\delta_*(X)$ is a UMPU test.*

Proof. Because $\delta_*(X)$ is uniformly at least as powerful as $\delta(X) \equiv \alpha$, $\delta_*(X)$ is unbiased, and hence $\delta_*(X)$ is a UMPU test. ∎

Use of Theorem 7.3, when it applies, is one of the simplest ways of determining a UMPU test, or given a test, to show that it is UMPU. This theorem has immediate applications in tests of hypotheses in exponential families. Theorem 6.4 in MS2 summarizes those results.

Similar UMPU tests remain so in the presence of nuisance parameters. ********** more on Neyman structure, similarity

7.2.5 UMP Invariant (UMPI) Tests

We generally want statistical procedures to be invariant to various transformations of the problem. For example, if the observables X are transformed in some way, it should be possible to transform a "good" test for a certain hypothesis in some obvious way so that the test remains "good" using the transformed data. (This of course means that the hypothesis is also transformed.)

To address this issue more precisely, we consider transformation groups $\mathcal{G}$, $\overline{\mathcal{G}}$,**** fix notation and $\mathcal{G}^*$, defined and discussed beginning on page 280.

We are often able to define optimal tests under the restriction of invariance.

A test δ is said to be invariant under G, whose domain is the sample space $\mathcal{X}$, if for all $x \in \mathcal{X}$ and $g \in G$,

$$\delta(g(x)) = \delta(x). \tag{7.24}$$

(This is just the definition of an invariant function, equation (0.1.103).)

We seek most powerful invariant tests. (They are invariant because the accept/reject decision does not change.) Because of the meaning of "invariance"

in this context, the most powerful invariant test is uniformly most powerful (UMPI), just as we saw in the case of the equivariant minimum risk estimator. The procedure for finding UMPI (or just MPI) tests is similar to the procedure used in the estimation problem. For a given class of transformations, we first attempt to characterize the form of ϕ, and then to determine the most powerful test of that form. Because of the relationship of invariant functions to a maximal invariant function, we may base our procedure on a maximal invariant function.

As an example, consider the group G of translations, for $x = (x_1, \ldots, x_n)$:

$$g(x) = (x_1 + c, \ldots, x_n + c).$$

Just as before, we see that for $n > 1$, the set of differences

$$y_i = x_i - x_n \quad \text{for } i = 1, \ldots, n-1,$$

is invariant under G. This function is also maximal invariant. For x and $\tilde{x}$, let $y(x) = y(\tilde{x})$. So we have for $i = 1, \ldots, n-1$,

$$\begin{aligned} \tilde{x_i} - \tilde{x_n} &= x_i - x_n \\ &= (x_i + c) - (x_n + c) \\ &= g(x), \end{aligned}$$

and therefore the function is maximal invariant. Now, suppose we have the sample $X = (X_1, \ldots, X_n)$ and we wish to test the hypothesis that the density of X is $p_0(x_1 - \theta, \ldots, x_n - \theta)$ versus the alternative that it is $p_1(x_1 - \theta, \ldots, x_n - \theta)$. This testing problem is invariant under the group G of translations, with the induced group of transformations $\overline{G}$ of the parameter space (which are translations also). Notice that there is only one orbit of $\overline{G}$, the full parameter space. The most powerful invariant test will be based on $Y = (X_1 - X_n, \ldots, X_{n-1} - X_n)$. The density of Y under the null hypothesis is given by

$$\int p_0(y_1 + z, \ldots, y_{n-1} + z, z)\mathrm{d}z,$$

and the density of Y under the alternate hypothesis is similar. Because both densities are independent of θ, we have two simple hypotheses, and the Neyman-Pearson lemma gives us the UMP test among the class of invariant tests. The rejection criterion is

$$\frac{\int p_1(y_1 + u, \ldots, y_n + u)\mathrm{d}u}{\int p_0(y_1 + u, \ldots, y_n + u)\mathrm{d}u} > c,$$

for some c.

As we might expect, there are cases in which invariant procedures do not exist. For $n = 1$ there are no invariant functions under G in the translation example above. In such situations, obviously, we cannot seek UMP invariant tests.

7.2.6 Equivariance, Unbiasedness, and Admissibility

In some problems, the principles of invariance and unbiasedness are completely different; and in some cases, one may be relevant and the other totally irrelevant. In other cases there is a close connection between the two.

For the testing problem, the most interesting relationship between invariance and unbiasedness is that if a unique up to sets of measure zero UMPU test exists, and a UMPI test up to sets of measure zero exists, then the two tests are the same up to sets of measure zero:

Theorem 7.4
equivalence of UMPI and UMPU

Proof.
∎

Admissibility of a statistical procedure means that there is no procedure that is at least as "good" as the given procedure everywhere, and better than the given procedure somewhere. In the case of testing "good" means "powerful", and, of course, everything depends on the level of the test.

A UMPU test is admissible, but a UMPI test is not necessarily admissible.

7.2.7 Asymptotic Tests

We develop various asymptotic tests based on asymptotic distributions of tests and test statistics. For example, the asymptotic distribution of a maximum of a likelihood is a chi-squared and the ratio of two is asymptotically an F.

We assume a family of distributions $\mathcal{P}$, a sequence of statistics $\{\delta_n\}$ based on a random sample $X_1, \ldots, X_n$. In hypothesis testing, the standard setup is that we have an observable random variable with a distribution in the family $\mathcal{P}$. Our hypotheses concern a specific member $P \in \mathcal{P}$. We want to test

$$H_0 : P \in \mathcal{P}_0 \qquad \text{versus} \qquad H_1 : P \in \mathcal{P}_1,$$

where $\mathcal{P}_0 \subseteq \mathcal{P}$, $\mathcal{P}_1 \subseteq \mathcal{P}$, and $\mathcal{P}_0 \cap \mathcal{P}_1 = \emptyset$.

We consider a sequence of tests $\{\delta_n\}$, with power function $\beta(\delta_n, P)$.

For use of asymptotic approximations for hypothesis testing, we first need the concepts of asymptotic significance and limiting size, as discussed on page 314. These concepts apply to the asymptotic behavior of the test under the null hypothesis. We also must consider the consistency and uniform consistency, which concern the asymptotic behavior of the test under the alternative hypothesis. These properties ensure that the the propbability of a type II error goes to zero. We may also be interested in a different type of asymptotic behavior under the null hypothesis, in which we reqiure that the propbability of a type I error go to zero. The concept of Chernoff consistency is relevant here.

Definition 7.2 (Chernoff consistency)
The sequence of tests $\{\delta_n\}$ with power function $\beta(\delta(X_n), P)$ is *Chernoff-consistent* for the test iff δ_n is consistent and furthermore,

$$\lim_{n\to\infty} \beta(\delta(X_n), P) = 0 \,\forall P \in \mathcal{P}_0. \tag{7.25}$$

∎

7.3 Likelihood Ratio Tests, Wald Tests, and Score Tests

We see that the Neyman-Pearson Lemma leads directly to use of the ratio of the likelihoods in constructing tests. Now we want to generalize this approach and to study the properties of tests based on that ratio.

There are two types of tests that arise from likelihood ratio tests. These are called Wald tests and score tests. Score tests are also called Rao test or Lagrange multiplier tests.

The Wald tests and score tests are asymptotically equivalent. They are consistent under the Le Cam regularity conditions, and they are Chernoff-consistent if α is chosen so that as $n \to \infty$, $\alpha \to 0$ and $\chi^2_{r,\alpha_n} \in \mathrm{o}(n)$.

7.3.1 Likelihood Ratio Tests

Although as we have emphasized, the likelihood is a function of the distribution rather than of the random variable, we want to study its properties under the distribution of the random variable. Using the idea of the ratio as in the test (7.12) of $H_0 : \theta \in \Theta_0$, but inverting that ratio and including both hypotheses in the denominator, we define the *likelihood ratio* as

$$\lambda(X) = \frac{\sup_{\theta\in\Theta_0} L(\theta; X)}{\sup_{\theta\in\Theta} L(\theta; X)}. \tag{7.26}$$

The test, similarly to (7.12), rejects H_0 if $\lambda(X) \le c_\alpha$, where c_α is some value in $[0, 1]$. Tests such as this are called *likelihood ratio tests*. (We should note that there are other definitions of a likelihood ratio; in particular, in TSH3 its denominator is the sup over the alternative hypothesis. If the alternative hypothesis does not specify $\Theta - \Theta_0$, such a definition requires specification of both H_0, and H_1; whereas (7.26) requires specification only of H_0. Also, the direction of the inequality depends on the ratio; it may be inverted — compare the ratios in (7.12) and (7.26).)

The likelihood ratio may not exist, but if it is well defined, clearly it is in the interval $[0, 1]$, and values close to 1 provide evidence that the null hypothesis is true, and values close to 0 provide evidence that it is false.

If there is no c_α such that

$$\Pr(\lambda(X) \le c_\alpha | H_0) = \alpha,$$

then it is very unlikely that the likelihood ratio test is UMP. In such cases c_α is chosen so that $\Pr(\lambda(X) \leq c_\alpha | H_0) < \alpha$ and a randomization procedure is used to raise the probability of rejection.

Example 7.7 Likelihood ratio test in the exponential family (continuation of Example 7.1)

We have observations $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{exponential}(\theta)$. The likelihood is

$$L(\theta, ; x) = \theta^{-n} \mathrm{e}^{-n\bar{x}/\theta} \mathrm{I}_{]0,\infty[}(\theta).$$

Suppose as before, we wish to test

$$H_0 : \theta \leq \theta_0 \qquad \text{versus} \qquad H_1 : \theta > \theta_0.$$

From equation (7.26), we have

$$\begin{aligned} \lambda(X) &= \frac{\max_{0<\theta\leq\theta_0} L(\theta; X)}{\max_{0<\theta} L(\theta; X)} \\ &= ********** \end{aligned}$$

From the analysis in Example 7.1, we know that ********** ∎

Asymptotic Likelihood Ratio Tests

Some of the most important properties of LR tests are asymptotic ones.

There are various ways of using the likelihood to build practical tests. Some are asymptotic tests that use MLEs (or RLEs).

Regularity Conditions

The interesting asymptotic properties of LR tests depend on the Le Cam regularity conditions, which go slightly beyond the Fisher information regularity conditions. (See page 169.)

These are the conditions to ensure that superefficiency can only occur over a set of Lebesgue measure 0 (Theorem 5.5, page 422), the asymptotic efficiency of RLEs (Theorem 6.5, page 482), and the chi-squared asymptotic distribution of the likelihood ratio (Theorem 7.5 below).

Asymptotic Significance of LR Tests

We consider a general form of the null hypothesis,

$$H_0 : R(\theta) = 0 \tag{7.27}$$

versus the alternative

$$H_1 : R(\theta) \neq 0, \tag{7.28}$$

for a continuously differentiable function $R(\theta)$ from $\mathbb{R}^k$ to $\mathbb{R}^r$. (The notation of MS2, $H_0 : \theta = g(\vartheta)$ where ϑ is a $(k-r)$-vector, although slightly different, is equivalent.)

Theorem 7.5
assuming the Le Cam regularity conditions, says that under H_0,

$$-2\log(\lambda_n) \xrightarrow{\mathrm{d}} \chi^2_r,$$

where χ^2_r *is a random variable with a chi-squared distribution with* r *degrees of freedom and* r *is the number of elements in* $R(\theta)$. *(In the simple case,* r *is the number of equations in the null hypothesis.)*

Proof.
∎

This allows us to determine the asymptotic significance of an LR test. It is also the basis for constructing asymptotically correct confidence sets, as we discuss beginning on page 551.

7.3.2 Wald Tests

For the hypostheses (7.27) and (7.28), the *Wald test* uses the test statistic

$$W_n = \left(R(\hat{\theta})\right)^{\mathrm{T}} \left(\left(S(\hat{\theta})\right)^{\mathrm{T}} \left(I_n(\hat{\theta})\right)^{-1} S(\hat{\theta})\right)^{-1} R(\hat{\theta}), \tag{7.29}$$

where $S(\theta) = \partial R(\theta)/\partial\theta$ and $I_n(\theta)$ is the Fisher information matrix, and these two quantities are evaluated at an MLE or RLE $\hat{\theta}$. The test rejects the null hypothesis when this value is large.

Notice that for the simple hypothesis $H_0 : \theta = \theta_0$, $S(\theta) = 1$, and so this simplifies to

$$(\hat{\theta} - \theta_0)^{\mathrm{T}} I_n(\hat{\theta})(\hat{\theta} - \theta_0). \tag{7.30}$$

An asymptotic test can be constructed because $W_n \xrightarrow{\mathrm{d}} Y$, where $Y \sim \chi^2_r$ and r is the number of elements in $R(\theta)$. This is proved in Theorem 6.6 of MS2, page 434.

The test rejects at the α level if $W_n > \chi^2_{r,1-\alpha}$, where $\chi^2_{r,1-\alpha}$ is the $1-\alpha$ quantile of the chi-squared distribution with r degrees of freedom. (Note that MS2 denotes this quantity as $\chi^2_{r,\alpha}$.)

7.3.3 Score Tests

A related test is the Rao *score test*, sometimes called a *Lagrange multiplier test*. It is based on a MLE or RLE $\tilde{\theta}$ under the restriction that $R(\theta) = 0$ (whence the Lagrange multiplier), and rejects H_0 when the following is large:

$$R_n = (s_n(\tilde{\theta}))^{\mathrm{T}} \left(I_n(\tilde{\theta})\right)^{-1} s_n(\tilde{\theta}), \tag{7.31}$$

where $s_n(\theta) = \partial l_L(\theta)/\partial\theta$, and is called the score function (see page 244).

An asymptotic test can be constructed because $R_n \xrightarrow{\mathrm{d}} Y$, where $Y \sim \chi^2_r$ and r is the number of elements in $R(\theta)$. This is proved in Theorem 6.6 (ii) of MS2.

The test rejects at the α level if $R_n > \chi^2_{r,1-\alpha}$, where $\chi^2_{r,1-\alpha}$ is the $1-\alpha$ quantile of the chi-squared distribution with r degrees of freedom.

7.3.4 Examples

Example 7.8 tests in a binomial model
*****

$H_0 : \pi = \pi_0$ versus $H_0 : \pi \neq \pi_0$

Wald – uses estimated values

$$W_n = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}}$$

problems when $x = 0$

asymptotically valid of course, but not good for finite (especially for small) n

Score – uses hypothesized values as well as estimated values

$$R_n = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

no problems when $x = 0$

usually better than Wald for finite (especially for small) n ∎

Example 7.9 tests in a linear model
Consider a general regression model:

$$X_i = f(z_i, \beta) + \epsilon_i, \quad \text{where } \epsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2). \tag{7.32}$$

For given $k \times r$ matrix L, we want to test

$$H_0 : L\beta = \beta_0. \tag{7.33}$$

Let X be the sample (it's an n-vector). Let Z be the matrix whose rows are the z_i.

The log likelihood is

$$\log \ell(\beta; X) = c(\sigma^2) - \frac{1}{2\sigma^2}(X - f(Z,\beta))^{\mathrm{T}}(X - f(Z,\beta)).$$

The MLE is the LSE, $\hat{\beta}$.

Let $\tilde{\beta}$ be the maximizer of the log likelihood under the restriction $L\beta = \beta_0$.

The likelihood ratio is the same as the difference in the log likelihoods.

The maximum of the unrestricted log likelihood (minus a constant) is the minimum of the residuals:

$$\frac{1}{2\sigma^2}(X - f(Z,\hat{\beta}))^{\mathrm{T}}(X - f(Z,\hat{\beta})) = \frac{1}{2\sigma^2}\mathrm{SSE}(\hat{\beta})$$

and likewise, for the restricted:

$$\frac{1}{2\sigma^2}(X - f(Z,\tilde{\beta}))^{\mathrm{T}}(X - f(Z,\tilde{\beta})) = \frac{1}{2\sigma^2}\mathrm{SSE}(\tilde{\beta}).$$

Now, the difference,

$$\frac{\mathrm{SSE}(\hat{\beta}) - \mathrm{SSE}(\tilde{\beta})}{\sigma^2},$$

has an asymptotic $\chi^2(r)$ distribution. (Note that the 2 goes away.)

We also have that

$$\frac{\mathrm{SSE}(\hat{\beta})}{\sigma^2}$$

has an asymptotic $\chi^2(n-k)$ distribution.

So for the likelihood ratio test we get an F-type statistic:

$$\frac{(\mathrm{SSE}(\hat{\beta}) - \mathrm{SSE}(\tilde{\beta}))/r}{\mathrm{SSE}(\hat{\beta})/(n-k)}. \tag{7.34}$$

Use unrestricted MLE $\hat{\beta}$ and consider $L\hat{\beta} - \beta_0$.

$$\mathrm{V}(\hat{\beta}) \to \left(\mathrm{J}_{f(\hat{\beta})}^{\mathrm{T}}\mathrm{J}_{f(\hat{\beta})}\right)^{-1}\sigma^2,$$

and so

$$\mathrm{V}(L\hat{\beta}) \to L\left(\mathrm{J}_{f(\hat{\beta})}^{\mathrm{T}}\mathrm{J}_{f(\hat{\beta})}\right)^{-1}L^{\mathrm{T}}\sigma^2,$$

where $\mathrm{J}_{f(\hat{\beta})}$ is the $n \times k$ Jacobian matrix.

Hence, we can write an asymptotic $\chi^2(r)$ statistic as

$$(L\hat{\beta} - \beta_0)^{\mathrm{T}}\left(L\left(\mathrm{J}_{f(\hat{\beta})}^{\mathrm{T}}\mathrm{J}_{f(\hat{\beta})}\right)^{-1}L^{\mathrm{T}}s^2\right)^{-1}(L\hat{\beta} - \beta_0)$$

We can form a Wishart-type statistic from this.

If $r = 1$, L is just a vector (the linear combination), and we can take the square root and from a "pseudo t":

$$\frac{L^{\mathrm{T}}\hat{\beta} - \beta_0}{s\sqrt{L^{\mathrm{T}}\left(\mathrm{J}_{f(\hat{\beta})}^{\mathrm{T}}\mathrm{J}_{f(\hat{\beta})}\right)^{-1}L}}.$$

Get MLE with the restriction $L\beta = \beta_0$ using a Lagrange multiplier, λ of length r.

Minimize

$$\frac{1}{2\sigma^2}(X - f(Z,\beta))^{\mathrm{T}}(X - f(Z,\beta)) + \frac{1}{\sigma^2}(L\beta - \beta_0)^{\mathrm{T}}\lambda.$$

Differentiate and set $= 0$:

$$\begin{aligned} -\mathrm{J}_{f(\hat{\beta})}^{\mathrm{T}}(X - f(Z,\hat{\beta})) + L^{\mathrm{T}}\lambda &= 0 \\ L\hat{\beta} - \beta_0 &= 0. \end{aligned}$$

$\mathrm{J}_{f(\hat{\beta})}^{\mathrm{T}}(X - f(Z,\hat{\beta}))$ is called the *score* vector. It is of length k.

Now $\mathrm{V}(X - f(Z,\hat{\beta})) \to \sigma^2 I_n$, so the variance of the score vector, and hence, also of $L^{\mathrm{T}}\lambda$, goes to $\sigma^2 \mathrm{J}_{f(\beta)}^{\mathrm{T}}\mathrm{J}_{f(\beta)}$.
(Note this is the true β in this expression.)

Estimate the variance of the score vector with $\tilde{\sigma}^2 \mathrm{J}_{f(\tilde{\beta})}^{\mathrm{T}}\mathrm{J}_{f(\tilde{\beta})}$,

where $\tilde{\sigma}^2 = \mathrm{SSE}(\tilde{\beta})/(n-k+r)$.

Hence, we use $L^{\mathrm{T}}\tilde{\lambda}$ and its estimated variance.

Get

$$\frac{1}{\tilde{\sigma}^2}\tilde{\lambda}^{\mathrm{T}}L\left(\mathrm{J}_{f(\tilde{\beta})}^{\mathrm{T}}\mathrm{J}_{f(\tilde{\beta})}\right)^{-1}L^{\mathrm{T}}\tilde{\lambda} \tag{7.35}$$

It is asymptotically $\chi^2(r)$.

This is the Lagrange multiplier form.

Another form:

Use $\mathrm{J}_{f(\tilde{\beta})}^{\mathrm{T}}(X - f(Z,\tilde{\beta}))$ in place of $L^{\mathrm{T}}\tilde{\lambda}$.

Get

$$\frac{1}{\tilde{\sigma}^2}(X - f(Z,\tilde{\beta}))^{\mathrm{T}}\mathrm{J}_{f(\tilde{\beta})}\left(\mathrm{J}_{f(\tilde{\beta})}^{\mathrm{T}}\mathrm{J}_{f(\tilde{\beta})}\right)^{-1}\mathrm{J}_{f(\tilde{\beta})}^{\mathrm{T}}(X - f(Z,\tilde{\beta})) \tag{7.36}$$

This is the score form. Except for the method of computing it, it is the same as the Lagrange multiplier form.

This is the SSReg in the AOV for a regression model. ∎

Example 7.10 an anomalous score test

Morgan et al. (2007) illustrate some interesting issues using a simple example of counts of numbers of stillbirths in each of a sample of litters of laboratory animals. They suggest that a zero-inflated Poisson is an appropriate model. This distribution is an ω mixture of a point mass at 0 and a Poisson distribution. The CDF (in a notation we will use often later) is

$$P_{0,\omega}(x|\lambda) = (1-\omega)P(x|\lambda) + \omega \mathrm{I}_{[0,\infty[}(x),$$

where $P(x)$ is the Poisson CDF with parameter λ.

(Write the PDF (under the counting measure). Is this a reasonable probability model? What are the assumptions? Do the litter sizes matter?)

If we denote the number of litters in which the number of observed stillbirths is i by n_i, the log-likelihood function is

$$l(\omega,\lambda) = n_0 \log\left(\omega + (1-\omega)\mathrm{e}^{-\lambda}\right) + \sum_{i=1}^{\infty} n_i \log(1-\omega) - \sum_{i=1}^{\infty} n_i\lambda + \sum_{i=1}^{\infty} i n_i \log(\lambda) + c.$$

Suppose we want to test the null hypothesis that $\omega = 0$.

The score test has the form

$$s^{\mathrm{T}} J^{-1} s,$$

where s is the score vector and J is either the observed or the expected information matrix. For each we substitute $\omega = 0$ and $\lambda = \hat{\lambda}_0$, where $\hat{\lambda}_0 = \sum_{i=1}^{\infty} i n_i / n$ with $n = \sum_{i=0}^{\infty} n_i$, which is the MLE when $\omega = 0$.

Let

$$n_+ = \sum_{i=1}^{\infty} n_i$$

and

$$d = \sum_{i=0}^{\infty} i n_i.$$

The frequency of 0s is important. Let

$$f_0 = n_0/n.$$

Taking the derivatives and setting $\omega = 0$, we have

$$\frac{\partial l}{\partial \omega} = n_0 \mathrm{e}^{\lambda} - n,$$

$$\frac{\partial l}{\partial \lambda} = -n + d/\lambda,$$

$$\frac{\partial^2 l}{\partial \omega^2} = -n - n_0 \mathrm{e}^{2\lambda} + n_0 \mathrm{e}^{\lambda},$$

$$\frac{\partial^2 l}{\partial \omega \lambda} = n_0 \mathrm{e}^{\lambda},$$

and

$$\frac{\partial^2 l}{\partial \lambda^2} = -d/\lambda^2.$$

So, substituting the observed data and the restricted MLE, we have observed information matrix

$$O(0,\hat{\lambda}_0) = n \begin{bmatrix} 1 + f_0 \mathrm{e}^{2\hat{\lambda}_0} - 2 f_0 \mathrm{e}^{\hat{\lambda}_0} & -f_0 \mathrm{e}^{\hat{\lambda}_0} \\ -f_0 \mathrm{e}^{\hat{\lambda}_0} & 1/\hat{\lambda}_0 \end{bmatrix}.$$

Now, for the expected information matrix when $\omega = 0$, we first observe that $\mathrm{E}(n_0) = n\mathrm{e}^{-\lambda}$, $\mathrm{E}(d) = n\lambda$, and $\mathrm{E}(n_+) = n(1-\mathrm{e}^{-\lambda})$; hence

$$I(0, \hat{\lambda}_0) = n \begin{bmatrix} e^{\hat{\lambda}_0} - 1 & -1 \\ -1 & 1/\hat{\lambda}_0 \end{bmatrix}.$$

Hence, the score test statistic can be written as

$$\kappa(\hat{\lambda}_0)(n_0 e^{\hat{\lambda}_0} - n)^2,$$

where $\kappa(\hat{\lambda}_0)$ is the (1,1) element of the inverse of either $O(0, \hat{\lambda}_0)$ or $I(0, \hat{\lambda}_0)$.

Inverting the matrices (they are 2×2), we have as the test statistic for the score test, either

$$s_I = \frac{n e^{-\hat{\lambda}_0}(1-\theta)^2}{1 - e^{-\hat{\lambda}_0} - \hat{\lambda}_0 e^{-\hat{\lambda}_0}}$$

or

$$s_O = \frac{n e^{-\hat{\lambda}_0}(1-\theta)^2}{e^{-\hat{\lambda}_0} + \theta - 2\theta e^{-\hat{\lambda}_0} \theta^2 \hat{\lambda}_0 e^{-\hat{\lambda}_0}},$$

where $\theta = f_0 e^{\hat{\lambda}_0}$, which is the ratio of the observed proportion of 0 counts to the estimated probability of a zero count under the Poisson model. (If n_0 is actually the number expected under the Poisson model, then $\theta = 1$.)

Now consider the actual data reported by Morgan et al. (2007) for stillbirths in each litter of a sample of 402 litters of laboratory animals.

No. stillbirths	0	1	2	3	4	5	6	7	8	9	10	11
No. litters	314	48	20	7	5	2	2	1	2	0	0	1

For these data, we have $n = 402$, $d = 185$, $\hat{\lambda}_0 = 0.4602$, $e^{-\hat{\lambda}_0} = 0.6312$, and $\theta = 1.2376$.

What is interesting is the difference in s_I and s_O.

In this particular example, if all n_i for $i \geq 1$ are held constant at the observed values, but different values of n_0 are considered, as n_0 increases the ratio s_I/s_O increases from about 1/4 to 1 (when the n_0 is the expected number under the Poisson model; i.e., $\theta = 1$), and then decreases, actually becoming negative (around $n_0 = 100$).

This example illustrates an interesting case. The score test is inconsistent because the observed information generates negative variance estimates at the MLE under the null hypothesis. (The score test can also be inconsistent if the expected likelihood equation has spurious roots.) ∎

7.4 Nonparametric Tests

7.4.1 Permutation Tests

For $i = 1, 2$, given the random $X_{i1}, \ldots, X_{in_i}$ from a a distribution with continuous CDF F_i, we wish to test

$$H_0 : F_1 = F_2 \quad \text{versus} \quad H_0 : F_1 \neq F_2.$$

$\Pi(X)$, where $X = \{X_{ij} : i = 1, 2; j = 1, \ldots, n_i\}$

$\Pi(\{X_1, \ldots, X_k\})$, for any Π, where $\Pi(A)$ denotes a permutation of the elements of the set A.

****

7.4.2 Sign Tests and Rank Tests

We have a sample $X_1, \ldots, X_n$ and the associated ranks of the absolute values, $\widetilde{R}(X)$. We denote the subvector of $\widetilde{R}(X)$ that corresponds to positive values of X as $R_+(X)$, and let n_* be the number of positive values of X. We let $R_+^{\mathrm{o}}(X)$ be the vector of the elements of $R_+(X)$ in increasing order. We let J be a continuous and strictly increasing function on $[0, 1]$, and let

$$W(R_+^{\mathrm{o}}) = J(R_{+1}^{\mathrm{o}}/n) + \cdots + J(R_{+n_*}^{\mathrm{o}}/n) \tag{7.37}$$

7.4.3 Goodness of Fit Tests

Kolmogorov-Smirnov (KS) Test

If P_1 and P_2 are CDFs, the L_∞ norm of their difference is called the *Kolmogorov distance* between the two distributions; that is, the Kolmogorov distance between two CDFs P_1 and P_2, written as $\rho_{\mathrm{K}}(P_1, P_2)$, is defined as

$$\rho_{\mathrm{K}}(P_1, P_2) = \sup |P_1 - P_2|. \tag{7.38}$$

Because a CDF is bounded by 0 and 1, it is clear that

$$\rho_{\mathrm{K}}(P_1, P_2) \leq 1, \tag{7.39}$$

and if the supports $\mathcal{S}_1$ and $\mathcal{S}_2$ are such that $x_1 \in \mathcal{S}_1$ and $x_2 \in \mathcal{S}_2$ implies $x_1 \leq x_2$ with at least one $x_1 \in \mathcal{S}_1$ is less than some $x_2 \in \mathcal{S}_2$, then $\rho_{\mathrm{K}}(P_1, P_2) = 1$.

If one or both of P_1 and P_2 are ECDFs we can compute the Kolmogorov distance fairly easily using the order statistics.

*** distribution

Cramér von Mises Test

7.4.4 Empirical Likelihood Ratio Tests

7.5 Multiple Tests

In many applications, we test several hypotheses using only one set of observations. For example in the one-way fixed-effects AOV model

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad i = 1, \ldots, m, \quad j = 1, \ldots, n,$$

the usual hypothesis of interest is

$$H_0 : \alpha_1 = \cdots = \alpha_m = 0.$$

This can be thought of as several separate hypotheses, $H_{0_i} : \alpha_i = 0$; and the researcher is not just interested in whether all $\alpha_i = 0$. Which effects are larger, or which set of effects differ for other effects, and similar questions would be of interest.

In other types of applications, different populations may be tested in the same way and in some tests the null hypothesis is rejected and in others it is not. An example of such a situation is in seeking to identify human genes associated with a particular condition, say a certain type of illness. A common way of doing this is by use of DNA microarrays. The observations come from a set of subjects with the condition and from another set of subjects without the condition. A large number of genes from each subject are compared across the two groups. An hypothesis of "no difference" is to be tested for each gene. What is the probability that a gene will erroneously be identified as different across the two groups; that is, what is the probability that a type I error occurs in such a situation?

In such cases of multiple testing, the question of errors is not so straightforward. We begin by considering a single error within a *family* of tests. The family is any group of tests that we want to consider together. In the one-way model, the family would be the multiple comparisons among the effects. The probability of a type I error in any test in the family is called the *family wise error rate*, or FWER.

There are several ways to measure the errors. Letting m be the number of tests, F be the number of false positives, T be the number of true positives, and $S = F + T$ be the total number of "discoveries", or rejected null hypotheses, we define various error measures. The family wise error rate, as above, is

$$\text{FWER} = \Pr(F \geq 1).$$

The *per comparison error rate* is

$$\text{PCER} = \text{E}(F)/m.$$

The *false discovery rate* is

$$\text{FDR} = \text{E}(F/S).$$

The *false nondiscovery rate* is

$$\text{FNR} = \text{E}(T/S).$$

The Benjamini-Hochberg (BH) method for controlling FDR works as follows. First, order the m p-values from the tests: $p_1 \leq \cdots \leq p_m$. Then determine a threshold value for rejection by finding the largest integer j such that $p_j \leq j\alpha/m$. Finally, reject any hypothesis whose p-value is smaller than or equal to p_j. Benjamini and Hochberg (1995) prove that this procedure is guaranteed to force FDR $\leq \alpha$. Genovese and Wasserman (2002), however, showed that this procedure does not minimize FNR subject to FDR $\leq \alpha$ for a given α.

Example 7.11 Variable selection in a linear regression model
One of the most common instances of multiple hypotheses testing is in the variable selection problem in linear regression. ∎

7.6 Sequential Tests

In the simplest formulation of statistical hypothesis testing, corresponding to the setup of the Neyman-Pearson lemma, we test a given hypothesized distribution versus another given distribution. After setting some ground rules regarding the probability of falsely rejecting the null hypothesis, and then determining the optimal test in the case of simple hypotheses, we determined more general optimal tests in cases for which they exist, and for other cases, we determined optimal tests among classes of tests that had certain desirable properties. In some cases, the tests involved regions within the sample space in which the decision between the two hypotheses was made randomly; that is, based on a random process over and above the randomness of the distributions of interest.

Another logical approach to take when the data generated by the process of interest does not lead to a clear decision is to decide to take more observations. Recognizing at the outset that this is a possibility, we may decide to design the test as a sequential procedure. We take a small number of observations, and if the evidence is strong enough either to accept the null hypothesis or to reject it, the test is complete and we make the appropriate decision. On the other hand, if the evidence from the small sample is not strong enough, we take some additional observations and perform the test again. We repeat these steps as necessary.

Sequential procedures always present issues that may affect the inference. There are various kinds of sequential procedures. Example 7.12, which revisits an inference problem in a Bernoulli distribution that has been considered in Examples 3.12 and 6.1, sets up an experiment in which Bernoulli random variables are to be observed until a specified number of successes are observed. Although in some sense this is a sequential procedure (and it does raise questions about the principles underlying our statistical inference), the stopping rule is not dependent on the inference.

In a sequential testing procedure, at any point, the question of whether or not to continue observing random variables depends on the inference that could be made at that point. If the hypothesis can be rejected or if it is very unlikely that it can be rejected, the decision is made, and the test is terminated; otherwise, the test continues. When we refer to a "sequential test", this is the type of situation we have in mind.

7.6.1 Sequential Probability Ratio Tests

Let us again consider the test of a simple null hypothesis against a simple alternative. Thinking of the hypotheses in terms of a parameter θ that indexes these two PDFs by θ_0 and θ_1, for a sample of size n, we have the likelihoods associated with the two hypotheses as $L_n(\theta_0; x)$ and $L_n(\theta_1; x)$. The best test indicates that we should reject if

$$\frac{L_n(\theta_0; x)}{L_n(\theta_1; x)} \leq k,$$

for some appropriately chosen k.

define and show optimality

7.6.2 Sequential Reliability Tests

7.7 The Likelihood Principle and Tests of Hypotheses

*** introduce; refer to likelihood in N-P

Tests of Hypotheses that Depend on the Data-Generating Process

***

Example 7.12 Sampling in a Bernoulli distribution; p-values and the likelihood principle revisited
In Examples 3.12 and 6.1, we have considered the family of Bernoulli distributions that is formed from the class of the probability measures $P_\pi(\{1\}) = \pi$ and $P_\pi(\{0\}) = 1 - \pi$ on the measurable space $(\Omega = \{0, 1\}, \mathcal{F} = 2^\Omega)$. Suppose now we wish to test

$$H_0 : \pi \geq 0.5 \qquad \text{versus} \qquad H_1 : \pi < 0.5.$$

As we indicated in Example 3.12 there are two ways we could set up an experiment to make inferences on π. One approach is to take a random sample of size n, $X_1, \ldots, X_n$ from the Bernoulli(π), and then use some function of that sample as an estimator. An obvious statistic to use is the number of 1's in the sample, that is, $T = \sum X_i$. To assess the performance of an estimator

using T, we would first determine its distribution and then use the properties of that distribution to decide what would be a good estimator based on T.

A very different approach is to take a sequential sample, $X_1, X_2, \ldots$, until a fixed number t of 1's have occurred. This yields N, the number of trials until t 1's have occurred.

The distribution of T is binomial with parameters n and π; its PDF is

$$p_T(t\,;\,n,\pi) = \binom{n}{t}\pi^t(1-\pi)^{n-t}, \quad t = 0, 1, \ldots, n. \tag{7.40}$$

The distribution of N is the negative binomial with parameters t and π, and its PDF is

$$p_N(n\,;\,t,\pi) = \binom{n-1}{t-1}\pi^t(1-\pi)^{n-t}, \quad n = t, t+1, \ldots. \tag{7.41}$$

Suppose we do this both ways. We choose $n = 12$ for the first method and $t = 3$ for the second method. Now, suppose that for the first method, we observe $T = 3$ and for the second method, we observe $N = 12$. The ratio of the likelihoods satisfies equation (6.4); that is, it does not involve π, so by the likelihood principle, we should make the same conclusions about π.

Let us now compute the respective p-values for a one-sided test. For the binomial setup we get $p = 0.073$ (using the R function `pbinom(3,12,0.5)`), but for the negative binomial setup we get $p = 0.033$ (using the R function `1-pnbinom(8,3,0.5)` in which the first argument is the number of "failures" before the number of "successes" specified in the second argument). The p-values are different, and in fact, if we had decided to perform the test at the $\alpha = 0.05$ significance level, in one case we would reject the null hypothesis and in the other case we would not. ∎

Further comments on Example 7.12

The problem in a significance test as we have just described is determining what is "more extreme". In the sampling design that specifies a stopping rule based on n that leads to a binomial distribution, "more extreme" means $T \in \{0, \ldots, t\}$, and in a data-generating process that depends on a specified number of 1's, "more extreme" means $N \in \{n, n+1, \ldots\}$.

Notice the key element in the example: in one case, the experiment that was conducted to gather information is completely independent of what is observed, while in the other case the experiment itself depended on what is observed. The latter type of experiment is often a type of Markov process in which there is a stopping time as in equation (1.262).

This example illustrates a basic conflict between the likelihood principle and significance testing. Observance of the likelihood principle while making inferences about a probability distribution leads us to ignore the overall data-generating process. The likelihood principle states that only the observed data are relevant.

Monte Carlo simulation would be an appropriate way to study this situation. The data could be obtained by a process that involves a stopping rule, and the tests could be performed in a manner that ignores the process. Whether or not the test is valid could be assessed by evaluating the p-value under the null hypothesis.

You are asked to explore this issue in Exercise 7.5. ∎

Evidence Supporting Hypotheses

Example 7.4 is a good illustration of the Neyman-Pearson solution to a simple hypothesis testing problem. We might look at this problem in a slightly different context, however, as suggested on page 318. This particular example, in fact, is used in Royall (1997) to show how the data actually provide evidence that might contradict our decision based on the Neyman-Pearson hypothesis testing approach.

Example 7.13 Critique of the hypothesis test of Example 7.4
Let's consider the Bernoulli distribution of Example 7.4 and the two hypotheses regarding the parameter π, $H_0 : \pi = 1/4$ and $H_1 : \pi = 3/4$. The test is based on x, the total number of 1s in n trials. When $n = 30$, the Neyman-Pearson testing procedure at the level $\alpha = 0.05$ rejects H_0 in favor of H_1 if $x \geq 13$.

Looking at the problem of choosing H_0 or H_1 based on the evidence in the data, however, we might ask what evidence $x = 13$ provides. The likelihood ratio in equation 7.16 is

$$\frac{L(P_1, x)}{L(P_0, x)} = 1/81,$$

which would seem to be rather compelling evidence in favor of H_0 over H_1.

An additional problem with the test in Example 7.4 is the manner in which a decision is made if $x = 12$. In order to achieve an exact size of $\alpha = 0.05$, a randomization procedure that does not depend on the evidence of the data is required. This kind of randomization procedure does not seem to be a reasonable way to make a statistical decision. ∎

The alternative approach to hypothesis testing involves a comparison of the evidence from the data in favor of each of the competing hypotheses.

This approach is similar to the use of the Bayes factor discussed in Section 4.5.3.

7.8 Confidence Sets

For statistical confidence sets, the basic problem is to use a random sample X from an unknown distribution P to determine a random subfamily $A(X)$ of a given family of distributions $\mathcal{P}$ such that

$$\Pr_P(\mathcal{P}_S \ni P) \geq 1-\alpha \quad \forall P \in \mathcal{P}, \tag{7.42}$$

for some given $\alpha \in]0,1[$. The set $\mathcal{P}_S$ is called a $1-\alpha$ *confidence set* or *confidence set*. The "confidence level" is $1-\alpha$, so we sometimes call it a "level $1-\alpha$ confidence set". Notice that α is given a priori. We call

$$\inf_{P\in\mathcal{P}} \Pr_P(\mathcal{P}_S \ni P) \tag{7.43}$$

the *confidence coefficient* of $\mathcal{P}_S$.

If the confidence coefficient of $\mathcal{P}_S$ is $> 1-\alpha$, then $\mathcal{P}_S$ is said to be a *conservative* $1-\alpha$ confidence set.

We generally wish to determine a region with a given confidence coefficient, rather than with a given significance level.

If the distributions are characterized by a parameter θ in a given parameter space Θ an equivalent $1-\alpha$ confidence set for θ is a random subset Θ_S such that

$$\Pr_\theta(\Theta_S \ni \theta) \geq 1-\alpha \quad \forall \theta \in \Theta. \tag{7.44}$$

The basic paradigm of statistical confidence sets was described in Section 3.5.2, beginning on page 296. We first review some of those basic ideas, starting first with simple interval confidence sets. Then in Section 7.9 we discuss optimality of confidence sets.

As we have seen in other problems in statistical inference, it is often not possible to develop a procedure that is *uniformly* optimal. As with the estimation problem, we can impose restrictions, such as unbiasedness or equivariance.

We can define optimality in terms of a global averaging over the family of distributions of interest. If the the global averaging is considered to be a true probability distribution, then the resulting confidence intervals can be interpreted differently, and it can be said that the probability that the distribution of the observations is in some fixed family is some stated amount. The HPD Bayesian credible regions discussed in Section 4.6.2 can also be thought of as optimal sets that address similar applications in which confidence sets are used.

Because determining an exact $1-\alpha$ confidence set requires that we know the exact distribution of some statistic, we often have to form approximate confidence sets. There are three common ways that we do this as discussed in Section 3.1.4. In Section 7.10 we discuss asymptotic confidence sets, and in Section 7.11, bootstrap confidence sets.

Our usual notion of a confidence interval relies on a frequency approach to probability, and it leads to the definition of a $1-\alpha$ confidence interval for the (scalar) parameter θ as the random interval $[T_L, T_U]$, that has the property

$$\Pr(T_L \leq \theta \leq T_U) = 1-\alpha.$$

This is also called a $(1-\alpha)100\%$ confidence interval. The interval $[T_L, T_U]$ is not uniquely determined.

The concept extends easily to vector-valued parameters. A simple extension would be merely to let T_L and T_U, and let the confidence set be hyperrectangle defined by the cross products of the intervals. Rather than taking vectors T_L and T_U, however, we generally define other types of regions; in particular, we often take an ellipsoidal region whose shape is determined by the covariances of the estimators.

A realization of the random interval, say $[t_L, t_U]$, is also called a confidence interval. Although it may seem natural to state that the "probability that θ is in $[t_L, t_U]$ is $1 - \alpha$", this statement can be misleading unless a certain underlying probability structure is assumed.

In practice, the interval is usually specified with respect to an estimator of θ, $T(X)$. If we know the sampling distribution of $T - \theta$, we may determine c_1 and c_2 such that

$$\Pr(c_1 \leq T - \theta \leq c_2) = 1 - \alpha;$$

and hence

$$\Pr(T - c_2 \leq \theta \leq T - c_1) = 1 - \alpha.$$

If either T_L or T_U is infinite or corresponds to a bound on acceptable values of θ, the confidence interval is one-sided. For two-sided confidence intervals, we may seek to make the probability on either side of T to be equal. This is called an *equal-tail* confidence interval. We may, rather, choose to make $c_1 = -c_2$, and/or to minimize $|c_2 - c_1|$ or $|c_1|$ or $|c_2|$. This is similar in spirit to seeking an estimator with small variance.

Prediction Sets

We often want to identify a set in which a future observation on a random variable has a high probability of occurring. This kind of set is called a *prediction set*.

For example, we may assume a given sample $X_1, \ldots, X_n$ is from a $\mathrm{N}(\mu, \sigma^2)$ and we wish to determine a measurable set $C(X)$ such that for a future observation X_{n+1}

$$\inf_{P \in \mathcal{P}} \Pr_P(X_{n+1} \in C(X)) \geq 1 - \alpha.$$

More generally, instead of X_{n+1}, we could define a prediction interval for any random variable Y.

The difference in this and a confidence set for μ is that there is an additional source of variation. The prediction set will be larger, so as to account for this extra variation.

We may want to separate the statements about X_{n+1} or Y and $C(X)$. A *tolerance set* attempts to do this.

Given a sample X, a measurable set $S(X)$, and numbers δ and α in $]0, 1[$, if

$$\inf_{P \in \mathcal{P}} (\Pr_P(Y \in S(X) | X) \geq \delta) \geq 1 - \alpha,$$

then $S(X)$ is called a δ-tolerance set for Y with confidence level $1 - \alpha$.

Randomized confidence Sets

For discrete distributions, as we have seen, sometimes to achieve a test of a specified size, we had to use a randomized test.

Confidence sets may have exactly the same problem – and solution – in forming confidence sets for parameters in discrete distributions. We form *randomized confidence sets*. The idea is the same as in randomized tests, and we will discuss randomized confidence sets in the context of hypothesis tests below.

Pivotal Functions

A straightforward way to form a confidence interval is to use a function of the sample that also involves the parameter of interest, but that does not involve any nuisance parameters. This kind of function is called a pivotal function. The confidence interval is then formed by separating the parameter from the sample values, as in Example 3.24 on page 298.

Example 7.14 Confidence Interval for a Quantile
***distribution free ∎

For a given parameter and family of distributions there may be multiple pivotal values. For purposes of statistical inference, such considerations as unbiasedness and minimum variance may guide the choice of a pivotal value to use.

Approximate Pivot Values

It may not be possible to identify a pivotal quantity for a particular parameter. In that case, we may seek an approximate pivot. A function is asymptotically pivotal if a sequence of linear transformations of the function is pivotal in the limit as $n \to \infty$.

*** nuisance parameters ***** find consistent estimator

If the distribution of T is known, c_1 and c_2 can be determined. If the distribution of T is not known, some other approach must be used. A common method is to use some numerical approximation to the distribution. Another method is to use bootstrap samples from the ECDF.

Relation to Acceptance Regions of Hypothesis Tests

A test at the α level has a very close relationship with a $1-\alpha$ level confidence set.

When we test the hypothesis $H_0 : \theta \in \Theta_{H_0}$ at the α level, we form a critical region for a test statistic or rejection region for the values of the observable X. This region is such that the probability that the test statistic is in it is $\leq \alpha$.

For any given $\theta_0 \in \Theta$, consider the nonrandomized test T_{θ_0} for testing the simple hypothesis $H_0 : \theta = \theta_0$, against some alternative H_1. We let $A(\theta_0)$ be the set of all x such that the test statistic is not in the critical region; that is, $A(\theta_0)$ is the acceptance region.

Now, for any θ and any value x in the range of X, we let

$$C(x) = \{\theta \ : \ x \in A(\theta)\}.$$

For testing $H_0 : \theta = \theta_0$ at the α significance level, we have

$$\sup \Pr(X \notin A(\theta_0) \mid \theta = \theta_0) \leq \alpha;$$

that is,

$$1 - \alpha \leq \inf \Pr(X \in A(\theta_0) \mid \theta = \theta_0) = \inf \Pr(C(X) \ni \theta_0 \mid \theta = \theta_0).$$

This holds for any θ_0, so

$$\begin{aligned} \inf_{P \in \mathcal{P}} \Pr_P(C(X) \ni \theta) &= \inf_{\theta_0 \in \Theta} \inf \Pr_P(C(X) \ni \theta_0 \mid \theta = \theta_0) \\ &\geq 1 - \alpha. \end{aligned}$$

Hence, $C(X)$ is a $1 - \alpha$ level confidence set for θ.

If the size of the test is α, the inequalities are equalities, and so the confidence coefficient is $1 - \alpha$.

For example, suppose $Y_1, Y_2, \ldots, Y_n$ is a random sample from a $\mathrm{N}(\mu, \sigma^2)$ distribution, and $\overline{Y}$ is the sample mean.

To test $H_0 : \mu = \mu_0$, against the universal alternative, we form the test statistic

$$T(X) = \frac{\sqrt{n(n-1)}\,(\overline{Y} - \mu_0)}{\sqrt{\sum \left(Y_i - \overline{Y}\right)^2}}$$

which, under the null hypothesis, has a Student's t distribution with $n - 1$ degrees of freedom.

An acceptance region at the α level is

$$\left[t_{(\alpha/2)},\ t_{(1-\alpha/2)}\right],$$

and hence, putting these limits on $T(X)$ and inverting, we get

$$\left[\overline{Y} - t_{(1-\alpha/2)}\, s/\sqrt{n}, \quad \overline{Y} - t_{(\alpha/2)}\, s/\sqrt{n}\right],$$

which is a $1 - \alpha$ level confidence interval.

The test has size α and so the confidence coefficient is $1 - \alpha$.

Randomized Confidence Sets

To form a $1 - \alpha$ confidence level set, we form a nonrandomized confidence set (which may be null) with $1 - \alpha_1$ confidence level, with $0 \leq \alpha_1 \leq \alpha$, and then we define a random experiment with some event that has a probability of $\alpha - \alpha_1$.

7.9 Optimal Confidence Sets

Just as we refer to a test as being "valid" if the significance level of the test is not exceeded (that is, if the probability of rejecting a true null hypothesis is bounded by the level of significance), we refer to a a confidence interval that has a probability of at least the confidence level as being "correct".

We often evaluate a confidence set using a family of distributions that *does not include the true parameter*.

For example, "accuracy" refers to the (true) probability of the set including an incorrect value. A confidence that is more accurate has a smaller probability of including a distribution that did not give rise to the sample. This is a general way relates to the size of the confidence set.

Size of Confidence Sets

The "volume" (or "length") of a confidence set is the Lebesgue measure of the set:

$$\mathrm{vol}(\Theta_S) = \int_{\Theta_S} \mathrm{d}\tilde{\theta}.$$

This may not be finite.

If the volume is finite, we have (Theorem 7.6 in MS2)

$$\mathrm{E}_\theta(\mathrm{vol}(\Theta_S)) = \int_{\theta \neq \tilde{\theta}} \Pr_\theta(\Theta_S \ni \tilde{\theta}) \mathrm{d}\tilde{\theta}.$$

We see this by a simple application of Fubini's theorem to handle the integral over the product space, and then an interchange of integration:

Want to minimize volume (if appropriate; i.e., finite.)

7.9.1 Most Accurate Confidence Set

Accuracy of Confidence Sets

**********Want to maximize accuracy.???????????

Confidence sets can be thought of a a family of tests of hypotheses of the form $\theta \in H_0(\tilde{\theta})$ versus $\theta \in H_1(\tilde{\theta})$. A confidence set of size $1-\alpha$ is equivalent to a critical region $S(X)$ such that

$$\Pr\left(S(X) \ni \tilde{\theta}\right) \geq 1-\alpha \quad \forall\, \theta \in H_0\left(\tilde{\theta}\right).$$

The power of the related tests is just

$$\Pr\left(S(X) \ni \tilde{\theta}\right)$$

for any θ. In testing hypotheses, we are concerned about maximizing this for $\theta \in H_1(\tilde{\theta})$.

This is called the *accuracy* of the confidence set, and so in this terminology, we seek the *most accurate* confidence set, and, of course, the *uniformly most accurate* confidence region. Similarly to the case of UMP tests, the uniformly most accurate confidence region may or may not exist.

The question of existence of uniformly most accurate confidence intervals also depends on whether or not there are nuisance parameters. Just as with UMP tests, in the presence of nuisance parameters, usually uniformly most accurate confidence intervals do not exist. (We must add other restrictions on the intervals, as we see below.) The nonexistence of uniformly most accurate confidence sets can also be addressed by imposing unbiasedness.

Uniformly most accurate $1-\alpha$ level set:
$\Pr_\theta(\Theta_S \ni \tilde{\theta})$ is minimum among all $1-\alpha$ level sets and $\forall \tilde{\theta} \neq \theta$.

This definition of UMA may not be so relevant in the case of a one-sided confidence interval.

If $\widetilde{\Theta}$ is a subset of Θ that does not include θ, and

$$\Pr_\theta(\Theta_S \ni \tilde{\theta}) \leq \Pr_\theta(\Theta_{\widetilde{S}} \ni \tilde{\theta})$$

for any $1-\alpha$ level set $\Theta_{\widetilde{S}}$ and $\forall \tilde{\theta} \in \widetilde{\Theta}$, then Θ_S is said to be $\widetilde{\Theta}$-uniformly most accurate.

A confidence set formed by inverting a nonrandomized UMP test is UMA.

We see this easily from the definitions of UMP and UMA. (This is Theorem 7.4 in MS2.)

7.9.2 Unbiased Confidence Sets

With tests, sometimes no UMP exists, and hence we added a criterion, such as unbiasedness or invariance.

Likewise, sometimes we cannot form a UMA confidence interval, so we add some criterion.

We define unbiasedness in terms of a subset $\widetilde{\Theta}$ that does not include the true θ.

A $1-\alpha$ level confidence set $C(X)$ is said to be $\widetilde{\Theta}$-unbiased if

$$\Pr_\theta(\Theta_S \ni \tilde{\theta}) \leq 1-\alpha \quad \forall \tilde{\theta} \in \widetilde{\Theta}.$$

If $\widetilde{\Theta} = \{\theta\}^c$, we call the set *unbiased*.

A $\widetilde{\Theta}$-unbiased set that is uniformly more accurate ("more" is defined similarly to "most") than any other $\widetilde{\Theta}$-unbiased set is said to be a *uniformly most accurate unbiased* (UMAU) set. **********

The concept of unbiasedness in tests carries over immediately to confidence sets. A family of confidence sets of size $1-\alpha$ is said to be *unbiased* if

$$\Pr\left(S(X) \ni \tilde{\theta}\right) \leq 1-\alpha \quad \forall\, \theta \in H_1\left(\tilde{\theta}\right).$$

In the case of nuisance parameters θ_u, unbiasedness means that this holds for all values of the nuisance parameters. In this case, similar regions and Neyman structure also are relevant, just as in the case of testing.

Volume of a Confidence Set

If there are no nuisance parameters, the expected volume of a confidence set is usually known a priori, e.g., for μ in $\mathrm{N}(\mu, 1)$.

What about a confidence set for μ in $\mathrm{N}(\mu, \sigma^2)$, with σ^2 unknown?

The expected length is proportional to σ, and can be very long. (This is a consequence of the fact that two normal distributions $\mathrm{N}(\mu_1, \sigma^2)$ and $\mathrm{N}(\mu_2, \sigma^2)$ become indistinguishable as $\sigma \to \infty$.

The length of the confidence interval is inversely proportional to $\sqrt{n}$. How about a sequential procedure?

A Sequential Procedure for a Confidence Set

Let $X_1, X_2, \ldots \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$.

Fix n_0. Let $\bar{x}_0 = \sum_{i=1}^{n_0} x_i/n_0$ and $s_0^2 = \sum_{i=1}^{n_0}(x_i - \bar{x}_0)^2/(n_0 - 1)$.

Now, for measurable function of s, a and b, for $n \geq n_0$, let

$$a_1 = \cdots = a_{n_0} = a$$

and

$$a_{n_0+1} = \cdots = a_n = b.$$

Then

$$Y = \frac{\sum_{i=1}^{n} a_i(X_i - \mu)}{\sqrt{S_0^2 \sum_{i=1}^{n} a_i^2}}$$

has a Student's t distribution with $n_0 - 1$ df.

Controlling the Volume

Now compute s^2 from an initial sample of size n_0. Let c be a given positive constant. Now choose $n - n_0$ additional observations where

$$n = \max\left(n_0 + 1, \left[\frac{s^2}{c}\right] + 1\right).$$

Then there exists numbers $a_1, \ldots, a_n$ with $a_1 = \cdots = a_{n_0}$ and $a_{n_0+1} = \cdots = a_n$ such that $\sum_{i=1}^{n} a_i = 1$ and $\sum_{i=1}^{n} a_i^2 = c/s^2$.

And so (from above), $\sum_{i=1}^{n} a_i(X_i - \mu)/\sqrt{c}$ has a Student's t distribution with $n_0 - 1$ df.

Therefore, given $X_1, \ldots, X_{n_0}$ the expected length of the confidence interval can be controlled.

Example 7.15 Confidence Interval in Inverse Regression

∎

Consider $\mathrm{E}(Y|x) = \beta_0 + \beta_1 x$. Suppose we want to estimate the point at which $\beta_0 + \beta_1 x$ has a preassigned value; for example find dosage $x = -\beta_0/\beta_1$ at which $\mathrm{E}(Y|x) = 0$.

This is equivalent to finding the value $v = (x - \bar{x})/\sqrt{\sum(x_i - \bar{x})^2}$ at which

$$y = \gamma_0 + \gamma_1 v = 0.$$

So we want to find $v = -\gamma_0/\gamma_1$.

The most accurate unbiased confidence sets for $-\gamma_0/\gamma_1$ can be obtained from UMPU tests of the hypothesis $-\gamma_0/\gamma_1 = v_0$.

Acceptance regions of these tests are given by

$$\frac{|v_0 \sum v_i Y_i + \bar{Y}|/\sqrt{\frac{1}{n} + v_0^2}}{\sqrt{\left(\sum (Y_i - \bar{Y})^2 - (\sum v_i Y_i)^2\right)/(n-2)}} \leq c$$

where

$$\int_{-c}^{c} p(y)\mathrm{d}y = 1 - \alpha,$$

where p is the PDF of t with $n - 2$ df.

So square and get quadratic inequalities in v:

$$v^2 \left(c^2 s^2 - (\sum v_i Y_i)^2\right) - 2v\bar{Y} \sum v_i Y_i + \frac{1}{n}(c^2 x^2 - n\bar{Y}) \geq 0.$$

Now let $\underline{v}$ and $\overline{v}$ be the roots of the equation.

So the confidence statement becomes

$$\underline{v} \leq \frac{\gamma_0}{\gamma_1} \leq \overline{v} \quad \text{if} \quad \frac{|\sum v_i Y_i|}{s} > c,$$

$$\frac{\gamma_0}{\gamma_1} < \underline{v} \quad \text{or} \quad \frac{\gamma_0}{\gamma_1} > \overline{v} \quad \text{if} \quad \frac{|\sum v_i Y_i|}{s} < c,$$

and if $= c$, no solution.

If $y = \gamma_0 + \gamma_1 v$ is nearly parallel to the v-axis, then the intercept with the v-axis will be large in absolute value and its sign is sensitive to a small change in the angle.

Suppose in the quadratic that $n\bar{Y}^2 + (\sum v_i Y_i)^2 < c^2 s^2$, then there is no real solution.

For the confidence levels to remain valid, the confidence interval must be the whole real line.

7.9.3 Equivariant Confidence Sets

The connection we have seen between a $1 - \alpha$ confidence set $S(x)$, and the acceptance region of a α-level test, $A(\theta)$, that is

$$S(x) \ni \theta \quad \Leftrightarrow \quad x \in A(\theta),$$

can often be used to relate UMP invariant tests to best equivariant confidence sets.

Equivariance for confidence sets is defined similarly to equivariance in other settings.

Under the notation developed above, for the group of transformations G and the induced transformation groups G^* and $\overline{G}$, a confidence set $S(x)$ is *equivariant* if for all $x \in \mathcal{X}$ and $g \in G$,

$$g^*(S(x)) = S(g(x)).$$

The uniformly most powerful property of the test corresponds to uniformly minimizing the probability that the confidence set contains incorrect values, and the invariance corresponds to equivariance.

An equivariant set that is $\widetilde{\Theta}$-uniformly more accurate ("more" is defined similarly to "most") than any other equivariant set is said to be a *uniformly most accurate equivariant* (UMAE) set.

There are situations in which there do not exist confidence sets that have uniformly minimum probability of including incorrect values. In such cases, we may retain the requirement for equivariance, but impose some other criterion, such as expected smallest size (wrt Lebesgue measure) of the confidence interval.

7.10 Asymptotic Confidence sets

It is often difficult to determine sets with a specified confidence coefficient or significance level, or with other specified properties.

In such cases it may be useful to determine a set that "approximately" meets the specified requirements.

What does "approximately" mean?

- uses numerical approximations
- uses approximate distributions
- uses a random procedure
- uses asymptotics

We assume a random sample $X_1, \ldots, X_n$ from $P \in \mathcal{P}$

An *asymptotic significance level* of a confidence set $C(X)$ for $g(\theta)$ is $1-\alpha$ if

$$\liminf_n \Pr(C(X) \ni \theta) \geq 1-\alpha \quad \text{for any } P \in \mathcal{P}.$$

The *limiting confidence coefficient* of a confidence set $C(X)$ for θ is

$$\lim_n \inf_{P \in \mathcal{P}} \Pr(C(X) \ni \theta)$$

if it exists.

Example (MS2). Suppose $X_1, \ldots, X - n$ are iid from a distribution with CDF P_X and finite mean μ and variance σ^2. Suppose σ^2 is known, and we want to form a $1-\alpha$ level confidence interval for μ. Unless P_X is specified, we can only seek a confidence interval with asymptotic significance level $1-\alpha$. We have an asymptotic pivot $T(X,\mu) = (\overline{X} - \mu)/\sigma$, and $\sqrt{n}T$ has an asymptotic $\mathrm{N}(0,1)$ distribution. We then form an interval

$$\begin{aligned} C(X) &= (C_1(X),\ C_2(X)) \\ &= (\overline{X} - \sigma z_{1-\alpha/2}/\sqrt{:}n,\ \overline{X} + \sigma z_{1-\alpha/2}/\sqrt{:}n), \end{aligned}$$

where $z_{1-\alpha/2} = \Phi^{-1}(1-\alpha/2)$ and Φ is the $\mathrm{N}(0,1)$ CDF. Now consider $\Pr(\mu \in C(X))$. We have

Asymptotic Correctness and Accuracy

A confidence set $C(X)$ for θ is $1-\alpha$ *asymptotically correct* if

$$\lim_n \Pr(C(X) \ni \theta) = 1 - \alpha.$$

A confidence set $C(X)$ for θ is $1-\alpha$ l^{th}*-order asymptotically accurate* if it is $1-\alpha$ asymptotically correct and

$$\lim_n \Pr(C(X) \ni \theta) = 1 - \alpha + \mathrm{O}(n^{-l/2}).$$

Asymptotic Accuracy of Confidence sets

****

Constructing Asymptotic Confidence Sets

There are two straightforward ways of constructing good asymptotic confidence sets.

One is based on an *asymptotically pivotal function*, that is one whose limiting distribution does not depend on any parameters other than the one of the confidence set.

Another method is to invert the acceptance region of a test. The properties of the test carry over to the confidence set.

The likelihood yields statistics with good asymptotic properties (for testing or for confidence sets).

See Example 7.24.

Woodruff's interval

7.11 Bootstrap Confidence Sets

A method of forming a confidence interval for a parameter θ is to find a pivotal quantity that involves θ and a statistic T, $f(T,\theta)$, and then to rearrange the terms in a probability statement of the form

$$\Pr\left(f_{(\alpha/2)} \leq f(T,\theta) \leq f_{(1-\alpha/2)}\right) = 1-\alpha. \tag{7.45}$$

When distributions are difficult to work out, we may use bootstrap methods for estimating and/or approximating the percentiles, $f_{(\alpha/2)}$ and $f_{(1-\alpha/2)}$.

Basic Intervals

For computing confidence intervals for a mean, the pivotal quantity is likely to be of the form $T-\theta$.

The simplest application of the bootstrap to forming a confidence interval is to use the sampling distribution of T^*-T_0 as an approximation to the sampling distribution of $T-\theta$; that is, instead of using $f(T,\theta)$, we use $f(T^*,T_0)$, where T_0 is the value of T in the given sample.

The percentiles of the sampling distribution determine $f_{(\alpha/2)}$ and $f_{(1-\alpha/2)}$ in

$$\Pr\left(f_{(\alpha/2)} \leq f(T,\theta) \leq f_{(1-\alpha/2)}\right) = 1-\alpha.$$

Monte Carlo to Get Basic Intervals

If we cannot determine the sampling distribution of T^*-T, we can easily estimate it by Monte Carlo methods.

For the case $f(T,\theta) = T-\theta$, the probability statement above is equivalent to

$$\Pr\left(T-f_{(1-\alpha/2)} \leq \theta \leq T-f_{(\alpha/2)}\right) = 1-\alpha. \tag{7.46}$$

The $f_{(\pi)}$ may be estimated from the percentiles of a Monte Carlo sample of T^*-T_0.

Bootstrap-t Intervals

Methods of inference based on a normal distribution often work well even when the underlying distribution is not normal.

A useful approximate confidence interval for a location parameter can often be constructed using as a template the familiar confidence interval for the mean of a normal distribution,

$$\left[\overline{Y}-t_{(1-\alpha/2)}\,s/\sqrt{n},\quad \overline{Y}-t_{(\alpha/2)}\,s/\sqrt{n}\right],$$

where $t_{(\pi)}$ is a percentile from the Student's t distribution, and s^2 is the usual sample variance.

A confidence interval for any parameter constructed in this pattern is called a *bootstrap-t interval*. A bootstrap-t interval has the form

$$\left[T - \widehat{t}_{(1-\alpha/2)}\sqrt{\widehat{\mathrm{V}}(T)}, \quad T - \widehat{t}_{(\alpha/2)}\sqrt{\widehat{\mathrm{V}}(T)}\right]. \tag{7.47}$$

In the interval

$$\left[T - \widehat{t}_{(1-\alpha/2)}\sqrt{\widehat{\mathrm{V}}(T)}, \quad T - \widehat{t}_{(\alpha/2)}\sqrt{\widehat{\mathrm{V}}(T)}\right]$$

$\widehat{t}_{(\pi)}$ is the estimated percentile from the studentized statistic,

$$\frac{T^* - T_0}{\sqrt{\widehat{\mathrm{V}}(T^*)}}.$$

For many estimators T, no simple expression is available for $\widehat{\mathrm{V}}(T)$.

The variance could be estimated using a bootstrap. This bootstrap nested in the bootstrap to determine $\widehat{t}_{(\pi)}$ increases the computational burden multiplicatively.

If the underlying distribution is normal and T is a sample mean, the interval in expression (7.47) is an exact $(1-\alpha)100\%$ confidence interval of shortest length.

If the underlying distribution is not normal, however, this confidence interval may not have good properties. In particular, it may not even be of size $(1-\alpha)100\%$. An asymmetric underlying distribution can have particularly deleterious effects on one-sided confidence intervals.

If the estimators T and $\widehat{\mathrm{V}}(T)$ are based on sums of squares of deviations, the bootstrap-t interval performs very poorly when the underlying distribution has heavy tails. This is to be expected, of course. Bootstrap procedures can be no better than the statistics used.

Bootstrap Percentile Confidence Intervals

Given a random sample $(y_1, \ldots, y_n)$ from an unknown distribution with CDF P, we want an interval estimate of a parameter, $\theta = \Theta(P)$, for which we have a point estimator, T.

If T^* is a bootstrap estimator for θ based on the bootstrap sample $(y_1^*, \ldots, y_n^*)$, and if $G_{T^*}(t)$ is the distribution function for T^*, then the exact upper $1-\alpha$ confidence limit for θ is the value $t^*_{(1-\alpha)}$, such that $G_{T^*}(t^*_{(1-\alpha)}) = 1-\alpha$.

This is called the *percentile upper confidence limit*.

A lower limit is obtained similarly, and an interval is based on the lower and upper limits.

Monte Carlo for Bootstrap Percentile Confidence Intervals

In practice, we generally use Monte Carlo and m bootstrap samples to estimate these quantities.

The probability-symmetric bootstrap percentile confidence interval of size $(1-\alpha)100\%$ is thus

$$\left[t^*_{(\alpha/2)}, \quad t^*_{(1-\alpha/2)}\right],$$

where $t^*_{(\pi)}$ is the $[\pi m]^{\text{th}}$ order statistic of a sample of size m of T^*.

(Note that we are using T and t, and hence T^* and t^*, to represent estimators and estimates in general; that is, $t^*_{(\pi)}$ here does not refer to a percentile of the Student's t distribution.)

This percentile interval is based on the ideal bootstrap and may be estimated by Monte Carlo simulation.

Confidence Intervals Based on Transformations

Suppose that there is a monotonically increasing transformation g and a constant c such that the random variable

$$W = c(g(T^*) - g(\theta)) \tag{7.48}$$

has a symmetric distribution about zero. Here $g(\theta)$ is in the role of a mean and c is a scale or standard deviation.

Let H be the distribution function of W, so

$$G_{T^*}(t) = H\left(c(g(t) - g(\theta))\right) \tag{7.49}$$

and

$$t^*_{(1-\alpha/2)} = g^{-1}\left(g(t^*) + w_{(1-\alpha/2)}/c\right), \tag{7.50}$$

where $w_{(1-\alpha/2)}$ is the $(1-\alpha/2)$ quantile of W. The other quantile $t^*_{(\alpha/2)}$ would be determined analogously.

Instead of approximating the ideal interval with a Monte Carlo sample, we could use a transformation to a known W and compute the interval that way. Use of an exact transformation g to a known random variable W, of course, is just as difficult as evaluation of the ideal bootstrap interval. Nevertheless, we see that forming the ideal bootstrap confidence interval is equivalent to using the transformation g and the distribution function H.

Because transformations to approximate normality are well-understood and widely used, in practice, we generally choose g as a transformation to normality. The random variable W above is a standard normal random variable, Z. The relevant distribution function is Φ, the normal CDF. The normal approximations have a basis in the central limit property. Central limit approximations often have a bias of order $\mathrm{O}(n^{-1})$, however, so in small samples, the percentile intervals may not be very good.

Bias in Intervals Due to Bias in the Estimator

It is likely that the transformed statistic $g(T^*)$ in equation (7.48) is biased for the transformed θ, even if the untransformed statistic is unbiased for θ.

We can account for the possible bias by using the transformation

$$Z = c(g(T^*) - g(\theta)) + z_0,$$

and, analogous to equation (7.49), we have

$$G_{T^*}(t) = \Phi\left(c(g(t) - g(\theta)) + z_0\right).$$

The bias correction z_0 is $\Phi^{-1}\left(G_{T^*}(t)\right)$.

Bias in Intervals Due to Lack of Symmetry

Even when we are estimating θ directly with T^* (that is, g is the identity), another possible problem in determining percentiles for the confidence interval is the lack of symmetry of the distribution about z_0.

We would therefore need to make some adjustments in the quantiles instead of using equation (7.50) without some correction.

Correcting the Bias in Intervals

Rather than correcting the quantiles directly, we may adjust their levels.

For an interval of confidence $(1-\alpha)$, instead of $(t^*_{(\alpha/2)},\ t^*_{(1-\alpha/2)})$, we take

$$\left[t^*_{(\alpha_1)},\quad t^*_{(\alpha_2)}\right],$$

where the adjusted probabilities α_1 and α_2 are determined so as to reduce the bias and to allow for the lack of symmetry.

As we often do, even for a nonnormal underlying distribution, we relate α_1 and α_2 to percentiles of the normal distribution.

To allow for the lack of symmetry—that is, for a scale difference below and above z_0—we use quantiles about that point.

Efron (1987), who developed this method, introduced an "acceleration", a, and used the distance $a(z_0 + z_{(\pi)})$.

Using values for the bias correction and the acceleration determined from the data, Efron suggested the quantile adjustments

$$\alpha_1 = \Phi\left(\widehat{z}_0 + \frac{\widehat{z}_0 + z_{(\alpha/2)}}{1 - \widehat{a}(\widehat{z}_0 + z_{(\alpha/2)})}\right)$$

and

$$\alpha_2 = \Phi\left(\widehat{z}_0 + \frac{\widehat{z}_0 + z_{(1-\alpha/2)}}{1 - \widehat{a}(\widehat{z}_0 + z_{(1-\alpha/2)})}\right).$$

Use of these adjustments to the level of the quantiles for confidence intervals is called the bias-corrected and accelerated, or "BC_a", method.

This method automatically takes care of the problems of bias or asymmetry resulting from transformations that we discussed above.

Note that if $\widehat{a} = \widehat{z}_0 = 0$, then $\alpha_1 = \Phi(z_{(\alpha)})$ and $\alpha_2 = \Phi(z_{(1-\alpha)})$. In this case, the BC_a is the same as the ordinary percentile method.

The problem now is to estimate the bias correction z_0 and the acceleration a from the data.

Estimating the Correction

The bias-correction term z_0 is estimated by correcting the percentile near the median of the m bootstrap samples:

$$\widehat{z}_0 = \Phi^{-1}\left(\frac{1}{m}\sum_j \mathrm{I}_{]-\infty,T]}\left(T^{*j}\right)\right).$$

The idea is that we approximate the bias of the median (that is, the bias of a central quantile) and then adjust the other quantiles accordingly.

Estimating the Acceleration

Estimating a is a little more difficult. The way we proceed depends on the form the bias may take and how we choose to represent it.

Because one cause of bias may be skewness, Efron (1987) adjusted for the skewness of the distribution of the estimator in the neighborhood of θ.

The skewness is measured by a function of the second and third moments of T.

We can use the jackknife to estimate the second and third moments of T. The expression is

$$\widehat{a} = \frac{\sum\left(\mathrm{J}(T) - T_{(i)}\right)^3}{6\left(\sum\left(\mathrm{J}(T) - T_{(i)}\right)^2\right)^{3/2}}. \tag{7.51}$$

Bias resulting from other departures from normality, such as heavy tails, is not addressed by this adjustment.

There are R and S-Plus functions to compute BC_a confidence intervals.

Comparisons of Bootstrap-t and BC_a Intervals

It is difficult to make analytic comparisons between these two types of bootstrap confidence intervals.

In some Monte Carlo studies, it has been found that, for moderate and approximately equal sample sizes,the coverage of BC_a intervals is closest to

the nominal confidence level; however, for samples with very different sizes, the bootstrap-t intervals were better in the sense of coverage frequency.

Because of the variance of the components in the BC_a method, it generally requires relatively large numbers of bootstrap samples. For location parameters, for example, we may need $m = 1,000$.

Other Bootstrap Confidence Intervals

Another method for bootstrap confidence intervals is based on a delta method approximation for the standard deviation of the estimator.

This method yields *approximate bootstrap confidence*, or ABC, intervals.

Terms in the Taylor series expansions are used for computing $\widehat{a}$ and $\widehat{z}_0$ rather than using bootstrap estimates for these terms.

As with the BC_a method, bias resulting from other departures from normality, such as heavy tails, is not addressed.

There are R and S-Plus functions to compute ABC confidence intervals.

****

7.12 Simultaneous Confidence Sets

If $\theta = (\theta_1, \theta_2)$ a $1-\alpha$ level confidence set for θ is a region in $\mathbb{R}^2$, $C(X)$, such that $\Pr_\theta(C(X) \ni \theta) \geq 1-\alpha$.

Now consider the problem of separate intervals (or sets) in $\mathbb{R}^1$, $C_(X)$ and $C_2(X)$, such that $\Pr_\theta(C_1(X) \ni \theta_1 \quad \text{and} \quad C_2(X) \ni \theta_2) \geq 1-\alpha$.

These are called $1-\alpha$ simultaneous confidence intervals.

This is equivalent to $C(X) = C_1(X) \times C_2(X)$ in the case above. Or, in general $\times\, C_i(X)$.

(Of course, we may want to minimize expected area or some other geometric measures of $C(X)$.)

There are several methods. In linear models, many methods depend on contrasts, e.g., Scheffé's intervals or Tukey's intervals.

General conservative procedures depend depend on inequalities of probabilities of events.

7.12.1 Bonferroni's Confidence Intervals

A common conservative procedure called a Bonferroni method is based on the inequality

$$\Pr\left(\cup A_i\right) \leq \sum \Pr\left(A_i\right),$$

for any events $A_1, \ldots, A_k$. For each component of θ, θ_t, we choose α_t with $\sum \alpha_t = \alpha$, and we let $C_t(X)$ be a level $1-\alpha_t$ confidence interval. It is easy to see that these are of level $1-\alpha$ because

$$\begin{aligned}
\inf \Pr(C_t(X) \ni \theta_t \, \forall t) &= \Pr(\cap\{C_t(X) \ni \theta_t\}) \\
&= 1 - \Pr\left((\cap\{\theta_t \in C_t(X)\})^{\mathrm{c}}\right) \\
&= 1 - \Pr\left(\cup\{\theta_t \notin C_t(X)\}\right) \\
&\geq 1 - \sum \Pr\left(\{\theta_t \notin C_t(X)\}\right) \\
&\geq 1 - \sum \alpha_t \\
&= 1 - \alpha.
\end{aligned}$$

7.12.2 Scheffé's Confidence Intervals

7.12.3 Tukey's Confidence Intervals

Notes and Further Reading

Most of the material in this chapter is covered in MS2, Chapters 6 and 7, and in TSH3, Chapters 3, 4, and 5.

Foundations

p-values, Fisher; objective posterior probabilities of hypotheses, Jeffreys; testing with fixed error probabilities, Neyman.

Berger (2003)

Significance Tests and the Likelihood Principle

Example 7.12

Lindley and Phillips (1976), a Bayesian view

"Elementary statistics from an advanced standpoint".

suppose we have only the data from an experiment ...

Edwards (1992) example of measurements taken with defective or limited instruments (voltmeter)

Royall (1997) ... philosophy of science

Tests for Monotone Likelihood Ratio

Roosen and Hennessy (2004) tests for monotone likelihood ratio.

Types of Tests and Their Relationships to Each Other

Buse (1982) gives an interesting exposition of the three types of tests.

Verbeke and Molenberghs (2007) and Freedman (2007) discussed the example of Morgan et al. (Example 7.10), as well as other anomalies of a score test.

Sequential Tests

Wald (1945)

Because of the binary nature of statistical hypothesis testing, it is rather straightforward to add a third choice to obtain more data before making a decision. Although sequential tests seem natural, the idea of statistical inference that evolves through sequential sampling is not just limited to hypothesis testing. In any statistical procedure, some quantification of the uncertainty should be made. Based on the level of uncertainty, the statistician can choose to continue sampling. Wald (1947b) described sequential variations on the general decision-theoretic approach to statistical inference. See also Section 5.2 of Wald (1950).

Multiple Tests

Storey (2002) proposed use of the proportion of false positives for any hypothesis (feature) incurred, on average, when that feature defines the threshold value. The "q-value" can be calculated for each feature under investigation.

Storey (2003) Bayesian perspective

Exercises

7.1. The p-value is a random variable whose distribution depends on the test statistic and the state of nature. When the null hypothesis is true, it is often the case that the distribution of the p-value is $\mathrm{U}(0,1)$.
   a) State very clearly a set of conditions that ensures that under the null hypothesis, the distribution of the p-value is $\mathrm{U}(0,1)$. Given those conditions, prove that the distribution is $\mathrm{U}(0,1)$ under the null hypothesis.
   b) Give an example in which the distribution of the p-values is not uniform, even though the null hypothesis is true.

7.2. Prove expression (7.13).

7.3. In the statement of Theorem 7.1, we assume PDFs f_0 and f_1 both defined with respect to a common σ-finite measure μ. Does this limit the scope of the theorem; that is, might there be a situation in which we want to test between two distributions P_0 and P_1, yet there does not exist a σ-finite measure μ by which to define PDFs?

7.4. Consider a case of multiple testing in which the distribution of the p-values $p_1, \ldots, p_m$ of each of the m tests is $\mathrm{U}(0,1)$ under the null hypothesis. Now consider $\prod p_i$. Make a log transformation, and work out a chi-squared approximation that yields quantiles of the product.

7.5. Consider the data-generating process described in Example 7.12. In that process Bernoulli(π) results are observed until t 1's are observed, and the

number N of random Bernoulli variates is noted. Based on the observed value of N we wish to test the hypotheses

$$H_0 : \pi \geq 0.5 \qquad \text{versus} \qquad H_1 : \pi < 0.5.$$

using a test of size 0.05. A test with that exact size will require a random component.

a) Define such a test (based on the negative binomial distribution), and sketch its power curve.
b) Define a test for the same hypotheses and with the same size based on t being a realization of a binomial random variable with parameters N and π.
c) Now, suppose we repeat the experiment as described and we obtain observations $N_1, \ldots, N_m$. What is the mean size of m tests based on the binomial distributions as in Exercise 7.5b?

8

Nonparametric and Robust Inference

A major concern is how well the statistical model corresponds to the data-generating process. Analyses based on an inappropriate model are likely to yield misleading conclusions.

One approach is to develop procedures for inference based on a minimal set of assumptions about the underlying probability distribution. This leads to what we call *nonparametric inference*, and includes a wide range of procedures, such as the nonparametric tests discussed in Section 7.4. In this chapter we will discuss general methods in other areas of statistical inference.

Another approach is to consider the consequences on inferential procedures arising from differences in the model and the data-generating process. A major objective of the field of *robust statistics* is to identify or develop procedures that yield useful conclusions even when the data-generating process differs in certain ways from the statistical model. Such procedures are *robust* to departures within a certain class from the assumed model.

8.1 Nonparametric Inference

We have described statistical inference as the process of using observational data from a population that is in an assumed family of distributions $\mathcal{P}$ to identify a subfamily, $\mathcal{P}_H \subseteq \mathcal{P}$, that contains the population from which the data arose. If the assumed family of probability space is $(\Omega, \mathcal{F}, P_\theta)$ where the index on the probability measure is in some subset $\Theta \subseteq \mathbb{R}^d$ for some fixed positive integer d and θ fully determines the measure, we call θ the *parameter* and the statistical inference is *parametric inference*. Statistical inference is a process of identifying a sub parameter space $\Theta_H \subseteq \Theta$. For example, we may assume that a given sample $x_1, \ldots, x_n$ is taken independently from some member of a family of distributions

$$\mathcal{P} = \{\mathrm{N}(\mu, \sigma^2) \mid \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+\},$$

and statistical inference in this situation may lead us to place the population giving rise to the observed sample in the family

$$\mathcal{P}_H = \{\mathrm{N}(\mu, \sigma^2) \mid \mu \in [\mu_1, \mu_2], \sigma^2 \in \mathbb{R}_+\}$$

of distributions.

In a more general formulation of the inference problem, rather than identifying a parameter space, we may make only very general assumptions about the family of probability distributions. For example, we may assume only that they have Lebesgue PDFs, or that the PDFs are symmetric, or that all moments of order less than some integer k are finite. We may be interested in estimating, or performing tests of hypotheses about, certain distributional measures. In nonparametric inference, the distributional measures of interest are those that are likely to exist even for quite "pathological" distributions, such as the Cauchy distribution. The Cauchy distribution, for example does not have a mean or variance. It does, however, have a median and an interquartile range. Nonparametric inference often concerns such things as the median or interquartile range, rather than the mean or variance. The instantiation of the basic paradigm of statistical inference may be to assume a family of distributions

$$\mathcal{P} = \{\text{distributions with Lebesgue PDF}\},$$

and statistical inference leads us to place the population giving rise to the observed sample in the family

$$\mathcal{P}_H = \{\text{distributions with Lebesgue PDF with median greater than } 0\}$$

of distributions.

We can also express the statistical inference problem as beginning with a given family of distribution $\mathcal{P}$, and identify a subfamily based on values of some distributional measure expressed as a statistical function (see Section 1.1.9, beginning on page 51). The decision, for example, may be that the population at hand has a probability distribution in the family

$$\mathcal{P}_H = \{P \mid P \in \mathcal{P} \text{ and } \Upsilon(P) = \Upsilon_H\},$$

where Υ is a functional. In this case, the statistical inference has focused on the distributional measure $\Upsilon(P)$, which, of course, may be quite general.

The methods in nonparametric inference are often based on functionals of the ECDF. The strong convergence of the ECDF, as shown, for example, in the Glivenko-Cantelli Theorem 1.71 or by the Dvoretzky/Kiefer/Wolfowitz inequality (1.289), suggest that this approach will result in procedures with good asymptotic properties.

An example of nonparametric inference is the problem of testing that the distributions of two populations are the same versus the alternative that a

realization from one distribution is typically smaller (or larger) than a realization from the other distribution. A U-statistic involving two populations is The two-sample Wilcoxon statistic U (which happens to be a U-statistic) discussed in Example 5.22 could be used as a test statistic for this common problem in nonparametric inference. This U-statistic is an unbiased estimator of $\Pr(X_{11} \leq X_{21})$. If the distributions have similar shapes and differ primarily by a shift in location, U can be used as a test statistic for an hypothesis involving the medians of the two distributions instead of a two-sample t test for an hypothesis involving the means (and under the further assumptions that the distributions are normal with equal variances).

8.2 Inference Based on Order Statistics

8.2.1 Central Order Statistics

Asymptotic Properties

From equation (1.288) we have

$$\sqrt{n}(F_n(x) - F(x)) \xrightarrow{\mathrm{d}} \mathrm{N}\left(0, F(x)(1-F(x))\right).$$

Now, for a continuous CDF F with PDF f, consider a function $g(t)$ defined for $0 < t < 1$ by

$$g(t) = F^{-1}(t).$$

Then

$$g'(t) = \frac{1}{f(F^{-1}(t))},$$

and so, using the delta method, we have

$$\sqrt{n}(F^{-1}(F_n(x)) - x) \xrightarrow{\mathrm{d}} \mathrm{N}\left(0, \frac{F(x)(1-F(x))}{(f(x))^2}\right). \tag{8.1}$$

$F^{-1}(F_n(x))$ lies between the sample quantiles ***
$X_{(\lceil nF_n(x)\rceil)}$ fix notation ****

$$X_{(\lceil nF_n(x)\rceil)} - F^{-1}(F_n(x)) \xrightarrow{\text{a.s.}} 0$$

$$\sqrt{n}(X_{(\lceil nF_n(x)\rceil)} - x) \xrightarrow{\mathrm{d}} \mathrm{N}\left(0, \frac{F(x)(1-F(x))}{(f(x))^2}\right).$$

location family $F(x;\theta) = F(x-\theta;0)$ $F(0;0) = 1/2$ density $f(x;\theta)$ suppose $f(0;0) > 0$

$\widetilde{X}_n$ sample median

$$\sqrt{n}(\widetilde{X}_n - \theta) \xrightarrow{\mathrm{d}} \mathrm{N}\left(0, \frac{1}{4(f(0))^2}\right) \tag{8.2}$$

ARE median vs mean normal 0.637 t_3 1.62

8.2.2 Statistics of Extremes

*** tail index Because a PDF must integrate to 1 over the full range of real numbers, the PDF of a distribution whose range is infinite must approach 0 sufficiently fast in the tails, that is, as the argument of the PDF approaches $-\infty$ or ∞.

We will just consider the positive side of the distribution, even if the range is infinite on both sides. We can identify three general forms of the part of the PDF that determines the tail behavior (that is, the part of the PDF that contains the argument):

$$\mathrm{e}^{-|x|^\alpha} \quad \text{with } \alpha > 0, \tag{8.3}$$

as in the normal, exponential, and double exponential;

$$|x|^{-\alpha} \quad \text{with } \alpha > 1, \tag{8.4}$$

as in the Pareto (also the related "power distributions") and the Cauchy (with some additional terms); and

$$|x|^\alpha \mathrm{e}^{-|x|^\beta}, \tag{8.5}$$

as in the gamma and Weibull.

What happens as $x \to \infty$ determines whether the moments of the distribution are finite.

Consider first the form (8.3). We will call this an *exponential tail*. (We sometimes call it a "right exponential tail", because we are focusing only on the right side of the range.) Notice that for any $\alpha > 0$

$$E(|X|^k) = \int_0^\infty \gamma x^k \mathrm{e}^{-|x|^\alpha} \mathrm{d}x \; < \; \infty; \tag{8.6}$$

that is, all moments are finite.

We can see that some distributions with exponential tails have heavier tails than others. For example, the double exponential distribution has heavier tails than a normal distribution because

$$\mathrm{e}^{-|x|^2} \to 0$$

faster than

$$\mathrm{e}^{-|x|} \to 0.$$

Now consider the form (8.4). We will call this a *polynomial tail* or a *Pareto tail* because of the form of the Pareto PDF. We call α in expression (8.4) the *tail index* of the polynomial tail. (The tail index is also sometimes defined as this quantity minus 1.) The larger is the tail index the more rapidly the PDF will approach 0. The moments $E(|X|^k)$ will be finite only for $k < \alpha$. In the Pareto distribution, for example, the mean is finite only for $\alpha > 1$ and the variance is finite only for $\alpha > 2$.

***

8.3 Nonparametric Estimation of Functions

An interesting problem in statistics is estimation of a continuous function. In one common type of this problem, the function expresses a relationship between one set of variables that are the argument of the function and another set of variables that are the function value. In one special instance of this kind of problem, the argument of the is a single variable representing time. The function in such a case is called a time series model. Statistical inference for these kinds of problems is based on data of the form $(x_1, y_1), \ldots, (x_n, y_n)$, where x_i represents an observation on the argument of the function, and y_i represents an observation on the corresponding value of the function. Both x_i and y_i could be vectors. Often what is actually observed is x_i and y_i with some error, which may be assumed to be realizations of a random variable. In the case of a time series, x_i is a measure of time (and usually denoted as t_i).

In another important type of the problem of estimation of a continuous function, the function represents a probability density. Values of the argument of the function can be observed but the corresponding values of the function cannot be observed. A density cannot be observed directly. What is observed directly corresponds to the antiderivative of the function of interest. Figure 8.1 illustrates the difference in these two situations. The panel on the left shows observations that consist of pairs of values, the argument and the function value plus, perhaps, random noise. The panel on the right shows a "rug" of observations; there are no observed values that correspond to the probability density function being estimated.

In some common situations, the form of the function is assumed known and statistical inference involves the parameters that fully specify the function. This is the common situation in linear or nonlinear regression and in some problems in time series analysis. If the function is a parametric probability density function, the problem is the standard one of estimating the parameters.

In this chapter we consider the problem of nonparametric estimation of functions; that is, we do not assume that the form of the function is known. Our objective will not be to develop an expression for the function, but rather to develop a rule such that, given a value of the argument, an estimate of the value of the function is provided. This problem, whether nonparametric regression analysis or nonparametric density estimation, is generally difficult, and the statistical properties of an estimator of a function are more complicated than statistical properties of an estimator of a single parameter or even of a countable set of parameters.

The usual optimality properties that we use in developing a theory of estimation of a finite-dimensional parameter must be extended for estimation of a general function. As we will see, two of the usual desirable properties of point estimators, namely unbiasedness and maximum likelihood, cannot be attained in general by estimators of functions.

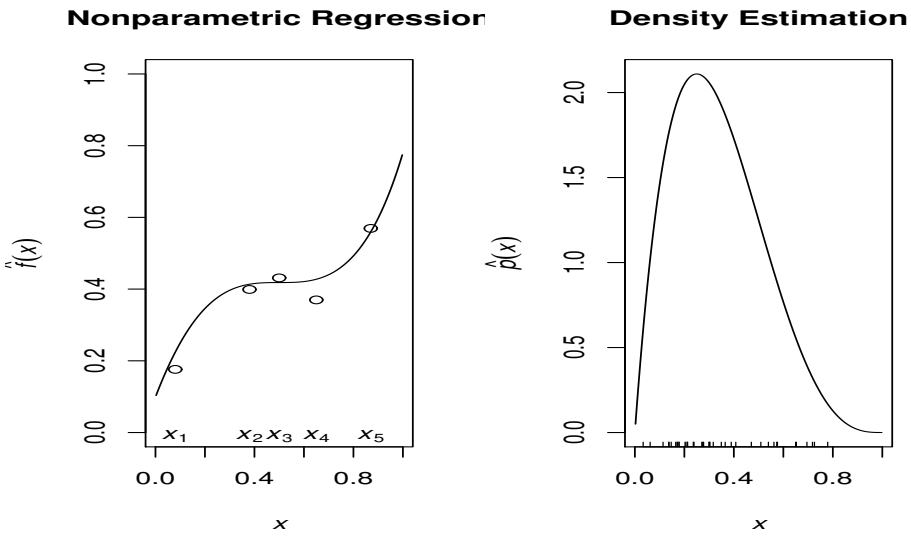


Figure 8.1. Nonparametric Function Estimation

Notation

We may denote a function by a single letter, f, for example, or by the function notation, $f(\cdot)$ or $f(x)$. When $f(x)$ denotes a function, x is merely a placeholder. The notation $f(x)$, however, may also refer to the value of the function at the point x. The meaning is usually clear from the context.

Using the common "hat" notation for an estimator, we use $\widehat{f}$ or $\widehat{f}(x)$ to denote the estimator of f or of $f(x)$. Following the usual terminology, we use the term "estimator" to denote a random variable, and "estimate" to denote a realization of the random variable. The hat notation is also used to denote an estimate, so we must determine from the context whether $\widehat{f}$ or $\widehat{f}(x)$ denotes a random variable or a realization of a random variable. The estimate or the estimator of the value of the function at the point x may also be denoted by $\widehat{f}(x)$. Sometimes, to emphasize that we are estimating the ordinate of the function rather than evaluating an estimate of the function, we use the notation $\widehat{f(x)}$. In this case also, we often make no distinction in the notation between the realization (the estimate) and the random variable (the estimator). We must determine from the context whether $\widehat{f}(x)$ or $\widehat{f(x)}$ denotes a random variable or a realization of a random variable. In most of the following discussion, the hat notation denotes a random variable that depends on the underlying random variable that yields the sample from which the estimator is computed.

Estimation or Approximation

There are many similarities in *estimation* of functions and *approximation* of functions, but we must be aware of the fundamental differences in the two problems. Estimation of functions is similar to other estimation problems: we are given a sample of observations; we make certain assumptions about the probability distribution of the sample; and then we develop estimators. The estimators are random variables, and how useful they are depends on properties of their distribution, such as their expected values and their variances. Approximation of functions is an important aspect of numerical analysis. Functions are often approximated to interpolate functional values between directly computed or known values. Functions are also approximated as a prelude to quadrature. Methods for estimating functions often use methods for approximating functions.

8.3.1 General Methods for Estimating Functions

In the problem of function estimation, we may have observations on the function at specific points in the domain, or we may have indirect measurements of the function, such as observations that relate to a derivative or an integral of the function. In either case, the problem of function estimation has the competing goals of providing a good fit to the observed data and predicting values at other points. In many cases, a smooth estimate satisfies this latter objective. In other cases, however, the unknown function itself is not smooth. Functions with different forms may govern the phenomena in different regimes. This presents a very difficult problem in function estimation, and it is one that we will not consider in any detail here.

There are various approaches to estimating functions. Maximum likelihood has limited usefulness for estimating functions because in general the likelihood is unbounded. A practical approach is to assume that the function is of a particular form and estimate the parameters that characterize the form. For example, we may assume that the function is exponential, possibly because of physical properties such as exponential decay. We may then use various estimation criteria, such as least squares, to estimate the parameter. An extension of this approach is to assume that the function is a mixture of other functions. The mixture can be formed by different functions over different domains or by weighted averages of the functions over the whole domain. Estimation of the function of interest involves estimation of various parameters as well as the weights.

Another approach to function estimation is to represent the function of interest as a linear combination of basis functions, that is, to represent the function in a series expansion. The basis functions are generally chosen to be orthogonal over the domain of interest, and the observed data are used to estimate the coefficients in the series.

It is often more practical to estimate the function value at a given point. (Of course, if we can estimate the function at any given point, we can effectively have an estimate at all points.) One way of forming an estimate of a function at a given point is to take the average at that point of a filtering function that is evaluated in the vicinity of each data point. The filtering function is called a kernel, and the result of this approach is called a kernel estimator.

In the estimation of functions, we must be concerned about the properties of the estimators at specific points and also about properties over the full domain. Global properties over the full domain are often defined in terms of integrals or in terms of suprema or infima.

Function Decomposition and Estimation of the Coefficients in an Orthogonal Expansion

We first do a PDF decomposition of the function of interest with the probability density function, p,

$$f(x) = g(x)p(x). \tag{8.7}$$

We have

$$\begin{aligned} c_k &= \langle f, q_k \rangle \\ &= \int_D q_k(x)g(x)p(x)\mathrm{d}x \\ &= \mathrm{E}(q_k(X)g(X)), \end{aligned} \tag{8.8}$$

where X is a random variable whose probability density function is p.

If we can obtain a random sample, $x_1, \ldots, x_n$, from the distribution with density p, the c_k can be unbiasedly estimated by

$$\widehat{c}_k = \frac{1}{n}\sum_{i=1}^{n} q_k(x_i)g(x_i).$$

The series estimator of the function for all x therefore is

$$\widehat{f}(x) = \frac{1}{n}\sum_{k=0}^{j}\sum_{i=1}^{n} q_k(x_i)g(x_i)q_k(x) \tag{8.9}$$

for some truncation point j.

The random sample, $x_1, \ldots, x_n$, may be an observed dataset, or it may be the output of a random number generator.

Kernel Methods

Another approach to function estimation and approximation is to use a *filter* or *kernel* function to provide local weighting of the observed data. This

approach ensures that at a given point the observations close to that point influence the estimate at the point more strongly than more distant observations. A standard method in this approach is to convolve the observations with a unimodal function that decreases rapidly away from a central point. This function is the filter or the kernel. A kernel has two arguments representing the two points in the convolution, but we typically use a single argument that represents the distance between the two points.

Some examples of univariate kernel functions are shown below.

$$\begin{array}{lll} \text{uniform:} & \kappa_{\mathrm{u}}(t) = 0.5, & \text{for } |t| \le 1, \\ \text{quadratic:} & \kappa_{\mathrm{q}}(t) = 0.75(1-t^2), & \text{for } |t| \le 1, \\ \text{normal:} & \kappa_{\mathrm{n}}(t) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-t^2/2}, & \text{for all } t. \end{array}$$

The kernels with finite support are defined to be 0 outside that range. Often multivariate kernels are formed as products of these or other univariate kernels.

In kernel methods, the locality of influence is controlled by a *window* around the point of interest. The choice of the size of the window is the most important issue in the use of kernel methods. In practice, for a given choice of the size of the window, the argument of the kernel function is transformed to reflect the size. The transformation is accomplished using a positive definite matrix, V, whose determinant measures the volume (size) of the window.

To estimate the function f at the point x, we first decompose f to have a factor that is a probability density function, p,

$$f(x) = g(x)p(x).$$

For a given set of data, $x_1, \ldots, x_n$, and a given scaling transformation matrix V, the kernel estimator of the function at the point x is

$$\widehat{f(x)} = (n|V|)^{-1} \sum_{i=1}^{n} g(x)\kappa\left(V^{-1}(x - x_i)\right). \tag{8.10}$$

In the univariate case, the size of the window is just the width h. The argument of the kernel is transformed to s/h, so the function that is convolved with the function of interest is $\kappa(s/h)/h$. The univariate kernel estimator is

$$\widehat{f(x)} = \frac{1}{nh} \sum_{i=1}^{n} g(x_i)\kappa\left(\frac{x - x_i}{h}\right).$$

8.3.2 Pointwise Properties of Function Estimators

The statistical properties of an estimator of a function at a given point are analogous to the usual statistical properties of an estimator of a scalar parameter. The statistical properties involve expectations or other properties of

random variables. In the following, when we write an expectation, $\mathrm{E}(\cdot)$, or a variance, $\mathrm{V}(\cdot)$, the expectations are usually taken with respect to the (unknown) distribution of the underlying random variable. Occasionally, we may explicitly indicate the distribution by writing, for example, $\mathrm{E}_p(\cdot)$, where p is the density of the random variable with respect to which the expectation is taken.

Bias

The bias of the estimator of a function value at the point x is

$$\mathrm{E}\big(\widehat{f}(x)\big) - f(x).$$

If this bias is zero, we would say that the estimator is unbiased at the point x. If the estimator is unbiased at every point x in the domain of f, we say that the estimator is pointwise unbiased. Obviously, in order for $\widehat{f}(\cdot)$ to be pointwise unbiased, it must be defined over the full domain of f.

Variance

The variance of the estimator at the point x is

$$\mathrm{V}\big(\widehat{f}(x)\big) = \mathrm{E}\left(\Big(\widehat{f}(x) - \mathrm{E}\big(\widehat{f}(x)\big)\Big)^2\right).$$

Estimators with small variance are generally more desirable, and an optimal estimator is often taken as the one with smallest variance among a class of unbiased estimators.

Mean Squared Error

The mean squared error, MSE, at the point x is

$$\mathrm{MSE}\big(\widehat{f}(x)\big) = \mathrm{E}\Big(\big(\widehat{f}(x) - f(x)\big)^2\Big). \tag{8.11}$$

The mean squared error is the sum of the variance and the square of the bias:

$$\begin{aligned}\mathrm{MSE}\big(\widehat{f}(x)\big) &= \mathrm{E}\Big(\big(\widehat{f}(x)\big)^2 - 2\widehat{f}(x)f(x) + \big(f(x)\big)^2\Big) \\ &= \mathrm{V}\big(\widehat{f}(x)\big) + \Big(\mathrm{E}\big(\widehat{f}(x)\big) - f(x)\Big)^2. \end{aligned} \tag{8.12}$$

Sometimes, the variance of an unbiased estimator is much greater than that of an estimator that is only slightly biased, so it is often appropriate to compare the mean squared error of the two estimators. In some cases, as we will see, unbiased estimators do not exist, so rather than seek an unbiased estimator with a small variance, we seek an estimator with a small MSE.

Mean Absolute Error

The mean absolute error, MAE, at the point x is similar to the MSE:

$$\mathrm{MAE}\big(\widehat{f}(x)\big) = \mathrm{E}\Big(\big|\widehat{f}(x) - f(x)\big|\Big). \tag{8.13}$$

It is more difficult to do mathematical analysis of the MAE than it is for the MSE. Furthermore, the MAE does not have a simple decomposition into other meaningful quantities similar to the MSE.

Consistency

Consistency of an estimator refers to the convergence of the expected value of the estimator to what is being estimated as the sample size increases without bound. A point estimator T_n, based on a sample of size n, is consistent for θ if

$$\mathrm{E}(T_n) \to \theta \quad \text{as } n \to \infty.$$

The convergence is stochastic, of course, so there are various types of convergence that can be required for consistency. The most common kind of convergence considered is weak convergence, or convergence in probability.

In addition to the type of stochastic convergence, we may consider the convergence of various measures of the estimator. In general, if m is a function (usually a vector-valued function that is an elementwise norm), we may define consistency of an estimator T_n in terms of m if

$$\mathrm{E}(m(T_n - \theta)) \to 0. \tag{8.14}$$

For an estimator, we are often interested in *weak convergence in mean square* or *weak convergence in quadratic mean*, so the common definition of consistency of T_n is

$$\mathrm{E}\big((T_n - \theta)^{\mathrm{T}}(T_n - \theta)\big) \to 0,$$

where the type of convergence is convergence in probability. Consistency defined by convergence in mean square is also called L_2 consistency. 2 consistence@L_2 consistency

If convergence does occur, we are interested in the rate of convergence. We define rate of convergence in terms of a function of n, say $r(n)$, such that

$$\mathrm{E}(m(T_n - \theta)) \in \mathrm{O}(r(n)).$$

A common form of $r(n)$ is n^{α}, where $\alpha < 0$. For example, in the simple case of a univariate population with a finite mean μ and finite second moment, use of the sample mean $\bar{x}$ as the estimator T_n, and use of $m(z) = z^2$, we have

$$\begin{aligned} \mathrm{E}(m(\bar{x} - \mu)) &= \mathrm{E}\big((\bar{x} - \mu)^2\big) \\ &= \mathrm{MSE}(\bar{x}) \\ &\in \mathrm{O}\big(n^{-1}\big). \end{aligned}$$

In the estimation of a function, we say that the estimator $\widehat{f}$ of the function f is *pointwise consistent* if

$$\mathrm{E}\big(\widehat{f}(x)\big) \to f(x) \tag{8.15}$$

for every x the domain of f. Just as in the estimation of a parameter, there are various kinds of pointwise consistency in the estimation of a function. If the convergence in expression (8.15) is in probability, for example, we say that the estimator is weakly pointwise consistent. We could also define other kinds of pointwise consistency in function estimation along the lines of other types of consistency.

8.3.3 Global Properties of Estimators of Functions

Often we are interested in some measure of the statistical properties of an estimator of a function over the full domain of the function. An obvious way of defining statistical properties of an estimator of an integrable function is to integrate the pointwise properties discussed in the previous section.

Global properties of a function or the difference between two functions are often defined in terms of a norm of the function or the difference.

For comparing $\widehat{f}(x)$ and $f(x)$, the L_p norm of the error is

$$\left(\int_D \left|\widehat{f}(x) - f(x)\right|^p \mathrm{d}x\right)^{1/p}, \tag{8.16}$$

where D is the domain of f. The integral may not exist, of course. Clearly, the estimator $\widehat{f}$ must also be defined over the same domain.

Three useful measures are the L_1 norm, also called the *integrated absolute error*, or IAE,

$$\mathrm{IAE}(\widehat{f}\,) = \int_D \left|\widehat{f}(x) - f(x)\right| \mathrm{d}x, \tag{8.17}$$

the square of the L_2 norm, also called the *integrated squared error*, or ISE,

$$\mathrm{ISE}(\widehat{f}\,) = \int_D \left(\widehat{f}(x) - f(x)\right)^2 \mathrm{d}x, \tag{8.18}$$

and the L_∞ norm, the *sup absolute error*, or SAE,

$$\mathrm{SAE}(\widehat{f}\,) = \sup \left|\widehat{f}(x) - f(x)\right|. \tag{8.19}$$

The L_1 measure is invariant under monotone transformations of the coordinate axes, but the measure based on the L_2 norm is not.

The L_∞ norm, or SAE, is the most often used measure in general function approximation. In statistical applications, this measure applied to two cumulative distribution functions is the *Kolmogorov distance*. The measure is not so useful in comparing densities and is not often used in density estimation.

Other useful measures of the difference in $\widehat{f}$ and f over the full range of x are the Kullback-Leibler measure and the Hellinger distance; see Section 0.1.9.

Integrated Bias and Variance

We now want to develop global concepts of bias and variance for estimators of functions. Bias and variance are statistical properties that involve expectations of random variables. The obvious global measures of bias and variance are just the pointwise measures integrated over the domain. In the case of the bias, of course, we must integrate the absolute value, otherwise points of negative bias could cancel out points of positive bias.

The estimator $\widehat{f}$ is pointwise unbiased if

$$\mathrm{E}\big(\widehat{f}(x)\big) = f(x) \quad \text{for all } x \in \mathbb{R}^d.$$

Because we are interested in the bias over the domain of the function, we define the *integrated absolute bias* as

$$\mathrm{IAB}\big(\widehat{f}\,\big) = \int_D \Big|\mathrm{E}\big(\widehat{f}(x)\big) - f(x)\Big| \,\mathrm{d}x \tag{8.20}$$

and the *integrated squared bias* as

$$\mathrm{ISB}\big(\widehat{f}\,\big) = \int_D \Big(\mathrm{E}\big(\widehat{f}(x)\big) - f(x)\Big)^2 \,\mathrm{d}x. \tag{8.21}$$

If the estimator is unbiased, both the integrated absolute bias and integrated squared bias are 0. This, of course, would mean that the estimator is pointwise unbiased almost everywhere. Although it is not uncommon to have unbiased estimators of scalar parameters or even of vector parameters with a countable number of elements, it is not likely that an estimator of a function could be unbiased at almost all points in a dense domain. ("Almost" means all except possibly a set with a probability measure of 0.)

The *integrated variance* is defined in a similar manner:

$$\begin{aligned} \mathrm{IV}\big(\widehat{f}\,\big) &= \int_D \mathrm{V}\big(\widehat{f}(x)\big) \,\mathrm{d}x \\ &= \int_D \mathrm{E}\Big(\big(\widehat{f}(x) - \mathrm{E}\big(\widehat{f}(x)\big)\big)^2\Big) \,\mathrm{d}x. \end{aligned} \tag{8.22}$$

Integrated Mean Squared Error and Mean Absolute Error

As we suggested above, global unbiasedness is generally not to be expected. An important measure for comparing estimators of funtions is, therefore, based on the mean squared error.

The *integrated mean squared error* is

$$\begin{aligned} \mathrm{IMSE}\big(\widehat{f}\,\big) &= \int_D \mathrm{E}\Big(\big(\widehat{f}(x) - f(x)\big)^2\Big) \,\mathrm{d}x \\ &= \mathrm{IV}\big(\widehat{f}\,\big) + \mathrm{ISB}\big(\widehat{f}\,\big) \end{aligned} \tag{8.23}$$

(compare equations (8.11) and (8.12)).

If the expectation integration can be interchanged with the outer integration in the expression above, we have

$$\begin{aligned}\mathrm{IMSE}(\widehat{f}) &= \mathrm{E}\left(\int_D \left(\widehat{f}(x) - f(x)\right)^2 \mathrm{d}x\right) \\ &= \mathrm{MISE}(\widehat{f}),\end{aligned}$$

the *mean integrated squared error*. We will assume that this interchange leaves the integrals unchanged, so we will use MISE and IMSE interchangeably.

Similarly, for the *integrated mean absolute error*, we have

$$\begin{aligned}\mathrm{IMAE}(\widehat{f}) &= \int_D \mathrm{E}\Big(\big|\widehat{f}(x) - f(x)\big|\Big)\,\mathrm{d}x \\ &= \mathrm{E}\left(\int_D \Big|\widehat{f}(x) - f(x)\Big|\,\mathrm{d}x\right) \\ &= \mathrm{MIAE}(\widehat{f}),\end{aligned}$$

the *mean integrated absolute error*.

Mean SAE

The *mean sup absolute error*, or MSAE, is

$$\mathrm{MSAE}(\widehat{f}) = \int_D \mathrm{E}\Big(\sup\big|\widehat{f}(x) - f(x)\big|\Big)\,\mathrm{d}x. \tag{8.24}$$

This measure is not very useful unless the variation in the function f is relatively small. For example, if f is a density function, $\widehat{f}$ can be a "good" estimator, yet the MSAE may be quite large. On the other hand, if f is a cumulative distribution function (monotonically ranging from 0 to 1), the MSAE may be a good measure of how well the estimator performs. As mentioned earlier, the SAE is the *Kolmogorov distance*. The Kolmogorov distance (and, hence, the SAE and the MSAE) does poorly in measuring differences in the tails of the distribution.

Large-Sample Statistical Properties

The pointwise consistency properties are extended to the full function in the obvious way. In the notation of expression (8.14), consistency of the function estimator is defined in terms of

$$\int_D \mathrm{E}\Big(m\big(\widehat{f}(x) - f(x)\big)\Big)\,\mathrm{d}x \to 0.$$

The estimator of the function is said to be *mean square consistent* or L_2 *consistent* if the MISE converges to 0; that is,

$$\int_D \mathrm{E}\Big(\big(\widehat{f}(x) - f(x)\big)^2\Big)\,\mathrm{d}x \;\to\; 0.$$

If the convergence is weak, that is, if it is convergence in probability, we say that the function estimator is weakly consistent; if the convergence is strong, that is, if it is convergence almost surely or with probability 1, we say the function estimator is strongly consistent.

The estimator of the function is said to be L_1 *consistent* if the mean integrated absolute error (MIAE) converges to 0; that is,

$$\int_D \mathrm{E}\Big(\Big|\widehat{f}(x) - f(x)\Big|\Big)\,\mathrm{d}x \;\to\; 0.$$

As with the other kinds of consistency, the nature of the convergence in the definition may be expressed in the qualifiers "weak" or "strong".

As we have mentioned above, the integrated absolute error is invariant under monotone transformations of the coordinate axes, but the L_2 measures are not. As with most work in L_1, however, derivation of various properties of IAE or MIAE is more difficult than for analogous properties with respect to L_2 criteria.

If the MISE converges to 0, we are interested in the rate of convergence. To determine this, we seek an expression of MISE as a function of n. We do this by a Taylor series expansion.

In general, if $\widehat{\theta}$ is an estimator of θ, the Taylor series for $\mathrm{ISE}(\widehat{\theta})$, equation (8.18), about the true value is

$$\mathrm{ISE}\big(\widehat{\theta}\big) = \sum_{k=0}^{\infty} \frac{1}{k!}\Big(\widehat{\theta} - \theta\Big)^k \mathrm{ISE}^{k'}(\theta), \tag{8.25}$$

where $\mathrm{ISE}^{k'}(\theta)$ represents the k^{th} derivative of ISE evaluated at θ.

Taking the expectation in equation (8.25) yields the MISE. The limit of the MISE as $n \to \infty$ is the *asymptotic mean integrated squared error*, AMISE. One of the most important properties of an estimator is the order of the AMISE.

In the case of an unbiased estimator, the first two terms in the Taylor series expansion are zero, and the AMISE is

$$\mathrm{V}(\widehat{\theta})\,\mathrm{ISE}''(\theta)$$

to terms of second order.

Other Global Properties of Estimators of Functions

There are often other properties that we would like an estimator of a function to possess. We may want the estimator to weight given functions in some particular way. For example, if we know how the function to be estimated,

f, weights a given function r, we may require that the estimate $\widehat{f}$ weight the function r in the same way; that is,

$$\int_D r(x)\widehat{f}(x)\mathrm{d}x = \int_D r(x)f(x)\mathrm{d}x.$$

We may want to restrict the minimum and maximum values of the estimator. For example, because many functions of interest are nonnegative, we may want to require that the estimator be nonnegative.

We may want to restrict the variation in the function. This can be thought of as the "roughness" of the function. A reasonable measure of the variation is

$$\int_D \left(f(x) - \int_D f(x)\mathrm{d}x\right)^2 \mathrm{d}x.$$

If the integral $\int_D f(x)\mathrm{d}x$ is constrained to be some constant (such as 1 in the case that $f(x)$ is a probability density), then the variation can be measured by the square of the L_2 norm,

$$\mathcal{S}(f) = \int_D \big(f(x)\big)^2\mathrm{d}x. \tag{8.26}$$

We may want to restrict the derivatives of the estimator or the smoothness of the estimator. Another intuitive measure of the roughness of a twice-differentiable and integrable univariate function f is the integral of the square of the second derivative:

$$\mathcal{R}(f) = \int_D \big(f''(x)\big)^2\mathrm{d}x. \tag{8.27}$$

Often, in function estimation, we may seek an estimator $\widehat{f}$ such that its roughness (by some definition) is small.

8.4 Semiparametric Methods and Partial Likelihood

In various contexts, we have considered estimation of probabilities of random variables being in specified intervals. This is estimation of a CDF evaluated at specified points. In this section, we will consider related problems of estimation of functional components of a CDF. These problems can be fully parametric or they can be *semiparametric*; that is, there are some "parameters", but the form of the PDF or CDF may not be fully specified.

We will focus on models of failure time data. The random variable of interest is the time to failure (from some arbitrary 0 time). We are interested in the distribution of the lifetimes of experimental units, for example, how long an electrical device will continue to operate. These problems may involve censoring, such as the setup in Example 6.3. A memoryless process, such as in

that example, is often unrealistic because the survival rate does not depend on the age of the experimental units. In other settings we may assume the rate does depend on the age, and so we may be interested in the conditional rate given that the units have survived for some given time. Alternatively, or additionally, we may assume that the survival rate depends on observable covariates. (Note that we will speak of "survival rate" sometimes, and "failure rate" sometimes.)

8.4.1 The Hazard Function

The hazard function measures the instantaneous rate of failure.

Definition 8.1 (hazard function)
Let F be a CDF with associated PDF f. The *hazard function* is defined at t, where $F(t) < 1$, as

$$\lambda(t) = f(t)/(1 - F(t)), \tag{8.28}$$

and the *cumulative hazard function* is

$$\Lambda(t) = \int_0^t \lambda(s)\mathrm{d}s. \tag{8.29}$$

■

In applications, the basic function is the survival function S instead of the CDF F. The survival function is the denominator in the hazard function; that is, $S(t) = 1 - F(t)$.

Note that if the CDF is absolutely continuous, the hazard function is the derivative of the log of the survival function, and we have

$$S(t) = \exp(-\Lambda(t)). \tag{8.30}$$

The common probability models for failure time data are the exponential, Weibull, log-normal, and gamma families. The hazard function generally is a function both of the time t, and of the parameters in the probability model. In the case of the exponential(θ) family, we have

$$\lambda(t) = \frac{1}{\theta}\mathrm{e}^{-t/\theta}/\mathrm{e}^{-t/\theta} = \frac{1}{\theta}.$$

In applications of with failure time data, the parameter is often taken to be $1/\theta$, and so the hazard rate is the parameter. In the exponential family the hazard function is a function only of the parameter in the probability model. It is constant with respect to time; this corresponds to the memoryless property. In other cases, the hazard rate may not be constant; see Exercise 8.3.

If $f(t)$ is interpreted as the instantaneous survival rate, then the hazard is the conditional survival rate, given survival to the point t.

Theorem 8.1
If λ is the hazard function associated with the random variable T, then

$$\lambda(t) = \lim_{\epsilon \downarrow 0} \epsilon^{-1} \Pr(t \leq T < t + \epsilon | T \geq t). \tag{8.31}$$

Proof. Exercise 8.5. ∎

In applications we often assume that the hazard function is affected by a p-vector of observable covariates x; that is, we have a function $\lambda(t, x, \theta, \beta)$, where I have written two sets of parameters, θ for those of the basic probability model (Weibull, for example), and β for parameters in a model of how the covariates x affect the hazard function. If the effects of the covariates are linear and additive, we may represent their overall effect by $\beta^{\mathrm{T}} x$, just as in the linear models discussed elsewhere in this book. It is unlikely that their effect is linear, but it is often the case that a function of the linear combination $\beta^{\mathrm{T}} x$, where β is a p-vector of unknown constants, seems to correspond well with observed data. In that case, we may write the conditional hazard function as $\lambda(t; \beta^{\mathrm{T}} x)$, suppressing other model parameters.

8.4.2 Proportional Hazards Models

In a very useful class of hazard functions, the effect of the covariates on the hazard function is multiplicative; that is, the function is of the form $\lambda(t, x, \beta) = \lambda_0(t)\phi(x, \beta)$, where $\lambda_0(t)$ is the "baseline" hazard function if there is no effect due to the covariates, and ϕ is some known function, and again we have suppressed other model parameters. Such models are called *proportional hazards models*.

Note that in a proportional hazards model, we can identify a baseline cumulative hazard function $\Lambda_0(t)$ based only on $\lambda_0(t)$, and furthermore, the survival function can be written as

$$1 - F(t) = \exp(\phi(x, \beta)\Lambda_0(t)). \tag{8.32}$$

This is an important property of proportional hazards models. The parameters for the effect of the covariates, that is, β, can be estimated by maximizing a partial likelihood without any consideration of the hazard function. The survival function is composed of a parametric component involving β and the component $\Lambda_0(t)$, which we may estimate without parameters.

*** estimate β using partial likelihood **** refer to Chapter 6.

*** estimate $\Lambda_0(t)$ nonparametrically. give both Breslow's estimator and Horowitz's estimator

If the effects of the covariates are linear and additive with respect to each other, we may represent the hazard function as $\lambda_0(t)\phi(\beta^{\mathrm{T}} x)$. A simple case may have the form $\lambda_0(t)(1 + \beta^{\mathrm{T}} x)$. This form, of course, would require some restriction on $\beta^{\mathrm{T}} x$, similar in some ways to the restriction that partially motivated the development of generalized linear models. Another form that does not require any restrictions is

$$\lambda(t;x) = \lambda_0(t)\mathrm{e}^{\beta^{\mathrm{T}}x}. \tag{8.33}$$

This model of the hazard function is called the Cox proportional hazards model.

8.5 Nonparametric Estimation of PDFs

********** Scott (1992) and Scott (2012)

There are obviously many connections between estimation of a CDF and the corresponding PDF. We have seen that the ECDF is a strongly consistent estimator of the CDF (Theorem 1.71), but it is not immediately obvious how to use the ECDF to construct an estimator of the PDF. Nevertheless, in many cases, an estimate of the PDF is more useful than an estimate of the CDF.

8.5.1 Nonparametric Probability Density Estimation

Estimation of a probability density function is similar to the estimation of any function, and the properties of the function estimators that we have discussed are relevant for density function estimators. A density function $p(y)$ is characterized by two properties:

- it is nonnegative everywhere;
- it integrates to 1 (with the appropriate definition of "integrate").

In this chapter, we consider several nonparametric estimators of a density; that is, estimators of a general nonnegative function that integrates to 1 and for which we make no assumptions about a functional form other than, perhaps, smoothness.

It seems reasonable that we require the density estimate to have the characteristic properties of a density:

- $\widehat{p}(y) \geq 0$ for all y;
- $\int_{\mathbb{R}^d} \widehat{p}(y)\,\mathrm{d}y = 1$.

A probability density estimator that is nonnegative and integrates to 1 is called a *bona fide* estimator.

Rosenblatt has shown that no unbiased bona fide estimator can exist for all continuous p. Rather than requiring an unbiased estimator that cannot be a bona fide estimator, we generally seek a bona fide estimator with small mean squared error or a sequence of bona fide estimators $\widehat{p}_n$ that are asymptotically unbiased; that is,

$$\mathrm{E}_p(\widehat{p}_n(y)) \to p(y) \quad \text{for all } y \in \mathbb{R}^d \text{ as } n \to \infty.$$

The Likelihood Function

Suppose that we have a random sample, $y_1, \ldots, y_n$, from a population with density p. Treating the density p as a variable, we write the likelihood functional as

$$L(p;\, y_1, \ldots, y_n) = \prod_{i=1}^{n} p(y_i).$$

The *maximum likelihood method* of estimation obviously cannot be used directly because this functional is unbounded in p. We may, however, seek an estimator that maximizes some modification of the likelihood. There are two reasonable ways to approach this problem. One is to restrict the domain of the optimization problem. This is called *restricted maximum likelihood*. The other is to *regularize* the estimator by adding a penalty term to the functional to be optimized. This is called *penalized maximum likelihood*.

We may seek to maximize the likelihood functional subject to the constraint that p be a bona fide density. If we put no further restrictions on the function p, however, infinite Dirac spikes at each observation give an unbounded likelihood, so a maximum likelihood estimator cannot exist, subject only to the restriction to the bona fide class. An additional restriction that p be Lebesgue-integrable over some domain D (that is, $p \in L^1(D)$) does not resolve the problem because we can construct sequences of finite spikes at each observation that grow without bound.

We therefore must restrict the class further. Consider a finite dimensional class, such as the class of step functions that are bona fide density estimators. We assume that the sizes of the regions over which the step function is constant are greater than 0.

For a step function with m regions having constant values, $c_1, \ldots, c_m$, the likelihood is

$$\begin{aligned} L(c_1, \ldots, c_m;\, y_1, \ldots, y_n) &= \prod_{i=1}^{n} p(y_i) \\ &= \prod_{k=1}^{m} c_k^{n_k}, \end{aligned} \tag{8.34}$$

where n_k is the number of data points in the k^{th} region. For the step function to be a bona fide estimator, all c_k must be nonnegative and finite. A maximum therefore exists in the class of step functions that are bona fide estimators.

If v_k is the measure of the volume of the k^{th} region (that is, v_k is the length of an interval in the univariate case, the area in the bivariate case, and so on), we have

$$\sum_{k=1}^{m} c_k v_k = 1.$$

We incorporate this constraint together with equation (8.34) to form the Lagrangian,

$$L(c_1, \ldots, c_m) + \lambda \left(1 - \sum_{k=1}^{m} c_k v_k\right).$$

Differentiating the Lagrangian function and setting the derivative to zero, we have at the maximum point $c_k = c_k^*$, for any λ,

$$\frac{\partial L}{\partial c_k} = \lambda v_k.$$

Using the derivative of L from equation (8.34), we get

$$n_k L = \lambda c_k^* v_k.$$

Summing both sides of this equation over k, we have

$$nL = \lambda,$$

and then substituting, we have

$$n_k L = nL c_k^* v_k.$$

Therefore, the maximum of the likelihood occurs at

$$c_k^* = \frac{n_k}{n v_k}.$$

The restricted maximum likelihood estimator is therefore

$$\begin{aligned} \widehat{p}(y) &= \frac{n_k}{n v_k}, \text{ for } y \in \text{region } k, \\ &= 0, \qquad \text{otherwise.} \end{aligned} \tag{8.35}$$

Instead of restricting the density estimate to step functions, we could consider other classes of functions, such as piecewise linear functions.

We may also seek other properties, such as smoothness, for the estimated density. One way of achieving other desirable properties for the estimator is to use a penalizing function to modify the function to be optimized. Instead of the likelihood function, we may use a penalized likelihood function of the form

$$L_{\mathrm{p}}(p;\, y_1, \ldots, y_n) = \prod_{i=1}^{n} p(y_i) \mathrm{e}^{-\mathcal{T}(p)}, \tag{8.36}$$

where $\mathcal{T}(p)$ is a transform that measures some property that we would like to minimize. For example, to achieve smoothness, we may use the transform $\mathcal{R}(p)$ of equation (8.27) in the penalizing factor. To choose a function $\hat{p}$ to maximize $L_{\mathrm{p}}(p)$ we would have to use some finite series approximation to $\mathcal{T}(\hat{p})$.

For densities with special properties there may be likelihood approaches that take advantage of those properties.

8.5.2 Histogram Estimators

Let us assume finite support D, and construct a fixed partition of D into a grid of m nonoverlapping bins T_k. (We can arbitrarily assign bin boundaries to one or the other bin.) Let v_k be the volume of the k^{th} bin (in one dimension, v_k is a length and in this simple case is often denoted h_k; in two dimensions, v_k is an area, and so on). The number of such bins we choose, and consequently their volumes, depends on the sample size n, so we sometimes indicate that dependence in the notation: $v_{n,k}$. For the sample $y_1, \ldots, y_n$, the histogram estimator of the probability density function is defined as

$$\begin{aligned}\widehat{p}_H(y) &= \sum_{k=1}^{m} \frac{1}{v_k} \frac{\sum_{i=1}^{n} \mathrm{I}_{T_k}(y_i)}{n} \mathrm{I}_{T_k}(y), \quad \text{for } y \in D,\\ &= 0, \quad \text{otherwise.}\end{aligned}$$

The histogram is the restricted maximum likelihood estimator (8.35).

Letting n_k be the number of sample values falling into T_k,

$$n_k = \sum_{i=1}^{n} \mathrm{I}_{T_k}(y_i),$$

we have the simpler expression for the histogram over D,

$$\widehat{p}_H(y) = \sum_{k=1}^{m} \frac{n_k}{n v_k} \mathrm{I}_{T_k}(y). \tag{8.37}$$

As we have noted already, this is a bona fide estimator:

$$\widehat{p}_H(y) \geq 0$$

and

$$\begin{aligned}\int_{\mathbb{R}^d} \widehat{p}_H(y)\mathrm{d}y &= \sum_{k=1}^{m} \frac{n_k}{n v_k} v_k\\ &= 1.\end{aligned}$$

Although our discussion generally concerns observations on multivariate random variables, we should occasionally consider simple univariate observations. One reason why the univariate case is simpler is that the derivative is a scalar function. Another reason why we use the univariate case as a model is because it is easier to visualize. The density of a univariate random variable is two-dimensional, and densities of other types of random variables are of higher dimension, so only in the univariate case can the density estimates be graphed directly.

In the univariate case, we assume that the support is the finite interval $[a, b]$. We partition $[a, b]$ into a grid of m nonoverlapping bins $T_k = [t_{n,k}, t_{n,k+1})$ where

$$a = t_{n,1} < t_{n,2} < \ldots < t_{n,m+1} = b.$$

The univariate histogram is

$$\widehat{p}_H(y) = \sum_{k=1}^{m} \frac{n_k}{n(t_{n,k+1} - t_{n,k})} \mathrm{I}_{T_k}(y). \tag{8.38}$$

If the bins are of equal width, say h (that is, $t_k = t_{k-1} + h$), the histogram is

$$\widehat{p}_H(y) = \frac{n_k}{nh}, \quad \text{for } y \in T_k.$$

This class of functions consists of polynomial splines of degree 0 with fixed knots, and the histogram is the maximum likelihood estimator over the class of step functions. Generalized versions of the histogram can be defined with respect to splines of higher degree. Splines with degree higher than 1 may yield negative estimators, but such histograms are also maximum likelihood estimators over those classes of functions.

The histogram as we have defined it is sometimes called a "density histogram", whereas a "frequency histogram" is not normalized by the n.

Some Properties of the Histogram Estimator

The histogram estimator, being a step function, is discontinuous at cell boundaries, and it is zero outside of a finite range. It is sensitive both to the bin size and to the choice of the origin.

An important advantage of the histogram estimator is its simplicity, both for computations and for analysis. In addition to its simplicity, as we have seen, it has two other desirable global properties:

- It is a bona fide density estimator.
- It is the unique maximum likelihood estimator confined to the subspace of functions of the form

$$\begin{aligned} g(t) &= c_k, \text{ for } t \in T_k, \\ &= 0, \text{ otherwise}, \end{aligned}$$

  and where $g(t) \geq 0$ and $\int_{\cup_k T_k} g(t)\,\mathrm{d}t = 1$.

Pointwise and Binwise Properties

Properties of the histogram vary from bin to bin. From equation (8.37), the expectation of the histogram estimator at the point y in bin T_k is

$$\mathrm{E}(\widehat{p}_H(y)) = \frac{p_k}{v_k}, \tag{8.39}$$

where

$$p_k = \int_{T_k} p(t)\,\mathrm{d}t \tag{8.40}$$

is the probability content of the k^{th} bin.

Some pointwise properties of the histogram estimator are the following:

- The **bias** of the histogram at the point y within the k^{th} bin is

$$\frac{p_k}{v_k} - p(y). \tag{8.41}$$

  Note that the bias is different from bin to bin, even if the bins are of constant size. The bias tends to decrease as the bin size decreases. We can bound the bias if we assume a regularity condition on p. If there exists γ such that for any $y_1 \neq y_2$ in an interval

$$|p(y_1) - p(y_2)| < \gamma\|y_1 - y_2\|,$$

  we say that p is Lipschitz-continuous on the interval, and for such a density, for any ξ_k in the k^{th} bin, we have

$$\begin{aligned} |\mathrm{Bias}(\widehat{p}_H(y))| &= |p(\xi_k) - p(y)| \\ &\leq \gamma_k\|\xi_k - y\| \\ &\leq \gamma_k v_k. \end{aligned} \tag{8.42}$$

- The **variance** of the histogram at the point y within the k^{th} bin is

$$\begin{aligned} \mathrm{V}\big(\widehat{p}_H(y)\big) &= \mathrm{V}(n_k)/(nv_k)^2 \\ &= \frac{p_k(1-p_k)}{nv_k^2}. \end{aligned} \tag{8.43}$$

  This is easily seen by recognizing that n_k is a binomial random variable with parameters n and p_k. Notice that the variance decreases as the bin size increases. Note also that the variance is different from bin to bin. We can bound the variance:

$$\mathrm{V}(\widehat{p}_H(y)) \leq \frac{p_k}{nv_k^2}.$$

  By the mean-value theorem, we have $p_k = v_k p(\xi_k)$ for some $\xi_k \in T_k$, so we can write

$$\mathrm{V}(\widehat{p}_H(y)) \leq \frac{p(\xi_k)}{nv_k}.$$

  Notice the tradeoff between bias and variance: *as h increases the variance, equation (8.43), decreases, but the bound on the bias, equation (8.42), increases.*

- The **mean squared error** of the histogram at the point y within the k^{th} bin is

$$\text{MSE}\big(\widehat{p}_H(y)\big) = \frac{p_k(1-p_k)}{nv_k^2} + \left(\frac{p_k}{v_k} - p(y)\right)^2. \tag{8.44}$$

  For a Lipschitz-continuous density, within the k^{th} bin we have

$$\text{MSE}\big(\widehat{p}_H(y)\big) \leq \frac{p(\xi_k)}{nv_k} + \gamma_k^2 v_k^2. \tag{8.45}$$

  We easily see that the histogram estimator is L_2 pointwise consistent for a Lipschitz-continuous density if, as $n \to \infty$, for each k, $v_k \to 0$ and $nv_k \to \infty$. By differentiating, we see that the minimum of the bound on the MSE in the k^{th} bin occurs for

$$h^*(k) = \left(\frac{p(\xi_k)}{2\gamma_k^2 n}\right)^{1/3}. \tag{8.46}$$

  Substituting this value back into MSE, we obtain the order of the optimal MSE at the point x,

$$\text{MSE}^*\big(\widehat{p}_H(y)\big) \in \text{O}\big(n^{-2/3}\big).$$

Asymptotic MISE (or AMISE) of Histogram Estimators

Global properties of the histogram are obtained by summing the binwise properties over all of the bins.

The expressions for the integrated variance and the integrated squared bias are quite complicated because they depend on the bin sizes and the probability content of the bins. We will first write the general expressions, and then we will assume some degree of smoothness of the true density and write approximate expressions that result from mean values or Taylor approximations. We will assume rectangular bins for additional simplification. Finally, we will then consider bins of equal size to simplify the expressions further.

First, consider the integrated variance,

$$\begin{aligned}
\text{IV}\big(\widehat{p}_H\big) &= \int_{\mathbb{R}^d} \text{V}(\widehat{p}_H(t))\,\text{d}t \\
&= \sum_{k=1}^{m} \int_{T_k} \text{V}(\widehat{p}_H(t))\,\text{d}t \\
&= \sum_{k=1}^{m} \frac{p_k - p_k^2}{nv_k} \\
&= \sum_{k=1}^{m} \left(\frac{1}{nv_k} - \frac{\sum p(\xi_k)^2 v_k}{n}\right) + \text{o}(n^{-1})
\end{aligned}$$

for some $\xi_k \in T_k$, as before. Now, taking $\sum p(\xi_k)^2 v_k$ as an approximation to the integral $\int (p(t))^2 \, \mathrm{d}t$, and letting $\mathcal{S}$ be the functional that measures the variation in a square-integrable function of d variables,

$$\mathcal{S}(g) = \int_{\mathbb{R}^d} (g(t))^2 \, \mathrm{d}t, \tag{8.47}$$

we have the integrated variance,

$$\mathrm{IV}\left(\widehat{p}_H\right) \approx \sum_{k=1}^{m} \frac{1}{n v_k} - \frac{\mathcal{S}(p)}{n}, \tag{8.48}$$

and the asymptotic integrated variance,

$$\mathrm{AIV}\left(\widehat{p}_H\right) = \sum_{k=1}^{m} \frac{1}{n v_k}. \tag{8.49}$$

The measure of the variation, $\mathcal{S}(p)$, is a measure of the roughness of the density because the density integrates to 1.

Now, consider the other term in the integrated MSE, the integrated squared bias. We will consider the case of rectangular bins, in which $h_k = (h_{k_1}, \ldots, h_{k_d})$ is the vector of lengths of sides in the k^{th} bin. In the case of rectangular bins, $v_k = \Pi_{j=1}^{d} h_{k_j}$.

We assume that the density can be expanded in a Taylor series, and we expand the density in the k^{th} bin about $\bar{t}_k$, the midpoint of the rectangular bin. For $\bar{t}_k + t \in T_k$, we have

$$p(\bar{t}_k + t) = p(\bar{t}_k) + t^{\mathrm{T}} \nabla p(\bar{t}_k) + \frac{1}{2} t^{\mathrm{T}} \mathrm{H}_p(\bar{t}_k) t + \cdots, \tag{8.50}$$

where $\mathrm{H}_p(\bar{t}_k)$ is the Hessian of p evaluated at $\bar{t}_k$.

The probability content of the k^{th} bin, p_k, from equation (8.40), can be expressed as an integral of the Taylor series expansion:

$$\begin{aligned} p_k &= \int_{\bar{t}_k + t \in T_k} p(\bar{t}_k + t) \, \mathrm{d}t \\ &= \int_{-h_{kd}/2}^{h_{kd}/2} \cdots \int_{-h_{k1}/2}^{h_{k1}/2} \left(p(\bar{t}_k) + t^{\mathrm{T}} \nabla p(\bar{t}_k) + \ldots \right) \mathrm{d}t_1 \cdots \mathrm{d}t_d \\ &= v_k p(\bar{t}_k) + \mathrm{O}\left(h_{k*}^{d+2}\right), \end{aligned} \tag{8.51}$$

where $h_{k*} = \min_j h_{kj}$. The bias at a point $\bar{t}_k + t$ in the k^{th} bin, after substituting equations (8.50) and (8.51) into equation (8.41), is

$$\frac{p_k}{v_k} - p(\bar{t}_k + t) = -t^{\mathrm{T}} \nabla p(\bar{t}_k) + \mathrm{O}\left(h_{k*}^2\right).$$

For the k^{th} bin the integrated squared bias is

$$\begin{aligned}
&\mathrm{ISB}_k(\widehat{p}_{\mathrm{H}}) \\
&= \int_{T_k} \left(\left(t^{\mathrm{T}} \nabla p(\bar{t}_k)\right)^2 - \mathrm{O}\left(h_{k*}^2\right) t^{\mathrm{T}} \nabla p(\bar{t}_k) + \mathrm{O}\left(h_{k*}^4\right) \right) \mathrm{d}t \\
&= \int_{-h_{kd}/2}^{h_{kd}/2} \cdots \int_{-h_{k1}/2}^{h_{k1}/2} \sum_i \sum_j t_{ki} t_{kj} \nabla_i p(\bar{t}_k) \nabla_j p(\bar{t}_k) \, \mathrm{d}t_1 \cdots \mathrm{d}t_d + \mathrm{O}\left(h_{k*}^{4+d}\right).
\end{aligned} \tag{8.52}$$

Many of the expressions above are simpler if we use a constant bin size, v, or $h_1, \ldots, h_d$. In the case of constant bin size, the asymptotic integrated variance in equation (8.49) becomes

$$\mathrm{AIV}\left(\widehat{p}_H\right) = \frac{m}{nv}. \tag{8.53}$$

In this case, the integral in equation (8.52) simplifies as the integration is performed term by term because the cross-product terms cancel, and the integral is

$$\frac{1}{12}(h_1 \cdots h_d) \sum_{j=1}^{d} h_j^2 \left(\nabla_j p(\bar{t}_k)\right)^2. \tag{8.54}$$

This is the asymptotic squared bias integrated over the k^{th} bin.

When we sum the expression (8.54) over all bins, the $\left(\nabla_j p(\bar{t}_k)\right)^2$ become $\mathcal{S}\left(\nabla_j p\right)$, and we have the asymptotic integrated squared bias,

$$\mathrm{AISB}\left(\widehat{p}_H\right) = \frac{1}{12} \sum_{j=1}^{d} h_j^2 \mathcal{S}\left(\nabla_j p\right). \tag{8.55}$$

Combining the asymptotic integrated variance, equation (8.53), and squared bias, equation (8.55), for the histogram with rectangular bins of constant size, we have

$$\mathrm{AMISE}\left(\widehat{p}_H\right) = \frac{1}{n(h_1 \cdots h_d)} + \frac{1}{12} \sum_{j=1}^{d} h_j^2 \mathcal{S}\left(\nabla_j p\right). \tag{8.56}$$

As we have seen before, smaller bin sizes increase the variance but decrease the squared bias.

Bin Sizes

As we have mentioned and have seen by example, the histogram is very sensitive to the bin sizes, both in appearance and in other properties. Equation (8.56) for the AMISE assuming constant rectangular bin size is often used as a guide for determining the bin size to use when constructing a histogram. This expression involves $\mathcal{S}\left(\nabla_j p\right)$ and so, of course, cannot be used

directly. Nevertheless, differentiating the expression with respect to h_j and setting the result equal to zero, we have the bin width that is optimal with respect to the AMISE,

$$h_{j*} = \mathcal{S}(\nabla_j p)^{-1/2} \left(6 \prod_{i=1}^{d} \mathcal{S}(\nabla_i p)^{1/2} \right)^{\frac{1}{2+d}} n^{-\frac{1}{2+d}}. \tag{8.57}$$

Substituting this into equation (8.56), we have the optimal value of the AMISE

$$\frac{1}{4} \left(36 \prod_{i=1}^{d} \mathcal{S}(\nabla_i p)^{1/2} \right)^{\frac{1}{2+d}} n^{-\frac{2}{2+d}}. \tag{8.58}$$

Notice that the optimal rate of decrease of AMISE for histogram estimators is that of $\mathrm{O}(n^{-\frac{2}{2+d}})$. Although histograms have several desirable properties, this order of convergence is not good compared to that of some other bona fide density estimators, as we will see in later sections.

The expression for the optimal bin width involves $\mathcal{S}(\nabla_j p)$, where p is the unknown density. An approach is to choose a value for $\mathcal{S}(\nabla_j p)$ that corresponds to some good general distribution. A "good general distribution", of course, is the normal with a diagonal variance-covariance matrix. For the d-variate normal with variance-covariance matrix $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$,

$$\mathcal{S}(\nabla_j p) = \frac{1}{2^{d+1} \pi^{d/2} \sigma_j^2 |\Sigma|^{1/2}}.$$

For a univariate normal density with variance σ^2,

$$\mathcal{S}(p') = 1/(4\sqrt{\pi}\sigma^3),$$

so the optimal constant one-dimensional bin width under the AMISE criterion is

$$3.49 \sigma n^{-1/3}.$$

In practice, of course, an estimate of σ must be used. The sample standard deviation s is one obvious choice. A more robust estimate of the scale is based on the sample interquartile range, r. The sample interquartile range leads to a bin width of $2rn^{-1/3}$.

The AMISE is essentially an L_2 measure. The L_∞ criterion—that is, the sup absolute error (SAE) of equation (8.19)—also leads to an asymptotically optimal bin width that is proportional to $n^{-1/3}$.

One of the most commonly used rules is for the number of bins rather than the width. Assume a symmetric binomial model for the bin counts, that is, the bin count is just the binomial coefficient. The total sample size n is

$$\sum_{k=0}^{m-1} \binom{m-1}{k} = 2^{m-1},$$

and so the number of bins is

$$m = 1 + \log_2 n.$$

Bin Shapes

In the univariate case, histogram bins may vary in size, but each bin is an interval. For the multivariate case, there are various possibilities for the shapes of the bins. The simplest shape is the direct extension of an interval, that is a hyperrectangle. The volume of a hyperrectangle is just $v_k = \prod h_{kj}$. There are, of course, other possibilities; any tessellation of the space would work. The objects may or may not be regular, and they may or may not be of equal size. Regular, equal-sized geometric figures such as hypercubes have the advantages of simplicity, both computationally and analytically. In two dimensions, there are three possible regular tessellations: triangles, squares, and hexagons.

For hyperrectangles of constant size, the univariate theory generally extends fairly easily to the multivariate case. The histogram density estimator is

$$\widehat{p}_H(y) = \frac{n_k}{nh_1h_2\cdots h_d}, \quad \text{for } y \in T_k,$$

where the h's are the lengths of the sides of the rectangles. The variance within the k^{th} bin is

$$\mathrm{V}(\widehat{p}_H(y)) = \frac{np_k(1-p_k)}{(nh_1h_2\cdots h_d)^2}, \quad \text{for } y \in T_k,$$

and the integrated variance is

$$\mathrm{IV}(\widehat{p}_H) \approx \frac{1}{nh_1h_2\cdots h_d} - \frac{\mathcal{S}(f)}{n}.$$

Other Density Estimators Related to the Histogram

There are several variations of the histogram that are useful as probability density estimators. The most common modification is to connect points on the histogram by a continuous curve. A simple way of doing this in the univariate case leads to the *frequency polygon*. This is the piecewise linear curve that connects the midpoints of the bins of the histogram. The endpoints are usually zero values at the midpoints of two appended bins, one on either side.

The *histospline* is constructed by interpolating knots of the empirical CDF with a cubic spline and then differentiating it. More general methods use splines or orthogonal series to fit the histogram.

As we have mentioned and have seen by example, the histogram is somewhat sensitive in appearance to the location of the bins. To overcome the problem of location of the bins, a density estimator that is the average of several histograms with equal bin widths but different bin locations can be used. This is called the *average shifted histogram*, or ASH. It also has desirable statistical properties, and it is computationally efficient in the multivariate case.

8.5.3 Kernel Estimators

Kernel methods are probably the most widely used technique for building nonparametric probability density estimators. They are best understood by developing them as a special type of histogram. The difference is that the bins in kernel estimators are centered at the points at which the estimator is to be computed. The problem of the choice of location of the bins in histogram estimators does not arise.

Rosenblatt's Histogram Estimator; Kernels

For the one-dimensional case, Rosenblatt defined a histogram that is shifted to be centered on the point at which the density is to be estimated. Given the sample $y_1, \ldots, y_n$, Rosenblatt's histogram estimator at the point y is

$$\widehat{p}_R(y) = \frac{\#\{y_i \text{ s.t. } y_i \in]y - h/2, \quad y + h/2]\,\}}{nh}. \tag{8.59}$$

This histogram estimator avoids the ordinary histogram's constant-slope contribution to the bias. This estimator is a step function with variable lengths of the intervals that have constant value.

Rosenblatt's centered histogram can also be written in terms of the ECDF:

$$\widehat{p}_R(y) = \frac{P_n(y + h/2) - P_n(y - h/2)}{h},$$

where, as usual, P_n denotes the ECDF. As seen in this expression, Rosenblatt's estimator is a centered finite-difference approximation to the derivative of the empirical cumulative distribution function (which, of course, is not differentiable at the data points). We could, of course, use the same idea and form other density estimators using other finite-difference approximations to the derivative of P_n.

Another way to write Rosenblatt's shifted histogram estimator over bins of length h is

$$\widehat{p}_R(y) = \frac{1}{nh} \sum_{i=1}^{n} \kappa\left(\frac{y - y_i}{h}\right), \tag{8.60}$$

where $\kappa(t) = 1$ if $|t| < 1/2$ and $= 0$ otherwise. The function κ is a kernel or filter. In Rosenblatt's estimator, it is a "boxcar" function, but other kernel functions could be used.

The estimator extends easily to the multivariate case. In the general kernel estimator, we usually use a more general scaling of $y - y_i$,

$$V^{-1}(y - y_i),$$

for some positive-definite matrix V. The determinant of V^{-1} scales the estimator to account for the scaling within the kernel function. The general kernel estimator is given by

$$\widehat{p}_K(y) = \frac{1}{n|V|} \sum_{i=1}^{n} \kappa\left(V^{-1}(y - y_i)\right), \tag{8.61}$$

where the function κ is called the *kernel*, and V is the *smoothing matrix*. The determinant of the smoothing matrix is exactly analogous to the bin volume in a histogram estimator. The univariate version of the kernel estimator is the same as Rosenblatt's estimator (8.60), but in which a more general function κ is allowed.

In practice, V is usually taken to be constant for a given sample size, but, of course, there is no reason for this to be the case, and indeed it may be better to vary V depending on the number of observations near the point y. The dependency of the smoothing matrix on the sample size n and on y is often indicated by the notation $V_n(y)$.

Properties of Kernel Estimators

The appearance of the kernel density estimator depends to some extent on the support and shape of the kernel. Unlike the histogram estimator, the kernel density estimator may be continuous and even smooth.

It is easy to see that if the kernel satisfies

$$\kappa(t) \geq 0, \tag{8.62}$$

and

$$\int_{\mathbb{R}^d} \kappa(t)\,\mathrm{d}t = 1 \tag{8.63}$$

(that is, if κ is a density), then $\widehat{p}_K(y)$ is a bona fide density estimator.

There are other requirements that we may impose on the kernel either for the theoretical properties that result or just for their intuitive appeal. It also seems reasonable that in estimating the density at the point y, we would want to emphasize the sample points near y. This could be done in various ways, but one simple way is to require

$$\int_{\mathbb{R}^d} t\kappa(t)\,\mathrm{d}t = 0. \tag{8.64}$$

In addition, we may require the kernel to be symmetric about 0.

For multivariate density estimation, the kernels are usually chosen as a radially symmetric generalization of a univariate kernel. Such a kernel can be formed as a product of the univariate kernels. For a product kernel, we have for some constant σ_κ^2,

$$\int_{\mathbb{R}^d} tt^{\mathrm{T}}\kappa(t)\,\mathrm{d}t = \sigma_\kappa^2 I_d, \tag{8.65}$$

where I_d is the identity matrix of order d. We could also impose this as a requirement on any kernel, whether it is a product kernel or not. This makes

the expressions for bias and variance of the estimators simpler. The spread of the kernel can always be controlled by the smoothing matrix V, so sometimes, for convenience, we require $\sigma_\kappa^2 = 1$.

In the following, we will assume the kernel satisfies the properties in equations (8.62) through (8.65).

The pointwise properties of the kernel estimator are relatively simple to determine because the estimator at a point is merely the sample mean of n independent and identically distributed random variables. The expectation of the kernel estimator (8.61) at the point y is the convolution of the kernel function and the probability density function,

$$\begin{aligned} \mathrm{E}\left(\widehat{p}_K(y)\right) &= \frac{1}{|V|}\int_{\mathbb{R}^d} \kappa\left(V^{-1}(y-t)\right) p(t)\,\mathrm{d}t \\ &= \int_{\mathbb{R}^d} \kappa(u) p(y-Vu)\,\mathrm{d}u, \end{aligned} \tag{8.66}$$

where $u = V^{-1}(y-t)$ (and, hence, $\mathrm{d}u = |V|^{-1}\mathrm{d}t$).

If we approximate $p(y-Vu)$ about y with a three-term Taylor series, using the properties of the kernel in equations (8.62) through (8.65) and using properties of the trace, we have

$$\begin{aligned} \mathrm{E}\left(\widehat{p}_K(y)\right) &\approx \int_{\mathbb{R}^d} \kappa(u)\left(p(y) - (Vu)^{\mathrm{T}}\nabla p(y) + \frac{1}{2}(Vu)^{\mathrm{T}}\mathrm{H}_p(y)Vu\right)\mathrm{d}u \\ &= p(y) - 0 + \frac{1}{2}\mathrm{trace}\left(V^{\mathrm{T}}\mathrm{H}_p(y)V\right). \end{aligned} \tag{8.67}$$

To second order in the elements of V (that is, to terms in $\mathrm{O}(|V|^2)$), the bias at the point y is therefore

$$\frac{1}{2}\mathrm{trace}\left(VV^{\mathrm{T}}\mathrm{H}_p(y)\right). \tag{8.68}$$

Using the same kinds of expansions and approximations as in equations (8.66) and (8.67) to evaluate $\mathrm{E}\left((\widehat{p}_K(y))^2\right)$ to get an expression of order $\mathrm{O}(|V|/n)$, and subtracting the square of the expectation in equation (8.67), we get the approximate variance at y as

$$\mathrm{V}\left(\widehat{p}_K(y)\right) \approx \frac{p(y)}{n|V|}\int_{\mathbb{R}^d} (\kappa(u))^2\,\mathrm{d}u,$$

or

$$\mathrm{V}\left(\widehat{p}_K(y)\right) \approx \frac{p(y)}{n|V|}\mathcal{S}(\kappa). \tag{8.69}$$

Integrating this, because p is a density, we have

$$\mathrm{AIV}\left(\widehat{p}_K\right) = \frac{\mathcal{S}(\kappa)}{n|V|}, \tag{8.70}$$

and integrating the square of the asymptotic bias in expression (8.68), we have

$$\mathrm{AISB}(\widehat{p}_K) = \frac{1}{4}\int_{\mathbb{R}^d} \left(\mathrm{trace}\left(V^{\mathrm{T}}\mathrm{H}_p(y)V\right)\right)^2 \mathrm{d}y. \tag{8.71}$$

These expressions are much simpler in the univariate case, where the smoothing matrix V is the smoothing parameter or window width h. We have a simpler approximation for $\mathrm{E}\left(\widehat{p}_K(y)\right)$ than that given in equation (8.67),

$$\mathrm{E}\left(\widehat{p}_K(y)\right) \approx p(y) + \frac{1}{2}h^2 p''(y)\int_{\mathbb{R}} u^2\kappa(u)\,\mathrm{d}u,$$

and from this we get a simpler expression for the AISB. After likewise simplifying the AIV, we have

$$\mathrm{AMISE}(\widehat{p}_K) = \frac{\mathcal{S}(\kappa)}{nh} + \frac{1}{4}\sigma_\kappa^4 h^4 \mathcal{R}(p), \tag{8.72}$$

where we have left the kernel unscaled (that is, $\int u^2\kappa(u)\,\mathrm{d}u = \sigma_K^2$).

Minimizing this with respect to h, we have the optimal value of the smoothing parameter

$$h^* = \left(\frac{\mathcal{S}(\kappa)}{n\sigma_\kappa^4\mathcal{R}(p)}\right)^{1/5}; \tag{8.73}$$

that is, the optimal bandwidth is $\mathrm{O}(n^{-1/5})$.

Substituting the optimal bandwidth back into the expression for the AMISE, we find that its optimal value in this univariate case is

$$\frac{5}{4}\mathcal{R}(p)(\sigma_\kappa\mathcal{S}(\kappa))^{4/5}\, n^{-4/5}. \tag{8.74}$$

The AMISE for the univariate kernel density estimator is thus in $\mathrm{O}(n^{-4/5})$. Recall that the AMISE for the univariate histogram density estimator is in $\mathrm{O}(n^{-2/3})$.

We see that the bias and variance of kernel density estimators have similar relationships to the smoothing matrix that the bias and variance of histogram estimators have. As the determinant of the smoothing matrix gets smaller (that is, as the window of influence around the point at which the estimator is to be evaluated gets smaller), the bias becomes smaller and the variance becomes larger. This agrees with what we would expect intuitively.

Kernel-Based Estimator of CDF

We can form an estimator of the CDF based on a kernel PDF estimator p_K by using

$$K(x) = \int_{-\infty}^{x} \kappa(t)\mathrm{d}t.$$

A kernel-based estimator of the CDF is

$$\widehat{P}_K(y) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{y - y_i}{h_n}\right).$$

Let us consider the convergence of $\{\widehat{P}_{Kn}\}$ to the true CDF P. We recall from the Glivenko-Cantelli theorem (Theorem 1.71) that the ECDF, P_n converges to P uniformly; that is, for any P, given ϵ and η, there exists an N independent of P such that $\forall n \geq N$

$$\Pr\left(\sup_{y\in\mathbb{R}} |P_n(y) - P(y)| \geq \epsilon\right) \leq \eta.$$

This does not hold for $\{\widehat{P}_{Kn}\}$. To see this, pick a point y_0 for which $P(y_0) > 0$. Now, assume that (1) for some i, $0 < K((y_0 - y_i)/h_n) < 1$, (2) for some $t \in]0, P(y_0)[$, $K^{-1}(t) < (y_0 - y_i)/h_n$, and (3) $h_n > 0 \forall n$. (If the kernel is a PDF and if the kernel density estimator is finite, then these conditions hold.)

********** Zieliński (2007)

Choice of Kernels

Standard normal densities have these properties described above, so the kernel is often chosen to be the standard normal density. As it turns out, the kernel density estimator is not very sensitive to the form of the kernel.

Although the kernel may be from a parametric family of distributions, in kernel density estimation, we do not estimate those parameters; hence, the kernel method is a nonparametric method.

Sometimes, a kernel with finite support is easier to work with. In the univariate case, a useful general form of a compact kernel is

$$\kappa(t) = \kappa_{rs}(1 - |t|^r)^s \mathrm{I}_{[-1,1]}(t),$$

where

$$\kappa_{rs} = \frac{r}{2\mathrm{B}(1/r, s+1)}, \quad \text{for } r > 0,\ s \geq 0,$$

and $\mathrm{B}(a, b)$ is the complete beta function.

This general form leads to several simple specific cases:

- for $r = 1$ and $s = 0$, it is the rectangular kernel;
- for $r = 1$ and $s = 1$, it is the triangular kernel;
- for $r = 2$ and $s = 1$ ($\kappa_{rs} = 3/4$), it is the "Epanechnikov" kernel, which yields the optimal rate of convergence of the MISE (see Epanechnikov, 1969);
- for $r = 2$ and $s = 2$ ($\kappa_{rs} = 15/16$), it is the "biweight" kernel.

If $r = 2$ and $s \to \infty$, we have the Gaussian kernel (with some rescaling).

As mentioned above, for multivariate density estimation, the kernels are often chosen as a product of the univariate kernels. The product Epanechnikov kernel, for example, is

$$\kappa(t) = \frac{d+2}{2c_d}(1 - t^{\mathrm{T}}t)\mathrm{I}_{(t^{\mathrm{T}}t \leq 1)},$$

where

$$c_d = \frac{\pi^{d/2}}{\Gamma(d/2+1)}.$$

We have seen that the AMISE of a kernel estimator (that is, the sum of equations (8.70) and (8.71)) depends on $\mathcal{S}(\kappa)$ and the smoothing matrix V. As we mentioned above, the amount of smoothing (that is, the window of influence) can be made to depend on σ_κ. We can establish an approximate equivalence between two kernels, κ_1 and κ_2, by choosing the smoothing matrix to offset the differences in $\mathcal{S}(\kappa_1)$ and $\mathcal{S}(\kappa_2)$ and in σ_{κ_1} and σ_{κ_2}.

Computation of Kernel Density Estimators

If the estimate is required at one point only, it is simplest just to compute it directly. If the estimate is required at several points, it is often more efficient to compute the estimates in some regular fashion.

If the estimate is required over a grid of points, a fast Fourier transform (FFT) can be used to speed up the computations.

8.5.4 Choice of Window Widths

An important problem in nonparametric density estimation is to determine the smoothing parameter, such as the bin volume, the smoothing matrix, the number of nearest neighbors, or other measures of locality. In kernel density estimation, the window width has a much greater effect on the estimator than the kernel itself does.

An objective is to choose the smoothing parameter that minimizes the MISE. We often can do this for the AMISE, as in equation (8.57) on page 588. It is not as easy for the MISE. The first problem, of course, is just to estimate the MISE.

In practice, we use cross validation with varying smoothing parameters and alternate computations between the MISE and AMISE.

In univariate density estimation, the MISE has terms such as $h^\alpha \mathcal{S}(p')$ (for histograms) or $h^\alpha \mathcal{S}(p'')$ (for kernels). We need to estimate the roughness of a derivative of the density.

Using a histogram, a reasonable estimate of the integral $\mathcal{S}(p')$ is a Riemann approximation,

$$\begin{aligned}\widehat{\mathcal{S}}(p') &= h\sum\left(\widehat{p}'(t_k)\right)^2 \\ &= \frac{1}{n^2h^3}\sum(n_{k+1}-n_k)^2,\end{aligned}$$

where $\widehat{p}'(t_k)$ is the finite difference at the midpoints of the k^{th} and $(k+1)^{\text{th}}$ bins; that is,

$$\widehat{p}'(t_k) = \frac{n_{k+1}/(nh) - n_k/(nh)}{h}.$$

This estimator is biased. For the histogram, for example,

$$\mathrm{E}(\widehat{\mathcal{S}}(p')) = \mathcal{S}(p') + 2/(nh^3) + \ldots$$

A standard estimation scheme is to correct for the $2/(nh^3)$ term in the bias and plug this back into the formula for the AMISE (which is $1/(nh) + h^2\mathcal{S}(r')/12$ for the histogram).

We compute the estimated values of the AMISE for various values of h and choose the one that minimizes the AMISE. This is called *biased cross validation* because of the use of the AMISE rather than the MISE.

These same techniques can be used for other density estimators and for multivariate estimators, although at the expense of considerably more complexity.

8.5.5 Orthogonal Series Estimators

A continuous real function $p(x)$, integrable over a domain D, can be represented over that domain as an infinite series in terms of a complete spanning set of real orthogonal functions $\{f_k\}$ over D:

$$p(x) = \sum_k c_k f_k(x). \tag{8.75}$$

The orthogonality property allows us to determine the coefficients c_k in the expansion (8.75):

$$c_k = \langle f_k, p\rangle. \tag{8.76}$$

Approximation using a truncated orthogonal series can be particularly useful in estimation of a probability density function because the orthogonality relationship provides an equivalence between the coefficient and an expected value. Expected values can be estimated using observed values of the random variable and the approximation of the probability density function. Assume that the probability density function p is approximated by an orthogonal series $\{q_k\}$ with weight function $w(y)$:

$$p(y) = \sum_k c_k q_k(y).$$

From equation (8.76), we have

$$\begin{aligned} c_k &= \langle q_k, p \rangle \\ &= \int_D q_k(y)p(y)w(y)\mathrm{d}y \\ &= \mathrm{E}(q_k(Y)w(Y)), \end{aligned} \tag{8.77}$$

where Y is a random variable whose probability density function is p.

The c_k can therefore be unbiasedly estimated by

$$\widehat{c}_k = \frac{1}{n}\sum_{i=1}^{n} q_k(y_i)w(y_i).$$

The orthogonal series estimator is therefore

$$\widehat{p}_S(y) = \frac{1}{n}\sum_{k=0}^{j}\sum_{i=1}^{n} q_k(y_i)w(y_i)q_k(y) \tag{8.78}$$

for some truncation point j.

Without some modifications, this generally is not a good estimator of the probability density function. It may not be smooth, and it may have infinite variance. The estimator may be improved by shrinking the $\widehat{c}_k$ toward the origin. The number of terms in the finite series approximation also has a major effect on the statistical properties of the estimator. Having more terms is not necessarily better. One useful property of orthogonal series estimators is that the convergence rate is independent of the dimension. This may make orthogonal series methods more desirable for higher-dimensional problems.

There are several standard orthogonal series that could be used. The two most commonly used series are the Fourier and the Hermite. Which is preferable depends on the situation.

The Fourier series is commonly used for distributions with bounded support. It yields estimators with better properties in the L_1 sense.

For distributions with unbounded support, the Hermite polynomials are most commonly used.

8.6 Perturbations of Probability Distributions

If a task in statistical inference begins with the assumption that the underlying family of probability distributions is $\mathcal{P}$, we use statistical methods that are optimized for this family. If, however, the true underlying family of probability distributions is $\mathcal{P}_\epsilon$, where $\mathcal{P}_\epsilon \neq \mathcal{P}$, our methods of inference may not be very good. If $\mathcal{P}_\epsilon \subseteq \mathcal{P}$, our methods are likely to be valid but suboptimal; if $\mathcal{P}_\epsilon \supseteq \mathcal{P}$, our methods are likely to be invalid; in more general cases when $\mathcal{P}_\epsilon \neq \mathcal{P}$, we may have no knowledge of how the method performs. Our objective is to identify methods that are likely to be "good" so long as $\mathcal{P}_\epsilon$ is "relatively close" to $\mathcal{P}$.

We often measure the difference in functions by a norm or pseudonorm functional (see Section 0.1.9 beginning on page 744). The measure of the difference is called a metric, or pseudometric (see Section 0.1.9).

Functionals of CDFs can be used as measures of the differences between two distributions. Because of the definition of a CDF, the functionals we use are true norms and true metrics.

In Section 7.4.3, we discussed ways of measuring the distance between two different distributions for the purpose of testing goodness of fit, that is, for testing an hypothesis that a given sample came from some specified distribution. Functionals used to measure differences between two distributions can also be used to evaluate the properties of statistical methods, especially if those methods are defined in terms of functionals.

Distances between Probability Distributions

In Section 0.1.9 beginning on page 747 we discuss various general measures of the difference between two functions. The difference in two probability distributions may be measured in terms of a distance between the cumulative distribution functions such as the Hellinger distance or the Kullback-Leibler measure as described on page 747, or it may be measured in terms of differences in probabilities or differences in expected values.

Because we use samples to make inferences about the distances between probability distributions, the measures of interest are usually taken between two ECDFs. If we measure the distance between probability distributions in terms of a distance between the cumulative distribution functions, we can compare the distance between the ECDFs from the samples. If the comparison is between a distribution of a sample and some family of distributions, we use the ECDF from the sample and a CDF from the family; If the comparison is between the distributions of two samples, we use the ECDFs from the samples.

It is important to note that even though the measure of the difference between two CDFs may be small, there may be very large differences in properties of the probability distributions. For example, consider the difference between the CDF of a standard Cauchy and a standard normal. The sup difference is about 0.1256. (It occurs near ± 1.85.) The sup dif between the ECDFs for samples of size 20 will often be between 0.2 and 0.3. (That is a significance level of between 0.83 and 0.34 on a KS test.)

The Kolmogorov Distance; An L_∞ Metric

On page 536, we defined the Kolmogorov distance between two CDFs P_1 and P_2, $\rho_K(P_1, P_2)$, as the L_∞ norm of the difference between the two CDFs:

$$\rho_K(P_1, P_2) = \sup |P_1 - P_2|. \tag{8.79}$$

The Lévy Metric

Another measure of the distance between two CDFs is the *Lévy distance*, defined for the CDFs P_1 and P_2 as

$$\rho_{\mathrm{L}}(P_1, P_2) = \inf\{h, \text{ s.t. } \forall x, P_1(x-h) - h \leq P_2(x) \leq P_1(x+h) + h\}. \quad (8.80)$$

Notice that h is both the deviation in the argument x and the deviation in the function values.

It can be shown that $\rho_{\mathrm{L}}(P_1, P_2)$ is a metric over the set of distribution functions. It can also be shown that for any CDFs P_1 and P_2,

$$\rho_{\mathrm{L}}(P_1, P_2) \leq \rho_{\mathrm{K}}(P_1, P_2). \quad (8.81)$$

Example 8.1 Kolmogorov and Lévy distances between distributions
Consider the $\mathrm{U}(0,1)$ distribution and the Bernoulli distribution with parameter $\pi = 0.3$, with CDFs P_1 and P_2 respectively. Figure 8.2 shows the CDFs and the Kolmogorov and Lévy distances between them.

We see that if h were any smaller and $x = 1$, then $P_1(x-h) - h$ would be greater than $P_2(x)$. On the other hand, we see that this value of h will satisfy the inequalities in the definition of the Lévy distance at any point x. ∎

The Wasserstein-Mallows Metric

Another useful measure of the distance between two CDFs is the *Mallows distance* or the *Wasserstein-Mallows distance*. This metric is also called by various other names, including the Renyi metric, and the "earth movers' distance". We will briefly describe this metric, but we will rarely use it in the following.

For the CDFs P_1 and P_2, with random variables X_1 having CDF P_1 and X_2 having CDF P_2, if $\mathrm{E}(\|X_1\|^p)$ and $\mathrm{E}(\|X_2\|^p)$ are finite, this distance is

$$\rho_{\mathrm{M}_p}(P_1, P_2) = \inf(\mathrm{E}(\|X_1 - X_2\|^p))^{1/p},$$

where the infimum is taken over all joint distributions $P(x_1, x_2)$ with marginals P_1 and P_2.

If P_1 and P_2 are univariate ECDFs based on the same number of observations, we have

$$\rho_{\mathrm{M}_p}(P_{n1}, P_{n2}) = \left(\frac{1}{n}\sum(|x_{1(i)} - x_{2(i)}|^p)\right)^{1/p}.$$

For a scalar-valued random variable X with CDF P, if $U \sim \mathrm{U}(0,1)$ then $P^{-1}(U) \overset{\mathrm{d}}{=} X$ (Corollary 1.7.1); that is,

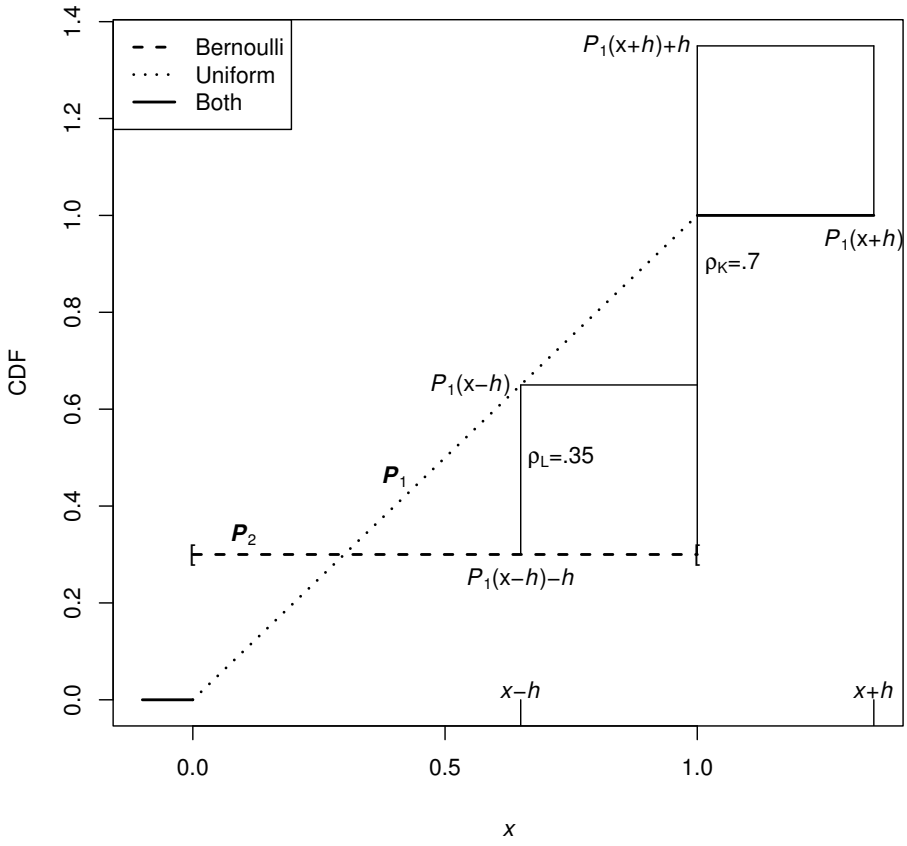


Figure 8.2. Two CDFs and the Kolmogorov and Lévy Distances between Them

$$\rho_{\mathrm{M}_p}(P_1, P_2) = (\mathrm{E}(\|P_1^{-1}(U) - P_2^{-1}(U)\|^p))^{1/p},$$

The first question we might consider given the definition of the Wasserstein-Mallows metric is whether the infimum exists, and then it is not clear whether this is indeed a metric. (The triangle inequality is the only hard question.) Bickel and Freedman (1981) answered both of these questions in the affirmative. The proof is rather complicated for vector-valued random variables; for scalar-valued random variables, there is a simpler proof in terms of the inverse CDF.

A Useful Class of Perturbations

In statistical applications using functionals defined on the CDF, we are interested in how the functional varies for "nearby" CDFs in the distribution function space.

A simple kind of perturbation of a given distribution is a mixture distribution with the given distribution as one of the components of the mixture. We often consider a simple type of function in the neighborhood of the CDF. This kind of CDF results from adding a single mass point to the given distribution. For a given CDF $P(x)$, we can define a simple perturbation as

$$P_{x_c,\epsilon}(x) = (1-\epsilon)P(x) + \epsilon \mathrm{I}_{[x_c,\infty[}(x), \tag{8.82}$$

where $0 \leq \epsilon \leq 1$. This is an ϵ-mixture family of distributions that we discussed on page 194. We will refer to the distribution with CDF P as the reference distribution. (The reference distribution is the distribution of interest, so I often refer to it without any qualification.)

A simple interpretation of the perturbation in equation (8.82) is that it is the CDF of a mixture of a distribution with CDF P and a degenerate distribution with a single mass point at x_c, which may or may not be in the support of the distribution. The extent of the perturbation depends on ϵ; if $\epsilon = 0$, the distribution is the reference distribution.

If the distribution with CDF P is continuous with PDF p, the PDF of the mixture is

$$\mathrm{d}P_{x_c,\epsilon}(x)/\mathrm{d}x = (1-\epsilon)p(x) + \epsilon\delta(x_c - x),$$

where $\delta(\cdot)$ is the Dirac delta function. If the distribution is discrete, the probability mass function has nonzero probabilities (scaled by $(1-\epsilon)$) at each of the mass points associated with P together with a mass point at x_c with probability ϵ.

The left-hand graph in Figure 8.3 shows the Lebesgue PDF of a continuous reference distribution (dotted line) and the PDF of an ϵ-mixture distribution (solid line together with the mass point at x_c). Over part of the support the PDF of the mixture is a Lebesgue PDF and over another part of the support (the single point) it is a probability mass function. The right-hand graph shows the corresponding CDFs. The reference distribution is a standard normal, $x_c = 1$, and $\epsilon = 0.3$. (Such a large value of ϵ was used so that the graphs would look better. In most applications when an ϵ-mixture distribution is assumed, the value of ϵ is much smaller, often of the order of .05.)

We will often analyze the sensitivity of statistical methods with respect to the perturbation of a reference distribution by x_c and ϵ.

Example 8.2 Kolmogorov and Lévy distances between standard normal and associated ϵ-mixture distribution
Consider the $\mathrm{N}(0,1)$ distribution and the associated ϵ-mixture distribution with $x_c = 1$ and $\epsilon = 0.1$. Figure 8.4 shows the CDFs and the Kolmogorov and Lévy distances between them.

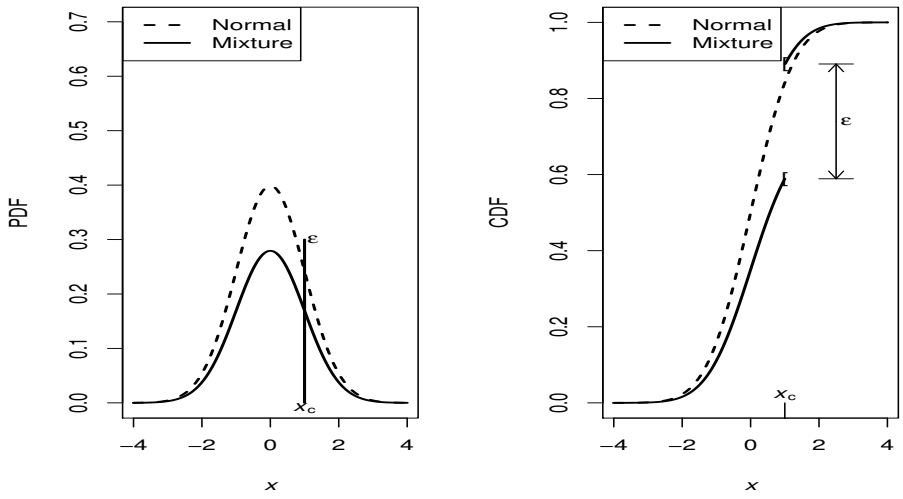


Figure 8.3. PDFs and the CDF of the ϵ-Mixture Distribution

The Kolmogorov distance is slightly less than ϵ. The Lévy distance is the length of a side of the square shown. (The square does not appear to be a square because the scales of the axes are different.)

Although by both measures, the distributions are quite "close" to each other, and increasing the value of x_c would not make these measures get larger, the effect on statistical inference about, say, the mean of the distribution could be quite large.

■

8.7 Robust Inference

Robust inference is concerned with methods that are not greatly affected by perturbations in the assumed family of distributions.

Functionals of the CDF and Estimators Based on Statistical Functions

While the cumulative distribution function is the most basic function for describing a probability distribution or a family of distributions, there are a number of other, simpler descriptors of probability distributions that are useful. Many of these are expressed as functionals of the CDF. For example, the

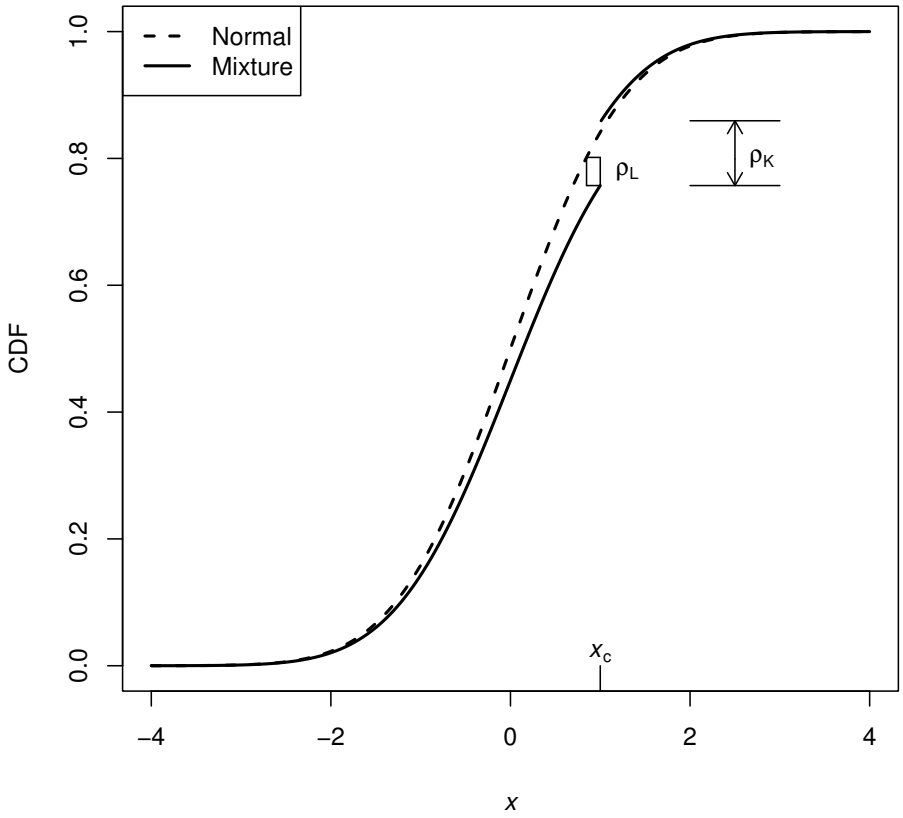


Figure 8.4. CDFs of a Standard Normal and Associated ϵ-Mixture Distribution and the Kolmogorov and Lévy Distances between Them

mean of a distribution, if it exists, may be written as the functional M of the CDF P:

$$M(P) = \int y \, \mathrm{d}P(y). \tag{8.83}$$

A natural way of estimating a distributional measure that is defined in terms of a statistical function of the CDF is to use the same statistical function on the ECDF. This leads us to plug-in estimators, as we discussed in Section 3.2.2 beginning on page 246.

Estimators based on statistical functions play major roles throughout nonparametric and semiparametric inference. They are also important in robust

statistics. In robustness studies, we first consider the sensitivity of the statistical function to perturbations in distribution functions. Statistical functions that are relatively insensitive to perturbations in distribution functions when applied to a ECDF should yield robust estimators.

These kinds of plug-in estimators should generally have good asymptotic properties *relative to the corresponding population measures* because of the global asymptotic properties of the ECDF.

Although the statistical functions we have considered have intuitive interpretations, the question remains as to what are the most useful distributional measures by which to describe a given distribution. In a simple case such as a normal distribution, the choices are obvious. For skewed distributions, or distributions that arise from mixtures of simpler distributions, the choices of useful distributional measures are not so obvious. A central concern in robust statistics is how a functional of a CDF behaves as the distribution is perturbed. If a functional is rather sensitive to small changes in the distribution, then one has more to worry about if the observations from the process of interest are contaminated with observations from some other process.

8.7.1 Sensitivity of Statistical Functions

******

One of the most interesting things about a function (or a functional) is how its value varies as the argument is perturbed. Two key properties are *continuity* and *differentiability*.

For the case in which the arguments are functions, the cardinality of the possible perturbations is greater than that of the continuum. We can be precise in discussions of continuity and differentiability of a functional Υ at a point (function) F in a domain $\mathcal{F}$ by defining another set $\mathcal{D}$ consisting of difference functions over $\mathcal{F}$; that is the set the functions $D = F_1 - F_2$ for $F_1, F_2 \in \mathcal{F}$.

Three kinds of functional differentials are defined on page 760.

Given a reference distribution P and an ϵ-mixture distribution $P_{x,\epsilon}$, a statistical function evaluated at $P_{x,\epsilon}$ compared to the function evaluated at P allows us to determine the effect of the perturbation on the statistical function. For example, we can determine the mean of the distribution with CDF $P_{x,\epsilon}$ in terms of the mean μ of the reference distribution to be $(1-\epsilon)\mu + \epsilon x$. This is easily seen by thinking of the distribution as a mixture. Formally, using the M in equation (8.83), we can write

$$\begin{aligned} M(P_{x,\epsilon}) &= \int y \, \mathrm{d}((1-\epsilon)P(y) + \epsilon \mathrm{I}_{[x,\infty[}(y)) \\ &= (1-\epsilon)\int y \, \mathrm{d}P(y) + \epsilon \int y\delta(y-x)\,\mathrm{d}y \\ &= (1-\epsilon)\mu + \epsilon x. \end{aligned} \tag{8.84}$$

For a discrete distribution we would follow the same steps using summations (instead of an integral of y times a Dirac delta function, we just have a point mass of 1 at x), and would get the same result.

The π quantile of the mixture distribution, $\Xi_\pi(P_{x,\epsilon}) = P^{-1}_{x,\epsilon}(\pi)$, is somewhat more difficult to work out. This quantile, which we will call q, is shown relative to the π quantile of the continuous reference distribution, y_π, for two cases in Figure 8.5. (In Figure 8.5, although the specifics are not important, the reference distribution is a standard normal, $\pi = 0.7$, so $y_\pi = 0.52$, and $\epsilon = 0.1$. In the left-hand graph, $x_1 = -1.25$, and in the right-hand graph, $x_2 = 1.25$.)

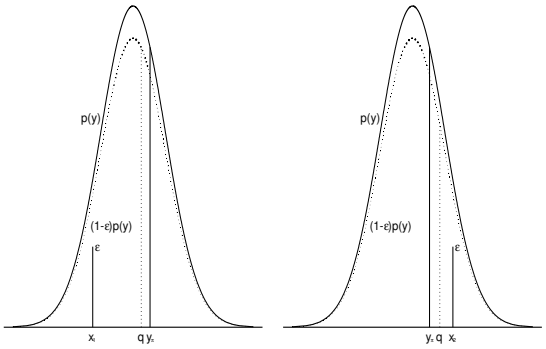


Figure 8.5. Quantile of the ϵ-Mixture Distribution

We see that in the case of a continuous reference distribution (implying P is strictly increasing),

$$P^{-1}_{x,\epsilon}(\pi) = \begin{cases} P^{-1}\left(\frac{\pi-\epsilon}{1-\epsilon}\right), & \text{for } (1-\epsilon)P(x) + \epsilon < \pi, \\ x, & \text{for } (1-\epsilon)P(x) \le \pi \le (1-\epsilon)P(x) + \epsilon, \\ P^{-1}\left(\frac{\pi}{1-\epsilon}\right), & \text{for } \pi < (1-\epsilon)P(x). \end{cases} \tag{8.85}$$

The conditions in equation (8.85) can also be expressed in terms of x and quantiles of the reference distribution. For example, the first condition is equivalent to $x < \frac{y_{\pi-\epsilon}}{1-\epsilon}$.

The Influence Function

The extent of the perturbation depends on ϵ, and so we are interested in the relative effect; in particular, the relative effect as ϵ approaches zero.

The *influence function* for the functional Υ and the CDF P, defined at x as

$$\phi_{\Upsilon,P}(x) = \lim_{\epsilon \downarrow 0} \frac{\Upsilon(P_{x,\epsilon}) - \Upsilon(P)}{\epsilon} \tag{8.86}$$

if the limit exists, is a measure of the sensitivity of the distributional measure defined by Υ to a perturbation of the distribution at the point x. The influence function is also called the influence curve, and denoted by IC.

The limit in equation (8.86) is the right-hand Gâteaux derivative of the functional Υ at P and x.

The influence function can also be expressed as the limit of the derivative of $\Upsilon(P_{x,\epsilon})$ with respect to ϵ:

$$\phi_{\Upsilon,P}(x) = \lim_{\epsilon \downarrow 0} \frac{\partial}{\partial \epsilon} \Upsilon(P_{x,\epsilon}). \tag{8.87}$$

This form is often more convenient for evaluating the influence function.

Some influence functions are easy to work out, for example, the influence function for the functional M in equation (8.83) that defines the mean of a distribution, which we denote by μ. The influence function for this functional operating on the CDF P at x is

$$\begin{aligned} \phi_{\mu,P}(x) &= \lim_{\epsilon \downarrow 0} \frac{M(P_{x,\epsilon}) - M(P)}{\epsilon} \\ &= \lim_{\epsilon \downarrow 0} \frac{(1-\epsilon)\mu + \epsilon x - \mu}{\epsilon} \\ &= x - \mu. \end{aligned} \tag{8.88}$$

We note that the influence function of a functional is a type of derivative of the functional, $\partial M(P_{x,\epsilon})/\partial \epsilon$. The influence function for other moments can be computed in the same way as the steps in equation (8.88).

Note that the influence function for the mean is unbounded in x; that is, it increases or decreases without bound as x increases or decreases without bound. Note also that this result is the same for multivariate or univariate distributions.

The influence function for a quantile is more difficult to work out. The problem arises from the difficulty in evaluating the quantile. As I informally described the distribution with CDF $P_{x,\epsilon}$, it is a mixture of some given distribution and a degenerate discrete distribution. Even if the reference distribution is continuous, the CDF of the mixture, $P_{x,\epsilon}$, does not have an inverse over the full support (although for quantiles we will write $P^{-1}_{x,\epsilon}$).

Let us consider a simple instance: a univariate continuous reference distribution, and assume $p(y_\pi) > 0$. We approach the problem by considering the PDF, or the probability mass function.

In the left-hand graph of Figure 8.5, the total probability mass up to the point y_π is $(1-\epsilon)$ times the area under the curve, that is, $(1-\epsilon)\pi$, plus the mass at x_1, that is, ϵ. Assuming ϵ is small enough, the π quantile of the ϵ-mixture distribution is the $\pi-\epsilon$ quantile of the reference distribution, or $P^{-1}(\pi-\epsilon)$. It is also the π quantile of the scaled reference distribution; that is, it is the value of the function $(1-\epsilon)p(x)$ that corresponds to the proportion π of the total probability $(1-\epsilon)$ of that component. Use of equation (8.85) directly in equation (8.86) is somewhat messy. It is more straightforward to differentiate $P^{-1}_{x_1,\epsilon}$ and take the limit as in equation (8.87). For fixed $x < y_\pi$, we have

$$\frac{\partial}{\partial\epsilon}P^{-1}\left(\frac{\pi-\epsilon}{1-\epsilon}\right) = \frac{1}{p\left(P^{-1}\left(\frac{\pi-\epsilon}{1-\epsilon}\right)\right)}\frac{(\pi-1)(1-\epsilon)}{(1-\epsilon)^2}.$$

Likewise, we take the derivatives for the other cases in equation (8.85), and then take limits. We get

$$\phi_{\Xi_\pi,P}(x) = \begin{cases} \dfrac{\pi-1}{p(y_\pi)}, & \text{for } x < y_\pi, \\ 0, & \text{for } x = y_\pi, \\ \dfrac{\pi}{p(y_\pi)}, & \text{for } x > y_\pi. \end{cases} \tag{8.89}$$

Notice that the actual value of x is not in the influence function; only whether x is less than, equal to, or greater than the quantile. Notice also that, unlike influence function for the mean, the influence function for a quantile is bounded; hence, a quantile is less sensitive than the mean to perturbations of the distribution. Likewise, quantile-based measures of scale and skewness, as in equations (1.115) and (1.116), are less sensitive than the moment-based measures to perturbations of the distribution.

The functionals L_J and M_ρ defined in equations (1.118) and (1.119), depending on J or ρ, can also be very insensitive to perturbations of the distribution.

The mean and variance of the influence function at a random point are of interest; in particular, we may wish to restrict the functional so that

$$\mathrm{E}(\phi_{\Upsilon,P}(X)) = 0$$

and

$$\mathrm{E}\left((\phi_{\Upsilon,P}(X))^2\right) < \infty.$$

8.7.2 Robust Estimators

If a distributional measure of interest is defined on the CDF as $\Upsilon(P)$, we are interested in the performance of the plug-in estimator $\Upsilon(P_n)$; specifically, we are interested in $\Upsilon(P_n) - \Upsilon(P)$. This turns out to depend crucially on the differentiability of Υ. If we assume Gâteaux differentiability, from equation (0.1.118), we can write

$$\begin{aligned}\sqrt{n}\,(\Upsilon(P_n) - \Upsilon(P)) &= \Lambda_P(\sqrt{n}(P_n - P)) + R_n \\ &= \frac{1}{\sqrt{n}} \sum_i \phi_{\Upsilon,P}(Y_i) + R_n\end{aligned}$$

where the remainder $R_n \to 0$.

We are interested in the stochastic convergence. First, we assume

$$\mathrm{E}(\phi_{\Upsilon,P}(X)) = 0$$

and

$$\mathrm{E}\left((\phi_{\Upsilon,P}(X))^2\right) < \infty.$$

Then the question is the stochastic convergence of R_n. Gâteaux differentiability does not guarantee that R_n converges fast enough. However, ρ-Hadamard differentiability, does imply that that R_n is in $\mathrm{o}_P(1)$, because it implies that norms of functionals (with or without random arguments) go to 0. We can also get that R_n is in $\mathrm{o}_P(1)$ by assuming Υ is ρ-Fréchet differentiable and that $\sqrt{n}\rho(P_n, P)$ is in $\mathrm{O}_P(1)$. In either case, that is, given the moment properties of $\phi_{\Upsilon,P}(X)$ and R_n is in $\mathrm{o}_P(1)$, we have by Slutsky's theorem (page 91),

$$\sqrt{n}\,(\Upsilon(P_n) - \Upsilon(P)) \xrightarrow{\mathrm{d}} \mathrm{N}(0, \sigma^2_{\Upsilon,P}),$$

where $\sigma^2_{\Upsilon,P} = \mathrm{E}\left((\phi_{\Upsilon,P}(X))^2\right)$.

For a given plug-in estimator based on the statistical function Υ, knowing $\mathrm{E}\left((\phi_{\Upsilon,P}(X))^2\right)$ (and assuming $\mathrm{E}(\phi_{\Upsilon,P}(X)) = 0$) provides us an estimator of the asymptotic variance of the estimator.

The influence function is also very important in leading us to estimators that are robust; that is, to estimators that are relatively insensitive to departures from the underlying assumptions about the distribution. As mentioned above, the functionals L_J and M_ρ, depending on J or ρ, can be very insensitive to perturbations of the distribution; therefore estimators based on them, called L-estimators and M-estimators, can be robust.

M-Estimators

L-Estimators

A class of L-estimators that are particularly useful are linear combinations of the order statistics. Because of the sufficiency and completeness of the order

statistics in many cases of interest, such estimators can be expected to exhibit good statistical properties.

Another class of estimators similar to the L-estimators are those based on ranks, which are simpler than order statistics. These are not sufficient – the data values have been converted to their ranks – nevertheless they preserve a lot of the information. The fact that they lose some information can actually work in their favor; they can be robust to extreme values of the data.

A functional to define even a simple linear combination of ranks is rather complicated. As with the L_J functional, we begin with a function J, which in this case we require to be strictly increasing, and also, in order to ensure uniqueness, we require that the CDF P be strictly increasing. The R_J functional is defined as the solution to the equation

$$\int J\left(\frac{P(y)+1-P(2R_J(P)-y)}{2}\right)\,\mathrm{d}P(y)=0. \tag{8.90}$$

A functional defined as the solution to this optimization problem is called an R_J *functional*, and an estimator based on applying it to a ECDF is called an R_J *estimator* or just an R-estimator.

Notes and Further Reading

Much of the material in this chapter is covered in MS2 Sections 5.1, 5.2, 5.3.

Nonparametric Statistics

There are a number of books on nonparametric methods of statistical inference. Most of the underlying theory for these methods is developed in the context of order statistics and ranks.

Failure Time Data and Survival Analysis

The analysis of failure time data has application in engineering reliability studies as well as in medical statistics. Many of the models used are parametric or semiparametric, but my brief discussion of it is included in this chapter on nonparametric methods. An indepth discussion of the theory and methods is given by Kalbfleisch and Prentice (2002).

Expansions of Functionals and Their Sensitivity to Perturbations

Small (2010)

Robust Statistics

The books by Staudte and Sheather (1990) and Huber and Ronchetti (2009) and the article by Davies and Gather (2012) provide a more complete coverage of the general topic of robust statistics.

The Influence Function

Davies and Gather (2012) discuss and give several examples of this kind of perturbation to study the sensitivity of a functional to perturbations of the CDF at a given point x.

Adaptive Procedures

Although the idea of adapting the statistical methods to the apparent characteristics of the data is appealing, there are many practical problems to contend with. Hogg (1974) and Hogg and Lenth (1984) review many of the issues and discuss several adaptive procedures for statistical inference.

Exercises

8.1. Consider the problem of estimating the function $f(x) = \theta^{-1}\mathrm{e}^{-x/\theta}\mathrm{I}_{\mathbb{R}_+}(x)$ based on a random sample of size n from a population with PDF $f(x)$. Let $\widehat{f}$ be an estimator of f that is the given functional form of f with the sample mean in place of λ.
   a) What is the bias and the variance of $\widehat{f}$ at the point x?
   b) What is the asymptotic mean squared error, AMSE, of $\widehat{f}$ at the point x?

8.2. Integrated measures in a parametric problem.
Consider the $\mathrm{U}(0,\theta)$ distribution, with θ unknown. The true probability density is $p(x) = 1/\theta$ over $(0,\theta)$ and 0 elsewhere. Suppose we have a sample of size n and we estimate the density as $\widehat{p}(x) = 1/x_{(n)}$ over $(0, x_{(n)})$ and 0 elsewhere, where $x_{(n)}$ is the maximum order statistic.
   a) Determine the integrated squared error, ISE, of $\widehat{p}(x)$.
   b) Determine (that is, write an explicit expression for) the integrated squared bias, ISB, of $\widehat{p}(x)$.
   c) Determine the mean integrated squared error, MISE, of $\widehat{p}(x)$.
   d) Determine the asymptotic mean integrated squared error, AMISE, of $\widehat{p}(x)$.

8.3. Determine the hazard function for
   (a) the Weibull(α,β) family;
   (b) the log-normal(μ,σ^2) family;
   (c) the gamma(α,β) family.

In each case, suggest a reparametrization that directly incorporates the hazard function, as in the example in the text when the θ parameter of the exponential family is replaced by $1/\lambda$.

8.4. Show that equations (8.30) and (8.32) are correct.

8.5. Prove Theorem 8.1.

0

Statistical Mathematics

Statistics is grounded in mathematics. Most of mathematics is important and it is difficult to identify the particular areas of mathematics, and at what levels, that must be mastered by statisticians. Of course, statistics is a large field, and statisticians working in different areas need different kinds and different levels of mathematical competence. One of the best general references for mathematical topics is PlanetMath, `http://planetmath.org/` (This is an outgrowth of MathWorld by Eric Weisstein.)

The purpose of this chapter is to provide some general mathematical background for the theory of probability and statistics. In Section 0.0 I start with some very basic material. This includes standard objects such as sets and various structures built onto sets. There are many standard methods we use in mathematical statistics. It may seem that many methods are ad hoc, but it is useful to identify common techniques and have a ready tool kit of methods with general applicability. There are many standard mathematical techniques that every statistician should have in a toolkit, and this section surveys several of them.

Beyond the general basics covered in Section 0.0, the statistician needs grounding in measure theory to the extent covered in Section 0.1, beginning on page 692, in stochastic calculus to the extent covered in Section 0.2, beginning on page 765, in linear algebra to the extent covered in Section 0.3, beginning on page 781, and in methods of optimization to the extent covered in Section 0.4 beginning on page 822.

The general development in this chapter is in the usual form of a mathematical development, moving from primitives to definitions to theorems, however, occasionally it is assumed that the reader already is somewhat familiar with such concepts as differentiation, integration, and mathematical expectation before we give their formal definitions. References within the chapter, therefore, may be either forward or backward. Although generally this chapter is meant to provide background for the other chapters, occasionally some material in this chapter depends on concepts from other chapters; in particu-

lar, Section 0.2 on stochastic processes depends on material in Chapter 1 on probability.

The attitudes and methods of mathematics pervade mathematical statistics. We study objects. These objects may be structures, such as groups and fields, or functionals, such as integrals, estimators, or tests. We want to understand the properties of these objects. We identify, describe, and name these properties in fixed statements, with labels, such as the "Neyman-Pearson Lemma", or the "Dominated Convergence Theorem". We identify limits to the properties of an object or boundary points on the characteristics of the object by means of "counterexamples".

Our understanding and appreciation of a particular object is enhanced by comparing the properties of the given object with similar objects. The properties of objects of the same class as the given object are stated in theorems (or "lemmas", or "corollaries", or "propositions" — unless you understand the difference, just call them all "theorems"; clearly, many otherwise competent mathematical statisticians have no idea what these English words mean). The hypotheses of the theorems define various classes of objects to which the conclusions of the theorems apply. Objects that do not satisfy all of the hypotheses of a given theorem provide us insight into these hypotheses. These kinds of objects are called counterexamples for the conclusions of the theorem. For example, the Lebesgue integral and the Riemann integral are similar objects. How are they different? First, we should look at the big picture: in the Lebesgue integral, we begin with a partitioning of the range of the function; in the Riemann integral, we begin with a partitioning of the domain of the function. What about some specific properties? Some important properties of the Lebesgue integral are codified in the Big Four Theorems: the bounded convergence theorem, Fatou's lemma, the (Lebesgue) monotone convergence theorem, and the dominated convergence theorem. None of these hold for the Riemann integral; that is, the Riemann integral provides counterexamples for the conclusions of these theorems. To understand these two types of integrals, we need to be able to prove the four theorems (they're related) for the Lebesgue integral, and to construct counterexamples to show that they do not hold for the Riemann integral. The specifics here are not as important as the understanding of the attitude of mathematics.

Notation

I must first of all point out a departure from the usual notation and terminology in regard to the real numbers. I use $\mathbb{R}$ to denote the scalar real number system in which the elements of the underlying set are singleton numbers. Much of the underlying theory is based on $\mathbb{R}$, but my main interest is usually $\mathbb{R}^d$, for some fixed positive integer d. The elements of the underlying set for $\mathbb{R}^d$ are d-tuples, or vectors. I sometimes emphasize the difference by the word "scalar" or "vector". I do not, however, distinguish in the notation for these elements from the notation for the singleton elements of $\mathbb{R}$; thus, the symbol

x may represent a scalar or a vector, and a "random variable" X may be a scalar random variable or a vector random variable.

This unified approach requires a generalized interpretation for certain functions and relational operators. Many of these functions and operators are interpreted as applying to the individual elements of a vector; for example, $|x|$, $|x|^p$, e^x, and $x < y$. If $x = (x_1, \ldots, x_d)$ and $y = (y_1, \ldots, y_d)$, then

$$|x| \stackrel{\text{def}}{=} (|x_1|, \ldots, |x_d|) \tag{0.1}$$

$$|x|^p \stackrel{\text{def}}{=} (|x_1|^p, \ldots, |x_d|^p) \tag{0.2}$$

$$\mathrm{e}^x \stackrel{\text{def}}{=} (\mathrm{e}^{x_1}, \ldots, \mathrm{e}^{x_d}) \tag{0.3}$$

and

$$x < y \iff x_1 < y_1, \ldots, x_d < y_d, \tag{0.4}$$

that is, these functions and relations are applied elementwise. For more complicated objects, such as matrices, the indicated operations may have different meanings.

There are, of course, other functions and operators that apply to combinations of the elements of a vector; for example,

$$\|x\|_p \stackrel{\text{def}}{=} \left(\sum_{i=1}^{d} |x_i|^p \right)^{1/p}. \tag{0.5}$$

This approach, however, results in some awkwardness in some contexts, such as for multiplication. If a and b are scalars, ab has a very simple meaning. Likewise, if a is a scalar and x is a vector, ax has a very simple meaning consistent with the elementwise operations defined above. Definition of multiplication of two vectors, however, is somewhat more complicated. First of all, the two operands must be vectors of the same length. There are three possibilities: the product of vectors x and y may be a scalar, a vector, or a matrix.

The scalar product is an inner product, also called the dot product, and is denoted as $\langle x, y \rangle$ or $x^{\mathrm{T}} y$. It is the sum of the elementwise products.

The vector product is the vector of elementwise products. (If we have the luxury of working only in 3-space, there is another useful vector product, often called the cross product, useful in physics.) We will rarely have occasion to use vector products of either type.

The matrix product is the outer product, denoted as xy^{T}, and defined as the matrix whose (i, j) element is $x_i y_j$.

In the following, and in all of my writing, I try to be very consistent in use of notation. Occasionally, I will mention alternative notation in common usage. A more complete coverage of the notation I use can be found beginning on page 857.

0.0 Some Basic Mathematical Concepts

We first need to develop some basic definitions and properties of sets. Given simple operations on sets, we expand the concepts to include various functions and structures over sets. The most important set and the one for which we define many functions is the set of reals, which I denote as $\mathbb{R}$ or $\mathbb{R}^d$. First, however, we begin with abstract sets, that is, sets whose elements are not restricted to lie in special mathematical structures.

In this section, I will often refer to objects or operations that are not defined in this section, for example derivatives and integrals and the associated operations. These will be defined and discussed rather carefully in Section 0.1.

In addition to understanding standard objects and their properties, we need to become familiar with the methods of mathematical statistics. There are many standard methods we use in mathematical statistics. It may seem that many methods are ad hoc, but it is useful to identify common techniques and have a ready tool kit of methods with general applicability. Monte Carlo methods play important roles in statistics, both for inference in applications and in the development of statistical theory. We discuss Monte Carlo methods in Section 0.0.7. Finally, in Section 0.0.9, I describe some examples that have general relevance and some standard mathematical techniques that every statistician should have in a toolkit.

0.0.1 Sets

We use the term *set* without a formal definition to denote a collection of things, called *elements* or *points*. If every element of a set A_2 is also in a set A_1, we say A_2 is a *subset* of A_1, and write $A_2 \subseteq A_1$. If A_2 is a subset of A_1, but A_1 is not a subset of A_2, we say A_2 is a *proper subset* of $A_1$and write $A_2 \subset A_1$.

Given sets A_1 and A_2, their *union*, written $A_1 \cup A_2$, is the set consisting of all elements that are in A_1 or A_2; and their *intersection*, written $A_1 \cap A_2$, is the set consisting of all elements that are in both A_1 and A_2. Obviously, both union and intersection operations are commutative: $A_1 \cup A_2 = A_2 \cup A_1$ and $A_1 \cap A_2 = A_2 \cap A_1$.

In working with sets, it is useful to define an *empty set*. This is the set that contains no elements. We often denote it as $\emptyset$.

The *cardinality* of a set is an indicator of how many elements the set contains. If the number of elements in a set is a finite integer, that number is the cardinality of the set. If the elements of a set can be put into a one-to-one correspondence with a sequence of positive integers, the set is said to be *countable*. If it is countable but its cardinality is not a finite integer, then the set is said to be *countably infinite*. Any interval of $\mathbb{R}$ is *uncountably infinite*. Its cardinality is said to be the cardinality of the continuum.

In any particular application, we can conceive of a set of "everything", or a "universe of discourse". In general, we call this the *universal set*. (Sometimes,

especially in applications of probability, we will call it the sample space.) If A is the universal set, then when we speak of the set A_1, we imply $A_1 \subseteq A$.

The concept of a universal set also leads naturally to the concept of the complement of a set. The *complement* of A_1, written A_1^c, is the set of all elements in the universal set A that are not in A_1, which we can also write as $A - A_1$. More generally, given the sets A_1 and A_2, we write $A_1 - A_2$ (some people write $A_1 \backslash A_2$ instead) to represent *difference* of A_1 and A_2; that is, the complement of A_2 in A_1, $A_1 - A_2 = A_1 \cap A_2^c$. If $A_2 \subseteq A_1$, the difference $A_1 - A_2$ is called the *proper difference*.

The *symmetric difference* of A_1 and A_2, written $A_1 \Delta A_2$, is the union of their differences:

$$A_1 \Delta A_2 \stackrel{\text{def}}{=} (A_1 - A_2) \cup (A_2 - A_1). \tag{0.0.1}$$

Obviously, the symmetric difference operation is commutative:

$$A_1 \Delta A_2 = A_2 \Delta A_1.$$

Two useful relationships, known as De Morgan's laws, are

$$(A_1 \cup A_2)^c = A_1^c \cap A_2^c \tag{0.0.2}$$

and

$$(A_1 \cap A_2)^c = A_1^c \cup A_2^c. \tag{0.0.3}$$

These two equations can be extended to countable unions and intersections:

$$(\cup_{i=1}^{\infty} A_i)^c = \cap_{i=1}^{\infty} A_i^c \tag{0.0.4}$$

and

$$(\cap_{i=1}^{\infty} A_i)^c = \cup_{i=1}^{\infty} A_i^c. \tag{0.0.5}$$

We often are interested in the "smallest" subset of the universal set A that has a given property. By the smallest subset we mean the intersection of all subsets of A with the given property.

Product Sets

The *cartesian product* (or *direct product* or *cross product*) of two sets A and B, written $A \times B$, is the set of all doubletons, (a_i, b_j), where $a_i \in A$ and $b_j \in B$. The set $A \times B$ is called a product set.

Obviously, $A \times B \neq B \times A$ unless $B = A$.

By convention, we have $\emptyset \times A = A \times \emptyset = \emptyset = \emptyset \times \emptyset$.

The concept of product sets can be extended to more than two sets in a natural way.

One statement of the Axiom of Choice is that the cartesian product of any non-empty collection of non-empty sets is non-empty.

Collections of Sets

Collections of sets are usually called "collections", rather than "sets". We usually denote collections of sets with upper-case calligraphic letters, e.g., $\mathcal{B}$, $\mathcal{F}$, etc.

The usual set operators and set relations are used with collections of sets, and generally have the same meaning. Thus if $\mathcal{F}_1$ is a collection of sets that contains the set A, we write $A \in \mathcal{F}_1$, and if $\mathcal{F}_2$ is also a collection of sets, we denote the collection of all sets that are in either $\mathcal{F}_1$, or $\mathcal{F}_2$ as $\mathcal{F}_1 \cup \mathcal{F}_2$.

The collection of all subsets of a given set is called the *power set* of the given set. An axiom of naive set theory postulates the existence of the power set for any given set. We denote the power set for a set S as 2^S.

Partitions; Disjoint Sets

A *partition* of a set S is a collection of disjoint subsets of S whose union is S. Partitions of sets play an important role.

A simple example is a partition of a set S that is the union two sets; $S = A_1 \cup A_2$. One partition of the union is just $\{A_1,\ A_2 - A_1\}$. Given any union of sets $\cup_{i=1} A_i$, we can obtain similar partitions. For a sequence of sets $\{A_n\}$ the following theorem gives a sequence of disjoint sets $\{D_n\}$ whose union is the same as that of $\{A_n\}$.

Theorem 0.0.1 *Let $A_1, A_2, \ldots$ be a sequence of subsets of the universal set A. Let the sequence of $\{D_n\}$ be defined as*

$$\begin{aligned} D_1 &= A_1 \\ D_n &= A_n - \cup_{i=1}^{n-1} A_{i-1} \quad \text{for } n = 2, 3, \ldots \end{aligned} \tag{0.0.6}$$

Then the sets in the sequence $\{D_n\}$ are disjoint, and

$$\cup_{i=1}^{\infty} D_i = \cup_{i=1}^{\infty} A_i.$$

Proof.
Let $\{D_n\}$ be as given; then $D_n = A_n \cap \left(\cap_{i=1}^{n-1} A_{i-1}^{\mathrm{c}}\right)$ and $D_n \subseteq A_n$. Now let n and m be distinct integers. Without loss, suppose $n < m$. Then

$$\begin{aligned} D_n \cap D_m &\subseteq A_n \cap D_m \\ &= A_n \cap A_m \cap \cdots \cap A_n^{\mathrm{c}} \cap \cdots \\ &= A_n \cap A_n^{\mathrm{c}} \cap \cdots \\ &= \emptyset. \end{aligned}$$

Also, because $D_i \subseteq A_i$,

$$\cup_{i=1}^{\infty} D_i \subseteq \cup_{i=1}^{\infty} A_i.$$

Now, let $x \in \cup_{i=1}^{\infty} A_i$. Then x must belong to a least one A_i. Let n be the smallest integer such that $x \in A_n$. Then $x \in D_n$ (by the definition of D_n), and so $x \in \cup_{i=1}^{\infty} D_i$. Hence,

$$\cup_{i=1}^{\infty} A_i \subseteq \cup_{i=1}^{\infty} D_i.$$

Because each is a subset of the other,

$$\cup_{i=1}^{\infty} D_i = \cup_{i=1}^{\infty} A_i.$$

∎

A partition of the full universal set S is formed by $\{D_n\}$ and $D_0 = S - \cup_{i=1}^{\infty} D_i$.

Sequences of nested intervals are important. The following corollary, which follows immediately from Theorem 0.0.1, applies to such sequences.

Corollary 0.0.1.1
Let $\{A_n\}$ be a sequence of sets such that $A_1 \subseteq A_2 \subseteq \ldots$. Let the sequence of $\{D_n\}$ be defined as

$$D_n = A_{n+1} - A_n \quad \text{for } n = 1, 2, \ldots \tag{0.0.7}$$

Then the sets in the sequence $\{D_n\}$ are disjoint, and

$$\cup_{i=1}^{\infty} D_i = \cup_{i=1}^{\infty} A_i.$$

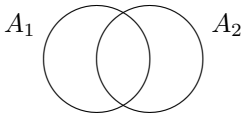
Notice that $\cup_{j=1}^{i} D_j = A_i$. (Notice also that there is an offset in the indices of the sequence (0.0.7) from those of the sequence (0.0.6).)

These ideas of partitioning a union of a sequence of sets may also be applied to an intersection of a sequence by using De Morgan's laws. For the intersection, instead of an increasing sequence, $A_1 \subseteq A_2 \subseteq \ldots$, our interest is usually in a decreasing sequence $A_1 \supset A_2 \supset \ldots$.

Another useful partition of the union $S = A_1 \cup A_2$ uses three sets:

$$\{A_1 - (A_1 \cap A_2),\ A_2 - (A_1 \cap A_2),\ (A_1 \cap A_2)\}. \tag{0.0.8}$$

We call this partitioning "disjointification", and expression (0.0.8) is called the "inclusion-exclusion formula".



Disjointification

Disjointification leads to the inclusion-exclusion formula for measures of sets that has common applications in deriving properties of measures (see, for example, equation (0.1.11) on page 707).

Covering Sets

A collection of sets $\mathcal{A}$ is said to *cover* a set S if $S \subseteq \cup_{A_i \in \mathcal{A}} A_i$.

Given a collection $\mathcal{A} = \{A_1, A_2, \ldots\}$ that covers a set S, a partition of S can be formed by removing some of the intersections of sets in $\mathcal{A}$. For example, if $S \subseteq A_1 \cup A_2$, then $\{A_1 \cap S,\ (A_2 \cap S) - (A_1 \cap A_2)\}$ is a partition of S.

It is often of interest to determine the "smallest" partition of the universal set formed by sets in a given collection that does not necessarily cover the universal set. (If the number of sets is finite, the smallest partition has the smallest number of sets; otherwise, we may be able to give meaning to "smallest" in terms of intersections of collections of sets. In some cases, there is no reasonable interpretation of "smallest".) For example, consider the collection $\mathcal{A} = \{A_1, A_2\}$. If neither A_1 nor A_2 is a subset of the other, then the partition

$$\{A_1 \cap A_2,\ A_1 - A_2,\ A_2 - A_1,\ (A_1 \cup A_2)^{\mathrm{c}}\}$$

consists of the "smallest" collection of subsets that can be identified with operations on A_1 and A_2, and whose union is the universal set.

If $A_1 \subseteq A_2$, then $A_1 - A_2 = \emptyset$, and so the smallest partition is

$$\{A_1,\ A_2 - A_1,\ A_2^{\mathrm{c}}\}.$$

Ordered Sets

A set A is said to be *partially ordered* if there exists a relation $\leq$ on $A \times A$ such that:

- $\forall a \in A,\ a \leq a$ (it is reflexive)
- for $a, b, c \in A,\ a \leq b,\ b \leq c \Rightarrow a \leq c$ (it is transitive)
- for $a, b \in A,\ a \leq b,\ b \leq a \Rightarrow a = b$ (it is antisymmetric)

The relation is called an *ordering*. A set together with a partial ordering is called a *poset*.

A set A is called *linearly ordered* (or *totally ordered*) if it is partially ordered and every pair of elements $a, b \in A$ can be compared with each other by the partial ordering relation. In that case, the relation is called a *linear ordering* or *total ordering*.

The real numbers $\mathbb{R}$ are linearly ordered using the common inequality relation. This relation can be used elementwise to define a partial ordering in a set $S \in \mathbb{R}^2$. Such an ordering, however, is not a linear ordering because, for example, $a = (1, 2)$ and $b = (2, 1)$ cannot be compared by that ordering.

A set A is called *well-ordered* if it is an ordered set for which every nonempty subset contains a smallest element. By the Axiom of Choice, every set (even $\mathbb{R}^d$) can be well-ordered. The positive integers are well-ordered by the usual inequality relation, but neither the set of all integers or the reals are well-ordered by the usual inequality relation.

For structures built on sets, such as a field (Definition 0.0.3), we may define different kinds of orderings that preserve the structure, as in Definition 0.0.4, for example.

0.0.2 Sets and Spaces

In Section 0.0.9 beginning on page 676, we discussed some basics of sets and operations on sets. Now we consider mathematical structures that are built on sets together with other objects or methods such as an operation on the given set or on subsets of the set. We often refer to these structures as "spaces".

Spaces

In any application it is generally useful to define some "universe of discourse" that is the set of all elements that will be considered in a given problem. Given a universe or universal set, which we often denote by the special symbol Ω (note the font), we then define various mathematical structures on Ω. These structures, or "spaces", are formed by specifying certain types of collections of subsets of Ω and/or by defining operations on the elements of Ω or on the subsets in the special collection of subsets. In probability and statistics, we will call the universal set the *sample space*.

Some of the general structures that we will find useful are *topological spaces*, which are defined in terms of the type of collection of subsets of the universal set, and *metric spaces* and *linear spaces*, which are defined in terms of operations on elements of the universal set. We will discuss these below, and then in Section 0.0.5, we will discuss some properties of the special spaces in which the universal set is the set of real numbers. In Section 0.1, we will discuss various types of collections of subsets of the universal set, and then for a particular type of collection, called a σ-field, we will discuss a special type of space, called a *measurable space*, and then, with the addition of a real-valued set function, we will define a *measure space*. A particular type of measure space is a *probability space*.

Topologies

One of the simplest structures based on the nonempty universal set Ω is a *topological space* or a *topology*, which is formed by any collection $\mathcal{T}$ of subsets of Ω with the following properties:

(t_1) $\emptyset, \Omega \in \mathcal{T}$, and
(t_2) $A, B \in \mathcal{T} \Rightarrow A \cap B \in \mathcal{T}$, and
(t_3) $\mathcal{A} \subseteq \mathcal{T} \Rightarrow \cup\{A : A \in \mathcal{A}\} \in \mathcal{T}$.

We denote a topological space by a double of the form $(\Omega, \mathcal{T})$.

We may use the term "topology" to denote either the space or the collection of subsets that defines it.

Properties of Ω that can be expressed in terms of a topology are called its *topological properties*. Without imposing any additional structure on a topological space, we can define several useful concepts, but because the collection of subsets that define a topology is arbitrary, many terms that relate to a

topology are too general for our use in developing a theory of probability for real-valued random variables.

Given a topological space $(\Omega, \mathcal{T})$, we can define a *subspace topology* as any set $S \subseteq \Omega$ together with the collection of subsets

$$\mathcal{T}_S = \{S \cap U \,|\, U \in \mathcal{T}\}.$$

Open and Closed Sets in a Topology

Let $(\Omega, \mathcal{T})$ be a topological space. Members of $\mathcal{T}$ are called *open sets*. A set $A \subseteq \Omega$ is said to be *closed* iff $\Omega \cap A^c \in \mathcal{T}$. Notice that this definition means that some sets, for example, $\emptyset$ and Ω, are both open and closed. Such sets are sometimes said to be *clopen*. Also, notice that some subsets of Ω are neither open nor closed.

For the set $A \subseteq \Omega$, the *closure* of A is the set

$$\overline{A} = \cap\{B \,:\, B \text{ is closed, and } A \subseteq B \subseteq \Omega\}.$$

(Notice that every $y \in A$ is a point of closure of A, and that A is closed iff $A = \overline{A}$.) For the set $A \subseteq \Omega$, the *interior* of A is the set

$$A^\circ = \cup\{U \,:\, U \text{ is open, and } U \subseteq A\}.$$

The *boundary* of the set $A \subseteq \Omega$ is the set

$$\partial A = \overline{A} \cap \overline{A^c}.$$

A set $A \subseteq \Omega$ such that $\overline{A} = \Omega$ is said to be *dense in* Ω.

A set is said to be *separable* if it contains a countable dense subset. Obviously, any countable set is itself separable. (This term is usually applied only to the universal set.)

Any subcollection $A_1, A_2, \ldots$ of $\mathcal{T}$ such that $\cup_i A_i = \Omega$ is called an *open cover* of Ω. A topological space for which each open cover contains a finite open cover is said to be *compact*. A set A in a topological space is said to be *compact* if each collection of open sets that covers A contains a finite subcollection of open sets that covers A.

The following properties of unions and intersections of open and closed sets are easy to show from the definitions:

- The intersection of a finite collection of open sets is open.
- The union of a countable collection of open sets is open.
- The union of a finite collection of closed sets is closed.
- The intersection of a countable collection of closed sets is closed.

Point Sequences in a Topology

A sequence $\{x_n\}$ in the topological space $(\Omega, \mathcal{T})$ is said to converge to the point x, or to have a limit x, if given any open set T containing x, there is an integer N such that $x_n \in T \; \forall n \geq N$.

A point x is said to be an *accumulation point* or *cluster point* of the sequence $\{x_n\}$ if given any open set T containing x and any integer N, there is an integer $n \geq N \;\ni x_n \in T$. This means that x is an accumulation point if $\{x_n\}$ has a subsequence that converges to x. In an arbitrary topological space, however, it is not necessarily the case that if x is an accumulation point of $\{x_n\}$ that there is a subsequence of $\{x_n\}$ that converges to x.

Neighborhoods Defined by Open Sets

Given a topological space $(\Omega, \mathcal{T})$, a *neighborhood of a point* $\omega \in \Omega$ is any set $U \in \mathcal{T}$ such that $x \in U$. Notice that Ω is a neighborhood of each point.

The space $(\Omega, \mathcal{T})$ is called a *Hausdorff space* iff each pair of distinct points of Ω have disjoint neighborhoods. For $x \in \Omega$ and $A \subseteq \Omega$, we say that x is a *limit point* of A iff for each neighborhood U of x, $U \cap \{x\}^{\mathrm{c}} \cap A \neq \emptyset$.

The topological space $(\Omega, \mathcal{T})$ is said to be *connected* iff there do not exist two disjoint open sets A and B such that $A \cup B = \Omega$. We can also speak of a subset of Ω as being connected, using this same condition.

In a space endowed with a *metric*, which we define below, open sets can be defined by use of the metric. The corresponding topology can then be defined as the collection of all open sets according to the definition of openness in that context, and the topological properties follow in the usual way. In a metric topological space, we can also define the topological properties in terms of the metric. These alternate definitions based on a metric are the ones we will use for sets of real numbers.

Metrics

A useful structure can be formed by introduction of a function that maps the product set $\Omega \times \Omega$ into the nonnegative reals. It is called a metric.

Definition 0.0.1 (metric)
Given a space Ω, a *metric* over Ω is a function ρ such that for $x, y, z \in \Omega$

- $\rho(x, y) \geq 0$
- $\rho(x, y) = 0$ if and only if $x = y$
- $\rho(x, y) = \rho(y, x)$
- $\rho(x, y) \leq \rho(x, z) + \rho(z, x)$.

∎

The structure (Ω, ρ) is called a *metric space*.

A common example of a metric space is the set $\mathbb{R}$ together with $\rho(x, y) = |x - y|$, where $|\cdot|$ denotes the ordinary absolute value.

The concept of a metric allows us to redefine the topological properties introduced above in terms of the metric. The definitions in terms of a metric are generally more useful, and also a metric allows us to define additional important properties, such as continuity. Rather than define special sets for general metric spaces here, we will discuss these sets and their properties in the context of $\mathbb{R}$ in Section 0.0.5.

Neighborhoods Defined by Metrics

We have defined neighborhoods in general topological spaces, but the concept of a metric allows us to give a more useful definition of a *neighborhood of a point* in a set. For a point $x \in \Omega$, a metric ρ on Ω, and any positive number ϵ, an *ϵ-neighborhood* of x, denoted by $\mathcal{N}_\rho(x, \epsilon)$, is the set of $y \in \Omega$ whose distance from x is less than ϵ; that is,

$$\mathcal{N}_\rho(x, \epsilon) \stackrel{\text{def}}{=} \{y \ : \ \rho(x, y) < \epsilon\}. \tag{0.0.9}$$

Notice that the meaning of a neighborhood depends on the metric, but in any case it is an open set, in the sense made more precise below. Usually, we assume that a metric is given and just denote the neighborhood as $\mathcal{N}(x, \epsilon)$ or, with the size in place of the metric, as $\mathcal{N}_\epsilon(x)$. We also often refer to some (unspecified) ϵ-neighborhood of x just as a "neighborhood" and denote it as $\mathcal{N}(x)$.

The concept of a neighborhood allows us to give a more meaningful definition of open sets and to define such things as continuity. These definitions are consistent with the definitions of the same concepts in a general topological space, as discussed above.

Theorem 0.0.2 *Every metric space is a Hausdorff space.*

Proof. Exercise. ∎

Open and Closed Sets in a Metric Space

The specification of a topology defines the open sets of the structure and consequently neighborhoods of points. It is often a more useful approach, however, first to define a metric, then to define neighborhoods as above, and finally to define open sets in terms of neighborhoods. In this approach, a subset G of Ω is said to be *open* if each member of G has a neighborhood that is contained in G.

Note that with each metric space (Ω, ρ), we can associate a topological space $(\Omega, \mathcal{T})$, where $\mathcal{T}$ is the collection of open sets in (Ω, ρ) that are defined

in terms of the metric. The topology provides the definition of a closed set, as above; that is, a set $A \subseteq \Omega$ is said to be *closed* iff $\Omega \cap A^c \in \mathcal{T}$, where $\mathcal{T}$ is the collection of open sets defined in terms of the metric. As with the definitions above for general topological spaces, some sets are both open and closed, and such sets are said to be *clopen*.

We note that $(\mathbb{R}, \rho)$ is a Hausdorff space because, given $x, y \in \mathbb{R}$ and $x \neq y$ we have $\rho(x, y) > 0$ and so $\mathcal{N}(x, \rho(x, y)/2)$ and $\mathcal{N}(y, \rho(x, y)/2)$ are disjoint open sets.

We also note that $\mathbb{R}$ is connected, as is any interval in $\mathbb{R}$. (Connectedness is a topological property that is defined on page 623.)

We will defer further discussion of openness and related concepts to page 645 in Section 0.0.5 where we discuss the real number system.

Relations and Functions

A *relation* is a set of doubletons, or pairs of elements; that is, a relation is a subset of a cartesian product of two sets. We use "relation" and "mapping" synonymously.

A *function* is a relation in which no two different pairs have the same first element.

To say that f is a function from Ω to Λ, written

$$f : \Omega \mapsto \Lambda,$$

means that for every $\omega \in \Omega$ there is a pair in f whose first member is ω. We use the notation $f(\omega)$ to represent the second member of the pair in f whose first member is ω, and we call ω the argument of the function. We call Ω the domain of the function and we call $\{\lambda | \lambda = f(\omega) \text{ for some } \omega \in \Omega\}$ the range of the function.

Variations include functions that are *onto*, meaning that for every $\lambda \in \Lambda$ there is a pair in f whose second member is λ; and functions that are *one-to-one*, often written as $1 : 1$, meaning that no two pairs have the same second member. A function that is one-to-one and onto is called a *bijection*.

A function f that is one-to-one has an inverse, written f^{-1}, that is a function from Λ to Ω, such that if $f(\omega_0) = \lambda_0$, then $f^{-1}(\lambda_0) = \omega_0$.

If $(a, b) \in f$, we may write $a = f^{-1}(b)$, although sometimes this notation is restricted to the cases in which f is one-to-one. If f is not one-to-one and if the members of the pairs in f are reversed, the resulting relation is not a function. We say f^{-1} does not exist; yet for convenience we may write $a = f^{-1}(b)$, with the meaning above.

If $A \subseteq \Omega$, the *image* of A, denoted by $f[A]$, or just by $f(A)$, is the set of all $\lambda \in \Lambda$ for which $\lambda = f(\omega)$ for some $\omega \in \Omega$. (The notation $f[A]$ is preferable, but we will often just use $f(A)$.) Similarly, if $\mathcal{C}$ is a collection of sets (see below), the notation $f[\mathcal{C}]$ denotes the collection of sets $\{f[C] : C \in \mathcal{C}\}$.

For the function f that maps from Ω to Λ, Ω is called the *domain* of the function and and $f[\Omega]$ is called the *range* of the function.

For a subset B of Λ, the *inverse image* or the *preimage* of B, denoted by $f^{-1}[B]$, or just by $f^{-1}(B)$, is the set of all $\omega \in \Omega$ such that $f(\omega) \in B$. The notation f^{-1} used in this sense must not be confused with the inverse function f^{-1} (if the latter exists). We use this notation for the inverse image whether or not the inverse of the function exists.

We also write $f[f^{-1}[B]]$ as $f \circ f^{-1}[B]$. The set $f[f^{-1}[B]]$ may be a proper subset of B; that is, there may be an element λ in B for which there is no $\omega \in \Omega$ such that $f(\omega) = \lambda$. In this case the inverse image of a set may not generate the set. If f is bijective, then $f[f^{-1}[B]] = B$.

We will discuss functions, images, and preimages further in Section 0.1.2, beginning on page 701.

Continuous Functions

A function f from the metric space Ω with metric ρ to the metric space Λ with metric τ is said to be *continuous* at the point $\omega_0 \in \Omega$ if for any $\epsilon > 0$ there is a $\delta > 0$ such that f maps $\mathcal{N}_\rho(\omega_0, \epsilon)$ into $\mathcal{N}_\tau(f(\omega_0), \delta)$. Usually, we assume that the metrics are given and, although they may be different, we denote the neighborhood without explicit reference to the metrics. Thus, we write $f[\mathcal{N}(\omega_0, \epsilon)] \subseteq \mathcal{N}(f(\omega_0), \delta)$.

We will discuss various types of continuity of real-valued functions over real domains in Section 0.1.5 beginning on page 720.

Sequences of Sets; lim sup and lim inf

De Morgan's laws (0.0.2) and (0.0.3) express important relationships between unions, intersections, and complements.

Two important types of unions and intersections of sequences of sets are called the *lim sup* and the *lim inf* and are defined as

$$\limsup_n A_n \stackrel{\text{def}}{=} \cap_{n=1}^{\infty} \cup_{i=n}^{\infty} A_i \tag{0.0.10}$$

and

$$\liminf_n A_n \stackrel{\text{def}}{=} \cup_{n=1}^{\infty} \cap_{i=n}^{\infty} A_i. \tag{0.0.11}$$

We sometimes use the alternative notation A^* or $\overline{\lim}_n$ for lim sup:

$$A^* \stackrel{\text{def}}{=} \overline{\lim}_n A_n \stackrel{\text{def}}{=} \limsup_n A_n, \tag{0.0.12}$$

and A_* or $\underline{\lim}_n$ for lim inf:

$$A_* \stackrel{\text{def}}{=} \underline{\lim}_n A_n \stackrel{\text{def}}{=} \liminf_n A_n. \tag{0.0.13}$$

We define convergence of a sequence of sets in terms of lim sup and lim inf.

The sequence of sets $\{A_n\}$ is said to *converge* if

$$\limsup_n A_n = \liminf_n A_n, \tag{0.0.14}$$

and this set is said to be the *limit* of the sequence, written simply as $\lim_n A_n$.

A sequence of sets $\{A_n\}$ is said to be *increasing* if

$$A_n \subseteq A_{n+1} \, \forall \, n, \tag{0.0.15}$$

and is said to be *decreasing* if

$$A_{n+1} \subseteq A_n \, \forall \, n. \tag{0.0.16}$$

In either case, the sequence is said to be *monotone*.

An increasing sequence $\{A_n\}$ converges to $\cup_{n=1}^{\infty} A_n$.

A decreasing sequence $\{A_n\}$ converges to $\cap_{n=1}^{\infty} A_n$.

Some Basic Facts about lim sup and lim inf

Two simple relationships that follow immediately from the definitions:

$$\limsup_n A_n \subseteq \cup_{i=k}^{\infty} A_i \quad \forall k \geq 1 \tag{0.0.17}$$

and

$$\cap_{i=k}^{\infty} A_i \subseteq \liminf_n A_n \quad \forall k \geq 1, \tag{0.0.18}$$

which of course leads to

$$\liminf_n A_n \subseteq \limsup_n A_n. \tag{0.0.19}$$

To see the latter fact directly, consider any $\omega \in \liminf_n A_n$:

$$\omega \in \cup_{n=1}^{\infty} \cap_{i=n}^{\infty} A_i \quad \Longleftrightarrow \quad \exists n \text{ such that } \forall i \geq n, \; \omega \in A_i,$$

so $\omega \in \limsup_n A_n$.

By De Morgan's laws, we have the useful relationships between lim sup and lim inf and sequences of complementary sets:

$$\left(\liminf_n A_n\right)^{\mathrm{c}} = \limsup_n A_n^{\mathrm{c}} \tag{0.0.20}$$

and

$$\left(\limsup_n A_n\right)^{\mathrm{c}} = \liminf_n A_n^{\mathrm{c}}. \tag{0.0.21}$$

We can interpret lim sup and lim inf in intuitive terms. From the definition $\limsup_n A_n = \cap_{n=1}^{\infty} \cup_{i=n}^{\infty} A_i$, we have immediately

$$\limsup_n A_n = \{\omega \,|\, \omega \in A_i \text{ for infinitely many } i\}, \tag{0.0.22}$$

and

$$\omega \in \limsup_n A_n \iff \forall n \, \exists i \geq n \ni \omega \in A_i. \tag{0.0.23}$$

From the definition $\liminf_n A_n = \cup_{n=1}^{\infty} \cap_{i=n}^{\infty} A_i$, similarly we have

$$\liminf_n A_n = \{\omega \,|\, \omega \in A_i \text{ for all but a finite number of } i\}, \tag{0.0.24}$$

and

$$\omega \in \liminf_n A_n \iff \exists n \ni \forall i \geq n, \omega \in A_i. \tag{0.0.25}$$

While at first glance, equations (0.0.22) and (0.0.24) may seem to say the same thing, they are very different, and in fact characterize lim sup and lim inf respectively. They could therefore be used as definitions of lim sup and lim inf. Exercise 0.0.3 asks you to prove these two equations.

Examples

Example 0.0.1 (Alternating-constant series)
Consider the alternating-constant series of abstract sets:

$$A_{2n} = B \text{ and } A_{2n+1} = C.$$

Then

$$\limsup_n A_n = B \cup C$$

and

$$\liminf_n A_n = B \cap C.$$

∎

Example 0.0.2 (Alternating series of increasing and decreasing intervals)

Now let the sample space be $\mathbb{R}$, and consider the intervals

$$A_{2n} =]-n, n[\text{ and } A_{2n+1} =]0, 1/n[.$$

Then

$$\limsup_n A_n = \mathbb{R}$$

and

$$\liminf_n A_n = \emptyset.$$

∎

Example 0.0.3
Now, again in $\mathbb{R}$, consider the sequence of intervals

$$A_n = \begin{cases} \left]\frac{1}{n}, \ \frac{3}{4} - \frac{1}{n}\right[& \text{for } n = 1, 3, 5, \ldots \\ \left]\frac{1}{4} - \frac{1}{n}, \ 1 + \frac{1}{n}\right[& \text{for } n = 2, 4, 6, \ldots \end{cases}$$

In this case,

$$\limsup_n A_n =]0, 1],$$

and

$$\liminf_n A_n = \left[\frac{1}{4}, \frac{3}{4}\right[.$$

■

0.0.3 Binary Operations and Algebraic Structures

In a given set S we may find it useful to define a *binary operation*; that is, a way of combining two elements of the set to form a single entity. If $x, y \in S$, we may denote the binary operation as $\circ$ and we denote the result of the combination of x and y under $\circ$ as $x \circ y$.

Given a set S and a binary operation $\circ$ defined on it, various properties of the operation may be of interest:

closure We say S is *closed* wrt $\circ$ iff

$$x, y \in S \Longrightarrow x \circ y \in S.$$

commutativity We say $\circ$ defined in S is *commutative* iff

$$x, y \in S \Longrightarrow x \circ y = y \circ x.$$

associativity We say $\circ$ defined in S is *associative* iff

$$x, y, z \in S \Longrightarrow x \circ (y \circ z) = (x \circ y) \circ z.$$

Groups

One of the most useful algebraic structures is a *group*, which is a set and an operation $(S, \circ)$ with special properties.

Definition 0.0.2 (group)
Let S be a nonempty set and let $\circ$ be a binary operation. The structure $(S, \circ)$ is called a *group* if the following conditions hold.

- $x_1, x_2 \in S \Rightarrow x_1 \circ x_2 \in S$ (closure);
- $\exists e \in S \ni \forall x \in S, \ e \circ x = x$ (identity);

- $\forall\, x \in S\ \exists\, x^{-1} \in S \ni x^{-1} \circ x = e$ (inverse);
- $x_1, x_2, x_3 \in S \Rightarrow x_1 \circ (x_2 \circ x_3) = (x_1 \circ x_2) \circ x_3$ (associativity). ∎

If a group contains only one element, it is called a *trivial group*, and obviously is not of much interest.

Notice that the binary operation need not be commutative, but some specific pairs commute under the operation. In particular, we can easily see that $x \circ e = e \circ x$ and $x \circ x^{-1} = x^{-1} \circ x$ (exercise). We can also see that e is unique, and for a given x, x^{-1} is unique (exercise).

If the binary operation in the group $(S, \circ)$ is commutative, then the group is called a *commutative group* or an *Abelian group*.

A simple example of an Abelian group is the set of all real numbers together with the binary operation of ordinary addition.

We often denote a group by a single symbol, say $\mathcal{G}$, for example, and we use the phrase $x \in \mathcal{G}$ to refer to an element in the set that is part of the structure $\mathcal{G}$; that is, if $\mathcal{G} = (S, \circ)$, then $x \in \mathcal{G} \Leftrightarrow x \in S$.

A very important group is formed on a set of transformations.

Example 0.0.4 Group of bijections
Let X be a set and let $\mathcal{G}$ be a set of bijective functions g on X. For $g_1, g_2 \in \mathcal{G}$, let $g_1 \circ g_2$ represent function composition; that is, for $x \in X$, $g_1 \circ g_2(x) = g_1(g_2(x))$. We see that

- $g_1 \circ g_2$ is a bijection on X, so $\mathcal{G}$ is closed with respect to $\circ$;
- the function $g_e(x) = x$ is a bijection, and $\forall\, g \in \mathcal{G}$, $g_e \circ g = g$ (identity);
- $\forall\, g \in \mathcal{G}\ \exists\, g^{-1} \in \mathcal{G} \ni g^{-1} \circ g = g_e$ (inverse);
- $g_1, g_2, g_3 \in \mathcal{G} \Rightarrow g_1 \circ (g_2 \circ g_3) = (g_1 \circ g_2) \circ g_3$ (associativity); that is, $\forall\, x \in X$, either grouping of the operations is $g_1(g_2(g_3(x)))$.

Therefore, $(\mathcal{G}, \circ)$ is a group. The group of bijections in general is not Abelian. ∎

In this Example 0.0.4, we began with bijective functions and showed that they formed a group. We could have begun with a group of functions (and obviously they must be onto functions), and showed that they must be 1:1 (Exercise 0.0.5).

Subgroups and Generating Sets

Any subset of the set on which the group is defined that is closed and contains the identity and all inverses forms a group with the same operation as the original group. This subset together with the operation is called a *subgroup*. We use the standard terminology of set operations for operations on groups.

A set G_1 together with an operation $\circ$ defined on G_1 *generates* a group $\mathcal{G}$ that is the smallest group $(G, \circ)$ such that $G_1 \subseteq G$. If $\mathcal{G}_1$ and $\mathcal{G}_2$ are groups over G_1 and G_2 with a common operation $\circ$, the group generated by G_1 and G_2 is $(G, \circ)$, where G is the smallest set containing G_1 and G_2 so that $(G, \circ)$

is a group. Notice that the G may contain elements that are in neither G_1 nor G_2.

Homomorphisms

It is often of interest to consider relationships between two groups. The simplest and most useful is a *morphism*, or *homomorphism*. (The two words are synonymous; I will generally use the latter.) Given two groups $\mathcal{G} = (G, \circ)$ and $\mathcal{G}^* = (G^*, \diamond)$, a homomorphism from $\mathcal{G}$ to $\mathcal{G}^*$ is a function f from G to G^* such that for $g_1, g_2 \in G$,

$$f(g_1 \circ g_2) = f(g_1) \diamond f(g_2).$$

Often in applications, G and G^* are sets of the same kind of objects, say functions for example, and the operations $\circ$ and $\diamond$ are the same, say function composition.

If the homomorphism from $\mathcal{G}$ to $\mathcal{G}^*$ is a bijection, then the homomorphism is called an *isomorphism*, and since it has an inverse, we say the two groups are isomorphic. Isomorphic groups of transformations are the basic objects underlying the concept of equivariant and invariant statistical procedures.

Structures with Two Binary Operators

Often it is useful to define two different binary operators, say $+$ and $\circ$ over the same set. An important type of relationship between the two operators is called distributivity:

distributivity We say $\circ$ is *distributive* over $+$ in S iff

$$x, y, z \in S \Longrightarrow x \circ (y + z) = (x \circ y) + (x \circ z) \in S.$$

Another very useful algebraic structure is a *field*, which is a set and two binary operations $(S, +, \circ)$ with special properties.

Definition 0.0.3 (field)
Let S be a set with two distinct elements and let $+$ and $\circ$ be binary operations. The structure $(S, +, \circ)$ is called a *field* if the following conditions hold.

(f_1) $x_1, x_2 \in S \Rightarrow x_1 + x_2 \in S$ (closure of $+$);
(f_2) $\exists 0 \in S \ni \forall x \in S,\ 0 + x = x$ (identity for $+$);
(f_3) $\forall x \in S \exists -x \in S \ni -x \circ x = e$ (inverse wrt $+$);
(f_4) $x_1, x_2 \in S \Rightarrow x_1 + x_2 = x_2 + x_1$ (commutativity of $+$);
(f_5) $x_1, x_2, x_3 \in S \Rightarrow x_1 \circ (x_2 \circ x_3) = (x_1 \circ x_2) \circ x_3$ (associativity of $+$).
(f_6) $x_1, x_2 \in S \Rightarrow x_1 \circ x_2 \in S$ (closure of $\circ$);
(f_7) $\exists 1 \in S \ni \forall x \in S,\ 1 \circ x = x$ (identity for $\circ$);
(f_8) $\forall x \neq 0 \in S \exists x^{-1} \in S \ni x^{-1} \circ x = 1$ (inverse wrt $\circ$);
(f_9) $x_1, x_2 \in S \Rightarrow x_1 \circ x_2 = x_2 \circ x_1$ (commutativity of $\circ$);

(f_{10}) $x_1, x_2, x_3 \in S \Rightarrow x_1 \circ (x_2 \circ x_3) = (x_1 \circ x_2) \circ x_3$ (associativity of $\circ$).
(f_{11}) $x_1, x_2, x_3 \in S \Rightarrow x_1 \circ (x_2 + x_3) = (x_1 \circ x_2) + (x_1 \circ x_3)$ (distributivity of $\circ$ over $+$).

∎

Although a field is a structure consisting of a set and two binary operations, $(S, +, \circ)$, we often refer to the field by the name of the set, S, in this case.

Notice that the word "field" is also used in a slightly different sense to refer to a structure consisting of a collection of sets and the unary operation of set complementation and the binary operation of set union (see Definition 0.1.3 on page 693).

For x_1, x_2 in the field $(S, +, \circ)$, we will adopt the notation $x_1 - x_2$ to mean the element $d \in S$, such that $d + x_2 = x_1$, and if $x_2 \neq 0$, we will adopt the notation x_1/x_2 to mean the element $q \in S$, such that $qx_2 = x_1$.

If x is an element in a field $(S, +, \circ)$, and n is an integer (not necessarily an element of the field), then by the expression nx we mean "$x + \cdots + x$" n times if n is positive, we mean "0" if n is zero, and we mean "$-x - \cdots - x$" n times if n is negative. Thus, we have a multiplication-type operation between an integer and any element of a field, and the operation is closed within the field. Also, by the expression x^n we mean "$x \circ \cdots \circ x$" n times if n is positive, we mean "1" if n is zero, and we mean "$(1/x) \circ \cdots \circ (1/x)$" n times if n is negative.

A *ring* is a structure having all properties of a field except, possibly, no inverses wrt the second operation, that is, lacking property (f_8), and possibly having only one element (that is, a ring can be trivial). Sometimes a ring in which the second operation is commutative (as we have required and as is always required for a field) is called a *commutative ring*, and a structure that may possibly lack that commutativity is called a ring.

The concept of homomorphisms between fields or between rings follows similar development as that of homomorphisms between groups.

Examples

The set of integers together with the operations of ordinary addition and multiplication form a ring, but do not form a field.

The set of rational numbers (which are numbers that can be represented in the form a/b, where a and b are integers) is a field. Other common examples of fields are the sets of all real numbers, of all complex numbers, and of all p-adic numbers, each together with the binary operations of ordinary addition and multiplication.

For our purposes, the most important of these is the field of real numbers. We denote this field as $\mathbb{R}$; that is, $\mathbb{R} = (S, +, \circ)$, where S is the set of all real numbers, $+$ is ordinary addition, and $\circ$ is ordinary multiplication. (Of course, we also use the symbol $\mathbb{R}$ to denote just the set of all real numbers.)

While I generally attempt to give due attention to $\mathbb{R}^d$ for $d = 1, 2, \ldots$, because for $d > 1$, there is only one useful operation that takes $\mathbb{R}^d \times \mathbb{R}^d$ into $\mathbb{R}^d$ (that is, addition), there is no useful field defined on $\mathbb{R}^d$ for $d > 1$.

Note that the extended reals $\overline{\mathbb{R}}$ together with the extensions of the definitions of $+$ and $\circ$ for $-\infty$ and ∞ (see page 640) is not a field. (Operations for some pairs are not defined, and additional elements fail to have inverses.)

In at least two areas of statistics (design of experiments and random number generation), fields with a finite number of elements are useful. A finite field is sometimes called a *Galois field*. We often denote a Galois field with m elements as $\mathbb{G}(m)$.

A field can be defined by means of an addition table and a multiplication table. For example a $\mathbb{G}(5)$ over the set $\{0, 1, 2, 3, 4\}$ together with the "addition" operation $+$ with identity 0 and the "multiplication" operation $\circ$ with identity 1 can be defined by giving the tables below.

Table 0.1. Operation Tables for $\mathbb{G}(5)$

$+$	0	1	2	3	4
0	0	1	2	3	4
1	1	2	3	4	0
2	2	3	4	0	1
3	3	4	0	1	2
4	4	0	1	2	3

$\circ$	0	1	2	3	4
0	0	0	0	0	0
1	0	1	2	3	4
2	0	2	4	1	3
3	0	3	1	4	2
4	0	4	3	2	1

Because the tables are symmetric, we know that the operations are commutative. We see that each row in the addition table contains the additive identity; hence, each element has an additive inverse. We see that each row in the multiplication table, except for the row corresponding to additive identity contains the multiplicative identity; hence, each element except for the additive identity has a multiplicative inverse.

Also we see, for example, that the additive inverse of 3, that is, -3, is 1, and the multiplicative inverse of 3, that is, 3^{-1}, is 2.

In fields whose elements are integers, the smallest positive integer k such that $k \circ 1 = 0$ is called the *characteristic* of the field. It is clear that no such integer exists for an infinite field, so we define the characteristic of an infinite field to be 0.

The number of elements of any finite field, called its *order*, is of the form p^n, where p is a prime number and n is a positive integer; conversely, for every prime number p and positive integer n, there exists a finite field with p^n elements. (See Hewitt and Stromberg (1965) for a proof.) In the case of a field of order p^n, p is the characteristic of the field. The term "order" has other meanings in regard to fields, as we see below.

Ordered Fields

We have defined ordered sets in Section 0.0.1 in terms of the existence of a binary relation. We now define ordered fields in terms of the field operations.

Definition 0.0.4 (ordered field)
A field S is said to be *ordered* if there is a subset P of S such that

- $P \cap (-P) = \emptyset$;
- $P \cup \{0\} \cup (-P) = S$;
- $x, y \in P \Rightarrow x + y, x \circ y \in P$.

∎

Notice that this is only a partial ordering; it is neither a linear ordering nor a well-ordering. Applying this definition to the real numbers, we can think of P as the positive reals, and the elements of $-P$ as the negative reals. Notice that 1 must be in P (because $a \in P \Rightarrow a^2 \in P$ and $b \in -P \Rightarrow b^2 \in P$) and so an ordered field must have characteristic 0; in particular, a finite field cannot be ordered.

We define the binary relations "$\leq$", "$<$","$\geq$", and "$>$" in the ordered field S in a way that is consistent with our previous use of those symbols. For example, for $x, y \in S$, $x < y$ or $y > x$ implies $y - x \in P$.

We now define a stronger ordering.

Definition 0.0.5 (Archimedean ordered field)
An ordered field S defined by the subset P is said to be *Archimedean ordered* if for all $x \in S$ and all $y \in P$, there exists a positive integer n such that $ny > x$. ∎

An Archimedean ordered field must be dense, in the sense that we can find an element between any two given elements.

Theorem 0.0.3
Let S be an Archimedean ordered field, and let $x, y \in S$ such that $x < y$. Then there exists integers m and n such that $m/n \in S$ and

$$x < \frac{m}{n} < y.$$

Proof.
Exercise. ∎

The practical application of Theorem 0.0.3 derives from its implication that there exists a rational number between any two real numbers.

0.0.4 Linear Spaces

An interesting class of spaces are those that have a closed commutative and associative addition operation for all elements, an additive identity, and each

element has an additive inverse (that is, the spaces are Abelian groups under that addition operation), and for which we define a multiplication of real numbers and elements of the space. Such spaces are called linear spaces, and we will formally define them below. Linear spaces are also called vector spaces, especially if the elements are n-tuples of real numbers, called real vectors.

We denote the addition operation by "+", the additive identity by "0", and the multiplication of a real number and an element of the space by juxtaposition. (Note that the symbol "+" also denotes the addition in $\mathbb{R}$ and "0" also denotes the additive identity in $\mathbb{R}$.)

Definition 0.0.6 (linear space)
A structure $(\Omega, +)$ in which the operator $+$ is commutative and associative, and for which the the multiplication of a real number a and an element of Ω x is defined as a closed operation in Ω (whose value is denoted as ax) is called a *linear space* if for any $x, y \in \Omega$ and any $a, b \in \mathbb{R}$,

- $a(x + y) = ax + ay$,
- $(a + b)x = ax + bx$,
- $(ab)x = a(bx)$,
- $1x = x$, where 1 is the multiplicative identity in $\mathbb{R}$.

∎

The "axpy operation", $ax+y$, is the fundamental operation in linear spaces.

The two most common linear spaces that we will encounter are those whose elements are real vectors and those whose elements are real-valued functions. The linear spaces consisting of real vectors are subspaces of $\mathbb{R}^d$, for some integer $d \geq 1$.

Linear spaces, however, may be formed from various types of elements. The elements may be sets, for example. In this case, $x \in \Omega$ would mean that $x = \{x_i \,|\, i \in \mathcal{I}\}$, "+" may represent set union, and "0" may be $\emptyset$. The multiplication between real numbers and elements of the space could be defined in various ways; if the sets in Ω are sets of real numbers, then "ax" above could be interpreted as ordinary multiplication of each element: $ax = \{ax_i \,|\, i \in \mathcal{I}\}$.

Two synonyms for "linear spaces" are "vector spaces" and "linear manifolds", although "vector space" is often often restricted to linear spaces over $\mathbb{R}^d$, and "linear manifold" is used differently by different authors.

Linear Combinations, Linear Independence, and Basis Sets

Given $x_1, x_2, \ldots \in \Omega$ and $c_1, c_2, \ldots \in \mathbb{R}$, $\sum_i c_i x_i$ is called a *linear combination*.

A set of elements $x_1, x_2, \ldots \in \Omega$ are said to be *linearly independent* if $\sum_i c_i x_i = 0$ for $c_1, c_2, \ldots \in \mathbb{R}$ implies that $c_1 = c_2 = \cdots = 0$.

Given a linear space Ω and a set $B = \{b_i\}$ of linearly independent elements of Ω if for any element $x \in \Omega$, there exist $c_1, c_2, \ldots \in \mathbb{R}$ such that $x = \sum_i c_i b_i$, then B is called a *basis set* of Ω.

Subsets of Linear Spaces

Interesting subsets of a given linear space Ω are formed as $\{x\,;\, x \in \Omega, g(x) = 0\}$, for example, the plane in $\mathbb{R}^3$ defined by $c_1x_1 + c_2x_2 + c_3x_3 = c_0$ for some constants $c_0, c_1, c_2, c_3 \in \mathbb{R}$ and where $(x_1, x_2, x_3) \in \mathbb{R}^3$.

Many subsets of $\mathbb{R}^d$ are of interest because of their correspondence to familiar geometric objects, such as lines, planes, and structures of finite extent such as cubes, and spheres. An object in higher dimensions that is analogous to a common object in $\mathbb{R}^3$ is often called by the name of the three-dimensional object preceded by "hyper-"; for example, hypersphere.

A subset of a linear space may or may not be a linear space. For example, a hyperplane is a linear space only if it goes through the origin. Other subsets of a linear space may have very little in common with a linear space, for example, a hypersphere.

A hyperplane in $\mathbb{R}^d$ is often of interest in statistics because of its use as a model for how a random variable is affected by covariates. Such a linear manifold in $\mathbb{R}^d$ is determined by d points that have the property of affine independence. A set of elements $x_1, \ldots, x_d \in \Omega$ are said to be *affinely independent* if $x_2 - x_1, \ldots, x_d - x_1$ are linearly independent. Affine independence in this case insures that the points do not lie in a set of less than $d-1$ dimensions.

Inner Products

We now define a useful real-valued binary function on linear spaces.

Definition 0.0.7 (inner product)
If Ω is a linear space, an *inner product* on Ω is a real-valued function, denoted by $\langle x, y\rangle$ for all x and y in Ω, that satisfies the following three conditions for all x, y, and z in Ω.

1. Nonnegativity and mapping of the identity:
   if $x \neq 0$, then $\langle x, x\rangle > 0$ and $\langle 0, x\rangle = \langle x, 0\rangle = \langle 0, 0\rangle = 0$.
2. Commutativity:
   $\langle x, y\rangle = \langle y, x\rangle$.
3. Factoring of scalar multiplication in inner products:
   $\langle ax, y\rangle = a\langle x, y\rangle$ for real a.
4. Relation of vector addition to addition of inner products:
   $\langle x + y, z\rangle = \langle x, z\rangle + \langle y, z\rangle$.

∎

Inner products are often called dot products, although "dot product" is often used to mean a specific inner product.

A linear space together with an inner product, $(\Omega, \langle\cdot,\cdot\rangle)$, is called an *inner product space*.

A useful property of inner products is the *Cauchy-Schwarz inequality*:

$$\langle x, y\rangle \leq \langle x, x\rangle^{\frac{1}{2}} \langle y, y\rangle^{\frac{1}{2}}. \tag{0.0.26}$$

The proof of this is a classic:
first form the nonnegative polynomial in t,

$$0 \leq \langle tx + y, tx + y\rangle = \langle x, x\rangle t^2 + 2\langle x, y\rangle t + \langle y, y\rangle,$$

and then, because there can be at most one real root in t, require that the discriminant

$$(2\langle x, y\rangle)^2 - 4\langle x, x\rangle\langle y, y\rangle$$

be nonpositive. Finally, rearrange to get inequality (0.0.26).

Inner product spaces give rise to the very useful concept of *orthogonality*. If $(\Omega, \langle\cdot,\cdot\rangle)$ is an inner product space, the elements $x, y \in \Omega$ if $\langle x, y\rangle = 0$ we write $x \perp y$ and say that x and y are *orthogonal*. There are many instances in which we mention orthogonality, but we defer more discussion to Section 0.3.

Norms

We now define a useful function from a linear space to the nonnegative reals.

Definition 0.0.8 (norm)
If Ω is a linear space, a *norm* on Ω is a function, denoted by $\|\cdot\|$, from Ω to $\overline{\mathbb{R}}_+$ that satisfies the following three conditions for all x and y in Ω.

1. Nonnegativity and mapping of the identity:
   if $x \neq 0$, then $\|x\| > 0$, and $\|0\| = 0$
2. Relation of scalar multiplication to real multiplication:
   $\|ax\| = |a|\,\|x\|$ for real a
3. Triangle inequality:
   $\|x + y\| \leq \|x\| + \|y\|$

∎

A linear space together with a norm, $(\Omega, \|\cdot\|)$, is called a *normed linear space*.

Normed linear spaces give rise to the very useful concept of projections.

Definition 0.0.9 (projection)
Let $(\Omega, \|\cdot\|)$ be a normed linear space and let $(\Lambda, \|\cdot\|)$ be a normed linear subspace; that is, $(\Lambda, \|\cdot\|)$ is a normed linear space and $\Lambda \subseteq \Omega$. For $x \in \Omega$, the *projection* of x onto Λ is x_p such that

$$\|x - x_p\| = \inf_{y\in\Lambda} \|x - y\|.$$

∎

We can show that such a x_p exists and that it is unique. (See, for example, Bachman and Narici (2000), Section 10.4.) There are many instances in which we mention projections. They arise commonly in linear algebra and we will discuss them in more detail in Section 0.3. Projections are also important in probability and statistics. We discuss them in this context in Section 1.5.3 on page 115.

Pseudonorms

If, in the first condition in Definition 0.0.8, the requirement if $x \neq 0$, then $\|x\| > 0$ is replaced by

1. if $x \neq 0$, then $\|x\| \geq 0$ and $\|0\| = 0$,

the resulting function is called a *pseudonorm*, or a *seminorm*. (Here, I am considering these two terms to be synonyms. Some people use one or the other of these terms as I have defined it and use the other term for some different functional.)

In some linear spaces whose elements are sets, we define a real nonnegative "measure" function on subsets of the space (see Section 0.1). A set whose measure is 0 is negligible for most purposes of interest In some cases, however, $\emptyset$ is not the only set with measure 0. In such contexts, the requirement on a norm that $\|x\| = 0 \Rightarrow x = \emptyset$ is generally too restrictive to be useful. One alternative is just to use a pseudonorm in such a context, but the pseudonorm is not restrictive enough to be very useful. The best way out of this dilemma is to define equivalence classes of sets such that two sets whose difference has measure zero are equivalent. In such equivalence classes the sets are said to be the same "almost everywhere". (See page 710 for a more careful development of these ideas.) Following this approach, we may qualify the implication implicit in the first condition in the list above by "almost everywhere"; that is, $\|x\| = 0$ implies that the measure of x is 0.

Norms and Metrics Induced by an Inner Product

A norm can be defined simply in terms of an inner product.

Theorem 0.0.4
Let $\langle \cdot, \cdot \rangle$ be an inner product on Ω. Then $\sqrt{\langle x, x \rangle}$, for x in Ω, is a norm on Ω.

Proof.
Exercise. (Using the properties of $\langle x, x \rangle$ given in Definition 0.0.7, show that $\sqrt{\langle x, x \rangle}$ satisfies the properties of Definition 0.0.8. The only one that is nontrivial is the triangle inequality; Exercise 0.0.10.) ∎

Given the inner product, $\langle \cdot, \cdot \rangle$, $\|x\| = \sqrt{\langle x, x \rangle}$, is called the norm induced by that inner product.

A metric can be defined simply in terms of a norm.

Theorem 0.0.5
Let $\|\cdot\|$ be a norm on Ω. Then $\|x-y\|$, for $x, y \in \Omega$, is a metric on Ω.

Proof.
Exercise 0.0.11. (This is easily seen from the definition of metric, Definition 0.0.1, on page 623.) ∎

This metric is said to be induced by the norm $\|\cdot\|$. A pseudonorm induces a *pseudometric* in the same way.

Notice also that we have a sort of converse to Theorem 0.0.5: that is, a metric induces a norm. Given the metric $\rho(x, y)$, $\|x\| = \rho(x, 0)$ is a norm. (Here, in general, 0 is the additive identity of the linear space.)

The terms "normed linear space" and "linear metric space" are therefore equivalent, and we will use them interchangeably.

Countable Sequences and Complete Spaces

Countable sequences of elements of a linear metric space, $\{x_i\}$, for $i = 1, 2, \ldots$, are often of interest. The limit of the sequence, that is, $\lim_{i\to\infty} x_i$, is of interest. The first question, of course, is whether it exists. An oscillating sequence such as $-1, +1, -1, +1, \ldots$ does not have a limit. Likewise, a divergent sequence such as $1, 2, 3, \ldots$ does not have a limit. If the sequence has a finite limit, we say the sequence *converges*, but the next question is whether it converges to a point in the given linear space.

Let $A = \{x_i\}$ be a countable sequence of elements of a linear metric space with metric $\rho(\cdot, \cdot)$. If for every $\epsilon > 0$, there exists a constant n_ϵ such that

$$\rho(x_n, x_m) < \epsilon \quad \forall m, n > n_\epsilon, \tag{0.0.27}$$

then A is called a *Cauchy sequence*. Notice that the definition of a Cauchy sequence depends on the metric.

A sequence of elements of a linear metric space converges only if it is a Cauchy sequence.

A normed linear space is said to be *complete* if every Cauchy sequence in the space converges to a point in the space.

A complete normed linear space is called a *Banach space*.

A Banach space whose metric arises from an inner product is called a *Hilbert space*. Equivalently, a Hilbert space is the same as a complete inner product space.

Given a metric space (Ω, ρ), the space (Ω_c, ρ_c) is called the *completion of* (Ω, ρ) if

- Ω is dense in Ω_c
- $\rho_c(x, y) = \rho(x, y)\ \forall x, y \in \Omega$.

For any given metric space, the completion exists. We show this by first defining an equivalence class of sequences to consist of all sequences $\{x_n\}$ and

$\{y_m\}$ such that $\rho(x_n, y_m) \to 0$, and then defining ρ_c over pairs of equivalence classes.

In the foregoing we have assumed very little about the nature of the point sets that we have discussed. Lurking in the background, however, was the very special set of *real numbers*. Real numbers are involved in the definition of a linear space no matter the nature of the elements of the space. Real numbers are also required in the definitions of inner products, norms, and metrics over linear spaces no matter the nature of the elements of the spaces.

We now turn to the special linear space, the real number system, and provide specific instantiations of some the concepts discussed previously. We also introduce additional concepts, which could be discussed in the abstract, but whose relevance depends so strongly on the real number system that abstraction would be a distraction.

0.0.5 The Real Number System

The most important sets we will work with are sets of real numbers or product sets of real numbers. The set of real numbers together with ordinary addition and multiplication is a *field* (page 631); it is a *Hausdorff space* (page 623); it is a *complete metric space* (page 639); it is a *connected topological space* (page 623); it is a *Banach space* (page 639); it is a *Hilbert space* (with the Euclidean norm) (page 639); et cetera, et cetera.

We denote the full set of real numbers, that is, the "reals", by $\mathbb{R}$, the set of positive real numbers by $\mathbb{R}_+$, and the set of negative real numbers by $\mathbb{R}_-$. For a positive integer d, we denote the product set $\prod_{i=1}^{d} \mathbb{R}$ as $\mathbb{R}^d$. The reals themselves and these three sets formed from them are all clopen.

The Extended Reals

The reals do not include the two special elements ∞ and $-\infty$, but we sometimes speak of the "extended reals", which we denote and define by

$$\overline{\mathbb{R}} \stackrel{\text{def}}{=} \mathbb{R} \cup \{-\infty, \infty\}. \tag{0.0.28}$$

This notation may seem to imply that $\overline{\mathbb{R}}$ is the closure of $\mathbb{R}$, and to some extent, that is the case; however, we should recall that

$$\mathbb{R} =]-\infty, \infty[$$

is itself a closed set. (It is also open; that is, it is clopen.)

We define operations with the two elements ∞ and $-\infty$ as follows (where "$\times$" means multiplication):

- $\forall x \in \mathbb{R},\ -\infty < x < \infty$
- $\forall x \in \mathbb{R},\ x \pm \infty = \pm\infty + x = \pm\infty$

- $\forall x \in \mathbb{R} \cup \{\infty\},\ x \times \pm\infty = \pm\infty \times x = \pm\infty$
- $\forall x \in \mathbb{R} \cup \{-\infty\},\ x \times \pm\infty = \pm\infty \times x = \mp\infty$
- $\forall x \in \mathbb{R},\ x/\pm\infty = 0$
- $\forall x \in \mathbb{R}_+,\ \pm\infty/x = \pm\infty$
- $\forall x \in \mathbb{R}_-,\ \pm\infty/x = \mp\infty$.

Other operations may or may not be defined, depending on the context. For example, often in probability theory or measure theory, we may use the definition

- $0 \times \pm\infty = \pm\infty \times 0 = 0$.

In numerical mathematics, we may use the definition

- $\forall x \in \mathbb{R},\ x/0 = \text{sign}(x)\infty$.

The other cases,

$$\infty - \infty, \quad \pm\infty/\pm\infty, \quad \pm\infty/\mp\infty$$

are almost aways undefined; that is, they are considered *indeterminate forms*.

Notice that the extended reals is not a field, but in general, all of the laws of the field $\mathbb{R}$ hold so long as the operations are defined.

The finite reals without ∞ and $-\infty$, are generally more useful. By not including the infinities in the reals, that is, by working with the field of reals, we often make the discussions simpler.

Important Subsets of the Reals

We denote the full set of integers by $\mathbb{Z}$, and the set of positive integers by $\mathbb{Z}_+$. The positive integers are also called the natural numbers. Integers are reals and so $\mathbb{Z} \subseteq \mathbb{R}$ and $\mathbb{Z}_+ \subseteq \mathbb{R}_+$. We often seem implicitly to include ∞ as an integer in expressions such as $\sum_{i=1}^{\infty}$. Such an expression, however, should be interpreted as $\lim_{n\to\infty} \sum_{i=1}^{n}$.

The set of numbers that can be represented in the form a/b, where $a, b \in \mathbb{Z}$, are the rational numbers. The rationals can be mapped to the integers; hence the set of rationals is countable.

The closure (page 622) of the set of rational numbers is $\mathbb{R}$; hence, the rationals are *dense in* $\mathbb{R}$ (page 622). Because the rationals are countable and are dense in $\mathbb{R}$, $\mathbb{R}$ is *separable* (page 622).

Norms and Metrics on the Reals

The simplest and most commonly used norms on $\mathbb{R}^d$ are the L_p *norms*, which for $p \geq 1$ and $x \in \mathbb{R}^d$ are defined by

$$\|x\|_p = \left(\sum_{i=1}^{d} |x_i|^p\right)^{1/p}. \tag{0.0.29}$$

For this to be a norm, for any $x, y \in \mathbb{R}^d$ and $1 \leq p$, we must have

$$\left(\sum_{i=1}^{d} |x_i + y_i|^p\right)^{1/p} \leq \left(\sum_{i=1}^{d} |x_i|^p\right)^{1/p} + \left(\sum_{i=1}^{d} |y_i|^p\right)^{1/p}. \tag{0.0.30}$$

This inequality is called Minkowski's inequality, and hence the norm itself is sometimes called the Minkowski norm.

A straightforward proof of Minkowski's inequality is based on Hölder's inequality. Hölder's inequality, using the same notation as above, except with the requirement $1 < p$, is

$$\sum_{i=1}^{d} |x_i y_i| \leq \left(\sum_{i=1}^{d} |x_i|^p\right)^{1/p} + \left(\sum_{i=1}^{d} |y_i|^{p/(p-1)}\right)^{(p-1)/p}. \tag{0.0.31}$$

To prove Hölder's inequality, we use the concavity property of the log function. For $0 < a_1, a_2$ and $1 < p$, because $1/p + (p-1)/p = 1$, the concavity of the log function yields

$$\log(a_1)/p + \log(a_2)(p-1)/p \leq \log(a_1/p + a_2(p-1)/p),$$

or

$$a_1^{1/p} a_2^{(p-1)/p} \leq a_1/p + a_2(p-1)/p. \tag{0.0.32}$$

Now, if $|x_i| > 0$ and $|y_i| > 0$, let

$$a_1 = \frac{|x_i|^p}{\sum_{i=1}^{d} |x_i|^p}, \quad a_2 = \frac{|y_i|^{p/(p-1)}}{\sum_{i=1}^{d} |y_i|^{p/(p-1)}}.$$

Substituting these in equation (0.0.32) and then summing over i, we have

$$\sum_{i=1}^{d} \left(\frac{|x_i|^p}{\sum_{i=1}^{d} |x_i|^p}\right)^{1/p} \left(\frac{|y_i|^{p/(p-1)}}{\sum_{i=1}^{d} |y_i|^{p/(p-1)}}\right)^{(p-1)/p}$$

$$\leq$$

$$\sum_{i=1}^{d} \left(\frac{1}{p} \frac{|x_i|^p}{\sum_{i=1}^{d} |x_i|^p} + \frac{p-1}{p} \frac{|y_i|^{p/(p-1)}}{\sum_{i=1}^{d} |y_i|^{p/(p-1)}}\right)$$

$$= 1.$$

Writing the summands in the numerators as $|x_i|$ and $|y_i|$ and simplifying the resulting expression, we have Hölder's inequality for the case where all x_i and y_i are nonzero. We next consider the case where some are zero, and see that the same proof holds; we just omit the zero terms. For the case that all x_i and/or y_i are zero, the inequality holds trivially.

For Minkowski's inequality, we first observe that it holds if $p = 1$, and it holds trivially if either $\sum_{i=1}^d |x_i| = 0$ or $\sum_{i=1}^d |y_i| = 0$. For $p > 1$, we first obtain from the left side of inequality (0.0.30)

$$\begin{aligned}\sum_{i=1}^d |x_i + y_i|^p &= \sum_{i=1}^d |x_i + y_i||x_i + y_i|^{p-1} \\ &\leq \sum_{i=1}^d |x_i||x_i + y_i|^{p-1} + \sum_{i=1}^d |y_i||x_i + y_i|^{p-1}.\end{aligned}$$

We next apply Hölder's inequality, and simplify under the assumption that neither $\sum_{i=1}^d |x_i| = 0$ nor $\sum_{i=1}^d |y_i| = 0$.

Because Hölder's inequality implies Minkowski's inequality, the L_p norm is also called the Hölder norm. (As with many things in mathematics, the implications of eponyms are ambiguous. Otto Hölder stated what we call Hölder's inequality in an open publication before Hermann Minkowski published what we call Minkowski's inequality. But who knows?)

We see from the definition that $\|x\|_p$ is a nonincreasing function in p, that is,

$$1 \leq p_1 < p_2 \Rightarrow \|x\|_{p_1} \geq \|x\|_{p_2}. \tag{0.0.33}$$

(Exercise.)

In the case of $d = 1$, the L_p is just the absolute value for any p; that is, for $x \in \mathbb{R}$, $\|x\| = |x|$.

Because

$$\lim_{p\to\infty} \|x\|_p = \max(\{|x_i|\}), \tag{0.0.34}$$

we formally define the L_∞ norm of the vector $x = (x_1, \ldots, x_d)$ as $\max(\{|x_i|\})$. It is clear that it satisfies the properties of a norm.

Metrics are often defined in terms of norms of differences (Theorem 0.0.5). Because the L_p norm is a common norm on $\mathbb{R}^d$, a common metric on $\mathbb{R}^d$ is the L_p metric, $\rho_p(x, y) = \|x - y\|_p$.

The L_p metric is also called the Minkowski metric or the Minkowski distance.

The most common value of p is 2, in which case we have the Euclidean metric (or Euclidean distance):

$$\|x - y\|_2 = \left(\sum_{i=1}^d (x_i - y_i)^2\right)^{1/2}. \tag{0.0.35}$$

We often write the Euclidean distance between x and y as $\|x - y\|$.

Any of these metrics allows us to define neighborhoods and open sets, and to define convergence of sequences.

Ordering the Reals

The field of real numbers $\mathbb{R}$ is complete and Archimedean ordered; in fact, the reals can be defined as a (any) complete Archimedean ordered field. The ordering properties correspond to our intuitive understanding of ordering.

The more useful ordering of the reals is the linear ordering that results from the usual inequality relation. On the other hand, $\mathbb{R}^d$ cannot be linearly ordered in an intuitive or universally useful way. An ordering of $x, y \in \mathbb{R}^d$ based on notions of $x < y$ or $x \leq y$ are rarely useful in statistical applications (see Section 0.0.1 on page 620). Most useful orderings are based on relations between $\|x\|$ and $\|y\|$, but their usefulness depends on the application. (See Gentle (2009), pages 538 to 549, for some discussion of this.)

In order to simplify the following discussion, we will focus on subsets of $\mathbb{R}$. Because the reals are not a well-ordered set using the usual inequality relation, we find it convenient to define limiting maxima and minima based on that relation.

For $X \subseteq \mathbb{R}$, the *supremum* of X or *least upper bound* of X, is the number

$$x^* = \sup(X), \tag{0.0.36}$$

defined by

$$\forall x \in X, x \leq x^*$$

and

$$\forall \epsilon > 0,\ \exists x \in X \ni x > x^* - \epsilon.$$

The *infimum* of X, that is, the *greatest lower bound* of X, is written as

$$x_* = \inf(X) \tag{0.0.37}$$

and is defined similarly, with the inequalities reversed.
Examples:

- Let $A = \{x\}$. Then $\sup(A) = \inf(A) = x$.
- Let $A = \{x \,|\, x = 1/i, i \in \mathbb{Z}_+\}$. Then $\sup(A) = 1$ and $\inf(A) = 0$. Notice that $\inf(A) \notin A$.
- Let $A = \{x \,|\, i, i \in \mathbb{Z}_+\}$. Then $\sup(A) = \infty$ and $\inf(A) = 1$. Alternatively, we may say that $\sup(A)$ does not exist. In any event, $\sup(A) \notin A$.

An important fundamental property of the reals is that every bounded set of reals has a supremum that is a real number. This property is often called *Dedekind completeness*. In the usual axiomatic development of the reals, Dedekind completeness is an axiom.

The *maximum* of a well-ordered set is the largest element of the set, if it exists; likewise, the *minimum* of a well-ordered set is the smallest element of the set, if it exists. The maximum and/or the minimum may not exist if the set has an infinite number of elements. This can happen in two ways: one, the set may have no bound; and another, the bound may not be in the set.

Sets of Reals; Open, Closed, Compact

For sets of reals we can redefine various topological properties in a way that is simpler but yet consistent with our discussion of general topologies on page 621. These concepts yield important kinds of sets, such as open, closed, compact, and convex sets (see page 658).

A set A of reals is called *open* if for each $x \in A$, there exists a $\delta > 0$ such that for each y with $\|x - y\| < \delta$ belongs to A.

If A is a set of reals and if for a given $x \in A$, there exists a $\delta > 0$ such that for each y with $\|x - y\| < \delta$ belongs to A, then x is called an *interior point* of A.

We denote the set of all interior points of A as

$$A^\circ$$

and call it the *interior* of A. Clearly A° is open, and, in fact, it is the union of all open subsets of A.

A real number (vector) x is called a *point of closure* of a set A of real numbers (vectors) if for every $\delta > 0$ there exists a y in A such that $\|x-y\| < \delta$. (Notice that every $y \in A$ is a point of closure of A.)

We denote the set of points of closure of A by

$$\overline{A},$$

and a set A is called *closed* if $A = \overline{A}$. (This is the same definition and notation as for any topology.)

The *boundary* of the set A, denoted ∂A, is the set of points of closure of A that are not interior points of A; that is,

$$\partial A = \overline{A} - A^\circ. \tag{0.0.38}$$

As we have mentioned above, in any topological space, a set A is said to be *compact* if each collection of open sets that covers A contains a finite subcollection of open sets that covers A. Compactness of subsets of the reals can be characterized simply, as in the following theorem.

Theorem 0.0.6 (Heine-Borel theorem)
A set of real numbers is compact iff it is closed and bounded.

For a proof of this theorem, see a text on real analysis, such as Hewitt and Stromberg (1965). Because of the Heine-Borel theorem, we often take closed and bounded as the definition of a compact set of reals.

Intervals in $\mathbb{R}$

A very important type of set is an *interval* in $\mathbb{R}$, which is a connected subset of $\mathbb{R}$. Intervals are the basis for building important structures on $\mathbb{R}$.

We denote an *open interval* with open square brackets; for example $]a, b[$ is the set of all real x such that $a < x < b$. We denote a *closed interval* with closed square brackets; for example $[a, b]$ is the set of all real x such that $a \leq x \leq b$. We also have "half-open" or "half-closed" intervals, with obvious meanings.

The main kinds of intervals have forms such as
$]-\infty, a[, \quad]-\infty, a], \quad]a, b[, \quad [a, b], \quad]a, b], \quad [a, b[, \quad]b, \infty[, \quad$ and $[b, \infty[$.
Of these,

- $]-\infty, a[, \quad]a, b[, \quad$ and $]b, \infty[$ are open;
- $[a, b]$ is closed and $\overline{]a, b[} = [a, b]$;
- $]a, b]$ and $[a, b[$ are neither (they are "half-open");
- $]-\infty, a]$ and $[b, \infty[$ are closed, although in a special way that sometimes requires special treatment.

A finite closed interval is compact (by the Heine-Borel theorem); but an open or half-open interval is not, as we see below.

The following facts for real intervals are special cases of the properties of unions and intersections of open and closed sets we listed above, which can be shown from the definitions:

- $\cap_{i=1}^{n}]a_i, b_i[=]a, b[$ (that is, some open interval)
- $\cup_{i=1}^{\infty}]a_i, b_i[$ is an open set
- $\cup_{i=1}^{n} [a_i, b_i]$ is a closed set
- $\cap_{i=1}^{\infty} [a_i, b_i] = [a, b]$ (that is, some closed interval).

Two types of interesting intervals are

$$\left]a - \frac{1}{i},\ b + \frac{1}{i}\right[\tag{0.0.39}$$

and

$$\left[a + \frac{1}{i},\ b - \frac{1}{i}\right]. \tag{0.0.40}$$

Sequences of intervals of these two forms are worth remembering because they illustrate interesting properties of intersections and unions of infinite sequences. Infinite intersections and unions behave differently with regard to collections of open and closed sets. For finite intersections and unions we know that $\cap_{i=1}^{n}]a_i, b_i[$ is an open interval, and $\cup_{i=1}^{n} [a_i, b_i]$ is a closed set.

First, observe that

$$\lim_{i\to\infty} \left]a - \frac{1}{i},\ b + \frac{1}{i}\right[= [a, b] \tag{0.0.41}$$

and

$$\lim_{i\to\infty} \left[a + \frac{1}{i},\ b - \frac{1}{i}\right] = [a, b]. \tag{0.0.42}$$

Now for finite intersections of the open intervals and finite unions of the closed intervals, that is, for finite k, we have

$$\bigcap_{i=1}^{k} \left]a - \frac{1}{i},\ b + \frac{1}{i}\right[\quad \text{is open}$$

and

$$\bigcup_{i=1}^{k} \left[a + \frac{1}{i},\ b - \frac{1}{i}\right] \quad \text{is closed.}$$

Infinite intersections and unions behave differently with regard to collections of open and closed sets. With the open and closed intervals of the special forms, for infinite intersections and unions, we have the important facts:

$$\bigcap_{i=1}^{\infty} \left]a - \frac{1}{i},\ b + \frac{1}{i}\right[= [a, b] \tag{0.0.43}$$

and

$$\bigcup_{i=1}^{\infty} \left[a + \frac{1}{i},\ b - \frac{1}{i}\right] =]a, b[. \tag{0.0.44}$$

These equations follow from the definitions of intersections and unions. To see equation (0.0.44), for example, we note that $a \in \cup A_i$ iff $a \in A_i$ for some i; hence, if $a \notin A_i$ for any i, then $a \notin \cup A_i$.

Likewise, we have

$$\begin{aligned} \bigcup_{i=1}^{\infty} \left[a + \frac{1}{i},\ b\right] &= \bigcap_{i=1}^{\infty} \left]a,\ b + \frac{1}{i}\right[\\ &=]a, b]. \end{aligned} \tag{0.0.45}$$

From this we see that

$$\lim_{n\to\infty} \bigcup_{i=1}^{n} \left[a + \frac{1}{i},\ b - \frac{1}{i}\right] \neq \bigcup \lim_{i\to\infty} \left[a + \frac{1}{i},\ b - \frac{1}{i}\right]. \tag{0.0.46}$$

Equations (0.0.44) and (0.0.45) for $]a, b[$ and $]a, b]$ above show that open intervals and half-open intervals are not compact, because no finite collection of sets in the unions cover the intervals.

Intervals in $\mathbb{R}^d$

"Intervals" in $\mathbb{R}^d$ are merely product sets of intervals in $\mathbb{R}$; that is, they are hyperrectangles. Because of the possibilities of the types of endpoints of the intervals in $\mathbb{R}$, the intervals in $\mathbb{R}^d$ cannot always be specified using "[,]" "], [", and so on. In more restrictive cases in which all of the intervals in $\mathbb{R}$ are of

the same types, we use the same notation as above to indicate the product sets. For example, given the vectors $a = (a_1, \ldots, a_d)$ and $b = (b_1, \ldots, b_d)$, we could write $]a, b]$ with the meaning

$$]a, b] =]a_1, b_1] \times \cdots \times]a_d, b_d].$$

Sequences of Reals, Limit Points, and Accumulation Points

Because of the Heine-Borel theorem, bounded sequence in $\mathbb{R}^d$ are of interest. (A sequence $\{x_n\}$ is *bounded* if there exists an $M \in \mathbb{R}^d \ni x_n \leq M \ \forall n$.) We let $\{x_n\} = x_1, x_2, \ldots$ be a sequence in $\mathbb{R}^d$. The Bolzano-Weierstrass theorem (see below) states that every bounded sequence has a convergent subsequence; that is, there is a subsequence that has a limit point. A point $x \in \mathbb{R}^d$ is called an accumulation point, or a cluster point, of $\{x_n\}$ if there is a subsequence $\{x_{n_i}\}$ that converges to x.

An important property of the reals is that a sequence of reals converges if and only if it is a Cauchy sequence. (The "if" part means that the reals are complete.) This is sometimes called the "Cauchy criterion". See Exercise 0.0.6 for an outline of a proof or see Hewitt and Stromberg (1965).

Notice that the field of rationals is not complete for we can form a Cauchy sequence in the rationals that does not converge to a rational number. For example consider the rational sequence

$$\begin{aligned} x_1 &= 1 \\ x_n &= x_{n-1}/2 + 1/x_{n-1}, \quad n = 2, 3, \ldots \end{aligned}$$

which is a Cauchy sequence (in the Euclidean metric). The sequence, however, converges to $\sqrt{2}$, which is not in the rationals. The field of the reals can, in fact, be considered to be a completion of the rationals in the Euclidean norm.

A linear space formed on $\mathbb{R}^d$ with the usual addition operation for vectors together with any metric is a *Banach space*. If the metric is taken to be the L_2 metric then that linear space is a *Hilbert space*.

Monotone Sequences in $\mathbb{R}$

We will now limit the discussion to subsets and sequences in $\mathbb{R}$. This allows us to use a simple definition of monotonicity, based on the linear ordering of the reals.

A useful theorem tells us that if a bounded sequence is monotone, then the sequence must converge:

Theorem 0.0.7 (monotone convergence of sequence in $\mathbb{R}$)
Let $x_1 \leq x_2 \leq \cdots$ be a sequence in $\mathbb{R}$. This sequence has a finite limit iff the sequence is bounded.

Proof.
First, obviously if $\lim_{n\to\infty} x_n = x < \infty$, then every $x_n \leq x$, and the sequence is bounded.

Now we assume $x_1 \leq x_2 \leq \cdots x < \infty$. By Dedekind completeness of the reals, $x^* = \sup(\{x_n\})$ exists and is finite. For every $\epsilon > 0$, there exists x_N such that $x_N > x^* - \epsilon$ because otherwise $x^* - \epsilon$ would be an upper bound less than $\sup(\{x_n\})$. So now, because $\{x_n\}$ is increasing, $\forall n > N$ we have

$$|x^* - x_n| = x^* - x_n \leq x^* - x_N < \epsilon;$$

therefore, the limit of $\{x_n\}$ is $\sup(\{x_n\})$. ∎

Theorem 0.0.8 (Bolzano-Weierstrass)
Every bounded sequence in $\mathbb{R}$ *has a convergent subsequence.*

Proof.
This theorem is an immediate result of the following lemma. ∎

Lemma 0.0.8.1
Every sequence in $\mathbb{R}$ *has a monotone subsequence.*

Proof.
Define a *dominant term* (or a *peak*) as a term x_j in a sequence $\{x_n\}$ such that $x_j > x_{j+1}, x_{j+2}, \ldots$. The number of dominant terms must either be finite or infinite.

Suppose the number is infinite. In that case, we can form a subsequence that contains only those dominant terms $\{x_{n_i}\}$. This sequence is (strictly) monotonic decreasing, $x_{n_i} > x_{n_{i+1}} > \cdots$.

On the other hand, suppose the number of dominant terms is finite. If there is a dominant term, consider the next term after the last dominant term. Call it x_{n_1}. If there are no dominant terms, $n_1 = 1$. Now, since x_{n_1} is not a dominant term, there must be another term, say x_{n_2}, that is no greater than x_{n_1}, and since that term is not dominant, there must be a term x_{n_3} that is no greater than x_{n_2}, and so on. The sequence $\{x_{n_i}\}$ formed in this way is monotonic nondecreasing. ∎

The Bolzano-Weierstrass theorem is closely related to the Heine-Borel theorem. Either can be used in the proof of the other. A system for which the Bolzano-Weierstrass theorem holds is said to have the Bolzano-Weierstrass property.

Some properties of sequences or subsequences in $\mathbb{R}$ that we discuss that depend on $x_i \leq x_j$ can often be extended easily to sequences in $\mathbb{R}^d$ using the partial ordering imposed by applying the $\mathbb{R}$ ordering element by element. For example, if $\{x_n\}$ is a sequence in $\mathbb{R}^d$, then Lemma 0.0.8.1 could first be applied to the first element in each vector x_n to form a subsequence based on the first element, and then applied to the second element in each vector in this subsequence to form a subsubsequence, and then applied to the third element of each vector in the subsubsequence, and so on. The definition of

the order in $\mathbb{R}^d$ may require some redefinition of the order that arises from applying the $\mathbb{R}$ order directly (see Exercise 0.0.7).

Theorem 0.0.9 is an alternate statement of the Bolzano-Weierstrass theorem.

Theorem 0.0.9 (Bolzano-Weierstrass (alternate))
Every bounded sequence in $\mathbb{R}$ *has an accumulation point.*

We will prove this statement of the theorem directly, because in doing so we are led to the concept of a largest accumulation point, which has more general uses.

Proof.
For the bounded sequence $\{x_n\}$, define the set of real numbers

$$S = \{x \,|\, \text{there are infinitely many } x_n > x\}.$$

Let $x^* = \sup(S)$. Because the sequence is bounded, x^* is finite. By definition of S and $\sup(S)$, for any $\epsilon > 0$, only there are only finitely many x_n such that $x_n \geq x^* + \epsilon$, but there are infinitely many x_n such that $x_n \geq x^* - \epsilon$, so there are infinitely many x_n in the interval $[x^* - \epsilon, x^* + \epsilon]$.

Now for $i = 1, 2, \ldots$, consider the intervals $I_i = [x^* - 1/i, x^* + 1/i]$ each of which contains infinitely many x_n, and form a monotone increasing sequence $\{n_i\}$ such that $x_{n_i} \in I_i$. (Such a sequence is not unique.) Now use the sequence $\{n_i\}$ to form a subsequence of $\{x_n\}$, $\{x_{n_i}\}$. The sequence $\{x_{n_i}\}$ converges to x^*; which is therefore an accumulation point of $\{x_n\}$. ∎

lim sup and lim inf

Because of the way S was defined in the proof of Theorem 0.0.9, the accumulation point $x^* = \sup(S)$ is the largest accumulation point of $\{x_n\}$. The largest accumulation point of a sequence is an important property of that sequence. We call the largest accumulation point of the sequence $\{x_n\}$ the *limit superior* of the sequence and denote it as $\limsup_n x_n$. If the sequence is not bounded from above, we define $\limsup_n x_n$ as ∞. We have

$$\limsup_n x_n = \lim_n \sup_{k \geq n} x_k. \tag{0.0.47}$$

We see that

$$\limsup_n x_n = \sup\{x \,|\, \text{there are infinitely many } x_n > x\}, \tag{0.0.48}$$

which is a characterization of lim sup for any nonincreasing real point sequence $\{x_n\}$. (Compare this with equation (0.0.22) on page 628.)

Likewise, for a bounded sequence, we define the smallest accumulation point of the sequence $\{x_n\}$ the *limit inferior* of the sequence and denote it as $\liminf_n x_n$. If the sequence is not bounded from below, we define $\liminf_n x_n$ as $-\infty$. We have

$$\liminf_n x_n = \lim_n \inf_{k \geq n} x_k. \tag{0.0.49}$$

We have

$$\liminf_n x_n = \inf\{x \,|\, \text{there are infinitely many } x_n < x\}. \tag{0.0.50}$$

The properties of lim sup and lim inf of sequences of sets discussed on page 627 have analogues for lim sup and lim inf of sequences of points.

For a bounded sequence $\{x_n\}$, it is clear that

$$\liminf_n x_n \leq \limsup_n x_n, \tag{0.0.51}$$

and $\{x_n\}$ converges iff $\liminf_n x_n = \limsup_n x_n$, and in that case we write the quantity simply as $\lim_n x_n$.

We also have

$$\limsup_n x_n = \inf_n \sup_{k \geq n} x_k \tag{0.0.52}$$

and

$$\liminf_n x_n = \sup_n \inf_{k \geq n} x_k. \tag{0.0.53}$$

The triangle inequalities also hold:

$$\limsup_n (x_n + y_n) \leq \limsup_n x_n + \limsup_n y_n. \tag{0.0.54}$$

$$\liminf_n (x_n + y_n) \geq \liminf_n x_n + \liminf_n y_n. \tag{0.0.55}$$

Common Sequences of Reals

There are some forms of sequences that arise often in applications; for example, $x_n = 1/n$ or $x_n = 1 + c/n$. Having lists of convergent sequences or convergent series (see Sections 0.0.5 and 0.0.9) can be a useful aid in work in mathematics.

A useful limit of sequences of reals that we will encounter from time to time is

$$\lim_{n \to \infty} \left(1 + \frac{c}{n}\right)^n = \mathrm{e}^c. \tag{0.0.56}$$

We can prove this easily using some simple properties of the logarithm function, which we define as $L(t) = \int_1^t (1/x)\mathrm{d}x$ for $t > 0$. We first observe that L is continuous and increasing, $L(1) = 0$, that L' exists at 1, $L'(1) = 1$, and $nL(x) = L(x^n)$. For a fixed constant $c \neq 0$ we can write the derivative at 1 as

$$\lim_{n \to \infty} \frac{L(1 + c/n) - L(1)}{c/n} = 1,$$

which, because $L(1) = 0$, we can rewrite as $\lim_{n\to\infty} L((1 + c/n)^n) = c$. Since L is continuous and increasing $\lim_{n\to\infty}(1 + c/n)^n$ exists and is the value of x such that $L(x) = c$; that is, it is e^c.

A related limit for a function $g(n)$ that has the limit $\lim_{n\to\infty} = b$ is

$$\lim_{n\to\infty}\left(1 + \frac{cg(n)}{n}\right)^n = \mathrm{e}^{bc}, \tag{0.0.57}$$

which can be shown easily by use of the limit above, and the bounds

$$\left(1 + \frac{c(b-\epsilon)}{n}\right)^n \le \left(1 + \frac{cg(n)}{n}\right)^n \le \left(1 + \frac{c(b+\epsilon)}{n}\right)^n,$$

for $c > 0$ and any $\epsilon > 0$, which arise from the bounds $b - \epsilon < g(n) < b + \epsilon$ for n sufficiently large. Taking limits, we get

$$\mathrm{e}^{c(b-\epsilon)} \le \lim_{n\to\infty}\left(1 + \frac{cg(n)}{n}\right)^n \le \mathrm{e}^{c(b+\epsilon)},$$

and since ϵ was arbitrary, we have the desired conclusion under the assumption that $c > 0$. We get the same result (with bounds reversed) for $c < 0$.

Another related limit is for a function $g(n)$ that has the limit $\lim_{n\to\infty} g(n) = 0$, and constants b and c with $c \neq 0$ is

$$\lim_{n\to\infty}\left(1 + \frac{c}{n} + \frac{g(n)}{n}\right)^b n = \mathrm{e}^{bc}. \tag{0.0.58}$$

The Rate of Convergence; Big O and Little o Notation

We are often interested in how quickly one sequence of real numbers or of real-valued functions converges to another sequence. We will distinguish two types limiting behavior, and the rate of convergence is measured by asymptotic ratios of the sequences.

We consider a class determined by the rate of a given sequence $\{a_n\} \in \mathbb{R}^d$. We identify another sequence $\{b_n\} \in \mathbb{R}^d$ as belonging to the order class if its rate of convergence is similar. If these are sequences of functions, we assume a common domain for all functions in the sequences and our comparisons of $a_n(x)$ and $b_n(x)$ are for all points x in the domain. We refer to one type of limiting behavior as "big O" and to another type as "little o".

Big O, written $\mathrm{O}(a_n)$.

$\{b_n\} \in \mathrm{O}(a_n)$ means there exists some fixed finite c such that $\|b_n\| \le c\|a_n\|\ \forall n$.

In particular, $b_n \in \mathrm{O}(1)$ means b_n is bounded.

Little o, written $\mathrm{o}(a_n)$.

$\{b_n\} \in \mathrm{o}(a_n)$ means $\|b_n\|/\|a_n\| \to 0$ as $n \to \infty$.

In particular, $b_n \in \mathrm{o}(1)$ means $b_n \to 0$.

Instead of "$\{b_n\} \in \mathrm{O}(a_n)$" or "$\{b_n\} \in \mathrm{o}(a_n)$", most people write "$\{b_n\} = \mathrm{O}(a_n)$" or "$\{b_n\} = \mathrm{o}(a_n)$". (From this, we could deduce such nonsense as $n = n^2$, since it is clear, in this notation that $n = \mathrm{O}(n^2)$ and $n^2 = \mathrm{O}(n^2)$.) I do not like this level of imprecision in notation. I so sometimes abuse this notation slightly, for example, by referring to a sequence as "being $\mathrm{O}(f(n))$" rather than as "being in the order class $\mathrm{O}(f(n))$". In one very common case, I abuse the notation in this way. As most people, I may use $\mathrm{O}(f(n))$ to represent some unspecified scalar or vector $x \in \mathrm{O}(f(n))$ in the case of a convergent series, for example,

$$s = f_1(n) + \cdots + f_k(n) + \mathrm{O}(f(n)),$$

where $f_1(n), \ldots, f_k(n)$ are constants. I also use $\mathrm{o}(f(n))$ to represent some unspecified scalar or vector $x \in \mathrm{o}(f(n))$ in special case of a convergent series, as above:

$$s = f_1(n) + \cdots + f_d(n) + \mathrm{o}(f(n)).$$

We often write $b_n \in \mathrm{O}(a_n)$ or $b_n \in \mathrm{O}(a_n)$ instead of $\{b_n\} \in \mathrm{O}(a_n)$ or $\{b_n\} \in \mathrm{O}(a_n)$.

We sometimes omit the arguments of functions; for example, we may write $f \in \mathrm{O}(g)$, with the understanding that the limits are taken with respect to the arguments.

The defining sequence a_n is often a simple expression in n; for examples, $a_n = n^{-2}$, $a_n = n^{-1}$, $a_n = n^{-1/2}$, and so on, or $a_n = n$, $a_n = n\log(n)$, $a_n = n^2$, and so on. We have

$$\mathrm{O}(1) \subseteq \mathrm{O}(n^{-2}) \subseteq \mathrm{O}(n^{-1}) \subseteq \mathrm{O}(n) \subseteq \mathrm{O}(n\log(n)) \subseteq \mathrm{O}(n^2) \quad \text{etc.} \tag{0.0.59}$$

Our interests in these orders are generally different for decreasing functions of n than for increasing sequences in n. The former are often used to measure how quickly an error rate goes to zero, and the latter are often used to evaluate the speed of an algorithm as the problem size grows. In either case, it is important to recognize that the order expressed in big O (or little o) is a lower bound, as indicated in expression (0.0.59). (There are variations on the big O concept referred to as "big Ω" and "big Θ" that specify upper bounds and two-sided bounds.)

Some additional properties of big O classes are the following.

$$b_n \in \mathrm{O}(a_n),\ d_n \in \mathrm{O}(c_n) \Longrightarrow b_n d_n \in \mathrm{O}(a_n c_n), \tag{0.0.60}$$

$$b_n \in \mathrm{O}(a_n),\ d_n \in \mathrm{O}(c_n) \Longrightarrow b_n + d_n \in \mathrm{O}(\|a_n\|\|c_n\|), \tag{0.0.61}$$

$$\mathrm{O}(ca_n) = \mathrm{O}(a_n) \quad \text{for constant } c. \tag{0.0.62}$$

The proofs of these are left as exercises. Similar results hold for little o classes.

In probability and statistics, the sequences may involve random variables, in which case we may need to distinguish the type of convergence. If the convergence is almost sure, there is little difference whether or not the sequence

involves random variables. Weak convergence, however, results in different types of measures for the rate of convergence, and we find the related concepts of big O in probability, O_P, and little o in probability, o_P, useful; see page 83.

Sums of Sequences of Reals

Sums of countable sequences of real numbers $\{x_i\}$, for $i = 1, 2, \ldots$, are often of interest. A sum of a countable sequence of real numbers is called a (real) *series*. The usual question is what is $\lim_{n\to\infty} \sum_{i=1}^{n} x_i$. If this limit is finite, the series is called *convergent*, or the series is said to *converge*; otherwise, the series is called *divergent*, or the series is said to *diverge*. If $\lim_{n\to\infty} \sum_{i=1}^{n} |x_i|$ is finite, the series is called *absolutely convergent*, or the series is said to *converge* absolutely.

We often simply write $\sum x_i$ to mean $\lim_{n\to\infty} \sum_{i=1}^{n} x_i$, and often we restrict the meaning of "series" to refer to this limit; however, I may occasionally use the word "series" to refer to a sum of a finite number of elements.

A useful way to investigate sums of sequences of reals is by use of partial sums. When we are interested in $\sum x_i$, we form the partial sum,

$$S_k = \sum_{i=1}^{k} x_i,$$

where k is some integer. Clearly, assuming the x_is are finite, S_k is finite. The use of partial sums can be illustrated by considering the geometric series, which is the sum of the geometric progression, $a, ar, ar^2, \ldots$. Let

$$S_k = \sum_{i=0}^{k} ar^i.$$

Multiplying both sides by r and subtracting the resulting equation, we have

$$(1 - r)S_k = a(1 - r^{k+1}),$$

which yields for the partial sum

$$S_k = a\frac{1 - r^{k+1}}{1 - r}.$$

This formula is useful for finite sums, but its main use is for the series. If $|r| < 1$, then

$$\sum_{i=0}^{\infty} ar^i = \lim_{k\to\infty} S_k = \frac{a}{1 - r}.$$

If $|r| > 1$, then the series diverges.

Another important fact about series, called Kronecker's lemma, is useful in proofs of theorems about sums of independent random variables, such as the strong law of large numbers (Theorem 1.52, page 103):

Theorem 0.0.10 (Kronecker's Lemma) *Let* $\{x_i \mid i = 1, 2, \ldots\}$ *and* $\{a_i \mid i = 1, 2, \ldots\}$ *be sequences of real numbers such that* $\sum_{i=1}^{\infty} x_i$ *exists (and is finite), and* $0 < a_1 \leq a_2 \leq \ldots$ *and* $a_n \to \infty$. *Then*

$$\lim_{n\to\infty} \frac{1}{a_n} \sum_{i=1}^{n} a_i x_i = 0.$$

Proof. Form the partial sums in x_i, S_k and S_n, with $k < n$. We have

$$\frac{1}{a_n} \sum_{i=1}^{n} a_i x_i = S_n - \frac{1}{a_n} \sum_{i=1}^{n-1} (a_{i+1} - a_i) S_k.$$

Let $s = \sum_{i=1}^{\infty} x_i$, and for any $\epsilon > 0$, let N be such that for $n > N$, $|S_n - s| < \epsilon$. We can now write the left-hand side of the equation above as

$$\begin{aligned}
&S_n - \frac{1}{a_n} \sum_{i=1}^{N-1} (a_{i+1} - a_i) S_k - \frac{1}{a_n} \sum_{i=N}^{n-1} (a_{i+1} - a_i) S_k \\
&\quad = S_n - \frac{1}{a_n} \sum_{i=1}^{N-1} (a_{i+1} - a_i) S_k - \frac{1}{a_n} \sum_{i=N}^{n-1} (a_{i+1} - a_i) s - \frac{1}{a_n} \sum_{i=N}^{n-1} (a_{i+1} - a_i)(S_k - s) \\
&\quad = S_n - \frac{1}{a_n} \sum_{i=1}^{N-1} (a_{i+1} - a_i) S_k - \frac{a_n - a_N}{a_n} s - \frac{1}{a_n} \sum_{i=N}^{n-1} (a_{i+1} - a_i)(S_k - s) \qquad .
\end{aligned}$$

Now, consider $\lim_{n\to\infty}$. The first term goes to s, which cancels with the third term. The second term goes to zero (because the sum is a fixed value). Since the sequence $\{a_i\}$ is nondecreasing, the last term is bounded by $\frac{a_n - a_N}{a_n}\epsilon$, which is less than or equal to ϵ, which was any positive number. ∎

In Section 0.0.9 beginning on page 677, we list some additional ways of determining whether or not a series converges.

Real Functions

Real-valued functions over real domains are some of the most important mathematical objects. Here we will discuss some of their simpler characteristics. In Section 0.1.5 we will consider some properties in more detail and in Sections 0.1.6 through 0.1.13 we will consider important operations on real functions.

We will often consider the domain of a function to be an interval $[a, b]$, or if the domain is in $\mathbb{R}^k$, to be a rectangle $[a_1, b_1] \times \cdots \times [a_k, b_k]$, and many concepts relating to a finite partitioning P of that domain. The partition may be defined by the sets $\{I_i : i \in P\}$, or especially in the case of $[a, b]$ in $\mathbb{R}$, by $(a = x_0, x_1, \ldots, x_n = b)$.

Important properties of functions include *continuity*, *differentiability*, *integrability*, and *shape*. The first three of these properties, which are defined in terms of limits, are essentially dichotomous, but they have various levels depending on whether they hold over certain subdomains of the function.

Taylor's Theorem

One of the most important and useful facts in analysis is Taylor's theorem. We state the theorem here for scalar-valued real functions of a scalar real variable, but similar results hold for more general functions.

Theorem 0.0.11 (Taylor's theorem)
Let f be a function defined on $D \subseteq \mathbb{R}$, let n be a positive integer, suppose that the $(n-1)^{\text{th}}$ derivative of f is continuous on the interval $[x_0, x] \subseteq D$, and suppose that the n^{th} derivative of f exists on the interval $]x_0, x[$. Then,

$$f(x) = f(x_0)+(x-x_0)f'(x_0)+\frac{(x-x_0)^2}{2!}f''(x_0)+\cdots+\frac{(x-x_0)^{n-1}}{(n-1)!}f^{(n-1)}(x_0)+R_n, \tag{0.0.63}$$

where the remainder $R_n = \frac{(x-x_0)^n}{n!}f^{(n)}(\xi)$ with $x_0 < \xi < x$.

The proof starts with the identity

$$f(x) = f(x_0) + \int_{x_0}^{x} f'(t)\mathrm{d}t,$$

and then proceeds iteratively by integrating by parts. This expression also suggests another form of the remainder. It clearly can be expressed as an integral over t of $f^{(n)}(x-t)^{n-1}/n!$.

It is not necessary to restrict x_0 and x as we have in the statement of Taylor's theorem above. We could, for example, require that D is an interval and $x_0, x \in D$.

Notice that for $n = 1$, Taylor's theorem is the mean-value theorem (Theorem 0.0.19).

The properties of the remainder term, R_n, are important in applications of Taylor's theorem.

Taylor Series and Analytic Functions

As $n \to \infty$, if the n^{th} derivative of f continues to exist on the open interval, and if the remainder R_n goes to zero, then we can use Taylor's theorem to represent the function in terms of a *Taylor series*. (The common terminology has a possessive in the name of the theorem and a simple adjective in the name of the series.)

As indicated in the comments above about Taylor's theorem, the points x and x_0 must be in a closed interval over which the derivatives exist and have finite values. The Taylor series is

$$f(x) = f(x_0) + (x-x_0)f'(x_0) + \frac{(x-x_0)^2}{2!}f''(x_0) + \cdots \tag{0.0.64}$$

A Taylor series in the form of equation (0.0.64) is said to be an expansion of f about x_0.

Note that equation (0.0.64) implies that the function is infinitely differentiable and that the Taylor series converges to the function value. If equation (0.0.64) for all real x and x_0, the function f said to be *analytic*. If the

Examples

The are many examples of functions for which the Taylor series expansion does not hold. First of all, it only holds for infinitely differentiable functions.

For infinitely differentiable functions, the question is whether the remainder R_n goes to zero; that is, does the series converge, and if so, does it converge to $f(x)$ and if so is the convergence over the whole domain?

Example 0.0.5 (Convergence only at a single point)
Consider

$$f(x) = \sum_{n=0}^{\infty} \mathrm{e}^{-n} \cos n^2 x.$$

The function is infinitely differentiable but the Taylor series expansion about 0 converges only for $x = x_0$. (Exercise.) ∎

Example 0.0.6 (Convergence only over a restricted interval)
Consider

$$f(x) = \frac{1}{1+x^2}.$$

The function is infinitely differentiable but the Taylor series expansion about 0 converges only for $|x| < 1$. (Exercise.) ∎

Example 0.0.7 (Convergence to the wrong value)
Consider

$$f(x) = \begin{cases} \mathrm{e}^{-1/x^2} & \text{for } x \neq 0 \\ 0 & \text{for } x = 0. \end{cases}$$

The function is infinitely differentiable and $f^n(0) = 0$ for all n. A Taylor series expansion about 0 converges to 0, but $f(x) \neq 0$ if $x \neq 0$. ∎

Functions of Bounded Variation

Another important class of real functions consists of those of *bounded variation*. The function f defined over $[a, b]$ is said to be of bounded variation on $[a, b]$ if there exists a number M such that for any partition $(a = x_0, x_1, \ldots, x_n = b)$

$$\sum_{i=1}^{n} |\Delta f_i| \leq M, \tag{0.0.65}$$

where $\Delta f_i = f(x_i) - f(x_{i-1})$.

A sufficient condition for a function f to be of bounded variation on $[a, b]$ is that it is continuous on $[a, b]$ and its derivative exists and is bounded on $]a, b[$.

Convexity

Another useful concept for real sets and for real functions of real numbers is *convexity*.

Definition 0.0.10 (convex set)
A set $A \subseteq \mathbb{R}^d$ *is convex iff for* $x, y \in A$, $\forall a \in [0,1]$, $ax + (1-a)y \in A$. ∎

Intervals, rectangles, and hyperrectangles are convex.

Theorem 0.0.12
If A *is a set of real numbers that is convex, then both* $\overline{A}$ *and* A° *are convex.*

Proof.
Exercise; just use the definitions. ∎

Definition 0.0.11 (convex function)
A function $f : D \subseteq \mathbb{R}^d \mapsto \mathbb{R}$, where D is convex, is *convex* iff for $x, y \in D$, $\forall a \in [0,1]$,

$$f(ax + (1-a)y) \leq af(x) + (1-a)f(y).$$

A function is *strictly convex* if the inequality above is strict. ∎

Definition 0.0.12 (concave function)
A function f is *(strictly) concave* iff $-f$ is (strictly) convex.

A useful theorem that characterizes convexity of twice differentiable functions is the following

Theorem 0.0.13
If the function f *is twice differentiable over an open convex set* D*, then* f *is convex iff the Hessian,* H_f*, is nonnegative definite at all points in* D*. Iff it is positive definite,* f *is strictly convex.*

For a proof of this theorem, see a text on continuous optimization, such as Griva et al. (2009).

Theorem 0.0.14
The composition of a convex function and a convex function is convex.

Proof.
Let f and g be any convex functions for which $f \circ g$ is defined. Now let a be any real number in $[0,1]$. Then $f \circ g(ax + (1-a)y) \leq f(ag(x) + (1-a)g(y)) \leq af \circ g(x) + (1-a)f \circ g(y)$. ∎

Subharmonic Functions

Convexity of a function is defined in terms of the average of the function at two points, compared to the function at the average of the two points. We can extend that basic idea to the average of the function over a sphere compared to the function at the sphere. (The average of the function over a sphere is defined in terms of the ratio of a measure of the function image to a measure of the surface of the sphere. The measures are integrals.)

Definition 0.0.13 (subharmonic function)
A function $f : D \subseteq \mathbb{R}^d \mapsto \mathbb{R}$, where D is convex, is *subharmonic* over D, iff for every point $x_0 \in D$ and for every $r > 0$, the average of f over the surface of the sphere $S_r(x_0) = \{x : \|x - x_0\| = r\}$ is greater than or equal to $f(x_0)$. ∎

Definition 0.0.14 (superharmonic function)
A function f is *superharmonic* if $-f$ is subharmonic. ∎

Definition 0.0.15 (harmonic function)
A function is *harmonic* if it is both superharmonic and subharmonic. ∎

In one dimension, a subharmonic function is convex and a superharmonic function is concave.

A useful theorem that characterizes harmonicity of twice differentiable functions is the following:

Theorem 0.0.15 *If the function f is twice differentiable over an open convex set D, then f is subharmonic iff the Laplacian, $\nabla^2 f$, (which is just the trace of H_f) is nonnegative at all points in D. The function is harmonic if the Laplacian is 0, and superharmonic if the Laplacian is nonpositive.*

Proof. Exercise. ∎

The relatively simple Laplacian operator captures curvature only in the orthogonal directions corresponding to the principal axes; if the function is twice differentiable everywhere, however, this is sufficient to characterize the (sub-, super-) harmonic property. These properties are of great importance in multidimensional loss functions.

Harmonicity is an important concept in potential theory. It arises in field equations in physics. The basic equation $\nabla^2 f = 0$, which implies f is harmonic, is called Laplace's equation. Another basic equation in physics is $\nabla^2 f = -c\rho$, where $c\rho$ is positive, which implies f is superharmonic. This is called Poisson's equation, and is the basic equation in a potential (electrical, gravitational, etc.) field. A superharmonic function is called a potential for this reason. These PDE's, which are of the elliptical type, govern the diffusion of energy or mass as the domain reaches equilibrium. Laplace's equation represents a steady diffusion and Poisson's equation models an unsteady diffusion, that is, diffusion with a source or sink.

Example 0.0.8
Consider $f(x) = \exp\left(\sum_{j=1}^k x_j^2\right)$. This function is twice differentiable, and we have

$$\nabla^2 \exp\left(\sum_{j=1}^k x_j^2\right) = \sum_{i=1}^k (4x_i^2 - 2) \exp\left(\sum_{j=1}^k x_j^2\right).$$

The exponential term is positive, so the condition depends on $\sum_{i=1}^k (4x_i^2 - 2)$. If $\sum_{i=1}^k x_i^2 < 1/2$, it is superharmonic; if $\sum_{i=1}^k x_i^2 = 1/2$, it is harmonic; if $\sum_{i=1}^k x_i^2 > 1/2$, it is subharmonic. ∎

0.0.6 The Complex Number System

The complex number system, $\mathbb{C}$, can be developed most directly from the real number system by first introducing an element, i, defined as

$$\mathrm{i}^2 = -1.$$

We call i the *imaginary unit*, and define the field $\mathbb{C}$ on the set $\mathbb{R} \times \mathbb{R}$ as the set $\{x + \mathrm{i}y \ ; \ x, y \in \mathbb{R}\}$, together with the operations of addition and multiplication for x and y as in $\mathbb{R}$.

The complex number system is very important in mathematical statistics in three areas: transformations (see Section 0.0.9); transforms, as in characteristic functions (see Section 1.1.7) and Fourier transforms, including "discrete" Fourier transforms; and eigenanalysis. In this section, we will just state some of the important properties of the complex number system.

Operations and Elementary Functions

The operations of complex addition and multiplication can be defined directly in terms of the real operations. For example, if for $z_1, z_2 \in \mathbb{C}$ with $z_1 = x_1 + \mathrm{i}y_1$ and $z_2 = x_2 + \mathrm{i}y_2$, then we denote and define addition by

$$z_1 + z_2 = x_1 + x_2 + \mathrm{i}(y_1 + y_2),$$

and multiplication by

$$z_1 z_2 = (x_1 x_2 - y_1 y_2) + \mathrm{i}(x_1 y_2 + x_2 y_1).$$

Both the sum and the product are in $\mathbb{C}$.

Other operations such as powers and division can be defined in terms of the expected results. For example, for an integer k and $z \in \mathbb{C}$, z^k is the element of $\mathbb{C}$ that would result from the appropriate number of multiplications, and $z^{1/k}$ is the element of $\mathbb{C}$ such that if it is raised to the k^{th} power would yield z. We also note $\mathrm{i}^{-1} = -\mathrm{i}$.

The operation $z^{1/k}$ causes us to recognize another difference in the real and the complex number systems. With the reals, if k is even the operation is defined only over a subset of $\mathbb{R}$ and while there are two real numbers that could be considered as results of the operation, we define the operation so that it yields a single value. In the case of the complex number system, the operation of taking the k^{th} root is defined over all $\mathbb{C}$. The main point, however, is that the operation may yield multiple values and there may not be an immediately obvious one to call the result. In complex analysis, we develop ways of dealing with multi-valued operations and functions.

There are other operations and functions whose definitions do not arise from simple extensions of the corresponding operation or function on the reals. For example, we define "modulus" as an extension of the absolute value: for $z = x + \mathrm{i}y$, we define $|z|$ as $\sqrt{x^2 + y^2}$.

Other examples are the exponentiation operation, the functions $\log(\cdot)$, $\sin(\cdot)$, and so on. The standard way of defining the elementary functions is through analytic continuation of a Taylor series expansion into $\mathbb{C}$. For example, for $z \in \mathbb{C}$, we may define $\exp(z)$ in terms of the convergent series

$$\mathrm{e}^z = 1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + \frac{z^4}{4!} + \cdots \tag{0.0.66}$$

The ratio test can be used to show that this is convergent over all $\mathbb{C}$ (exercise).

The definition of e^z can be used to define the exponentiation operation $z_1^{z_2}$ in general.

Complex Conjugates

For $z = x + \mathrm{i}y \in \mathbb{C}$, we define define the *complex conjugate* as $x - \mathrm{i}y$, and denote it as $\overline{z}$. We define the modulus of z as $\sqrt{z\overline{z}}$, and denote it as $|z|$. It is clear that $|z|$ is real and nonnegative, and it corresponds to the absolute value of x if $z = x + 0\mathrm{i}$.

We have some simple relationships for complex conjugates:

Theorem 0.0.16
For all $z, z_1, z_2 \in \mathbb{C}$, we have

- $\overline{\overline{z}} = z$,
- $\overline{z_1 + z_2} = \overline{z_1} + \overline{z_2}$,
- $\overline{z_1 z_2} = \overline{z_1}\,\overline{z_2}$.

Proof. Exercise. ∎

Euler's Formula

One of the most useful facts is given in *Euler's formula*, for a real number x:

$$\mathrm{e}^{\mathrm{i}x} = \cos(x) + \mathrm{i}\sin(x). \tag{0.0.67}$$

This relationship can be derived in a number of ways. A straightforward method is to expand $\mathrm{e}^{\mathrm{i}x}$ in a Taylor series about 0 and then reorder the terms:

$$\begin{aligned}\mathrm{e}^{\mathrm{i}x} &= 1 + \mathrm{i}x + \frac{(\mathrm{i}x)^2}{2!} + \frac{(\mathrm{i}x)^3}{3!} + \frac{(\mathrm{i}x)^4}{4!} + \cdots \\ &= \left(1 - \frac{x^2}{2!} + \frac{x^4}{4!} + \frac{x^6}{6!} + \cdots\right) + \mathrm{i}\left(x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots\right) \\ &= \cos(x) + \mathrm{i}\sin(x).\end{aligned}$$

We can do this because the series are absolutely convergent for $x \in \mathbb{R}$.

Euler's formula has a number of applications and special cases. For examples,

$$\cos(x) = \frac{\mathrm{e}^{\mathrm{i}x} + \mathrm{e}^{-\mathrm{i}x}}{2},$$

$$\mathrm{e}^{\mathrm{i}\pi} = -1,$$

and

$$\left|\mathrm{e}^{\mathrm{i}x}\right| = 1.$$

Other Properties of $\mathrm{e}^{\mathrm{i}x}$

The function $\mathrm{e}^{\mathrm{i}x}$, where $x \in \mathbb{R}$, has a number of interesting properties that we can derive from the Taylor expansion with integral remainder:

$$\mathrm{e}^{\mathrm{i}x} = \sum_{k=0}^{n} \frac{(\mathrm{i}x)^k}{k!} + \frac{\mathrm{i}^{n+1}}{n!}\int_0^x (x-t)^n \mathrm{e}^{\mathrm{i}t}\mathrm{d}t. \tag{0.0.68}$$

Now, following the ideas suggested for the proof of Theorem 0.0.11 (Taylor's theorem), starting with $n = 1$, evaluation of $\int_0^x (x-t)^n \mathrm{e}^{\mathrm{i}t}\mathrm{d}t$ by integration by parts, and then recognizing the form of the resulting integrals, we let n be a positive integer and get

$$\int_0^x (x-t)^n \mathrm{e}^{\mathrm{i}t}\mathrm{d}t = \frac{x^{n+1}}{n+1} + \frac{\mathrm{i}}{n+1}\int_0^x (x-t)^{n+1}\mathrm{e}^{\mathrm{i}t}\mathrm{d}t. \tag{0.0.69}$$

This gives another form of the remainder in the Taylor series:

$$\mathrm{e}^{\mathrm{i}x} = \sum_{k=0}^{n} \frac{(\mathrm{i}x)^k}{k!} + \frac{\mathrm{i}^n}{(n-1)!}\int_0^x (x-t)^n (\mathrm{e}^{\mathrm{i}t} - 1)\mathrm{d}t. \tag{0.0.70}$$

Now, we see that the remainder in equation (0.0.68) is bounded from above by

$$\frac{|x|^{n+1}}{(n+1)!},$$

and the remainder in equation (0.0.70) is bounded from above by

$$\frac{2|x|^n}{n!}.$$

Hence, we have a bound on the difference in e^{ix} and its approximation by the truncated series:

$$\left| e^{ix} - \sum_{k=0}^{n} \frac{(ix)^k}{k!} \right| \leq \min\left(\frac{2|x|^n}{n!}, \frac{|x|^{n+1}}{(n+1)!} \right). \tag{0.0.71}$$

Ordering the Complex Numbers

We have discussed various orderings of sets (see Section 0.0.1), and for the $\mathbb{R}$ we have seen that there is a useful linear ordering based on the usual inequality relation. This simple linear ordering is not possible for the complex numbers.

Orderings of sets also sometimes carry over to a relation on a field (see Section 0.0.3). While the reals constitute an Archimedean ordered field, there can be no ordering on the complex field.

Theorem 0.0.17
The field $\mathbb{C}$ cannot be ordered.

Proof. In order to show by contradiction that this must be the case, assume the existence in $\mathbb{C}$ of a subset P as in Definition 0.0.4 on page 634. Now $i \neq 0$, so $i \in P$ or $i \in -P$. But $i \notin P$ because $i \circ i \in -P$; furthermore $i \notin -P$ because $i \circ i \in P$; hence, there can be no such P as required in the definition of an ordered field. ∎

0.0.7 Monte Carlo Methods

Monte Carlo methods involve sampling, usually artificially, in the sense that the samples are generated on the computer. To sample from any given distribution, we generally begin with samples from a $U(0,1)$ distribution (or an approximate $U(0,1)$ distribution, in the sense that the samples are generated on the computer). A raw sample of uniforms, $U_1, U_2, \ldots$, is transformed into a sequence $\{X_j\}$ of (pseudo)random variables from a distribution of interest.

We often want the sequence $\{X_j\}$ to be iid. As part of the transformation process, however, we may use a sequence $\{Y_i\}$ that has internal dependencies.

The simplest type of transformation makes use of the inverse of the CDF of the random variable of interest.

Inverse CDF Transformation

Assume that the CDF of the distribution of interest is F_X, and further, suppose that F_X is continuous and strictly monotone.

In that case, if X is a random variable with CDF F_X, then $U = F_X(X)$ has a U$(0, 1)$ distribution.

In the inverse CDF method, we transform each U_i to an X_i by

$$X_i = F_X^{-1}(U_i).$$

If F_X is not continuous or strictly monotone, we can modify this transformation slightly.

Acceptance/Rejection

Often it is not easy to invert the CDF. In the case of Bayesian inference the posterior distribution may be known only proportionally. First, let us consider the problem in which the CDF is known fully.

We will transform an iid sequence $\{U_i\}$ of uniforms into an iid sequence $\{X_j\}$ from a distribution that has a probability density $p(\cdot)$.

We use an intermediate sequence $\{Y_k\}$ from a distribution that has a probability density $g(\cdot)$. (It could also be the uniform distribution.)

Further, suppose for some constant c that $h(x) = cg(x)$ is such that $h(x) \geq p(x)$.

1. Generate a variate y from the distribution having pdf g.
2. Generate independently a variate u from the uniform(0,1) distribution.
3. If $u \leq p(y)/h(y)$, then accept y as the variate, otherwise, reject y and return to step 1.

See Figure 0.1.

To see that the accepted ys have the desired distribution, first let X be the random variable delivered. For any x, because Y (from the density g) and U are independent, we have

$$\begin{aligned}
\Pr(X \leq x) &= \Pr\left(Y \leq x \,|\, U \leq \frac{p(Y)}{cg(Y)}\right) \\
&= \frac{\int_{-\infty}^{x} \int_{0}^{p(t)/cg(t)} g(t)\,\mathrm{d}s\,\mathrm{d}t}{\int_{-\infty}^{\infty} \int_{0}^{p(t)/cg(t)} g(t)\,\mathrm{d}s\,\mathrm{d}t} \\
&= \int_{-\infty}^{x} p(t)\,\mathrm{d}t,
\end{aligned}$$

the distribution function corresponding to p. Differentiating this quantity with respect to x yields $p(x)$.

Obviously, the closer $cg(x)$ is to $p(x)$, the faster the acceptance/rejection algorithm will be, if we ignore the time required to generate y from the dominating density g. A good majorizing function would be such that the l is almost as large as k.

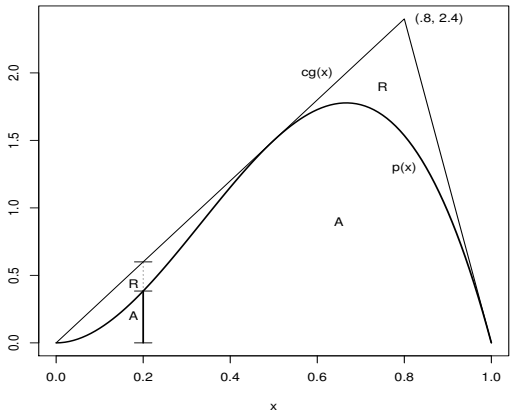


Figure 0.1. Acceptance/Rejection

Often, g is chosen to be a very simple density, such as a uniform or a triangular density. When the dominating density is uniform, the acceptance/rejection method is similar to the "hit-or-miss" method of Monte Carlo quadrature.

Variations of Acceptance/Rejection

There are many variations of the basic acceptance/rejection.

One is called *transformed rejection*. In the transformed acceptance/rejection method, the steps of the algorithm are combined and rearranged slightly.

There are various ways that acceptance/rejection can be used for discrete distributions.

It is clear from the description of the algorithm that the acceptance/rejection method also applies to multivariate distributions. (The uniform random number is still univariate, of course.)

Use of Dependent Random Variables

The methods described above use a sequence of iid variates from the majorizing density. It is also possible to use a sequence from a conditional majorizing density.

A method using a nonindependent sequence is called a Metropolis method, and there are variations of these, with their own names.

There are two related cases:

Suppose $\{X_j : j = 0, 1, 2, \ldots\}$ is such that for $j = 1, 2, \ldots$ we know the conditional distributions of $X_j|X_0, \ldots, X_{j-1}$.

Alternatively, suppose we know the functional form (up to the normalizing constant) of the joint density of $X_1, X_2, \ldots, X_k$, and that we know the distribution of at least one $X_i|X_j (i \neq j)$.

Markov Chain Monte Carlo

A Markov chain is a stochastic process $X_0, X_1, \ldots$ in which the conditional distribution of X_t given $X_0, X_1, \ldots, X_{t-1}$ is the same as the conditional distribution of X_t given only X_{t-1}. An aperiodic, irreducible, positive recurrent Markov chain is associated with a *stationary distribution* or *invariant distribution*, which is the limiting distribution of the chain. See Section 1.6.3 beginning on page 126 for description of these terms.

If the density of interest, p, is the density of the stationary distribution of a Markov chain, correlated samples from the distribution can be generated by simulating the Markov chain.

This appears harder than it is.

A Markov chain is the basis for several schemes for generating random samples. The interest is not in the sequence of the Markov chain itself.

The elements of the chain are accepted or rejected in such a way as to form a different chain whose stationary distribution or limiting distribution is the distribution of interest.

Convergence

An algorithm based on a stationary distribution of a Markov chain is an *iterative method* because a sequence of operations must be performed until they *converge*; that is, until the chain has gone far enough to wash out any transitory phase that depends on where we start.

Several schemes for assessing convergence have been proposed. For example, we could use multiple starting points and then use an ANOVA-type test to compare variances within and across the multiple streams.

The Metropolis Algorithm

For a distribution with density p, the Metropolis algorithm, introduced by Metropolis et al. (1953) generates a random walk and performs an acceptance/rejection based on p evaluated at successive steps in the walk.

In the simplest version, the walk moves from the point y_i to a candidate point $y_{i+1} = y_i + s$, where s is a realization from $U(-a, a)$, and accepts y_{i+1} if

$$\frac{p(y_{i+1})}{p(y_i)} \geq u,$$

where u is an independent realization from $U(0, 1)$.

This method is also called the "heat bath" method because of the context in which it was introduced.

The random walk of Metropolis et al. is the basic algorithm of *simulated annealing*, which is currently widely used in optimization problems.

If the range of the distribution is finite, the random walk is not allowed to go outside of the range.

Example 0.0.9 Simulation of the von Mises Distribution with the Metropolis Algorithm

Consider, for example, the von Mises distribution, with density,

$$p(x) = \frac{1}{2\pi I_0(c)} e^{c\cos(x)}, \quad \text{for } -\pi \leq x \leq \pi,$$

where I_0 is the modified Bessel function of the first kind and of order zero.

The von Mises distribution is an easy one to simulate by the Metropolis algorithm. This distribution is often used by physicists in simulations of lattice gauge and spin models, and the Metropolis method is widely used in these simulations.

It is not necessary to know the normalizing constant, because it is canceled in the ratio. The fact that all we need is a nonnegative function that is proportional to the density of interest is an important property of this method.

If $c = 3$, after a quick inspection of the amount of fluctuation in p, we may choose $a = 1$. The R statements below implement the Metropolis algorithm to generate $n - 1$ deviates from the von Mises distribution.

Notice the simplicity of the algorithm in the R code. We did not need to determine a majorizing density, nor even evaluate the Bessel function that is the normalizing constant for the von Mises density.

```
n <- 1000
x <- rep(0,n)
a <-1
c <-3
yi <-3
j <-0
i <- 2
while (i < n) {
  i <- i + 1
  yip1 <- yi + 2*a*runif(1)- 1
  if (yip1 < pi & yip1 > - pi) {
  if (exp(c*(cos(yip1)-cos(yi))) > runif(1)) yi <- yip1
    else yi <- x[i-1]
  x[i] <- yip1
  }
}
```

A histogram is not affected by the sequence of the output in a large sample.

The Markov chain samplers generally require a "burn-in" period; that is, a number of iterations before the stationary distribution is achieved.

In practice, the variates generated during the burn-in period are discarded.

The number of iterations needed varies with the distribution, and can be quite large, sometimes several hundred.

The von Mises example is unusual; no burn-in is required. In general, convergence is much quicker for univariate distributions with finite ranges such as this one.

It is important to remember what convergence means; it does *not* mean that the sequence is independent from the point of convergence forward. The deviates are still from a Markov chain.

The Metropolis-Hastings Algorithm

The Metropolis-Hastings algorithm uses a more general chain for the acceptance/rejection step.

To generate deviates from a distribution with density p_X it uses deviates from a Markov chain with density $g_{Y_{t+1}|Y_t}$. The conditional density $g_{Y_{t+1}|Y_t}$ is chosen so that it is easy to generate deviates from it.

0. Set $k = 0$.
1. Choose $x^{(k)}$ in the range of p_X. (The choice can be arbitrary.)
2. Generate y from the density $g_{Y_{t+1}|Y_t}(y|x^{(k)})$.
3. Set r:
$$r = p_X(y)\frac{g_{Y_{t+1}|Y_t}(x^{(k)}|y)}{p_X(x^{(k)})g_{Y_{t+1}|Y_t}(y|x^{(k)})}$$
4. If $r \geq 1$, then
   4.a. set $x^{(k+1)} = y$;
   otherwise
   4.b. generate u from uniform(0,1) and
      if $u < r$, then
         4.b.i. set $x^{(k+1)} = y$,
      otherwise
         4.b.ii. set $x^{(k+1)} = x^{(k)}$.
5. If convergence has occurred, then
   5.a. deliver $x = x^{(k+1)}$;
   otherwise
   5.b. set $k = k + 1$, and go to step 2.

Compare the Metropolis-Hastings algorithm with the basic acceptance/rejection method.

The majorizing function in the Metropolis-Hastings algorithm is

$$\frac{g_{Y_{t+1}|Y_t}(x|y)}{p_X(x)\, g_{Y_{t+1}|Y_t}(y|x)}.$$

r is called the "Hastings ratio", and step 4 is called the "Metropolis rejection". The conditional density, $g_{Y_{t+1}|Y_t}(\cdot|\cdot)$ is called the "proposal density" or the "candidate generating density". Notice that because the majorizing function contains p_X as a factor, we only need to know p_X to within a constant of proportionality. As we have mentioned already, this is an important characteristic of the Metropolis algorithms.

As with the acceptance/rejection methods with independent sequences, the acceptance/rejection methods based on Markov chains apply immediately to multivariate random variables.

We can see why this algorithm works by using the same method as we used to analyze the acceptance/rejection method; that is, determine the CDF and differentiate.

The CDF is the probability-weighted sum of the two components corresponding to whether the chain moved or not. In the case in which the chain does move, that is, in the case of acceptance, for the random variable Z whose realization is y, we have

$$\begin{aligned}\Pr(Z \le x) &= \Pr\left(Y \le x \,\middle|\, U \le p(Y)\frac{g(x_i|Y)}{p(x_i)g(Y|x_i)}\right)\\ &= \frac{\int_{-\infty}^{x}\int_{0}^{p(t)g(x_i|t)/(p(x_i)g(t|x_i))} g(t|x_i)\,\mathrm{d}s\,\mathrm{d}t}{\int_{-\infty}^{\infty}\int_{0}^{p(t)g(x_i|t)/(p(x_i)g(t|x_i))} g(t|x_i)\,\mathrm{d}s\,\mathrm{d}t}\\ &= \int_{-\infty}^{x} p_X(t)\,\mathrm{d}t.\end{aligned}$$

Gibbs Sampling

An iterative method, somewhat similar to the use of marginals and conditionals, can also be used to generate multivariate observations. It was first used for a a Gibbs distribution (Boltzmann distribution), and so is called the *Gibbs method*.

In the Gibbs method, after choosing a starting point, the components of the d-vector variate are generated one at a time conditionally on all others.

If p_X is the density of the d-variate random variable X, we use the conditional densities $p_{X_1|X_2,X_3,\cdots,X_d}$, $p_{X_2|X_1,X_3,\cdots,X_d}$, and so on.

At each stage the conditional distribution uses the most recent values of all the other components.

As with other MCMC methods, it may require a number of iterations before the choice of the initial starting point is washed out.

Gibbs sampling is often useful in higher dimensions. It depends on the convergence of a Markov chain to its stationary distribution, so a burn-in period is required.

0. Set $k = 0$.
1. Choose $x^{(k)} \in S$.

2. Generate $x_1^{(k+1)}$ conditionally on $x_2^{(k)}, x_3^{(k)}, \ldots, x_d^{(k)}$,
   Generate $x_2^{(k+1)}$ conditionally on $x_1^{(k+1)}, x_3^{(k)}, \ldots, x_d^{(k)}$,
   ...
   Generate $x_{d-1}^{(k+1)}$ conditionally on $x_1^{(k+1)}, x_2^{(k+1)}, \ldots, x_d^{(k)}$,
   Generate $x_d^{(k+1)}$ conditionally on $x_1^{(k+1)}, x_2^{(k+1)}, \ldots, x_{d-1}^{(k+1)}$.

3. If convergence has occurred, then
   3.a. deliver $x = x^{(k+1)}$;
   otherwise
   3.b. set $k = k + 1$, and go to step 2.

Example 0.0.10 Gibbs Sampling to Generate Independent Normals

Consider X_{t+1} normal with a mean of X_t and a variance of σ^2.

We will generate an iid sample from a standard normal distribution; that is, a normal with a mean of 0 and a variance of 1. In this example, the target distribution is simpler than the proposal.

We start with a x_0, chosen arbitrarily.

We take logs and cancel terms in the expression for r.

The following simple Matlab statements generate the sample.

```
x(1) = x0;
while i < n
  i = i + 1;
  yip1 = yi + sigma*randn;
  lr2 = yi^2 - yip1^2;
  if lr2 > 0
    yi = yip1;
  else
    u = rand;
    if lr2 > log(u)*2
      yi = yip1;
    else
      yi = x(i-1);
    end
  end
  x(i) = yi;
end
plot (x)
```

∎

There are several variations of the basic Metropolis-Hastings algorithm. Two common related methods are Gibbs sampling and hit-and-run sampling. Those methods are particularly useful in multivariate simulation.

Markov chain Monte Carlo has become one of the most important tools in statistics in recent years. Its applications pervade Bayesian analysis, as well

as many Monte Carlo procedures in the frequentist approach to statistical analysis.

Whenever a correlated sequence such as a Markov chain is used, variance estimation must be performed with some care. In the more common cases of positive autocorrelation, the ordinary variance estimators are negatively biased. The method of batch means or some other method that attempts to account for the autocorrelation should be used.

Convergence

Some of the most important issues in MCMC concern the rate of convergence, that is, the length of the burn-in, and the frequency with which the chain advances.

In many applications of simulation, such as studies of waiting times in queues, there is more interest in transient behavior than in stationary behavior.

This is usually not the case in use of MCMC methods. The stationary distribution is the only thing of interest.

The issue of convergence is more difficult to address in multivariate distributions. It is for multivariate distributions, however, that the MCMC method is most useful.

This is because the Metropolis-Hastings algorithm does not require knowledge of the normalizing constants, and the computation of a normalizing constant may be more difficult for multivariate distributions.

Various diagnostics have been proposed to assess convergence. Most of them use multiple chains in one way or another. Use of batch means from separate streams can be used to determine when the variance has stabilized. A cusum plot on only one chain to help to identify convergence.

Various methods have been proposed to speed up the convergence.

Methods of assessing convergence is currently an area of active research.

The question of whether convergence has practically occurred in a finite number of iterations is similar in the Gibbs method to the same question in the Metropolis-Hastings method.

In either case, to determine that convergence has occurred is not a simple problem.

Once a realization is delivered in the Gibbs method, that is, once convergence has been deemed to have occurred, subsequent realizations can be generated either by starting a new iteration with $k = 0$ in step 0, or by continuing at step 1 with the current value of $x^{(k)}$.

If the chain is continued at the current value of $x^{(k)}$, we must remember that the subsequent realizations are not independent.

Effects of Dependence

This affects variance estimates (second order sample moments), but not means (first order moments).

In order to get variance estimates we may use means of batches of subsequences or use just every m^{th} (for some $m > 1$) deviate in step 3. (The idea is that this separation in the sequence will yield subsequences or a systematic subsample with correlations nearer 0.)

If we just want estimates of means, however, it is best not to subsample the sequence; that is, the variances of the estimates of means (first order sample moments) using the full sequence is smaller than the variances of the estimates of the same means using a systematic (or any other) subsample (so long as the Markov chain is stationary.)

To see this, let $\bar{x}_i$ be the mean of a systematic subsample of size n consisting of every m^{th} realization beginning with the i^{th} realization of the converged sequence. Now, we observe that

$$|\mathrm{Cov}(\bar{x}_i, \bar{x}_j)| \leq \mathrm{V}(\bar{x}_l)$$

for any positive i, j, and l less than or equal to m. Hence if $\bar{x}$ is the sample mean of a full sequence of length nm, then

$$\begin{aligned}\mathrm{V}(\bar{x}) &= \mathrm{V}(\bar{x}_l)/m + \sum_{i\neq j;i,j=1}^{m} \mathrm{Cov}(\bar{x}_i, \bar{x}_j)/m^2 \\ &\leq \mathrm{V}(\bar{x}_l)/m + m(m-1)\mathrm{V}(\bar{x}_l)/m \\ &= \mathrm{V}(\bar{x}_l).\end{aligned}$$

In the Gibbs method the components of the d-vector are changed systematically, one at a time. The method is sometimes called *alternating conditional sampling* to reflect this systematic traversal of the components of the vector.

Ordinary Monte Carlo and Iterative Monte Carlo

The acceptance/rejection method can be visualized as choosing a subsequence from a sequence of iid realizations from the distribution with density g_Y in such a way the subsequence has density p_X.

iid from g_Y	y_i	y_{i+1}	y_{i+2}	y_{i+3}	$\cdots$	y_{i+k}	$\cdots$
accept?	no	yes	no	yes	$\cdots$	yes	$\cdots$
iid from p_X		x_j		x_{j+1}	$\cdots$	x_{j+l}	$\cdots$

A Markov chain Monte Carlo method can be visualized as choosing a subsequence from a sequence of realizations from a random walk with density $g_{Y_{i+1}|Y_i}$ in such a way that the subsequence selected has density p_X.

random walk	y_i	$y_{i+1} = y_i + s_{i+1}$	$y_{i+3} = y_{i+1} + s_{i+2}$	$y_{i+2} = y_{i+2} + s_{i+3}$	$\cdots$
accept?	no	yes	no	yes	$\cdots$
iid from p_X		x_j		x_{j+1}	$\cdots$

The general objective in Monte Carlo simulation is to calculate the expectation of some function g of a random variable X. In ordinary Monte Carlo simulation, the method relies on the fact that for independent, identically distributed realizations $X_1, X_2, \ldots$ from the distribution P of X,

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i) \to \mathrm{E}g(X)$$

almost surely as n goes to infinity. This convergence is a simple consequence of the law of large numbers.

In Monte Carlo simulation, the sample is simulated with a random number generator. When X is multivariate or a complicated stochastic process, however, it may be difficult or impossible to simulate independent realizations.

Monte Carlo Applications

Whether the random number generation is direct or iterative, there are generally two kinds of objectives in Monte Carlo applications. One is just to understand a probability distribution better. This may involve merely simulating random observations from the distribution and examining the distribution of the simulated sample.

The other main application of Monte Carlo methods is to evaluate some constant. No matter how complicated the problem is, it can always be formulated as the problem of evaluating a definite integral

$$\int_D f(x)\mathrm{d}x.$$

Using a PDF decomposition (0.0.95) $f(x) = g(x)p(x)$, by equation (0.0.96), we see that the evaluation of the integral is the same as the evaluation of the expected value of $g(X)$ where X is a random variable whose distribution has PDF p with support D.

The problem now is to estimate $\mathrm{E}(g(X))$. If we have a sample $x_1, \ldots, x_n$, the standard way of estimating $\mathrm{E}(g(X))$ is to use

$$\widehat{\mathrm{E}(g(X))} = \frac{1}{n}\sum_{i=1}^{n} g(x_i). \tag{0.0.72}$$

0.0.8 Mathematical Proofs

A mathematical system consists of a body of statements, which may be *definitions*, *axioms*, or *propositions*. A proposition is a conditional statement, which has the form "if A then B", or "$A \Rightarrow B$", where A and B are simple declarative statements or conditional statements. A conditional statement may be true or false, or neither. Our interest in mathematics is to establish the truth or

falsity of a conditional statement; that is, to prove or disprove the statement. A proposition that has a proof is sometimes called a "lemma", a "theorem", or a "corollary". While these terms have meanings, the meanings are rather vague or subjective, and many authors' usage of the different terms serves no purpose other than to annoy the reader. If a proposition has no known proof, it is sometimes called a "conjecture". We usually do not use the term "proposition" to refer to a conditional statement that has been disproved.

A declarative statement has one of two mutually exclusive states: "true" or "false". We denote the negation or falsification of the statement A by $\neg A$.

With a basic proposition, such as

$$A \Rightarrow B,$$

there are associated four related propositions:
contrapositive,

$$\neg B \Rightarrow \neg A,$$

inverse,

$$\neg A \Rightarrow \neg B,$$

converse,

$$B \Rightarrow A,$$

and contradiction,

$$\neg(A \Rightarrow B).$$

If a proposition is true, then its contraposition is also true, but its contradiction is not true. The inverse is the contrapositive of the converse. The contradiction of the contradiction of a proposition is the proposition. Within any mathematical system there are propositions which are neither true nor false.

There are various types of proofs for propositions. Some are "better" than others. (See Aigner and Ziegler (2010) for discussions of different types of proof.) The "best" proof of a proposition is a *direct* proof, which is a sequence of statements "if A then A_1, if $A_1 \ldots, \ldots$ then B", where each statement in the sequence is an axiom or a previously proven proposition. A direct proof is called *deductive*, because each of the steps after the first is deduced from the preceding step.

Occasionally, the Axiom of Choice is used in a proof. This axiom, which we encountered on page 617, is outside the usual axiomatic basis of much of mathematics. The Axiom of Choice basically says that given any collection of sets, *even an infinite collection*, it is possible to form a set consisting of exactly one element from each set in the collection. The Axiom of Choice is tautological for a finite collection.

Whenever the Axiom of Choice is used in a proof, that fact should be stated. Also, whenever an indirect method of proof is used, the type of the proof should be stated or described.

Two useful types of *indirect* proofs are *contradiction* and *induction*. Although proofs of these types often appear very clever, they lack the simple elegance of a direct proof.

In a proof of "$A \Rightarrow B$" by contradiction, we assume "A", and suppose "not B". Then we ultimately arrive at a conclusion that contradicts an axiom or a previously proven proposition. The means that the supposition "not B" cannot be true, and hence that "B" is true. The proof that the Vitali set is not Lebesgue-measurable uses contradiction as well as the Axiom of Choice (see Example 0.1.5 on page 718.)

A proof by induction may be appropriate when we can index a sequence of statements by $n \in \mathbb{Z}_+$, that is, S_n, and the statement we wish to prove is that S_n is true for all $n \geq m \in \mathbb{Z}_+$. We first show that S_m is true. (Here is where a proof by induction requires some care; this statement must be nontrivial; that is, it must be a legitimate member of the sequence of statements.) Then we show that for $n \geq m$, $S_n \Rightarrow S_{n+1}$, in which case we conclude that S_n is true for all $n \geq m \in \mathbb{Z}_+$.

As an example of mathematical induction, consider the statement that for any positive integer n,

$$\sum_{i=1}^{n} i = \frac{1}{2}n(n+1).$$

We use induction to prove this by first showing for $n = 1$,

$$1 = \frac{1}{2}(2).$$

Then we assume that for some $k > 1$,

$$\sum_{i=1}^{k} i = \frac{1}{2}k(k+1),$$

and consider $\sum_{i=1}^{k+1} i$:

$$\begin{aligned}\sum_{i=1}^{k+1} i &= \sum_{i=1}^{k} i + k + 1 \\ &= \frac{1}{2}k(k+1) + k + 1 \\ &= \frac{1}{2}(k+1)((k+1)+1).\end{aligned}$$

Hence, we conclude that the statement is true for any positive integer n.

Another useful type of deductive proof for "$A \Rightarrow B$" is a contrapositive proof; that is, a proof of "not $B \Rightarrow$ not A".

There are some standard procedures often used in proofs. If the conclusion is that two sets A and B are equal, show that $A \subseteq B$ and $B \subseteq A$. To do this (for the first one), choose any $x \in A$ and show $x \in B$. The same technique is used to show that two collections of sets, for example, two σ-fields, are equal.

To show that a sequence converges, use partial sums and an ϵ bound.

To show that a series converges, show that the sequence is a Cauchy sequence.

The standard procedures may not always work, but try them first. In the next section, I describe several facts that are often used in mathematical proofs.

0.0.9 Useful Mathematical Tools and Operations

In deriving results or in proving theorems, there are a number of operations that occur over and over. It is useful to list some of these operations so that they will more naturally come to mind when they are needed. The following subsections list mathematical operations that should be in fast memory. None of these should be new to the reader. In some cases, we mention a specific operation such as completing the square; in other cases, we mention a specific formula such as De Morgan's laws or the inclusion-exclusion formula.

Working with Abstract Sets

Two of the most useful relations are De Morgan's laws, equations (0.0.2) and (0.0.3), and their extensions to countable unions and intersections.

The inclusion-exclusion formula, equation (0.0.8), is particularly useful in dealing with collections of subsets of a sample space.

For a general sequence of sets $\{A_n\}$, the disjoint sequence (0.0.6) $\{D_n\}$ on page 618 that partitions their union is often useful.

If the sequence $\{A_n\}$ is increasing, that is, $A_1 \subseteq A_2 \subseteq \ldots$, the intersection is trivial, but the union $\cup_{n=1}^{\infty} A_n$ may be of interest. In that case, the disjoint sequence (0.0.7) $D_n = A_{n+1} - A_n$ may be useful. Conversely, if the sequence $\{A_n\}$ is decreasing, the union is trivial, but the intersection may be of interest. In that case, De Morgan's laws may be used to change the decreasing sequence into an increasing one.

Working with Real Sets

There are many useful properties of real numbers that simplify operations on them. Recognizing common sequences of reals as discussed beginning on page 651 or sequences of real intervals discussed beginning on page 645 will aid in solving many problems in mathematics. The sequences of intervals

$$O_i = \left]a - \frac{1}{i},\ b + \frac{1}{i}\right[\tag{0.0.73}$$

and

$$C_i = \left[a + \frac{1}{i},\ b - \frac{1}{i}\right]. \tag{0.0.74}$$

given in expressions (0.0.39) and (0.0.40) are worth remembering because

$$\cap_{i=1}^{\infty} O_i = [a, b],$$

that is, it is closed; and

$$\cup_{i=1}^{\infty} C_i =]a, b[,$$

that is, it is open. Note the nesting of the sequences; the sequence $\{O_i\}$ is decreasing and the sequence $\{C_i\}$ is decreasing.

Other sequences of real numbers that may be useful are nested intervals each of the form of O_i or C_i above, for example,

$$I_{ij} = \left]j + \frac{1}{i},\ j - \frac{1}{i}\right[. \tag{0.0.75}$$

These kinds of sequences may be used to form an interesting sequence of unions or intersections; for example,

$$U_j = \cup_{i=1}^{\infty} I_{ij}.$$

Two useful set of integers are the increasing sequence

$$A_i = \{1, \ldots, i\} \tag{0.0.76}$$

and the decreasing sequence

$$B_i = \mathbb{Z}_+ - \{1, \ldots, i\}. \tag{0.0.77}$$

Note that $\cup A_i = \mathbb{Z}_+$ and $\cap B_i = \emptyset$.

Working with Real Sequences and Series

It is helpful to be familiar with a few standard sequences and series such as those we mentioned on page 651 and in Section 0.0.5. A question that arises often is whether or not a given series of real numbers converges. We discussed that issue briefly in Section 0.0.5. Here we list some additional conditions are useful in addressing this question. In the following we write $\sum x_i$ to mean $\lim_{n\to\infty} \sum_{i=1}^{n} x_i$.

Comparison Tests

- If $\sum |y_i|$ converges and $|x_i| \leq |y_i|$ then $\sum x_i$ converges absolutely.
- If $\sum |y_i|$ diverges and $|x_i| \geq |y_i|$ then $\sum |x_i|$ diverges but $\sum x_i$ may converge.

Ratio Test

• If $\lim_{i\to\infty} |x_{i+1}|/|x_i| = \alpha$, then $\sum x_i$ converges absolutely if $\alpha < 1$ and diverges if $\alpha > 1$.

Root Test

• If $\lim_{i\to\infty} \sqrt[n]{|x_i|} = \alpha$, then $\sum x_i$ converges absolutely if $\alpha < 1$ and diverges if $\alpha > 1$.

Raabe's Test

• If $\lim_{n\to\infty} i\,(1 - |x_{i+1}|/|x_i|) = \alpha$, then $\sum x_i$ converges absolutely if $\alpha < 1$ and diverges if $\alpha > 1$.

Alternating Series Test

• If $x_i \geq 0$, $x_i \leq x_{i+1}$, and $\lim_{n\to\infty} x_i = 0$, then $\sum(-1)^i x_i$ converges.

Use of Standard Inequalities

Many mathematical inequalities lead to other interesting facts. I mention several useful inequalities in this chapter. In Appendix B I state versions of many of these in the setting of probabilities or expectations, and I also mention several additional ones in that appendix. Being familiar with these various inequalities (the more, the better!) helps one to prove or disprove other propositions.

The proofs of some of the standard inequalities themselves are templates of techniques that should be in the toolbox of mathematical statisticians. Two examples are the proof of the Cauchy-Schwarz inequality and the proof of Hölder's inequality. In each case, the main fact used may not have an obvious relationship with the inequality itself. For the Cauchy-Schwarz inequality, we use a simple and wellknown fact from the theory of equations. See page 637. For Hölder's inequality, we identify a relevant concave function and then use a property of such a function. See page 642 for a proof that uses a concave function, and see page 852 for a proof of a slightly different version of Hölder's inequality that uses the related convex function. The remembrance of how we get started on these two proofs can help when we are faced with a new proposition to prove or disprove.

Working with Real-Valued Functions

When dealing with general real-valued functions, it is often useful to decompose the function into its nonnegative part and its negative part. In this way, the function f is written as

$$f = f_+ - f_-.$$

An example of this technique is in the definition of the Lebesgue integral, Definition 0.1.41.

Use of Transformations

Many problems are simplified by use of transformations of the variables. Some useful transformations are those between trigonometric and exponential functions, such as Euler's formula,

$$\mathrm{e}^{\mathrm{i}(nx)} = \cos(nx) + \mathrm{i}\sin(nx), \tag{0.0.78}$$

for integer n and real x.

Euler's formula yields de Moivre's formula for multiples of angles,

$$(\cos(x) + \mathrm{i}\sin(x))^n = \cos(nx) + \mathrm{i}\sin(nx), \tag{0.0.79}$$

again for integer n and real x. (Note, in fact, that this formula does not hold for non-integer n.) There are many other formulas among the trigonometric functions that can be useful for transforming variables.

Another very useful class of transformations are those that take cartesian coordinates into circular systems. In two dimensions, the "polar coordinates" ρ and θ in terms of the cartesian coordinates x_1 and x_2 are

$$\begin{aligned} \rho &= \sqrt{x_1^2 + x_2^2} \\ \theta &= \begin{cases} 0 & \text{if} \quad x_1 = x_2 = 0 \\ \arcsin(x_2/\sqrt{x_1^2 + x_2^2}) & \text{if} \quad x_1 \geq 0 \\ \pi - \arcsin(x_2/\sqrt{x_1^2 + x_2^2}) & \text{if} \quad x_1 < 0 \end{cases} \end{aligned} \tag{0.0.80}$$

The extension of these kinds of transformations to higher dimensions is called "spherical coordinates".

Expansion in a Taylor Series

One of the most useful tools in analysis is the Taylor series expansion of a function about a point a. For a scalar-valued function of a scalar variable, it is

$$f(x) = f(a) + (x-a)f' + \frac{1}{2!}(x-a)^2 f'' + \cdots, \tag{0.0.81}$$

if the derivatives exist and the series is convergent. (The class of functions for which this is the case in some region that contains x and a is said to be *analytic* over that region; see page 656. An important area of analysis is the study of analyticity.)

In applications, the series is usually truncated, and we call the series with $k+1$ terms, the k^{th} order Taylor expansion.

For a function of m variables, it is a rather complicated expression:

$$f(x_1,\ldots,x_m) = \sum_{j=0}^{\infty}\left(\frac{1}{j!}\left(\sum_{k=1}^{m}(x_k-a_k)\frac{\partial}{\partial x_k}\right)^j f(x_1,\ldots,x_m)\right)_{(x_1,\ldots,x_m)=(a_1,\ldots,a_m)}. \tag{0.0.82}$$

The second order Taylor expansion for a function of an m-vector is the much simpler expression.

$$f(x) \approx f(a) + (x-a)^{\mathrm{T}}\nabla f(a) + \frac{1}{2}(x-a)^{\mathrm{T}}\mathrm{H}_f(a)(x-a), \tag{0.0.83}$$

where $\nabla f(a)$ is the vector of first derivatives evaluated at a and $\mathrm{H}_f(a)$ is the matrix of second second derivatives (the Hessian) evaluated at a. This is the basis for Newton's method in optimization, for example. Taylor expansions beyond the second order for vectors becomes rather messy (see the expression on the right side of the convergence expression (1.200) on page 95, for example).

Mean-Value Theorem

Two other useful facts from calculus are Rolle's theorem and the mean-value theorem, which we state here without proof. (Proofs are available in most texts on calculus.)

Theorem 0.0.18 (Rolle's theorem)
Assume the function $f(x)$ is continuous on $[a,b]$ and differentiable on $]a,b[$. If $f(a) = f(b)$, then there exists a point x_0 with $a < x_0 < b$ such that $f'(x_0) = 0$.

Theorem 0.0.19 (mean-value theorem)
Assume the function $f(x)$ is continuous on $[a,b]$ and differentiable on $]a,b[$. Then there exists a point x_0 with $a < x_0 < b$ such that

$$f(b) - f(a) = (b-a)f'(x_0).$$

Evaluation of Integrals

There are many techniques that are useful in evaluation of a definite integral. Before attempting to evaluate the integral, we should establish that the integral is finite. For example, consider the integral

$$\int_{-\infty}^{\infty} \exp(-t^2/2)\mathrm{d}t. \tag{0.0.84}$$

A technique for evaluation of this integral is to re-express it as an iterated integral over a product space. (See Exercise 0.0.22; this is an application of Fubini's theorem.) Before doing this, however, we might ask whether the integral is finite. Because $\exp(-t^2/2)$ decreases rapidly in the tails, there is a good chance that the integral is finite. We can see it directly, however, by observing that

$$0 \ < \ \exp(-t^2/2) \ < \ \exp(-|t|+1) \quad -\infty < t < \infty,$$

and

$$\int_{-\infty}^{\infty} \exp(-|t|+1)\mathrm{d}t = 2\mathrm{e}.$$

One of the simplest techniques for evaluation of an integral is to express the integral in terms of a known integral, as we discuss in Section 0.0.9.

Use of Known Integrals and Series

The standard families of probability distributions provide a compendium of integrals and series with known values.

Integrals

There are three basic continuous univariate distributions that every student of mathematical statistics should be familiar with. Each of these distributions is associated with an integral that is important in many areas of mathematics.

- over $\mathbb{R}$; the normal integral:

$$\int_{-\infty}^{\infty} \mathrm{e}^{-(x-\mu)^2/2\sigma^2}\mathrm{d}x = \sqrt{2\pi}\sigma, \tag{0.0.85}$$

  for $\sigma > 0$, and its multivariate extension,
- over $\mathbb{R}_+$; the gamma integral (called the complete gamma function):

$$\int_0^{\infty} \frac{1}{\gamma^\alpha} x^{\alpha-1}\mathrm{e}^{-x/\gamma}\mathrm{d}x = \Gamma(\alpha), \tag{0.0.86}$$

  for $\alpha, \gamma > 0$.
- over $]0,1[$; the beta integral (called the complete beta function):

$$\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\mathrm{d}x = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}, \tag{0.0.87}$$

  for $\alpha, \beta > 0$.

Multivariate integrals

Both the normal distribution and the beta distribution have important and straightforward multivariate extensions. These are associated with important multivariate integrals.

- over $\mathbb{R}^d$; Aitken's integral:

$$\int_{\mathbb{R}^d} \mathrm{e}^{-(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)/2}\,\mathrm{d}x = (2\pi)^{d/2}|\Sigma|^{1/2}, \tag{0.0.88}$$

  for positive definite Σ^{-1}.
- over $]0,1[^d$; Dirichlet integral:

$$\int_{]0,1[^d} \prod_{i=1}^{d} x_i^{\alpha_i-1}\left(1-\sum_{i=1}^{d} x_i\right)^{\alpha_{d+1}-1}\mathrm{d}x = \frac{\prod_{i=1}^{d+1}\Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{d+1}\alpha_i)}. \tag{0.0.89}$$

Series

There are four simple series that should also be immediately recognizable:

- over $0,\ldots,n$; the binomial series:

$$\sum_{x=0}^{n}\frac{\Gamma(n+1)}{\Gamma(x+1)\Gamma(n-x+1)}\pi^x(1-\pi)^{n-x} = 1, \tag{0.0.90}$$

  for $0<\pi<1$ and $n\geq 1$.
- over $\max(0,N-L+M),\ldots,\min(N,M)$; the hypergeometric series:

$$\sum_{x=\max(0,N-L+M)}^{\min(N,M)}\binom{M}{x}\binom{L-M}{N-x} = \binom{L}{n}, \tag{0.0.91}$$

  for $1\leq L$, $0\leq N\leq L$, and $0\leq M\leq L$.
- over $0,1,2,\ldots$; the geometric series:

$$\sum_{x=0}^{\infty}(1-\pi)^x = \pi^{-1} \tag{0.0.92}$$

  for $0<\pi<1$.
- over $0,1,2,\ldots$; the Poisson series:

$$\sum_{x=0}^{\infty}\frac{\theta^x}{x!} = \mathrm{e}^{\theta}, \tag{0.0.93}$$

  for $\theta>0$.

The beta integral and the binomial series have a natural connection through the relation

$$\frac{\Gamma(n+1)}{\Gamma(x+1)\Gamma(n-x+1)} = \binom{n}{x}. \tag{0.0.94}$$

The Dirichlet integral, which is a generalization of the beta, has a similar relation to the multinomial series, which is a generalization of the binomial.

For computing expected values or evaluating integrals or sums, the trick often is to rearrange the integral or the sum so that it is in the form of the original integrand or summand with different parameters.

As an example, consider the integral that is the q^{th} raw moment of a gamma(α, β) random variable:

$$\int_0^\infty \frac{1}{\Gamma(\alpha)\beta^\alpha} x^q x^{\alpha-1} \mathrm{e}^{-x/\beta} \mathrm{d}x.$$

We use the known value of the integral of the density:

$$\int_0^\infty \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} \mathrm{e}^{-x/\beta} \mathrm{d}x = 1.$$

So

$$\begin{aligned}
\int_0^\infty \frac{1}{\Gamma(\alpha)\beta^\alpha} x^q x^{\alpha-1} \mathrm{e}^{-x/\beta} \mathrm{d}x &= \int_0^\infty \frac{1}{\Gamma(\alpha)} \frac{\Gamma(q+\alpha)\beta^q}{\Gamma(q+\alpha)\beta^{q+\alpha}} x^{(q+\alpha)-1} \mathrm{e}^{-x/\beta} \mathrm{d}x \\
&= \frac{\Gamma(q+\alpha)\beta^q}{\Gamma(\alpha)} \int_0^\infty \frac{1}{\Gamma(q+\alpha)\beta^{q+\alpha}} x^{(q+\alpha)-1} \mathrm{e}^{-x/\beta} \mathrm{d}x \\
&= \frac{\Gamma(q+\alpha)\beta^q}{\Gamma(\alpha)}
\end{aligned}$$

Another example is a series of the form

$$\sum_{x=0}^\infty x^q \, \theta^x \, \frac{\mathrm{e}^{-\theta}}{x!}.$$

We recognize in this the known series that corresponds to the probability function associated with the Poisson distribution:

$$\sum_{x=0}^\infty \theta^x \, \frac{\mathrm{e}^{-\theta}}{x!} = 1,$$

and realize that evaluation of the series involves a manipulation of x^q and $x!$. For $q = 1$, we have

$$\begin{aligned}
\sum_{x=0}^\infty x \, \theta^x \, \frac{\mathrm{e}^{-\theta}}{x!} &= \theta \sum_{x=1}^\infty \theta^{(x-1)} \, \frac{\mathrm{e}^{-\theta}}{(x-1)!} \\
&= \theta.
\end{aligned}$$

For $q = 2$, we form two sums so that we can get expressions involving the basic probability function:

$$\begin{aligned}\sum_{x=0}^{\infty} x^2\,\theta^x\,\frac{\mathrm{e}^{-\theta}}{x!} &= \sum_{x=2}^{\infty} x(x-1)\,\theta^x\,\frac{\mathrm{e}^{-\theta}}{x!} + \sum_{x=1}^{\infty} x\,\theta^x\,\frac{\mathrm{e}^{-\theta}}{x!} \\ &= \theta^2 \sum_{x=2}^{\infty} \theta^{(x-2)}\,\frac{\mathrm{e}^{-\theta}}{(x-2)!} + \theta \sum_{x=1}^{\infty} \theta^{(x-1)}\,\frac{\mathrm{e}^{-\theta}}{(x-1)!} \\ &= \theta^2 + \theta.\end{aligned}$$

The PDF Decomposition

It is often useful to decompose a given function f into a product of a PDF p and a function g:

$$f(x) = g(x)p(x). \tag{0.0.95}$$

An appropriate PDF depends on the domain of f, of course. For continuous functions over a finite domain, a scaled PDF of a beta distribution is often useful; for a domain of the form $[a, \infty[$, a shifted gamma works well; and for $]-\infty, \infty[$, a normal is often appropriate.

The PDF decomposition yields the relation

$$\int f(x)\,\mathrm{d}x = \mathrm{E}(g(X)), \tag{0.0.96}$$

where the expectation is taken wrt the distribution with PDF p.

The PDF decomposition is useful in Monte Carlo applications. When the PDF is chosen appropriately, the technique is called "importance sampling". The PDF decomposition is also used often in function estimation.

Completing the Square

Squared binomials occur frequently in statistical theory, often in a loss function or as the exponential argument in the normal density function. Sometimes in an algebraic manipulation, we have an expression of the form ax^2+bx, and we want an expression for this same quantity in the form $(cx+d)^2+e$, where e does not involve x. This form can be achieved by adding and subtracting $b^2/(4a)$, so as to have

$$ax^2 + bx = \left(\sqrt{a}x + b/(2\sqrt{a})\right)^2 - b^2/(4a). \tag{0.0.97}$$

We have a similar operation for vectors and positive definite matrices. If A is a positive definite matrix (meaning that $A^{-\frac{1}{2}}$ exists) and x and b are vectors, we can complete the square of $x^{\mathrm{T}}Ax + x^{\mathrm{T}}b$ in a similar fashion: we add and subtract $b^{\mathrm{T}}A^{-1}b/4$. This gives

$$\left(A^{\frac{1}{2}}x + A^{-\frac{1}{2}}b/2\right)^{\mathrm{T}} \left(A^{\frac{1}{2}}x + A^{-\frac{1}{2}}b/2\right) - b^{\mathrm{T}}A^{-1}b/4$$

or

$$\left(x + A^{-1}b/2\right)^{\mathrm{T}} A \left(x + A^{-1}b/2\right) - b^{\mathrm{T}}A^{-1}b/4. \tag{0.0.98}$$

This is a quadratic form in a linear function of x, together with a quadratic form $b^{\mathrm{T}}A^{-1}b/4$ that does not involve x.

The "Pythagorean Theorem" of Statistics

We often encounter a mean or an expectation taken over a squared binomial. In this case, it may be useful to decompose the squared binomial into a sum of squared binomials or an expectation of a squared binomial plus a single squared binomial:

$$\sum_{i=1}^{n}(x_i - m)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - m)^2, \tag{0.0.99}$$

where $\bar{x} = \sum x_i/n$, and

$$\mathrm{E}((X - m)^2) = \mathrm{E}((X - \mu)^2) + (\mu - m)^2, \tag{0.0.100}$$

where $\mu = \mathrm{E}(X)$.

This of course is also true for the d-vectors X, m, and μ:

$$\mathrm{E}(\|X - m\|_2) = \mathrm{E}(\|X - \mu\|_2) + \|\mu - m\|2, \tag{0.0.101}$$

where $\mu = \mathrm{E}(X)$.

Because the second term on the right-hand side in each of the equations above is positive, we can conclude that $\bar{x}$ and μ are the respective minimizers of the left-hand side, wrt a variable m.

Orthogonalizing Linearly Independent Elements of a Vector Space

Given a set of nonnull, linearly independent vectors, $x_1, x_2, \ldots$, it is easy to form orthonormal vectors, $\tilde{x}_1, \tilde{x}_2, \ldots$, that span the same space. This can be done with respect to any inner product and the norm defined by the inner product. The most common inner product for vectors of course is $\langle x_i, x_j\rangle = x_i^{\mathrm{T}} x_j$, and the Euclidean norm, $\|x\| = \sqrt{\langle x, x\rangle}$, which we often write without the subscript.

$$\begin{aligned}
\tilde{x}_1 &= \frac{x_1}{\|x_1\|} \\
\tilde{x}_2 &= \frac{(x_2 - \langle \tilde{x}_1, x_2\rangle \tilde{x}_1)}{\|x_2 - \langle \tilde{x}_1, x_2\rangle \tilde{x}_1)\|} \\
\tilde{x}_3 &= \frac{(x_3 - \langle \tilde{x}_1, x_3\rangle \tilde{x}_1 - \langle \tilde{x}_2, x_3\rangle \tilde{x}_2)}{\|x_3 - \langle \tilde{x}_1, x_3\rangle \tilde{x}_1 - \langle \tilde{x}_2, x_3\rangle \tilde{x}_2\|} \\
&\text{etc.}
\end{aligned}$$

These are called *Gram-Schmidt transformations*. These transformations also apply to other kinds of objects, such as functions, for which we can define an inner product. *(Note: the third expression above, and similar expressions for subsequent vectors may be numerically unstable. See Gentle (2007), pages 27–29 and 432, for a discussion of numerical issues.)*

Expansion in Basis Elements of a Vector Space

If $q_1, q_2, \ldots$ form an orthonormal basis for the real vector space $\mathcal{V}$, then $x \in \mathcal{V}$ can be expressed in the form

$$x = \sum c_i q_i, \tag{0.0.102}$$

where $c_1, c_2, \ldots$ are real numbers, called the Fourier coefficients with respect to the basis $(q_1, q_2, \ldots)$. (I have left off the limits, because the vector space may be infinite dimensional.) The Fourier coefficients satisfy

$$c_k = \langle x, q_k \rangle. \tag{0.0.103}$$

If the inner product arises from a norm (the L_2 norm!), then for any fixed j, the approximation to x

$$\tilde{x} = \sum_{i=1}^{j} c_i q_i,$$

where $c_1, \ldots, c_j$ are the Fourier coefficients is better than any other approximation of the form $\sum_{i=1}^{j} a_i q_i$ in the sense that

$$\left\| x - \sum_{i=1}^{j} c_i q_i \right\| \leq \left\| x - \sum_{i=1}^{j} a_i q_i \right\|. \tag{0.0.104}$$

Discrete Transforms

Operations on a vector can often be facilitated by first forming an inner product with the given vector and another specific vector that has an additional argument. This inner product is a function in the additional argument. The function is called a *transform* of the given vector. Because of the linearity of inner products, these are *linear transforms*.

One of the most useful transforms is the discrete Fourier transform (DFT), which is the weighted inner product of a given n vector with the vector

$$f(s) = (\mathrm{e}^{-2\pi \mathrm{i} s}, \mathrm{e}^{-4\pi \mathrm{i} s}, \ldots, \mathrm{e}^{-2n\pi \mathrm{i} s});$$

that is, for the n-vector x,

$$\begin{aligned} d(s) &= \frac{1}{\sqrt{n}} \langle x, f(s) \rangle \\ &= \frac{1}{\sqrt{n}} \sum_{t=1}^{\infty} x_t \mathrm{e}^{-2t\pi \mathrm{i} s} \\ &= \frac{1}{\sqrt{n}} \left(\sum_{t=1}^{\infty} x_t, \cos(-2t\pi s) - \mathrm{i} \sum_{t=1}^{\infty} x_t, \sin(-2t\pi s) \right). \end{aligned} \tag{0.0.105}$$

*** discuss uses

*** describe FFT

We will discuss continuous transforms and transforms generally in Section 0.1.12.

Differential Equations and Difference Equations

Many processes of interest can be modeled by differential equations or by difference equations. This is because in many cases the change in a system may be easy to observe, it may follow some physical law, or else its behavior can be related in a natural way to other observable events or measurable variables.

A *differential equation* is an equation that involves one or more derivatives. The variable(s) with respect to which the derivatives are taken are called "independent variable(s)". If all of the derivatives are taken with respect to the same variable, the equation is an *ordinary differential equation* or ODE; otherwise, it is a *partial differential equation* or PDE. A *solution* to a differential equation is an equation involving the variables of the differential equation that satisfies the differential equation identically, that is, for all values of the independent variables.

In ordinary differential equations, we generally denote a derivative by a prime on the variable being differentiated: y', y'', etc. A differential equation has an indefinite number of solutions; for example, the differential equation

$$y'' - y = 0,$$

in which x is the independent variable; that is, in which $y' \equiv \mathrm{d}y/\mathrm{d}x$ and $y'' \equiv \mathrm{d}y'/\mathrm{d}x$, has solutions

$$y = c_1 \mathrm{e}^x + c_2 \mathrm{e}^{-x},$$

where c_1 and c_2 are any constants.

***Define terms general solution initial value, boundary value particular solution order

*** types of ODEs order and degree separable

*** methods of solution

*** difference equations types, solutions

Optimization

Many statistical methods depend on maximizing something (e.g., MLE), or minimizing something, generally a risk (e.g., UMVUE, MRE) or something that has an intuitive appeal (e.g., squared deviations from observed values, "least squares").

First of all, when looking for an optimal solution, it is important to consider the problem carefully, and not just immediately differentiate something and set it equal to 0.

A practical optimization problem often has constraints of some kind.

$$\begin{aligned} \min_{\alpha} \quad & f(x, \alpha) \\ \text{s.t. } & g(x, \alpha) \leq b. \end{aligned}$$

If the functions are differentiable, and if the minimum occurs at an interior point, use of the Lagrangian is usually the way to solve the problem.

With the dependence on x suppressed, the Lagrangian is

$$L(\alpha, \lambda) = f(\alpha) + \lambda^{\mathrm{T}}(g(\alpha) - b).$$

Differentiating the Lagrangian and setting to 0, we have a system of equations that defines a stationary point, α_*.

For twice-differentiable functions, we check to insure that it is a minimum by evaluating the Hessian,

$$\nabla\nabla f(\alpha)\Big|_{\alpha=\alpha_*}.$$

If this is positive definite, there is a local minimum at α_*.

There are many other techniques useful in optimization problems, such as EM methods. We discuss various methods of optimization further in Section 0.4 beginning on page 822.

Some Useful Limits

There are a number of general forms of expressions involving fractions, exponentials, or trigonometric functions that occur frequently in limit operations. It is helpful to be familiar with some of these standard forms, so that when they arise in the course of a proof or derivation, we can quickly evaluate them and move on. I list some useful limits below, in no particular order.

$$\lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n = \mathrm{e}^x. \tag{0.0.106}$$

$$\lim_{h\to 0}\frac{1-\cos(hx)}{h^2} = \frac{1}{2}x^2. \tag{0.0.107}$$

Notes and References for Section 0.0

It is important that the student fully understand the concept of a mathematical proof. Solow (2003) discusses the basic ideas, and Aigner and Ziegler (2010), whose title comes from a favorite phrase of Paul Erdös, give many well-constructed proofs of common facts.

Gelbaum and Olmsted (1990, 2003) have remarked that mathematics is built on two types of things: theorems and counterexamples. Counterexamples help us to understand the principles in a way that we might miss if we only considered theorems. Counterexamples delimit the application of a theorem. They help us understand why each part of the hypothesis of a theorem is important.

The book by Romano and Siegel (1986) is replete with examples that illustrate the "edges" of statistical properties. Other books of this general type in various areas of mathematics are listed below.

Khuri (2003) describes the important facts and techniques in advanced calculus and other areas of applied mathematics that should be in every statistician's toolbox.

Sometimes a step in a proof or a derivation may seem to be a "trick", and the student may ask "how could I have thought of that?" Often these tricks involve thinking of an appropriate expansion or recognizing a convergent series. Jolley (1961) provides a useful compendium of convergent series.

A book that I encountered rather late in my life in mathematics is Graham et al. (1994). This beautiful book presents many "tricks", and puts them in a context in which they appear to be the natural thing to think of.

Exercises for Section 0.0

0.0.1. Prove Theorem 0.0.2.

0.0.2. Use De Morgan's laws to prove equation (0.0.20).

0.0.3. Prove equations (0.0.22) and (0.0.24).

0.0.4. Let $(S, \circ)$ be a group with identity e. Let x be any element of S. Prove:
   a) $x \circ e = e \circ x$.
   b) $x \circ x^{-1} = x^{-1} \circ x$.
   c) e is unique.
   d) for given x, x^{-1} is unique.

0.0.5. Let $(\mathcal{G}, \circ)$ be a group of functions on X and $\circ$ is function composition. Show that the functions must be bijections. (Compare Example 0.0.4.)

0.0.6. a) Show that a sequence $\{x_n\} \in \mathbb{R}^d$ that converges to $x \in \mathbb{R}^d$ is a Cauchy sequence. (In $\mathbb{R}^d$ convergence is usually defined in terms of the Euclidean metric, but that is not necessary.)
   b) Show that each Cauchy sequence is bounded.
   c) Show that if a Cauchy sequence has a subsequence that converges to x, then the original sequence converges to x.
   d) Finally, prove the **Cauchy criterion**: There is a number $x \in \mathbb{R}^d$ to which the sequence $\{x_n\} \in \mathbb{R}^d$ converges iff $\{x_n\}$ is a Cauchy sequence.

0.0.7. Consider the set $D \subseteq \mathbb{R}^2$ where $x_i \in D$ is

$$x_i = (i, 1/i).$$

Define a total order on D using the ordinary order relations in $\mathbb{R}$.

0.0.8. Prove Theorem 0.0.3.
Hint: Use the fact that the characteristic must be 0.

0.0.9. Using Definition 0.0.6, show that in the linear space S, for any $x \in S$,

$$0x = 0_s,$$

where 0_s is the additive identity in S.

0.0.10. Let $\langle \cdot, \cdot \rangle$ be an inner product on Ω. Show that for $x, y \in \Omega$,

$$\langle x + y, x + y \rangle \leq \langle x, x \rangle + \langle y, y \rangle.$$

0.0.11. Prove Theorem 0.0.5.

0.0.12. Prove statement (0.0.33).

0.0.13. Let $x \in \mathbb{R}^d$.

a) Prove

$$\lim_{p\to\infty} \|x\|_p = \max(\{|x_i|\}).$$

b) Prove that $h(x) = \max(\{|x_i|\})$ is a norm over $\mathbb{R}^d$.

0.0.14. Prove equations (0.0.51) through (0.0.55).

0.0.15. Prove statement (0.0.59).

0.0.16. Suppose $\|g(n)\|/\|f(n)\| \to c$ as $n \to \infty$, where c is a finite constant.

a) Show that $g(n) \in \mathrm{O}(f(n))$.

b) Suppose also that $g(n) \in \mathrm{O}(h(n))$. What can you say about $\mathrm{O}(f(n))$ and $\mathrm{O}(h(n))$?

0.0.17. a) Prove statements (0.0.60) through (0.0.62).

b) Prove statements (0.0.60) through (0.0.62) with little o in place of big O.

0.0.18. Why is equation (0.0.43) true?

0.0.19. Taylor series of real univariate functions.

a) Consider

$$f(x) = \sum_{n=0}^{\infty} \mathrm{e}^{-n} \cos(n^2 x).$$

Show that the function is infinitely differentiable but the Taylor series expansion about 0 converges only for $x = 0$.

b) Consider

$$f(x) = \frac{1}{1+x^2}.$$

Show that the function is infinitely differentiable but the Taylor series expansion about 0 converges only for $|x| < 1$.

c) Consider

$$f(x) = \begin{cases} \mathrm{e}^{-1/x^2} & \text{for } x \neq 0 \\ 0 & \text{for } x = 0. \end{cases}$$

(i) Show that the function is infinitely differentiable but that the Taylor series expansion about 0 does not converge to $f(x)$ if $x \neq 0$.

(ii) In the notation of equation (0.0.63), what is R_n in this case?

d) Consider

$$f(x) = \mathrm{e}^x.$$

Write out the Taylor series expansion about 0 and show that it converges to $f(x)$ $\forall x \in \mathbb{R}$.

0.0.20. Use the ratio test to show that the series in equation (0.0.66) is convergent.

0.0.21. Prove Theorem 0.0.16.

0.0.22. Evaluation of definite integrals.

a) Evaluate the integral (0.0.84):

$$I = \int_{-\infty}^{\infty} \exp(-x^2/2)\mathrm{d}x.$$

Hint: Write I^2 as the iterated integral in x and y with integrand $\exp(-(x^2+y^2)/2)$ and then change to polar coordinates.

b) Evaluate the integral

$$\int_0^{\infty} \frac{\sin(x)}{x}\mathrm{d}x,$$

interpreted in the Lebesgue sense as

$$\lim_{t\to\infty} \int_0^t \sin(x)/x \,\mathrm{d}x.$$

Hint: Write this as an integral in a product space and show that Fubini's theorem applies. See Billingsley (1995), page 235.

0.1 Measure, Integration, and Functional Analysis

Measure and integration and the probability theory built on those topics are major fields in mathematics. The objective of this section is just to get enough measure theory to support the probability theory necessary for a solid foundation in statistical inference.

Most of the early development in this section is for abstract objects; for each of these, however, there is a concrete instance that is relevant in probability theory. Althought the original development of these concepts generally involved real numbers and the ideas were later generalized, nowadays it is more satisfying to develop general ideas in the abstract, and then to specialize them to real numbers or to whatever structure is of interest.

We begin with abstract measurable spaces in Section 0.1.1 and then measures over general spaces in Section 0.1.3. A measurable space together with a measure is a *measure space*.

In Section 0.1.4, we discuss an important measure space, namely the reals, the Borel σ-field on the reals, and Lebesgue measure. In Section 0.1.5 we discuss real-valued functions over the reals. We then discuss integration and differentiation. In the ordinary calculus, differentiation is usually introduced before integration, and then the two are associated by means of the "Fundamental Theorem". In analysis, the order is usually reversed, and so in Section 0.1.6 we discuss integration of real functions, and then in Section 0.1.7 we define derivatives.

Sections 0.1.8 through 0.1.13 cover some basics of real functional analysis, including a calculus over functionals.

0.1.1 Basic Concepts of Measure Theory

Analysis depends heavily on a primitive concept of a set, or a collection of "elements". We also use "space" as a primitives, but it is usually just a set that may have some special properties. We generally avoid nesting "set"; that is, rather than a "set of sets", we speak of a collections of sets.

The Sample Space and Subsets of It

A sample space is a nonempty set. It is the "universe of discourse" in a given problem. It is often denoted by Ω. Interesting structures are built on a given set Ω by defining some important types of collections of subsets of Ω.

Special properties of collections of subsets of the sample space define π-systems, rings, fields or algebras, and Dynkin systems, or Sierpinski systems, λ-systems. The collection of subsets that constitute a general field is closed with respect to finite unions. A very important field is one in which the collection of subsets is also closed with respect to countable unions. This is called a σ-field.

We will now define these systems.

Definition 0.1.1 (π-system)
A nonempty collection of subsets, $\mathcal{P}$, is called a *π-system* iff

(π_1) $A, B \in \mathcal{P} \Rightarrow A \cap B \in \mathcal{P}$.

The name π-system comes from the condition that the collection includes products or intersections.

Definition 0.1.2 (ring)
A nonempty collection of subsets, $\mathcal{R}$, is called a *ring* iff

(r_1) $A, B \in \mathcal{R} \Rightarrow A \cup B \in \mathcal{R}$.
(r_2) $A, B \in \mathcal{R} \Rightarrow A - B \in \mathcal{R}$.

Definition 0.1.3 (field)
A collection of subsets, $\mathcal{F}$ is called a *field* iff

(a_1) $\Omega \in \mathcal{F}$, and
(a_2) $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$, and
(a_3) $A, B \in \mathcal{F} \Rightarrow A \cup B \in \mathcal{F}$.

A field of sets is also called an *algebra* of sets. (Compare the definition above with the definition of the algebraic structure given in Definition 0.0.3 on page 631.)

Notice that property (a_3) is equivalent to

(a_3') $A_1, A_2, \ldots A_n \in \mathcal{F} \Rightarrow \cup_{i=1}^n A_i \in \mathcal{F}$;

that is, $\mathcal{F}$ is closed under finite unions. The next systems we describe are closed under countable unions.

Notice that a field is nonempty by definition, although "nonempty" is not specified explicitly, as it is for a ring. A field contains at least one set, Ω, and because Ω is nonempty, it contains two sets, Ω and $\emptyset$.

Definition 0.1.4 (λ-system)
A collection of subsets, $\mathcal{L}$, is called a *λ-system* iff

(λ_1) $\Omega \in \mathcal{L}$, and
(λ_2) $A \in \mathcal{L} \Rightarrow A^c \in \mathcal{L}$, and
(λ_3) $A_1, A_2, \ldots \in \mathcal{L}$ and $A_i \cap A_j = \emptyset$ for $i \neq j \Rightarrow \cup_i A_i \in \mathcal{L}$.

The definitions of π-systems, rings, and fields have involved only finite numbers of sets. The name λ-system comes from its property that involves an operation on an infinite number of sets, that is, a limiting operation. A λ-system is also called a Dynkin system or a Sierpinski system.

We can see that the first and third properties of a λ-system imply that the second property is equivalent to

(λ_2') $A, B \in \mathcal{L}$ and $A \subseteq B \Rightarrow B - A \in \mathcal{L}$.

To see this, first assume the three properties that characterize a λ-system $\mathcal{L}$, and $A, B \in \mathcal{L}$ and $A \subseteq B$. We first see that this implies $B^c \in \mathcal{L}$ and so the disjoint union $A \cup B^c \in \mathcal{L}$. This implies that the complement $(A \cup B^c)^c \in \mathcal{L}$. But $(A \cup B^c)^c = B - A$; hence, we have the alternative property (λ_2'). Conversely, assume this alternative property together with the first property (λ_1). Hence, $A \in \mathcal{L} \Rightarrow \Omega - A \in \mathcal{L}$, but $\Omega - A = A^c$; that is, $A^c \in \mathcal{L}$.

In a similar manner, we can show that the third property is equivalent to

(λ_3') $A_1, A_2, \ldots \in \mathcal{L}$ with $A_1 \subseteq A_2 \subseteq \cdots \Rightarrow \cup_i A_i \in \mathcal{L}$.

Now, we define the most important type of system of collections of subsets:

Definition 0.1.5 (σ-field)
A collection of subsets, $\mathcal{F}$, of a given sample space, Ω, is called a *σ-field* iff

(σ_1) $\Omega \in \mathcal{F}$
(σ_2) $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$
(σ_3) $A_1, A_2, \ldots \in \mathcal{F} \Rightarrow \cup_i A_i \in \mathcal{F}$.

A σ-field is also called a σ-algebra or a a σ-ring. (Notice, however, that it is much more than a simple extension of a ring.)

A field with the properties (σ_1) and (σ_2), but with (σ_3) replaced with the property

(δ_3) $A_1, A_2, \ldots \in \mathcal{F} \Rightarrow \cap_i A_i \in \mathcal{F}$

is called a δ-field, a δ-algebra or a δ-ring. It is clear, however, that δ-field and σ-field are equivalent concepts. We will use the latter term.

The definition of a σ-field immediately implies that the field is closed with respect to set differences.

Theorem 0.1.1
Given the σ-field $\mathcal{F}$, if $A_1, A_2, \ldots \in \mathcal{F}$, then $\limsup_n A_n \in \mathcal{F}$ *and* $\liminf_n A_n \in \mathcal{F}$.

Proof. The conclusion follows because for any n, $\cup_{i=n}^{\infty} A_i \in \mathcal{F}$ and $\cap_{i=n}^{\infty} A_i \in \mathcal{F}$.

Notice that the definitions of a π-system and of a ring must specify that the collections are nonempty; the definitions of the other systems ensure that the collections are nonempty without saying so explicitly.

The exact definitions of these systems can be modified in various simple ways. For example, in the definitions of a field, a λ-system, and a σ-field the requirement that Ω be in the system could be replaced by the requirement that $\emptyset$ be in the system, because closure with respect to complementation guarantees the inclusion of Ω. (The requirement that Ω be in a λ-system, however, could not be replaced by the requirement that $\emptyset$ be in the system.) The closure property for unions in a field or a σ-field could be replaced by the requirement that the system be closed for intersections of the same kind as the unions.

Before concluding this subsection, we will define another type of collection of subsets that is often useful in statistics.

Definition 0.1.6 (σ-lattice)
A nonempty collection of subsets, $\mathcal{L}$, of a given sample space, Ω, is called a *σ-lattice* iff

(σ_{l1}) $A_1, A_2, \ldots \in \mathcal{L} \Rightarrow \cup_i A_i \in \mathcal{L}$
(σ_{l2}) $A_1, A_2, \ldots \in \mathcal{L} \Rightarrow \cap_i A_i \in \mathcal{L}$.

∎

The most useful of the systems we have defined is a σ-field, and in the following sections, we will focus our attention on σ-fields.

σ-Field Generated by a Collection of Sets

Given a sample space Ω and any collection $\mathcal{C}$ of subsets of Ω, the intersection of all σ-fields over Ω that contain $\mathcal{C}$ is called the σ-field generated by $\mathcal{C}$, and is denoted by

$$\sigma(\mathcal{C}).$$

It is the *minimal* σ-field that contains $\mathcal{C}$; that is, the "smallest" σ-field over Ω of which $\mathcal{C}$ is a subset.

Given $A_1, A_2, \ldots \subseteq \Omega$, we may use a similar notation to that above to refer to generated σ-fields. We use $\sigma(A_1)$ and $\sigma(A_1, A_2)$ to refer respectively to $\sigma(\{A_1\})$ and $\sigma(\{A_1, A_2\})$. That is, the argument in the operator $\sigma(\cdot)$ may be either a set or a collection of sets.

A σ-field can contain a very large number of subsets. If k is the maximum number of sets that partition Ω that can be formed by operations on the sets in $\mathcal{C}$, then the number of sets in the σ-field is 2^k. (What is the "largest" σ-field over Ω?)

Other special collections of subsets can also be generated by a given collection. For example, given a collection $\mathcal{C}$ of subsets of a sample space Ω, we can form a π-system by adding (only) enough subsets to make the collection

closed with respect to intersections. This system generated by $\mathcal{C}$ is the *minimal* π-system that contains $\mathcal{C}$. This π-system, denoted by $\pi(\mathcal{C})$, is the intersection of all π-systems that contain $\mathcal{C}$. Likewise, we define the λ-system generated by $\mathcal{C}$ as the minimal λ-system that contains $\mathcal{C}$, and we denote it by $\lambda(\mathcal{C})$.

Example 0.1.1 (σ-fields)

1. The "trivial σ-field" is $\{\emptyset, \Omega\}$.
2. For the sample space Ω, $\sigma(\{\Omega\})$ is the trivial σ-field, $\{\emptyset, \Omega\}$.
3. If $\mathcal{A} = \{A\}$ with respect to the sample space Ω, $\sigma(\mathcal{A}) = \{\emptyset, A, A^{\mathrm{c}}, \Omega\}$. If $A = \emptyset$ or $A = \Omega$, this is the trivial σ-field; otherwise, it is the second simplest σ-field.
4. If $\mathcal{A} = \{A_1, A_2\}$ and neither A_1 nor A_2 is a subset of the other, with respect to the sample space Ω, there are 4 "smallest" sets that partition $\mathcal{A}$. These are called *atoms*. They are

$$\{A_1 \cap A_2,\ A_1 - A_2,\ A_2 - A_1,\ (A_1 \cup A_2)^{\mathrm{c}}\}.$$

   Hence, there are $2^4 = 16$ sets in $\sigma(\mathcal{A})$. These can be written simply as the binary combinations of all above: (0000), (0001), (0010), ... Following this order, using the partition above, the sets are (after simplification):

$$\begin{aligned}\sigma(\mathcal{A}) = \{&\emptyset,\ (A_1 \cup A_2)^{\mathrm{c}},\ A_2 - A_1,\ A_1^{\mathrm{c}},\\ &A_1 - A_2,\ A_2^{\mathrm{c}},\ A_1 \Delta A_2,\ (A_1 \cap A_2)^{\mathrm{c}},\\ &A_1 \cap A_2,\ (A_1 \Delta A_2)^{\mathrm{c}},\ A_2,\ (A_1 - A_2)^{\mathrm{c}},\\ &A_1,\ (A_2 - A_1)^{\mathrm{c}},\ A_1 \cup A_2,\ \Omega\}.\end{aligned}$$

   Notice that $\sigma(\{A_1\}) \subseteq \sigma(\{A_1, A_2\})$.
5. If $\mathcal{A} = \{A_1, A_2\}$ and $A_1 \subseteq A_2$, with respect to the sample space Ω, there are 8 sets in $\sigma(\mathcal{A})$.
6. For the sample space Ω, the power set 2^{Ω} is a σ-field. It is the "largest" σ-field over Ω.

∎

Notice the notation in the example above. Why do we have the braces in $\sigma(\{\Omega\})$? We do often abuse the notation, however; if the argument of $\sigma(\dot{})$ is a singleton, we sometimes omit the braces. For example, the second most trivial σ-field is that generated by a single set, say A: $\sigma(A) = \{\emptyset, A, A^{\mathrm{c}}, \Omega\}$. We may also abuse the notation even further by writing a collection of subsets without putting them in braces. So, for example, $\sigma(A, B)$ may be used instead of $\sigma(\{A.B\})$.

Example 0.1.2 (A field that is not a σ-field)
Let Ω be a countably infinite set, and let $\mathcal{F}$ consist of all finite subsets of Ω along with all subsets of Ω whose complements are finite. We see immediately

that $\mathcal{F}$ is a field. To see that it is not a σ-field, all we must do is choose a set A that is countably infinite and has infinite complement. One such set can be constructed from a sequence $\omega_1, \omega_2, \ldots$, and let $A = \{\omega_1, \omega_3, \ldots\}$. Therefore for fixed i, the singleton $\{\omega_{2i-1}\} \in \mathcal{F}$, but $A \notin \mathcal{F}$ even though $A = \cup_i \{\omega_{2i-1}\}$. ∎

The sets in Example 0.1.2 are not specified exactly. How can we just give some property, such as finiteness, and then specify all sets with this property? While it may seem obvious that we can do this, in order to do it, we are relying on the Axiom of Choice (see pages 617 and 674, although even that fact is not obvious. Other interesting collections of sets similar to those in Example 0.1.2 can be formed. Instead of finiteness, we may focus on countability. We first define a related term: A set A is said to be *cocountable* iff A^c is countable.

Example 0.1.3 (A σ-field that does not contain certain sets)
Let Ω be the universal set, and let $\mathcal{F}$ consist of all countable and cocountable subsets of Ω. Then $\mathcal{F}$ is a σ-field on Ω (exercise).

If Ω is uncountable, then is contains a set A such that both A and A^c are countable. Such a set A (or, equivalently, A^c) is not in $\mathcal{F}$. ∎

Borel σ-Fields

There is a particularly interesting type of σ-field, called a Borel σ-field, that can be defined in topological spaces.

Definition 0.1.7 (Borel σ-field) (general)
Let $(\Omega, \mathcal{T})$ be a topological space. The σ-field generated by $\mathcal{T}$ is the *Borel σ-field* on $(\Omega, \mathcal{T})$. ∎

We often denote this Borel σ-field as $\mathcal{B}(\Omega, \mathcal{T})$ or as $\mathcal{B}(\Omega)$.

The most interesting topological space is the set of reals together with the class of open intervals, $(\mathbb{R}, \mathcal{C})$. We denote the Borel σ-field on this space as $\mathcal{B}(\mathbb{R})$ or just as $\mathcal{B}$.

Relations of σ-Fields to Other Structures

A σ-field is a π-system, a field, and a λ-system.

Theorem 0.1.2
A class that is both a π-system and a λ-system is a σ-field.

Proof. Because it is a λ-system, the class contains $\emptyset$ and is closed under formation of complements, and because it is a π-system, it is closed under finite intersections. It is therefore a field. Now, suppose that it contains sets A_i, for $i = 1, 2, \ldots$. The class then contains the sets $B_i = A_i \cap A_1^c \cap \cdots \cap A_{i-1}^c$, which are necessarily disjoint. Because it is a λ-system, it contains $\cup_i B_i$. But $\cup_i B_i = \cup_i A_i$, and since it contains $\cup_i A_i$ it is a σ-field. ∎

A useful fact is known as the π-λ *theorem*.

Theorem 0.1.3 (the π-λ theorem)
If $\mathcal{P}$ is a π-system and $\mathcal{L}$ is a λ-system, and if $\mathcal{P} \subseteq \mathcal{L}$, then

$$\sigma(\mathcal{P}) \subseteq \mathcal{L}.$$

The π-λ theorem is also called Dynkin's π-λ theorem or Sierpinski's π-λ theorem.

Proof. We use the given notation and assume the hypothesis. Let $\mathcal{L}_\mathcal{P}$ be the λ-system generated by $\mathcal{P}$; that is,

$$\mathcal{L}_\mathcal{P} = \lambda(\mathcal{P}).$$

$\mathcal{L}_\mathcal{P}$ is the intersection of every λ-system that contains $\mathcal{P}$, and it is contained in every λ-system that contains $\mathcal{P}$. Thus, we have

$$\mathcal{P} \subseteq \mathcal{L}_\mathcal{P} \subseteq \mathcal{L}.$$

It will now suffice to show that $\mathcal{L}_\mathcal{P}$ is also a π-system, because from the result above, if it is both a π-system and a λ-system it is a σ-field, and it contains $\mathcal{P}$ so it must be the case that $\sigma(\mathcal{P}) \subseteq \mathcal{L}_\mathcal{P}$ because $\sigma(\mathcal{P})$ is the minimal σ-field that contains $\mathcal{P}$.

Now define a collection of sets whose intersection with a given set is a member of $\mathcal{L}_\mathcal{P}$. For any set A, let

$$\mathcal{L}_A = \{B : A \cap B \in \mathcal{L}_\mathcal{P}\}.$$

Later in the proof, for some given set B, we use the symbol "$\mathcal{L}_B$" to denote the collection of sets whose intersection with B is a member of $\mathcal{L}_\mathcal{P}$.

If $A \in \mathcal{L}_\mathcal{P}$, then $\mathcal{L}_A$ is a λ-system, as we see by checking the conditions:

(λ_1) $A \cap \Omega = A \in \mathcal{L}_\mathcal{P}$ so $\Omega \in \mathcal{L}_A$

(λ_2') If $B_1, B_2 \in \mathcal{L}_A$ and $B_1 \subseteq B_2$, then $\mathcal{L}_\mathcal{P}$ contains $A \cap B_1$ and $A \cap B_2$, and hence contains the difference $(A \cap B_2) - (A \cap B_1) = A \cap (B_2 - B_1)$; that is, $B_2 - B_1 \in \mathcal{L}_A$.

(λ_3) If $B_1, B_2, \ldots \in \mathcal{L}_A$ and $B_i \cap B_j = \emptyset$ for $i \neq j$, then $\mathcal{L}_\mathcal{P}$ contains the disjoint sets $(A \cap B_1), (A \cap B_2), \ldots$ and hence their union $A \cap (\cup_i B_i)$, which in turn implies $\cup_i B_i \in \mathcal{L}_A$.

Now because $\mathcal{P}$ is a π-system,

$$\begin{aligned} A, B \in \mathcal{P} &\Rightarrow A \cap B \in \mathcal{P} \\ &\Rightarrow B \in \mathcal{L}_A \\ &\Rightarrow \mathcal{P} \subseteq \mathcal{L}_A \\ &\Rightarrow \mathcal{L}_\mathcal{P} \subseteq \mathcal{L}_A. \end{aligned}$$

(The last implication follows from the minimality of $\mathcal{L}_\mathcal{P}$ and because $\mathcal{L}_A$ is a λ-system containing $\mathcal{P}$.)

Using a similar argument as above, we have $A \in \mathcal{P}$ and $B \cap B \in \mathcal{L}_{\mathcal{P}}$ also imply $A \in \mathcal{L}_B$ (here $\mathcal{L}_B$ is in the role of $\mathcal{L}_A$ above) and we have

$$A \in \mathcal{L}_B \iff B \in \mathcal{L}_A.$$

Continuing as above, we also have $\mathcal{P} \subseteq \mathcal{L}_B$ and $\mathcal{L}_{\mathcal{P}} \subseteq \mathcal{L}_B$.

Now, to complete the proof, let $B, C \in \mathcal{L}_{\mathcal{P}}$. This means that $C \in \mathcal{L}_B$, which from the above means that $B \cap C \in \mathcal{L}_{\mathcal{P}}$; that is, $\mathcal{L}_{\mathcal{P}}$ is a π-system, which, as we noted above is sufficient to imply the desired conclusion: $\sigma(\mathcal{P}) \subseteq \mathcal{L}_{\mathcal{P}} \subseteq \mathcal{L}$. ∎

The π-λ theorem immediately implies that if $\mathcal{P}$ is a π-system then

$$\sigma(\mathcal{P}) = \lambda(\mathcal{P}). \tag{0.1.1}$$

Operations on σ-Fields

The usual set operators and set relations are used with collections of sets, and generally have the same meaning. If the collections of sets are σ-fields, the operation on the collections may not yield a collection that is a σ-field, however.

Theorem 0.1.4
Given σ-fields $\mathcal{F}_1$ and $\mathcal{F}_2$ defined with respect to a common sample space, the intersection, $\mathcal{F}_1 \cap \mathcal{F}_2$, is a σ-field.

Proof. Exercise. ∎

The union, $\mathcal{F}_1 \cup \mathcal{F}_2$, however, may not be a σ-field. A simple counterexample with $\Omega = \{a, b, c\}$ is

$$\mathcal{F}_1 = \{\emptyset, \{a\}, \{b, c\}, \Omega\} \quad \text{and} \quad \mathcal{F}_2 = \{\emptyset, \{b\}, \{a, c\}, \Omega\}. \tag{0.1.2}$$

The notation $\sigma(\mathcal{F}_1 \cup \mathcal{F}_2)$ refers to the smallest σ-field that contains all of the sets in either $\mathcal{F}_1$ or $\mathcal{F}_2$. For $\mathcal{F}_1$ and $\mathcal{F}_2$ as given above, we have

$$\sigma(\mathcal{F}_1 \cup \mathcal{F}_2) = \{\emptyset, \{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}, \Omega\}.$$

Sub-σ-Fields

A subset of a σ-field that is itself a σ-field is called a sub-σ-field.

Increasing sequences of σ-fields, $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots$, are often of interest, especially in stochastic processes.

Given a σ-field $\mathcal{F}$, an interesting sub-σ-field can be formed by taking a specific set B in $\mathcal{F}$, and forming its intersection with all of the other sets in $\mathcal{F}$. We often denote this sub-σ-field as $\mathcal{F}_B$:

$$\mathcal{F}_B = \{B \cap A : A \in \mathcal{F}\}. \tag{0.1.3}$$

It is an exercise to verify the three defining properties of a σ-field for $\mathcal{F}_B$.

By the definition of $\sigma(\mathcal{C})$ for a given collection $\mathcal{C}$ of subsets of Ω as the intersection of all σ-fields over Ω that contain $\mathcal{C}$, we have the trivial result

$$\sigma(\mathcal{C}_1) \subseteq \sigma(\mathcal{C}_2) \quad \text{if } \mathcal{C}_1 \subseteq \mathcal{C}_2. \tag{0.1.4}$$

Measurable Space: The Structure $(\Omega, \mathcal{F})$

If Ω is a sample space, and $\mathcal{F}$ is a σ-field over Ω, the double $(\Omega, \mathcal{F})$ is called a *measurable space*.

Measurable spaces are fundamental objects in our development of a theory of measure and its extension to probability theory.

Partitions of the sample space by sets in the σ-field are often useful. Notice that this is always possible, and, in fact, a finite partition always exists for if $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$, and $A \cap A^c = \emptyset$ and $A \cup A^c = \Omega$. This partitioning of the sample space, which is often called a decomposition of the sample space, is useful, especially in working with simple functions, as we will see later.

Notice that no *measure* is required for a measurable space. (We will define "measure" below, Definition 0.1.10. It is a scalar extended-real-valued nonnegative set function whose domain is a σ-field with the properties that the measure of the null set is 0 and the measure of the union of any collection of disjoint sets is the sum of the measures of the sets. A measurable space together with a measure forms a structure called a *measure space*.) We will consider measures and measure spaces in Section 0.1.3. For now, we continue discussions of measurability without reference to a specific measure.

Subspaces

Given a measurable space $(\Omega, \mathcal{F})$, and a set $B \in \mathcal{F}$, we have seen how to form a sub-σ-field $\mathcal{F}_B$. This immediately yields a sub measurable space $(B, \mathcal{F}_B)$, if we take the sample space to be $\Omega \cap B = B$.

Cartesian Products

The cartesian product of two sets A and B, written $A \times B$, is the set of all doubletons, (a_i, b_j), where $a_i \in A$ and $b_j \in B$. The cartesian product of two collections of sets is usually interpreted as the collection consisting of all possible cartesian products of the elements of each, e.g., if $\mathcal{A} = \{A_1, A_2\}$ and $\mathcal{B} = \{B_1, B_2\}$

$$\mathcal{A} \times \mathcal{B} = \{A_1 \times B_1,\ A_1 \times B_2,\ A_2 \times B_1,\ A_2 \times B_2\},$$

that is,

$$\{\{(a_{1i}, b_{1j}) \,|\, a_{1i} \in A_1, b_{1j} \in B_1\},\ \{(a_{1i}, b_{2j}) \,|\, a_{1i} \in A_1, b_{2j} \in B_2\},$$

$$\{(a_{2i}, b_{1j}) \,|\, a_{2i} \in A_2, b_{1j} \in B_1\},\ \{(a_{2i}, b_{2j}) \,|\, a_{2i} \in A_2, b_{2j} \in B_2\}\}.$$

The cartesian product of two collections of sets is not a very useful object, because, as we see below, important characteristics of the collections, such as being σ-fields do not carry over to the product.

Two measurable spaces $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$ can be used to form a *cartesian product measurable space* with sample space $\Omega_1 \times \Omega_2$. The product of the

σ-fields is not necessarily a σ-field. A simple counterexample is the same as we have used before with $\Omega = \{a, b, c\}$. Let

$$\mathcal{F}_1 = \{\{a\}, \{b, c\}, \emptyset, \Omega\} \quad \text{and} \quad \mathcal{F}_2 = \{\{b\}, \{a, c\}, \emptyset, \Omega\}.$$

The product $\mathcal{F}_1 \times \mathcal{F}_2$ contains 8 sets of doubletons, two of which are $\{(a, b)\}$ and $\{(b, b), (c, b)\}$; however, we see that their union $\{(a, b), (b, b), (c, b)\}$ is not a member of $\mathcal{F}_1 \times \mathcal{F}_2$; hence, $\mathcal{F}_1 \times \mathcal{F}_2$ is not a σ-field.

As another example, let $\Omega = \mathbb{R}$, let $\mathcal{F} = \sigma(\mathbb{R}_+) = \{\emptyset, \mathbb{R}_+, \mathbb{R} - \mathbb{R}_+, \mathbb{R}\}$, let $\mathcal{G}_1 = \sigma(\mathbb{R}_+ \times \mathbb{R}_+)$, and let $\mathcal{G}_2 = \sigma(\{F_i \times F_j \ : \ F_i, F_j \in \mathcal{F}\})$. We see that $\mathcal{G}_1 \neq \mathcal{G}_2$, because, for example, $\mathbb{R}_+ \times \mathbb{R}$ is in $\mathcal{G}_2$ but it is not in $\mathcal{G}_1$.

Definition 0.1.8 (cartesian product measurable space)
Given the measurable spaces $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$, we define the *cartesian product measurable space* as

$$(\Omega_1 \times \Omega_2, \sigma(\mathcal{F}_1 \times \mathcal{F}_2)).$$

■

As noted above, the collection $\sigma(\mathcal{F}_1 \times \mathcal{F}_2)$ is not the same as $\mathcal{F}_1 \times \mathcal{F}_2$.

Product measure spaces provide us the basis for developing a probability theory for vectors and multivariate distributions.

0.1.2 Functions and Images

A function is a set of ordered pairs such that no two pairs have the same first element. If (a, b) is an ordered pair in f, then a is called an argument of the function, b is called the corresponding value of the function, and we write $b = f(a)$. The set of all arguments of the function is called the domain of the function, and the set of all values of the function is called the range of the function. If the arguments of the function are sets, the function is called a set function.

We will be interested in a function, say f, that maps one measurable space $(\Omega, \mathcal{F})$ to another measurable space $(\Lambda, \mathcal{G})$. We may write $f : (\Omega, \mathcal{F}) \mapsto (\Lambda, \mathcal{G})$, or just $f : \Omega \mapsto \Lambda$ because the argument of the function is an element of Ω (in fact, *any* element of Ω) and the value of the function is an element of Λ. It may not be the case that all elements of Λ are values of f. If it is the case that for every element $\lambda \in \Lambda$, there is an element $\omega \in \Omega$ such that $f(\omega) = \lambda$, then the function is said to be "onto" Λ. Such a function is called *surjective*. (Any function is surjective with respect to its range.)

Note the convention that we are adopting here: the domain of the function is the sample space. If we wish to restrict the domain of definition of a function, we do so by redefining the sample space.

If $f : \Omega \mapsto \Lambda$ and $\forall x, y \in \Omega, f(x) = f(y) \Rightarrow x = y$, then the function is said to be *one-to-one*. If a function from Ω to Λ is one-to-one and surjective, it is said to be *bijective* and the function itself is called a *bijection*.

If $(a, b) \in f$, we may write $a = f^{-1}(b)$, although sometimes this notation is restricted to the cases in which f is one-to-one. (There are some subtleties here if f is not one-to-one. In that case, if the members of the pairs in f are reversed, the resulting set is not a function. We may then say f^{-1} does not exist; yet we may write $a = f^{-1}(b)$, with the meaning above. It is perhaps more appropriate to take $f^{-1}(b)$ to be an equivalence class, are if $f(a) = b$ to say that $a \in f^{-1}(b)$. We will not attempt to accommodate these subtleties, however.)

If $A \subseteq \Omega$, the *image* of A, denoted by $f[A]$, is the set of all $\lambda \in \Lambda$ for which $\lambda = f(\omega)$ for some $\omega \in A$. Likewise, if $\mathcal{C}$ is a collection of subsets of Ω, the *image* of $\mathcal{C}$, denoted by $f[\mathcal{C}]$, or just by $f(\mathcal{C})$, is the collection of all subsets of Λ that are images of the subsets of $\mathcal{C}$. (While I prefer the notation "$[\cdot]$" when the argument of the function is a set or a collection of sets — unless the function is a set function — in some cases I will do like most other people and just use the "$(\cdot)$", which actually applies more properly to an element.)

For a subset B of Λ, the *inverse image* or the *preimage* of B, denoted by $f^{-1}[B]$, is the set of all $\omega \in \Omega$ such that $f(\omega) \in B$. We also write $f[f^{-1}[B]]$ as $f \circ f^{-1}[B]$. The set $f[f^{-1}[B]]$ may be a proper subset of B; that is, there may be an element λ in B for which there is no $\omega \in \Omega$ such that $f(\omega) = \lambda$. If there is no element $\omega \in \Omega$ such that $f(\omega) \in B$, then $f^{-1}[B] = \emptyset$.

We see from the foregoing that $f^{-1}[\Lambda] = \Omega$, although it may be the case that $f[\Omega] \neq \Lambda$. Because f is defined at points in Ω (and only there), we see that $f[\emptyset] = \emptyset$ and $f^{-1}[\emptyset] = \emptyset$.

The following theorems state useful facts about preimages.

Theorem 0.1.5
Let $f : \Omega \mapsto \Lambda$. For $B \subseteq \Lambda$,

$$f^{-1}[B^c] = (f^{-1}[B])^c.$$

(Here, $B^c = \Lambda - B$, and $(f^{-1}[B])^c = \Omega - f^{-1}[B]$.)

Proof.
We see this in the standard way by showing that each is a subset of the other.

Let ω be an arbitrary element of Ω.

Suppose $\omega \in f^{-1}[B^c]$. Then $f(\omega) \in B^c$, so $f(\omega) \notin B$, hence $\omega \notin f^{-1}[B]$, and so $\omega \in (f^{-1}[B])^c$. We have $f^{-1}[B^c] \subseteq (f^{-1}[B])^c$.

Now suppose $\omega \in (f^{-1}[B])^c$. Then $\omega \notin f^{-1}[B]$, so $f(\omega) \notin B$, hence $f(\omega) \in B^c$, and so $\omega \in f^{-1}[B^c]$. We have $(f^{-1}[B])^c \subseteq f^{-1}[B^c]$. ∎

Theorem 0.1.6
Let $f : \Omega \mapsto \Lambda$, and let $A_1, A_2 \subseteq \Lambda$ with $A_1 \cap A_2 = \emptyset$. Then

$$f^{-1}[A_1] \cap f^{-1}[A_2] = \emptyset.$$

Proof. Exercise. ∎

Notice that the theorem does not hold in the other direction, unless f is bijective. That is, $B_1, B_2 \subseteq \Omega$ with $B_1 \cap B_2 = \emptyset$ does not imply that $f[B_1] \cap f[B_2] = \emptyset$.

Theorem 0.1.7
Let $f : \Omega \mapsto \Lambda$, and let $A_1, A_2, \ldots \subseteq \Lambda$. Suppose $(\cup_{i=1}^{\infty} A_i) \subseteq \Lambda$, then

$$f^{-1}[\cup_{i=1}^{\infty} A_i] = \cup_{i=1}^{\infty} f^{-1}(A_i).$$

Proof.
We see this as above; again, let λ be an arbitrary element of Λ.

Suppose $\lambda \in f^{-1}[\cup_{i=1}^{\infty} A_i]$. Then $f(\lambda) \in \cup_{i=1}^{\infty} A_i$, so for some j, $f(\lambda) \in A_j$ and $\lambda \in f^{-1}[A_j]$; hence $\lambda \in \cup_{i=1}^{\infty} f^{-1}[A_i]$. We have $f^{-1}[\cup_{i=1}^{\infty} A_i] \subseteq \cup_{i=1}^{\infty} f^{-1}[A_i]$.

Now suppose $\lambda \in \cup_{i=1}^{\infty} f^{-1}[A_i]$. Then for some j, $\lambda \in f^{-1}[A_j]$, so $f(\lambda) \in A_j$ and $f(\lambda) \in \cup_{i=1}^{\infty} A_i$; hence $\lambda \in f^{-1}[\cup_{i=1}^{\infty} A_i]$. We have $\cup_{i=1}^{\infty} f^{-1}[A_i] \subseteq f^{-1}[\cup_{i=1}^{\infty} A_i]$, and so the two sets are the same. ∎

It is worth noting a finite-union version of this result:

$$f^{-1}[A_1 \cup A_2] = f^{-1}[A_1] \cup f^{-1}[A_2].$$

For bijective functions, we have similar relationships for intersections and set differences, but in general, $f^{-1}[A_1 \cap A_2] \neq f^{-1}[A_1] \cap f^{-1}[A_2]$.

If $f : (\Omega, \mathcal{F}) \mapsto (\Lambda, \mathcal{G})$, the σ-fields in the measurable spaces determine certain properties of the function, the most important of which is measurability.

Measurable Functions

We have been discussing measurability without discussing a measure. We continue in this vein for one more concept; that of a measurable function. The importance of the concept of a measurable function from the measurable space $(\Omega, \mathcal{F})$ to the measurable space $(\Lambda, \mathcal{G})$ is that it allows a measure defined on $\mathcal{F}$ to be used immediately in $\mathcal{G}$. We discuss measure formally in Section 0.1.3.

Definition 0.1.9 (measurable function)
If $(\Omega, \mathcal{F})$ and $(\Lambda, \mathcal{G})$ are measurable spaces, and f is a mapping from Ω to Λ, with the property that $\forall A \in \mathcal{G}, f^{-1}[A] \in \mathcal{F}$, then f is a *measurable* function with respect to $\mathcal{F}$ and $\mathcal{G}$. It is also said to be measurable $\mathcal{F}/\mathcal{G}$. ∎

For a real-valued function, that is, a mapping from Ω to $\mathbb{R}$ with σ-field $\mathcal{B}(\mathbb{R})$, or in other cases where there is an "obvious" σ-field, we often just say that the function is measurable with respect to $\mathcal{F}$. In any event, the role of $\mathcal{F}$ is somewhat more important. We use the notation $f \in \mathcal{F}$ to denote the fact that f is measurable with respect to $\mathcal{F}$. (Note that this is an abuse of the notation, because f is not one of the sets in the collection $\mathcal{F}$.)

Given the measurable spaces $(\Omega, \mathcal{F})$ and $(\Lambda, \mathcal{G})$ and a mapping f from Ω to Λ, we also call call f a mapping from $(\Omega, \mathcal{F})$ to $(\Lambda, \mathcal{G})$.

Note that a measurable function $f(\cdot)$ does not depend on a measure. The domain of $f(\cdot)$ has no relationship to $\mathcal{F}$, except through the range of $f(\cdot)$ that happens to be in the subsets in $\mathcal{G}$.

σ-Field Generated by a Measurable Function

If f is a measurable function from $(\Omega, \mathcal{F})$ to $(\Lambda, \mathcal{G})$, then we can see that $f^{-1}[\mathcal{G}]$ is a sub-σ-field of $\mathcal{F}$ (Exercise 0.1.5). We call this the σ-field generated by f, and write it as $\sigma(f)$. Now we have a third type that the argument in the operator $\sigma(\cdot)$ may be. It can be either a set, a collection of sets, or a measurable function.

For measurable functions f and g from the same measurable space $(\Omega, \mathcal{F})$ to the same measurable space $(\Lambda, \mathcal{G})$ we may write $\sigma(f, g)$, with the meaning

$$\sigma(f, g) = \sigma\left(f^{-1}[\mathcal{G}] \cup g^{-1}[\mathcal{G}]\right). \tag{0.1.5}$$

As with σ-fields generated by collections of sets in equation (0.1.4), it is clear that

$$\sigma(f) \subseteq \sigma(f, g). \tag{0.1.6}$$

For measurable functions f and g from $(\Omega, \mathcal{F})$ to $(\Omega, \mathcal{F})$, it is clear (exercise) that

$$\sigma(g \circ f) \subseteq \sigma(f). \tag{0.1.7}$$

0.1.3 Measure

A measure is a scalar extended-real-valued nonnegative set function whose domain is a σ-field, with some useful properties, as stated next.

Definition 0.1.10 (measure)
Given a measurable space $(\Omega, \mathcal{F})$, a function ν defined on $\mathcal{F}$ is a *measure* if

1. $\nu(\emptyset) = 0$,
2. $\forall A \in \mathcal{F}$, $\nu(A) \in [0, \infty]$,
3. if $A_1, A_2, \ldots \in \mathcal{F}$ are disjoint, then

$$\nu(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \nu(A_i). \tag{0.1.8}$$

∎

An immediate generalization is a *vector measure*, which is a similar function whose range $\mathcal{R}$ is a Banach space and the series on the right of equation (0.1.8) is convergent in the of the Banach space.

Two generalizations of measure are signed measure and outer measure.

Definition 0.1.11 (signed measure)
Given a measurable space $(\Omega, \mathcal{F})$, a function σ defined on $\mathcal{F}$ is a *signed measure* if

1. $\sigma(\emptyset) = 0$,
2. $\forall A \in \mathcal{F}$, $\sigma(A) \in [-\infty, \infty]$,

3. if $A_1, A_2, \ldots \in \mathcal{F}$ are disjoint, then

$$\sigma(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \sigma(A_i). \tag{0.1.9}$$

∎

Definition 0.1.12 (outer measure)
An *outer measure* is an extended-real-valued function ν on the power set of a given sample space Ω with the properties

1. $\nu(\emptyset) = 0$,
2. $\forall A \subseteq \Omega$, $\nu(A) \in [0, \infty]$,
3. $\forall A, B \subseteq \Omega$, $A \subseteq B \Rightarrow \nu(A) \leq \nu(B)$,
4. if $A_1, A_2, \ldots \subseteq \Omega$, then

$$\nu(\cup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} \nu(A_i). \tag{0.1.10}$$

∎

An outer measure is useful when it is inconvenient to work with disjoint sets as required in the definition of a measure.

Properties of Measures

Several properties of a measure are derived immediately from Definition 0.1.10.

Theorem 0.1.8 (monotonicity)
Let ν be a measure with domain $\mathcal{F}$. If $A_1 \subseteq A_2 \in \mathcal{F}$, then $\nu(A_1) \leq \nu(A_2)$

Proof. Exercise. ∎

Theorem 0.1.9 (subadditivity)
Let ν be a measure with domain $\mathcal{F}$. If $A_1, A_2, \ldots \in \mathcal{F}$, then $\nu(\cup_i A_i) \leq \sum_i \nu(A_i)$.

Proof. Exercise. (*Hint:* Use the sequence (0.0.6) in Theorem 0.0.1. Notice that you must show that each D_i of that theorem is in $\mathcal{F}$.) ∎

Theorem 0.1.10 (continuity from below)
Let ν be a measure with domain $\mathcal{F}$. If $A_1 \subseteq A_2 \subseteq \ldots \in \mathcal{F}$, then $\nu(\cup_{i=1}^{\infty} A_i) = \lim_{i \to \infty} \nu(A_i)$.

Proof.
Let $\{D_n\}$ be the sequence of disjoint sets defined in equation (0.0.7); that is, $D_j = A_{j+1} - A_j$ and

$$\cup_{i=1}^{\infty} D_i = \cup_{i=1}^{\infty} A_i.$$

By the closure property for set differences in a σ-field, each D_j is in $\mathcal{F}$. We have

$$\begin{aligned}
\nu\left(\cup_{i=1}^{\infty} A_i\right) &= \nu\left(\cup_{i=1}^{\infty} D_i\right) \\
&= \sum_{i=1}^{\infty} \nu(D_i) \\
&= \lim_{i\to\infty} \sum_{j=1}^{i} \nu(D_j) \\
&= \lim_{i\to\infty} \nu\left(\cup_{j=1}^{i} D_j\right) \\
&= \lim_{i\to\infty} \nu(A_i).
\end{aligned}$$

∎

Sequences of nested intervals are important. We denote a sequence $A_1 \subseteq A_2 \subseteq \ldots$ with $A = \cup_{i=1}^{\infty} A_i$, as $A_i \nearrow A$. (This same notation is used for a sequence of real numbers x_i such that $x_1 \leq x_2 \cdots$ and $\lim x_i = x$, where we write $x_i \nearrow x$.)

Continuity from below is actually a little stronger than what is stated above, because the sequence of values of the measure is also monotonic: for $A_i \in \mathcal{F}, A_i \nearrow A \Rightarrow \nu(A_i) \nearrow \nu(A)$.

Although we defined the continuity from below, we could likewise define continuity from above for a sequence $A_1 \supset A_2 \supset \ldots \in \mathcal{F}$ in which $\nu(A_1) < \infty$. We let $A = \cap_{i=1}^{\infty} A_i$, and we denote this as $A_i \searrow A$. Continuity from above is the fact that for such a sequence $\nu(A_i) \searrow \nu(A)$. The proof uses methods similar to those of the proof of Theorem 0.1.10 along with De Morgan's laws with complementation being taken with respect to A_1. The condition that $\nu(A_1) < \infty$ is crucial, that is, certain measures may not be continuous from above (exercise). A probability measure, in which $\nu(\Omega) = 1$ (Definition 0.1.14), is continuous from above.

Without qualifying the property as "from below" or "from above", because both obtain, we can say the measure is continuous.

Notice that the definition of a measure does not preclude the possibility that the measure is identically 0. This often requires us to specify "nonzero measure" in order to discuss nontrivial properties. Another possibility, of course, would be just to specify $\nu(\Omega) > 0$ (remember $\Omega \neq \emptyset$ in a measurable space).

To evaluate $\nu(\cup_i A_i)$ we form disjoint sets by intersections. For example, we have $\nu(A_1 \cup A_2) = \nu(A_1) + \nu(A_2) - \nu(A_1 \cap A_2)$. This is an application of the simplest form of the inclusion-exclusion formula (see page 619). If there

are three sets, we take out all pairwise intersections and then add back in the triple intersection. We can easily extend this (the proof is by induction) so that, in general for $n \geq 4$, we have

$$\begin{aligned}\nu(\cup_i^n A_i) = &\sum_{1\leq i\leq n}\nu(A_i) \;-\; \sum_{1\leq i<j\leq n}\nu(A_i\cap A_j)\\ &+\sum_{1\leq i<j<k\leq n}\nu(A_i\cap A_j\cap A_k) \;-\cdots\\ &+(-1)^{n+1}\nu(A_1\cap\cdots\cap A_n).\end{aligned} \tag{0.1.11}$$

Some General Types of Measures

There are some types of measures that deserve special attention. Some of these are general classes of measures such as finite measures and Radon measures, and others are specific measures, such as the counting measure and the Lebesgue measure (see page 717). Some types are defined for any measurable space, and other types are defined only for measurable spaces with some additional structure, such as a topology.

Recall that measures take values in the extended nonnegative reals, $[0, \infty]$.

Definition 0.1.13 (finite measure)
A measure ν such that $\nu(\Omega) < \infty$ is called a *finite measure*. ∎

An important finite measure is a probability measure.

Definition 0.1.14 (probability measure)
A measure whose domain is a σ-field defined on the sample space Ω with the property that $\nu(\Omega) = 1$ is called a *probability measure*. We often use P to denote such a measure. ∎

Probability measures and their applications are discussed in Chapter 1.

Definition 0.1.15 (σ-finite measure)
A measure ν is *σ-finite* on $(\Omega, \mathcal{F})$ iff there exists a sequence $A_1, A_2, \ldots$ in $\mathcal{F}$ such that $\cup_i A_i = \Omega$ and $\nu(A_i) < \infty$ for all i. ∎

A finite measure is obviously σ-finite. In integration theory, many important results (for example Fubini's theorem and the Radon-Nikodym theorem) depend on the measures being σ-finite.

Definition 0.1.16 (complete measure)
A measure ν defined on the σ-field $\mathcal{F}$ is said to be *complete* if $A_1 \subseteq A \in \mathcal{F}$ and $\nu(A) = 0$ implies $A_1 \in \mathcal{F}$. ∎

Completeness of a measure means that all subsets of measurable sets with measure 0 and also measurable, and have measure 0. (For an A_1 in the definition above, clearly, $\nu(A_1) = 0$.)

Definition 0.1.17 (Radon measure)
In a topological measurable space $(\Omega, \mathcal{F})$, a measure μ such that for every compact set $B \in \mathcal{F}$, $\mu(B) < \infty$ is called a *Radon measure*. ∎

A Radon measure is σ-finite, although it is not necessarily finite.

Definition 0.1.18 (Haar invariant measures)
Let Ω be a group, and let $(\Omega, \mathcal{F})$ be a topological measurable space. For $B \in \mathcal{F}$ and $x \in \Omega$, let $Bx = \{yx : y \in B\}$ (where "xy" represents the element of Ω formed by the group operation on x and y), and let $xBx = \{xy : y \in B\}$.

a) Let μ_{r} be a measure such that for any $B \in \mathcal{F}$ and $x \in \Omega$, if $Bx \in \mathcal{F}$ then $\mu_{\mathrm{r}}(Bx) = \mu_{\mathrm{r}}(B)$. Then μ_{r} is said to be a *right invariant Haar measure*.
b) Let μ_{l} be a measure such that for any $B \in \mathcal{F}$ and $x \in \Omega$, if $xB \in \mathcal{F}$ then $\mu_{\mathrm{l}}(Bx) = \mu_{\mathrm{l}}(B)$. Then μ_{l} is said to be a *left invariant Haar measure*.
c) If μ is a right invariant Haar measure and a left invariant Haar measure, then μ is an *invariant Haar measure*. ∎

If Ω is Abelian, then both right and left invariant Haar measures are invariant Haar measures.

Some Useful Specific Measures

Definition 0.1.19 (Dirac measure)
Let $(\Omega, \mathcal{F})$ be a measurable space, let $A, B \in \mathcal{F}$, and let $\omega \in B$. The *Dirac measure* of A concentrated at ω, usually denoted by δ_ω, is defined as

$$\delta_\omega(A) = \begin{cases} 1 \text{ if } \omega \in A \\ 0 \text{ otherwise.} \end{cases} \tag{0.1.12}$$

∎

It is clear from Definition 0.1.10 that δ_ω is a measure, and further, it is a Radon measure (exercise).

Definition 0.1.20 (counting measure)
Let $(\Omega, \mathcal{F})$ be a measurable space, and assume that every $A \in \mathcal{F}$ is countable. The *counting measure* is defined as

$$\gamma(A) = \#(A), \tag{0.1.13}$$

where $\#(A) = \infty$ if A is countably infinite. ∎

The counting measure is σ-finite. (Notice that the counting measure is only defined for the case that Ω is countable.) If Ω is finite, the counting measure is finite.

If the sets of $\mathcal{F}$ are all countable the most common measure in applications is the counting measure. The counting measure is the most useful measure over the ring of integers $\mathbb{Z}$ with the σ-field $2^{\mathbb{Z}}$.

Other specific measures for metric spaces (in particular $\mathbb{R}$) are the Borel measure and the Lebesgue measure, which we will discuss on page 717.

Measure Space: The Structure $(\Omega, \mathcal{F}, \nu)$

If Ω is a sample space, $\mathcal{F}$ is a σ-field over Ω, and ν is a measure with domain $\mathcal{F}$, the triple $(\Omega, \mathcal{F}, \nu)$ is called a *measure space* (compare *measurable space*, above).

The elements in the measure space can be any kind of objects. They do not need to be numbers.

Definition 0.1.21 (complete measure space)
If the measure ν in the measure space $(\Omega, \mathcal{F}, \nu)$ is complete, then we say that the measure space is a *complete measure space*. ∎

A complete measure space is a Banach space if the norm in the Banach space is defined in terms of the measure of the measure space.

Definition 0.1.22 (probability space; event)
If P in the measure space $(\Omega, \mathcal{F}, P)$ is a probability measure, the triple $(\Omega, \mathcal{F}, P)$ is called a *probability space*. A set $A \in \mathcal{F}$ is called an "event". ∎

Restricted Measures and Sub-Measure Spaces

If $(\Omega, \mathcal{F})$ is a measurable space with measure ν, and $\mathcal{A} \subseteq \mathcal{F}$ is a sub-σ-field, then the function $\nu_{\mathcal{A}}$ that is the same as ν on $\mathcal{F}_B$ and undefined elsewhere is a measure. We say that $\nu_{\mathcal{A}}$ is the "measure ν restricted to $\mathcal{A}$".

Because $\nu_{\mathcal{A}}$ is a measure on $\mathcal{A}$, $(\Omega, \mathcal{A}, \nu_{\mathcal{A}})$ is a measure space. It is a sub measure space of $(\Omega, \mathcal{F}, \nu)$, and it corresponds in a natural way to all the usual subsetting operations.

If $(\Omega, \mathcal{F}, \nu)$ is a measure space, and for some set $B \in \mathcal{F}$, $(B, \mathcal{F}_B)$ is a sub measurable space as described above, then the function ν_B, which is the same as ν on $\mathcal{F}_B$ and undefined elsewhere, is a measure (Exercise 0.1.17), and $(B, \mathcal{F}_B, \nu_B)$ is a measure space.

We say that ν_B is the "measure ν restricted to $\mathcal{F}_B$".

Measurable Set

If ν is a measure with domain $\mathcal{F}$ then every set in $\mathcal{F}$ is said to be *ν-measurable*, or just *measurable*.

Note that unlike other terms above that involve "measurable", this term is defined in terms of a given measure.

As usual, we can get a better feel for a concept if we consider situations in which the concept does not apply; hence, we look for some set that is not measurable. We will consider a simple example of a nonmeasurable set, called a Vitali set, in Section 0.1.4.

Almost Everywhere (a.e.) and Negligible Sets

Given a measure space, $(\Omega, \mathcal{F}, \nu)$, a property that holds for all elements of $\mathcal{F}$ the σ-field with positive measure is said to hold *ν-almost everywhere*, or ν-a.e. This is also sometimes written as a.e.$[\nu]$. Also, when the measure is obvious, we often use the phrase almost everywhere or a.e. without explicit reference to the measure.

A set $A \in \mathcal{F}$ such that $\nu(A) = 0$ is said to be a *ν-negligible set*, or just a negligible set when the measure is obvious.

As we have defined it, "for all" or "everywhere" implies "almost everywhere". Almost everywhere is sometimes defined in such a way that it *requires* there to be a set in $\mathcal{F}$ with zero measure over which the property does not hold. Although it is only in such cases that the distinction between "for all" and "almost everywhere" is needed, it does not seem necessary to require the existence of a measurable set over which the property does not hold in order to define almost everywhere.

A property that holds a.e. with respect to a probability measure is said to hold *almost surely*, or a.s. (There is no essential difference in the two phrases.)

Support of a Measure

For a general measure space $(\Omega, \mathcal{F}, \nu)$, a "support" of the measure may be defined as any $A \in \mathcal{F}$ such that $\nu(A^c) = 0$. If the measure is finite, $A \in \mathcal{F}$ is a support iff $\nu(A) = \nu(\Omega)$. This definition, which is used by some authors (Billingsley (1995), for example), is not very useful in practice; in particular, it does not lead to a practical concept in probability distributions. A more useful definition of support of a measure requires restrictions on the measure space. If the measure space $(\Omega, \mathcal{F}, \nu)$ is a topological space or a space with a metric (that is, if points in the space have neighborhoods) *and* if ν is defined for some ϵ-neighborhood of every $\omega \in \Omega$, then we define the *topological support* of ν as

$$\mathrm{S}(\nu) = \{\omega \in \Omega \,|\, \nu(\mathcal{N}(\omega)) > 0\}. \tag{0.1.14}$$

We say ν is *concentrated* on $\mathrm{S}(\nu)$. The topological support is also called just the *support* or the *spectrum*.

The support of a measure, when it is defined, has some interesting properties, such as closure, but we will not pursue this topic here. We will define support of a probability distribution of a random variable later.

Relations of One Measure to Another

From the definition of measure, we see that the class of measures on a given measurable space $(\Omega, \mathcal{F})$ form a linear space; that is, if μ and ν are measures on $(\Omega, \mathcal{F})$ and $a \in \mathbb{R}$, then $a\mu + \nu$ is a measure on $(\Omega, \mathcal{F})$.

The are several kinds of relationships between measures on a given measurable space or related measurable spaces that are interesting. Some of these relationships are equivalence relations, but some are not symmetric. The interesting relationship are generally transitive, however, and so an ordering on the space of measures could be constructed.

Definition 0.1.23 (dominating measure; absolute continuity; equivalence)

Given measures ν and μ on the same measurable space, $(\Omega, \mathcal{F})$, if $\forall A \in \mathcal{F}$

$$\nu(A) = 0 \quad \Rightarrow \quad \mu(A) = 0,$$

then μ is said to be *dominated* by ν and we denote this by

$$\mu \ll \nu.$$

In this case we also say that μ is *absolutely continuous* with respect to ν.

If $\mu \ll \nu$ and $\nu \ll \mu$, then μ and ν are *equivalent*, and we write

$$\mu \equiv \nu.$$

∎

The definition says that $\mu \ll \nu$ iff every ν-negligible set is a μ-negligible set.

If μ is finite (that is, if $\mu(A) < \infty \, \forall A \in \mathcal{F}$), the absolute continuity of μ with respect to ν can be characterized by an ϵ-δ relationship as used in the definition of absolute continuity of functions (Definition 0.1.32): Given that μ is finite, μ is absolutely continuous with respect to ν iff for any $A \in \mathcal{F}$ and for any $\epsilon > 0$, there exists a δ such that

$$\nu(A) < \delta \quad \Rightarrow \quad \mu(A) < \epsilon.$$

Absolute continuity is a linear relationship; that is, if λ, μ, and ν are measures on $(\Omega, \mathcal{F})$ and $a \in \mathbb{R}$ then

$$\lambda \ll \nu \quad \text{and} \quad \mu \ll \nu \Longrightarrow (a\lambda + \mu) \ll \nu. \tag{0.1.15}$$

(Exercise.)

Definition 0.1.24 (singular measure)

Given measures ν and μ on the same measurable space, $(\Omega, \mathcal{F})$, if there exists two disjoint sets A and B in $\mathcal{F}$ such that $A \cup B = \Omega$ and for any measurable set $A_1 \subseteq A$, $\nu(A_1) = 0$, while for any measurable set $B_1 \subseteq B$, $\mu(B_1) = 0$ then the pair of measures ν and μ is said to be *singular* . We denote this property as

$$\nu \perp \mu.$$

∎

If $\nu \perp \mu$, it follows immediately from the definitions of singularity and of absolute continuity that neither ν nor μ can dominate the other.

Singular measures rely on, or equivalently, define, a partition of the sample space.

Singularity is a linear relationship; that is, if λ, μ, and ν are measures on $(\Omega, \mathcal{F})$ and $a \in \mathbb{R}$ then

$$\lambda \perp \nu \quad \text{and} \quad \mu \perp \nu \Longrightarrow (a\lambda + \mu) \perp \nu. \tag{0.1.16}$$

(Exercise.)

Induced Measure

If $(\Omega, \mathcal{F}, \nu)$ is a measure space, $(\Lambda, \mathcal{G})$ is a measurable space, and f is a function from Ω to Λ that is measurable with respect to $\mathcal{F}$, then the domain and range of the function $\nu \circ f^{-1}$ is $\mathcal{G}$ and it is a measure (Exercise 0.1.19).

The measure $\nu \circ f^{-1}$ is called an *induced measure* on $\mathcal{G}$. (It is induced from the measure space $(\Omega, \mathcal{F}, \nu)$.) An induced measure is also called a *pushforward measure*.

Completion of a Measure Space

Given the measure space $(\Omega, \mathcal{F}, \nu)$ and $A \in \mathcal{F}$ with $\nu(A) = 0$, if $A_1 \subseteq A$, it would seem reasonable to say that $\nu(A_1) = 0$. If $A_1 \notin \mathcal{F}$, however, $\nu(A_1)$ is not 0; it is not defined. We can form a measure space $(\Omega, \mathcal{F}_c, \nu_c)$ that is related to $(\Omega, \mathcal{F}, \nu)$, but which allows us to say that the measure of any subset of a zero-measure set in $\mathcal{F}$ has zero measure. We form $\mathcal{F}_c$ as the σ-field generated by $\mathcal{F}$ and $\mathcal{Z}$, where $\mathcal{Z}$ is the collection of all sets in $\mathcal{F}$ with ν-measure 0. Now define

$$\nu_c(A_1) = \inf\{\nu(A) \,|\, A_1 \subseteq A \in \mathcal{F}\}. \tag{0.1.17}$$

The measure space $(\Omega, \mathcal{F}_c, \nu_c)$ is complete. (Exercise.) It is called the *completion* of the measure space $(\Omega, \mathcal{F}, \nu)$. The construction above proves the existence of the completion of a measure space.

Notice that the completion of a measure space does not require addition of points to Ω; compare the completion of a metric space discussed on page 639.

Every $A \in \mathcal{F}_c$ constructed as above is of the form $B \cup C$ where $B \in \mathcal{F}$ and $C \in \mathcal{Z}$, and

$$\nu_c(B \cup C) = \nu(B). \tag{0.1.18}$$

(Exercise.)

Extensions of Measures

In applications we may have a measure space $(\Omega, \mathcal{F}_1, \nu)$ and wish to consider a different σ-field $\mathcal{F}_2$ over the same sample space and extend the measure to

$\sigma(\mathcal{F}_1, \mathcal{F}_2)$ while preserving its properties over $\mathcal{F}_1$. More generally we may have a measure defined on any collection of subsets $\mathcal{A}$ of Ω and wish to extend it to some σ-field of which $\sigma(\mathcal{A})$ is a subset while preserving the properties of the measure over $\mathcal{A}$. The Carathéodory extension theorem tells us not only that we can do this, but that the extension is unique so long as the measure on $\mathcal{A}$ is σ-finite.

Theorem 0.1.11 (Carathéodory extension theorem)
Given a collection $\mathcal{A}$ of subsets of a sample space Ω and a σ-finite measure ν_0 on $\mathcal{A}$. Then there exists a unique σ-finite measure ν on $\sigma(\mathcal{A})$ such that for any $A \in \mathcal{A}$, $\nu(A) = \nu_0(A)$.

This theorem is proved in Billingsley (1995) for probability measures on page 36, and for general σ-finite measures on page 166.

Product Measures

Given measure spaces $(\Omega_1, \mathcal{F}_1, \nu_1)$ and $(\Omega_2, \mathcal{F}_2, \nu_2)$, we define the cartesian product measure space as $(\Omega_1 \times \Omega_2, \sigma(\mathcal{F}_1 \times \mathcal{F}_2), \nu_1 \times \nu_2)$, where the product measure $\nu_1 \times \nu_2$ is defined on the product σ-field $\sigma(\mathcal{F}_1 \times \mathcal{F}_2)$ to have the property for $A_1 \in \mathcal{F}_1$ and $A_2 \in \mathcal{F}_2$

$$\nu_1 \times \nu_2(A_1 \times A_2) = \nu_1(A_1)\nu_2(A_2). \tag{0.1.19}$$

It can be shown that the measure with this property is unique, see Billingsley (1995), for example.

0.1.4 Sets in $\mathbb{R}$ and $\mathbb{R}^d$

First, recall some important definitions:

- A set A of real numbers is called *open* if for each $x \in A$, there exists a $\delta > 0$ such that for each y with $|x - y| < \delta$ belongs to A.
- A real number x is called a *point of closure* of a set A of real numbers if for every $\delta > 0$ there exists a y in A such that $|x - y| < \delta$. (Notice that every $y \in A$ is a point of closure of A.)
  We denote the set of points of closure of A by $\overline{A}$.
- A set A is called *closed* if $A = \overline{A}$.

Some simple facts follow:

- The intersection of a finite collection of open sets is open.
- The union of a countable collection of open sets is open.
- The union of a finite collection of closed sets is closed.
- The intersection of a countable collection of closed sets is closed.

Notice what is *not* said above (where we use the word "finite").

A very important type of set is an *interval* in $\mathbb{R}$. Intervals are the basis for building important structures on $\mathbb{R}$. All intervals are Borel sets. We discussed properties of real intervals and, in particular, sequences on real intervals beginning on page 645.

The Borel σ-Field on the Reals

On page 697, we have defined a Borel σ-field for a topological space as the σ-field generated by the topology, that is, by the collection of open sets that define the topology. In a metric space, such as $\mathbb{R}$, we define open sets in terms of the metric, and then we define a Borel σ-field as before in terms of those open sets. The most interesting topological space is the set of reals together with the class of open intervals, $(\mathbb{R}, \mathcal{C})$.

Definition 0.1.25 (Borel σ-field)
Let $\mathcal{C}$ be the collection of all open intervals in $\mathbb{R}$. The σ-field $\sigma(\mathcal{C})$ is called the *Borel σ-field over* $\mathbb{R}$, and is denoted by $\mathcal{B}(\mathbb{R})$. ∎

We often call this Borel σ-field over $\mathbb{R}$ just the Borel field, and denote it just by $\mathcal{B}$.

Borel Sets

Any set in $\mathcal{B}$ is called a Borel set. Such sets are said to be "Borel measurable", from the fact that they are λ-measurable, for the Lebesgue measure λ in equation (0.1.20).

Example 0.1.4 (Borel-measurable sets)
The following are all Borel-measurable sets.

1. $\mathbb{R}$
2. $\emptyset$
3. any countable set; in particular, any finite set, $\mathbb{Z}$, $\mathbb{Z}_+$ (the *natural numbers*), and the set of all rational numbers
4. hence, from the foregoing, the set of all irrational numbers (which is uncountable)
5. any interval, open, closed, or neither
6. the Cantor set
   The Cantor set is $\cap_{i=1}^{\infty} C_i$, where

   $$C_1 = [0, 1/3] \cup [2/3, 1], \quad C_2 = [0, 1/9] \cup [2/9, 1/3] \cup [2/3, 7/9] \cup [8/9, 1], \quad \ldots,$$

   We see that each of these is Borel, and hence, so is the intersection. A Cantor set has interesting properties; for example, its cardinality is the same as that of the interval $[0, 1]$, yet it is nowhere dense in $[0, 1]$. (The particular set described here is the Cantor ternary set; other similar sets are also called Cantor sets.)

7. the Smith-Volterra-Cantor set
   Instead of removing a fixed percentage of the subintervals at each stage, as in the case of a Cantor set, we can form a "fat" Cantor set by removing at each stage a decreasing percentage. The Smith-Volterra-Cantor set is formed by first removing the middle $1/4$ open subinterval from $[0,1]$ (that is, leaving the set $[0,3/8] \cup [5/8,1]$), then at the k^{th} stage, removing the middle 2^{-2k} open subintervals from each of the 2^{k-1} subintervals. The Smith-Volterra-Cantor set, as the Cantor set, has cardinality the same as that of the interval $[0,1]$ and yet is nowhere dense.
8. any union of any of the above ∎

So, are all subsets of $\mathbb{R}$ Borel sets?

No. Interestingly enough, the cardinality of $\mathcal{B}$ can be shown to be the same as that of $\mathbb{R}$, and the cardinality of the collection of all subsets of $\mathbb{R}$, that is, the cardinality of the power set, $2^{\mathbb{R}}$, is much larger – which means there are *many* subsets of $\mathbb{R}$ that are not Borel sets.

Equivalent Definitions of the Borel σ-Field

The facts that unions of closed sets may be open and that intersections of open intervals may be closed allow us to characterize the Borel σ-field $\mathcal{B}(\mathbb{R})$ in various ways. The canonical definition is that $\mathcal{B} = \sigma(\mathcal{C})$, where $\mathcal{C}$ is the collection of all finite open intervals. This is a simple version of the general definition of a Borel σ-field for a topological space. (In that definition, the generator is the collection of *all* open sets.) The following theorem list three other useful collections of subsets of $\mathbb{R}$ that generate the Borel σ-field.

Theorem 0.1.12
The Borel σ-field $\mathcal{B}(\mathbb{R})$ is generated by any of the following collections of subsets of $\mathbb{R}$:
(i) the collection of all finite closed intervals $[a,b]$ of $\mathbb{R}$,
(ii) the collection of all semi-infinite half-open intervals $]-\infty,b]$ of $\mathbb{R}$, and
(iii) the collection of all semi-infinite open intervals $]a,-\infty[$ of $\mathbb{R}$.

Proof.
To show that the σ-fields generated by two collections $\mathcal{C}$ and $\mathcal{D}$ are the same, we use the fact that a σ-field is closed with respect to countable intersections (remember the usual definition requires that it be closed with respect to countable unions) and then we show that
(1) $C \in \mathcal{C} \Rightarrow C \in \sigma(\mathcal{D})$ and
(2) $D \in \mathcal{D} \Rightarrow D \in \sigma(\mathcal{C})$.

Hence, to prove part (i), let $\mathcal{D}$ the collection of all finite closed intervals of $\mathbb{R}$.
(1) assume $D =$[a,b]$\in \mathcal{D}$. Now, consider the sequence of sets $B_i =]a-1/i,\ b+1/i[$. These open intervals are in $\mathcal{B}$, and hence,

$$\bigcap_{i=1}^{\infty}]a - 1/i,\ b + 1/i[= [a, b] \in \mathcal{B}.$$

Next,
(2) let $]a, b[$ be any set in the generator collection of $\mathcal{B}$, and consider the sequence of sets $D_i = [a+1/i,\ b-1/i]$, which are in $\mathcal{D}$. By definition, we have $\bigcup_{i=1}^{\infty}[a + 1/i,\ b - 1/i] =]a, b[\in \sigma(\mathcal{D})$

Proofs of parts (ii) and (iii) are left as exercises. ∎

Likewise, the collections of semi-infinite intervals of the form $]\infty, b[$ or $[a, \infty[$ generate $\mathcal{B}(\mathbb{R})$.

We also get the same Borel field $\mathcal{B}(\mathbb{R})$ by using the collection all open sets of $\mathbb{R}$, as in the general definition of a Borel σ-field for a topological space. (Exercise.)

The σ-Field $\mathcal{B}_{[0,1]}$

We are often interested in some subspace of $\mathbb{R}^d$, for example an interval (or rectangle). One of the most commonly-used intervals in $\mathbb{R}$ is $[0, 1]$.

For the sample space $\Omega = [0, 1]$, the most useful σ-field consists of the collection of all sets of the form $[0, 1] \cap B$, where $B \in \mathcal{B}(\mathbb{R})$. We often denote this σ-field as $\mathcal{B}_{[0,1]}$.

The σ-field formed in this way is the same as the σ-field generated by all open intervals on $[0, 1]$; that is, $\mathcal{B}([0, 1])$. (The reader should show this, of course.)

Product Borel σ-Fields

For the d-product measurable space generated by $(\mathbb{R}, \mathcal{B})$, the σ-field is $\sigma(\mathcal{B}^d)$, as stated in Definition 0.1.8, and so the measurable space of interest is $(\mathbb{R}^d, \sigma(\mathcal{B}^d))$.

As pointed out on page 701, the σ-field generated by the product of a number of σ-fields is not necessarily the same as the product of the σ-fields. It can be shown, however, that the σ-field $\sigma(\mathcal{B}^d)$ is $\mathcal{B}^d$. Furthermore, this is the σ-field that would result from definition 0.1.25 by extending the collection $\mathcal{C}$ of all open intervals in $\mathbb{R}$ to the collection $\mathcal{C}^d$ of all open intervals (or "hyperrectangles") in $\mathbb{R}^d$.

The product measurable space of interest, therefore, is $(\mathbb{R}^d, \mathcal{B}^d)$.

Measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$

The various types of measures we discussed beginning on page 707 may all be defined on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, but for this measurable space, the most common measure is the *Lebesgue measure*. Lebesgue measure is the extension (see page 712) of Borel measure, which we now define.

Definition 0.1.26 (Borel measure on $(\mathbb{R}, \mathcal{B})$)
The *Borel measure* is defined on $(\mathbb{R}, \mathcal{B})$ by the relation

$$\lambda(]a, b[) = b - a, \tag{0.1.20}$$

for given real numbers $a \leq b$. ∎

Lebesgue Measure on $\mathbb{R}$

Although for most purposes, $(\mathbb{R}, \mathcal{B})$ is the basic structure that we work with, as it turns out, $(\mathbb{R}, \mathcal{B})$ is not complete wrt the measure λ defined in equation (0.1.20).

*** describe extension *** add stuff on Carathéodory

*** Lebesgue σ-field

*** use λ to denote Lebesgue measure

It is clear from the definition of a measure (Definition 0.1.10) that λ is a measure, and because λ is a measure, we see that

$$\lambda([a, b]) = \lambda(]a, b[). \tag{0.1.21}$$

Although it is not finite, the Lebesgue measure is σ-finite, as can be seen from the sequence of open intervals $]-i, i[$.

The measurable space $(\mathbb{R}, \mathcal{B})$ is a topological space, and the Lebesgue measure is a Radon measure (exercise). Furthermore, along with addition, $\mathbb{R}$ is a group, and the Lebesgue measure is a Haar invariant measure wrt that group. (Note that the property of Haar invariance depends on the operation within a group). This latter fact is expressed by saying that Lebesgue measure is *translation invariant*.

It can be shown that any σ-finite translation invariant measure μ on $(\mathbb{R}, \mathcal{B})$ is equivalent to Lebesgue measure in the sense that there is a positive constant c, such that for any $A \in \mathcal{B}$, $\lambda(A) = c\mu(A)$.

The space over which the Lebesgue measure is defined is a linear space. The Lebesgue measure is translation invariant, as noted above, and it is also *scale equivariant*. Let $A \in \mathcal{B}$, and for $b, x \in \mathbb{R}$ with $b > 0$, let $bA + x = \{by + x \,|\, y \in A\}$. We have

$$\lambda(bA + x) = b\lambda(A). \tag{0.1.22}$$

A set $A \subseteq \mathbb{R}^d$ such that $\lambda(A) = 0$ is called a *null set* (whether or not $A \in \mathcal{B}^d$). It is clear that all countable sets are null sets. For example, the set of rational numbers has measure 0. An example of an uncountable set that is a null set is the Cantor set, as we see by computing the measure of what is taken out of the interval $[0, 1]$:

$$\sum_{k=1}^{\infty} 2^{k-1}3^{-k} = \frac{1}{3} + \frac{2}{3^2} + \frac{2^2}{3^3} + \cdots = 1.$$

This implies that the measure of what is left, that is, the nowhere dense (ternary) Cantor set is 0.

Another set that is uncountable but nowhere dense, is the Smith-Volterra-Cantor set. Its measure, however, is $1/2$, as we see by summing the measures of what is taken out.

$$\sum_{k=1}^{\infty} 2^{k-1} 2^{-2k} = \frac{1}{2}.$$

Singular Measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$

Two other measures that are often useful in $(\mathbb{R}, \mathcal{B})$ are the Dirac measure (equation (0.1.12)) and the counting measure (equation (0.1.13)).

It is easy to see from the definitions that both the Dirac measure δ and the counting measure γ are singular with respect to the Lebesgue measure:

$$\delta \perp \lambda \tag{0.1.23}$$

and

$$\gamma \perp \lambda. \tag{0.1.24}$$

Hence, neither of these measures is absolutely continuous with respect to the Lebesgue measure, and the Lebesgue measure is not absolutely continuous with respect to either of them.

Note, for example, $\delta_0(\{0\}) = 1$ but $\lambda(\{0\}) = 0$, while $\delta_0([1,2]) = 0$ but $\lambda([1,2]) = 1$.

A measure on $(\mathbb{R}, \mathcal{B})$ that is singular with respect to the Lebesgue measure is called simply a singular measure.

Example 0.1.5 (A non-Borel-measurable set)
A simple example of a non-Borel-measurable set, called a Vitali set, can be constructed using the Axiom of Choice. We begin by defining equivalence classes within $\mathbb{R}$ by making $x, y \in \mathbb{R}$ equivalent, written $x \sim y$, iff $x - y$ is rational. For each $x \in \mathbb{R}$, we identify the equivalence class E_x as $\{y \mid y \sim x\}$. The collection of these equivalence classes is a countable partition of $\mathbb{R}$. (Recall that the set of rationals is countable.) Now we form the *Vitali set* V by choosing exactly one member of each equivalence class in the interval $[0,1]$. Next we show that the Vitali set is nonmeasurable by contradiction. Let $q_1, q_2, \ldots$ represent the distinct (countable) rationals in $[-1,1]$, and form the disjoint countable sequence of sets $V_k = \{v + q_k \mid v \in V\}$. (Why are the sets disjoint?) Now, assume V is Lebesgue measurable (that is, "Borel measurable"). Because Lebesgue measure is translation invariant (equation (0.1.22)), if V is Lebesgue measurable, so is each V_k, and in fact, $\lambda(V_k) = \lambda(V)$. Note that

$$[0,1] \subseteq \bigcup_k V_k \subseteq [-1,2] \tag{0.1.25}$$

(why?), and so

$$1 \leq \lambda\left(\bigcup_k V_k\right) \leq 3.$$

We also have

$$\lambda\left(\bigcup_k V_k\right) = \sum_k \lambda(V_k) = \sum_k \lambda(V),$$

which must be either 0 or infinite, in either case contradicting

$$1 \leq \sum_k \lambda(V) \leq 3,$$

which follows only from the properties of measures and the assumption that V is Lebesgue measurable. We therefore conclude that V is not measurable. ∎

Borel Measurable Functions

We will now consider real-valued functions; that is, mappings into $\mathbb{R}^d$. The domains are not necessarily real-valued. We first identify two useful types of real-valued functions.

Definition 0.1.27 (indicator function)
The *indicator function*, denoted $\mathrm{I}_S(x)$ for a given set S, is defined by $\mathrm{I}_S(x) = 1$ if $x \in S$ and $\mathrm{I}_S(x) = 0$ otherwise. ∎

Notice that $\mathrm{I}_S^{-1}[A] = \emptyset$ if $0 \notin A$ and $1 \notin A$; $\mathrm{I}_S^{-1}[A] = S$ if $0 \notin A$ and $1 \in A$; $\mathrm{I}_S^{-1}[A] = S^{\mathrm{c}}$ if $0 \in A$ and $1 \notin A$; and $\mathrm{I}_S^{-1}[A] = \Omega$ if $0 \in A$ and $1 \in A$. Hence, $\sigma(\mathrm{I}_S)$ is the second most trivial σ-field we referred to earlier; i.e., $\sigma(S) = \{\emptyset, S, S^{\mathrm{c}}, \Omega\}$.

Definition 0.1.28 (simple function)
If $A_1, \ldots, A_k$ are measurable subsets of Ω and $a_1, \ldots, a_k$ are constant real numbers, a function φ is a *simple function* if for $\omega \in \Omega$,

$$\varphi(\omega) = \sum_{i=1}^{k} a_i \mathrm{I}_{A_i}(\omega), \tag{0.1.26}$$

where $\mathrm{I}_S(x)$ is the indicator function. ∎

Recall the convention for functions that we have adopted: the domain of the function is the sample space; hence, the subsets corresponding to constant values of the function form a finite partition of the sample space.

Definition 0.1.29 (Borel measurable function)
A measurable function from $(\Omega, \mathcal{F})$ to $(\mathbb{R}^d, \mathcal{B}^d)$ is said to be *Borel measurable* with respect to $\mathcal{F}$. ∎

A function that is Borel measurable is called a *Borel function*.

Simple functions are Borel measurable. (Exercise.)

The following theorem is useful because it allows us to build up any measurable real-valued function from a sequence of simple functions.

Theorem 0.1.13
Every measurable real-valued function can be represented at any point as the limit of a sequence of simple functions.

Proof. Let f be real and measurable. Now, if $f(\omega) \geq 0$, there exists a sequence $\{f_n\}$ of simple functions such that

$$0 \leq f_n(\omega) \nearrow f(\omega) \quad \text{a.e.},$$

and if $f(\omega) \leq 0$, there exists a sequence $\{f_n\}$ of simple functions such that

$$0 \geq f_n(\omega) \searrow f(\omega) \quad \text{a.e.}$$

The sequence is

$$f_n(\omega) = \begin{cases} -n & \text{if} \quad f(\omega) \leq -n, \\ -(k-1)2^{-n} & \text{if} \quad -k2^{-n} < f(\omega) \leq -(k-1)2^{-n}, \text{ for } 1 \leq k \leq n2^{-n}, \\ (k-1)2^{-n} & \text{if} \quad (k-1)2^{-n} < f(\omega) < k2^{-n}, \text{ for } 1 \leq k \leq n2^{-n}, \\ n & \text{if} \quad n \leq f(\omega). \end{cases}$$

∎

As a corollary of Theorem 0.1.13, we have that for a nonnegative random variable X, there exists a sequence of simple (degenerate) random variables $\{X_n\}$ such that

$$0 \leq X_n \nearrow X \quad \text{a.s.} \tag{0.1.27}$$

0.1.5 Real-Valued Functions over Real Domains

In the foregoing we have given special consideration to real-valued functions over arbitrary domains. In the following we consider real-valued functions over real domains. For such functions, we identify some additional properties, and then we define integrals and derivatives of real-valued functions over real domains.

In most practical purposes, two functions are "equal" if they are equal almost everywhere. For real-valued functions over real domains, almost everywhere usually means wrt Lebesgue measure, and when we use the phrase "almost everywhere" or "a.e." without qualification, that is what we mean.

Continuous Real Functions

Continuity is an important property of some functions. On page 626 we defined continuous functions in general topological spaces. For real functions on a real

domain, we equivalently define continuity in terms of the Euclidean distance between two points in the domain and the Euclidean distance between the corresponding function values.

Definition 0.1.30 (continuous function)
Let f be a real-valued function whose domain is a set $D \subseteq \mathbb{R}^d$. We say that f is *continuous at the point* $x \in D$ if f is defined in an open neighborhood of x, and given $\epsilon > 0, \exists \delta \ni \forall y \in D \ni \|x - y\| < \delta, \|f(x) - f(y)\| < \epsilon$. ∎

Here, the norms are the Euclidean norms. Notice that the order of $f(x)$ may be different from the order of x.

The δ in the definition may depend on x as well as on ϵ.

If f is continuous at each point in a subset of its domain, we say it is continuous on that subset. If f is continuous at each point in its domain, we say that f is *continuous*.

We have an immediate useful fact about continuous functions:

Theorem 0.1.14
If f is a continuous function, the inverse image f^{-1} of an open set is open.

Proof. Follows immediately from the definition. ∎

There are various types of continuity, and some examples will help to illustrate the differences.

Example 0.1.6 (the Dirichlet function;) *nowhere continuous function*
The indicator function of the rational numbers, called the Dirichlet function, is everywhere discontinuous. ∎

Example 0.1.7 (the Thomae function;) *continuous on irrationals, discontinuous on rationals*
Let $f(x)$ be defined as

$$f(x) = \begin{cases} 1 \text{ if } x = 0 \\ \frac{1}{q} \text{ if } x = \frac{p}{q} \text{ is rational,} \\ \quad \text{where } q \text{ is a positive integer and } p \text{ is relatively prime to } q \\ 0 \text{ if } x \text{ is irrational} \end{cases}$$

Then $f(x)$, called the Thomae function, is continuous at x if x is irrational and discontinuous at x if x is rational. ∎

We now consider three successively stronger types of continuity, and one modification of the strong type.

Definition 0.1.31 (uniformly continuous function)
Let f be a real-valued function whose domain includes a set $D \subseteq \mathbb{R}^d$. We say that f is *uniformly continuous over* D if, given $\epsilon > 0, \exists \delta \ni \forall x, y \in D$ with $\|x - y\| < \delta$,

$$\|f(x) - f(y)\| < \epsilon. \tag{0.1.28}$$

∎

Continuity is a point-wise property, while uniform continuity is a property for all points in some given set.

Example 0.1.8 *continuous but not uniformly continuous*
The function $f(x) = 1/x$ is continuous on $]0, \infty[$, but is not uniformly continuous over that interval. This function is, however, uniformly continuous over any closed and bounded subinterval of $]0, \infty[$. The Heine-Cantor theorem, in fact, states that any function that is continuous over a compact set is uniformly continuous over that set. ∎

If $\{x_n\}$ is a Cauchy sequence in the domain of a a uniformly continuous function f, then $\{f(x_n)\}$ is also a Cauchy sequence.

If a function f is uniformly continuous over a finite interval $]a, b[$, then f is bounded over $]a, b[$.

Definition 0.1.32 (absolutely continuous function)
Let f be a real-valued function defined on $[a, b]$ (its domain may be larger). We say that f is *absolutely continuous* on $[a, b]$ if, given $\epsilon > 0$, there exists a δ such that for every finite collection of nonoverlapping open rectangles $]x_i, y_i[\subseteq [a, b]$ with $\sum_{i=1}^{n} \|x_i - y_i\| < \delta$,

$$\sum_{i=1}^{n} \|f(x_i) - f(y_i)\| < \epsilon. \tag{0.1.29}$$

∎

(We defined absolute continuity of a measure with respect to another measure in Definition 0.1.23. Absolute continuity of a function f is a similar concept with respect to the Lebesgue measure over the domain and range of f.)

We also speak of local absolute continuity of functions in the obvious way.

If f is absolutely continuous over D, it is uniformly continuous on D, but the converse is not true.

Example 0.1.9 (the Cantor function) *uniformly continuous but not absolutely continuous*
The Cantor function, defined over the interval $[0, 1]$, is an example of a function that is continuous everywhere, and hence, uniformly continuous on that compact set, but not absolutely continuous. The Cantor function takes different values over the different intervals used in the construction of the Cantor set (see page 714). Let $f_0(x) = x$, and then for $n = 0, 1, \ldots$, let

$$\begin{aligned} f_{n+1}(x) &= 0.5 f_n(3x) \quad \text{for } 0 \le x < 1/3 \\ f_{n+1}(x) &= 0.5 \quad \text{for } 1/3 \le x < 2/3 \\ f_{n+1}(x) &= 0.5 + 0.5 f_n(3(x - 2/3)) \quad \text{for } 2/3 \le x \le 1. \end{aligned}$$

The Cantor function is

$$f(x) = \lim_{n\to\infty} f_n(x). \tag{0.1.30}$$

The Cantor function has a derivative of 0 almost everywhere, but has no derivative at any member of the Cantor set. (We define derivatives below, and in a more general way on page 739.) The properties of this function are discussed very carefully on pages 131–135 of Boas Jr. (1960), who uses a tertiary representation of the points in the set and a binary representation of the values of the function to demonstrate continuity and the derivatives or lack thereof. ∎

An absolutely continuous function is of bounded variation; it has a derivative almost everywhere; and if the derivative is 0 a.e., the function is constant.

A slightly stronger form of continuity is Lipschitz-continuity. It places an explicit bound on the amount by which the function can change.

Definition 0.1.33 (Lipschitz-continuous function)
Let f be a real-valued function whose domain is an interval $D \subseteq \mathbb{R}^d$. We say that f is *Lipschitz-continuous* if for any $y_1, y_2 \in D$ and $y_1 \neq y_2$, there exists γ such that

$$\|f(y_1) - f(y_2)\| \leq \gamma \|y_1 - y_2\|. \tag{0.1.31}$$

The smallest γ for which the inequality holds is called the *Lipschitz constant*. ∎

We also speak of local Lipschitz continuity in the obvious way.

Every Lipschitz-continuous function is absolutely continuous. Lipschitz continuity plays an important role in nonparametric function estimation.

The graph of a scalar-valued Lipschitz-continuous function f over $D \subseteq \mathbb{R}$ has the interesting geometric property that the entire graph of $f(x)$ lies between the lines $y = f(c) \pm \gamma(x - c)$ for any $c \in D$.

Example 0.1.10 *absolutely continuous but not Lipschitz continuous*
The function $f(x) = \sqrt{x}$ for $x \in [0, 1]$ is an example of a absolutely continuous everywhere on $[0, 1]$, but is not Lipschitz continuous on that set. (The problem with Lipschitz continuity occurs at $x = 0$.) ∎

Finally, a slight modification of Lipschitz-continuity yields another form of continuity called uniform Lipschitz-continuity of order α, or Hölder continuity of order α.

Definition 0.1.34 (Hölder-continuous function)
Let f be a real-valued function whose domain is an interval $D \subseteq \mathbb{R}^d$. We say that f is *Hölder-continuous of order α* where $\alpha > 0$, if for any $y_1, y_2 \in D$ and $y_1 \neq y_2$, there exists γ such that

$$\|f(y_1) - f(y_2)\| \leq \gamma \|y_1 - y_2\|^\alpha. \tag{0.1.32}$$

∎

We also speak of local Hölder continuity in the obvious way.

Depending on α, Hölder continuity may be stronger or weaker than Lipschitz continuity. For $\alpha < 1$, Hölder continuity does not guarantee differentiability, whereas uniform continuity, and a fortiori, Lipschitz continuity, does guarantee it, except on set of measure 0. (See Example 0.1.11.)

Differentiability; Derivatives of Functions

Continuity has to do with how function values change as the function argument changes. A continuous function does not have abrupt changes. Differentiability is a related concept that has to do with the rate of change. Here we define a very useful type of differentiation. We define derivatives in a more general way on page 739. Unlike continuity, here we will define differentiability only for functions defined on $\mathbb{R}$. The definition generalizes, but in $\mathbb{R}^d$ for $d > 1$ there are some additional important issues involving directions.

Definition 0.1.35 (differentiable function)
Let x be a point in $\mathbb{R}$ and let f be a real-valued function defined in an open neighborhood of x. We say that f is *differentiable at the point* x if the limit

$$\lim_{h \to 0} \frac{f(x+h) - f(x)}{h} \tag{0.1.33}$$

exists.

If the limit (0.1.33) exists, it is called the *derivative of* f *at the point* x and is denoted as f'. Wherever it exists, the derivative is a function, and we often denote it as $f'(x)$. ∎

Differentiability obviously depends on continuity, but does continuity guarantee differentiability?

Example 0.1.11 (the Weierstrass function) *continuous everywhere but differentiable nowhere*
The Weierstrass function, defined over the interval $[-2, 2]$, is an example of a function that is continuous everywhere but differentiable nowhere. The Weierstrass function is

$$f(x) = \sum_{n=0}^{\infty} a^n \cos(b^n x \pi), \tag{0.1.34}$$

where $0 < a < 1$ and b is a positive odd integer such that $ab > 1 + 3\pi/2$.

This example shows that Hölder continuity may not be sufficient to guarantee differentiability. The Weierstrass function is Hölder continuous for all orders $\alpha < 1$ (Exercise **??**). ∎

***Function that is continuous but not differentiable Weierstrass, Wiener

Uniform continuity is the weakest form that guarantees differentiability. A uniformly continuous function is differentiable almost everywhere. Even Lipschitz-continuity does not guarantee differentiability. For example $f(x) = |x|$ is Lipschitz continuous over $[-a, a]$, but it is not differentiable at $x = 0$.

Sequences of Functions; lim sup and lim inf

We now consider some properties of sequences of functions, $\{f_n\}$. We will limit our attention to Borel functions. Important properties of a sequence of functions are its lim sup and lim inf, which we define analogously to the meaning of lim sup and lim inf of a sequence of sets given in equations (0.0.10) and (0.0.11):

$$\limsup_n f_n \stackrel{\text{def}}{=} \inf_n \sup_{i \geq n} f_i \tag{0.1.35}$$

and

$$\liminf_n f_n \stackrel{\text{def}}{=} \sup_n \inf_{i \geq n} f_i. \tag{0.1.36}$$

We first consider the σ-fields generated by a sequence of functions. An important result is the following.

Theorem 0.1.15
Let $\{f_n\}$ be a sequence of Borel functions on a measurable space $(\Omega, \mathcal{F})$. Then (i)

$$\sigma(f_1, f_2, \ldots) = \sigma(\cup_{n=1}^{\infty} \sigma(f_n)) \tag{0.1.37}$$

$$= \sigma(\cup_{j=1}^{\infty} \sigma(f_1, \ldots, f_j)) \tag{0.1.38}$$

(ii)

$$\sigma(\limsup f_n) = \sigma(\cap_{n=1}^{\infty} \sigma(f_n, f_{n+1}, \ldots)). \tag{0.1.39}$$

Proof. *** ∎

We identify two types of convergence of functions, pointwise convergence and uniform convergence, with a distinction reminiscent of that between continuity and uniform continuity. We will then consider a stronger type of pointwise convergence, and show a relationship between strong pointwise convergence and uniform convergence.

Definition 0.1.36 (pointwise convergence)
Let $\{f_n\}$ be a sequence of real-valued function over a real domain D, and likewise let f be a real-valued function over D. We say $\{f_n\}$ *converges to f at the point $x \in D$* iff for

$$\forall \epsilon > 0, \exists N \ni n \geq N \Rightarrow \|f_n(x) - f(x)\| < \epsilon.$$

We write $f_n(x) \to f(x)$. ∎

The "N" in the definition may depend on x. In uniform convergence it does not.

Definition 0.1.37 (uniform convergence)
Let $\{f_n\}$ be a sequence of real-valued function over a real domain D, and likewise let f be a real-valued function over D. We say $\{f_n\}$ *converges uniformly* to f iff for

$$\forall \epsilon > 0, \exists N \ni n \geq N \Rightarrow \|f_n(x) - f(x)\| < \epsilon \, \forall x \in D.$$

We write $f_n \to f$ or $f_n(x) \overset{\text{uniformly}}{\to} f(x)$. ∎

Uniform convergence can also be limited to a subset of the domain, in which case we may call a function *locally uniformly convergent*.

Definition 0.1.38 (almost everywhere (pointwise) convergence)
*** We write $f_n(x) \overset{\text{a.e.}}{\to} f(x)$. ∎

Next, we consider a relationship between types convergence of functions. The most important and basic result is stated in the Severini-Egorov theorem, also called Egorov's or Egoroff's theorem.

Theorem 0.1.16 (Severini-Egorov theorem)
Let $\{f_n\}$ be a sequence of Borel functions on a measure space $(\Omega, \mathcal{F}, \nu)$. For any $A \in \mathcal{F}$ such that $\nu(A) < \infty$, suppose that $f_n(\omega) \to f(\omega) \forall \omega \in A$. Then

$$\forall \epsilon > 0, \exists B \subseteq A \text{ with } \nu(B) < \epsilon \ni f_n(\omega) \to f(\omega) \text{ on } A \cap B^c.$$

Proof. *** ∎

The Severini-Egorov theorem basically states that pointwise convergence almost everywhere on A implies the stronger uniform convergence everywhere except on some subset B of arbitrarily small measure. This type of convergence is also called *almost uniform convergence*.

This theorem is Littlewood's principle of real analysis that states that every convergent sequence of functions is "nearly" uniformly convergent (see Littlewood (1944)). We will encounter this principle again in connection with the monotone convergence theorem on page 733.

0.1.6 Integration

Integrals are some of the most important functionals of real-valued functions. Integrals and the action of integration are defined using measures. Although much of integration theory could be developed over abstract sets, we will generally assume that the domains of the functions are real and the functions are real-valued.

Integrals of nonnegative functions are themselves measures. There are various types of integrals, Lebesgue, Riemann, Riemann-Stieltjes, Ito, and so on. The most important in probability theory is the Lebesgue, and when we use the term "integral" without qualification that will be the integral meant. We begin with the definition and properties of the Lebesgue integral. We briefly discuss the Riemann integral in Section 0.1.6 on page 735, the Riemann-Stieltjes integral in Section 0.1.6 on page 736, and the Ito integral in Section 0.2.2 on page 775.

The Lebesgue Integral of a Function with Respect to a Given Measure: The Definition

An *integral of a function f with respect to a given measure ν*, if it exists, is a functional whose value is an average of the function weighted by the measure. It is denoted by $\int f \, \mathrm{d}\nu$. The function f is called the *integrand*.

The integral is defined over the sample space of a given measure space, say $(\Omega, \mathcal{F}, \nu)$. This is called the *domain* of the integral. We often may consider integrals over different domains formed from a sub measure space, $(D, \mathcal{F}_D, \nu)$ for some set $D \in \mathcal{F}$, as described above. We often indicate the domain explicitly by notation such as $\int_D f \, \mathrm{d}\nu$.

If the domain is a real interval $[a, b]$, we often write the restricted interval as $\int_a^b f \, \mathrm{d}\nu$. If ν is the Lebesgue measure, this integral is the same as the integral over the open interval $]a, b[$.

We also write an integral in various equivalent ways. For example if the integrand is a function of real numbers and our measure is the Lebesgue measure, we may write the integral over the interval $]a, b[$ as $\int_a^b f(x) \, \mathrm{d}x$.

We build the definition of an integral of a function in three steps: first for nonnegative simple functions, then for nonnegative Borel functions, and finally for general Borel functions.

Definition 0.1.39 (integral of a nonnegative simple function)
If f is a simple function defined as $f(\omega) = \sum_{i=1}^k a_i \mathrm{I}_{A_i}(\omega)$, where the A_is are measurable with respect to ν, then

$$\int f \, \mathrm{d}\nu = \sum_{i=1}^k a_i \nu(A_i). \tag{0.1.40}$$

∎

Note that a simple function over measurable A_is is necessarily measurable.

What about the case in which $\nu(A_i) = \infty$, as for example when the domain of f is the real line and ν is the Lebesgue measure? We adopt the convention that

$$\infty + \infty = \infty,$$
$$\infty \cdot 0 = 0 \cdot \infty = 0,$$

and for $c > 0$,

$$c \cdot \infty = c + \infty = \infty,$$

and so the integral (0.1.40) is always defined, although it may be infinite.

We define the integral of a nonnegative Borel function in terms of the supremum of a collection of simple functions.

Definition 0.1.40 (integral of a nonnegative Borel function)
Let f be a nonnegative Borel function with respect to ν on Ω, and let S_f be the collection of all nonnegative simple functions such that

$$\varphi \in S_f \Rightarrow \varphi(\omega) \leq f(\omega) \quad \forall\, \omega \in \Omega$$

The integral of f with respect to ν is

$$\int f \,\mathrm{d}\nu = \sup \left\{ \int \varphi \,\mathrm{d}\nu \;\middle|\; \varphi \in S_f \right\}. \tag{0.1.41}$$

∎

Another way of stating this definition in the measure space $(\Omega, \mathcal{F}, \nu)$ is to consider various finite partitions of Ω using sets in $\mathcal{F}$. (See the discussion on page 700.) If $\{A_i\}$ is such a partition, we form the sum

$$\sum_i \inf_{\omega \in A_i} f(\omega)\nu(A_i), \tag{0.1.42}$$

in which we adopt the conventions above so that if, in any addend in expression (0.1.42), either factor is 0, then the addend is 0. The definition in equation (0.1.41) is therefore equivalent to

$$\int f \,\mathrm{d}\nu = \sup_{\text{all partitions}} \sum_i \inf_{\omega \in A_i} f(\omega)\nu(A_i) \tag{0.1.43}$$

and so again the integral (0.1.41) is always defined, although it may be infinite.

Now consider general Borel functions. For a general Borel function f, we form two nonnegative Borel functions f_+ and f_- such that $f = f_+ - f_-$:

$$f_+(\omega) = \max\{f(\omega), 0\}$$

$$f_-(\omega) = \max\{-f(\omega), 0\}.$$

Definition 0.1.41 (integral of a general Borel function)
The integral of f with respect to ν is the difference of the integrals of the two nonnegative functions:

$$\int f \,\mathrm{d}\nu = \int f_+ \,\mathrm{d}\nu - \int f_- \,\mathrm{d}\nu, \tag{0.1.44}$$

so long as either $\int f_+ \,\mathrm{d}\nu$ or $\int f_- \,\mathrm{d}\nu$ is finite (because $\infty - \infty$ is not defined).
∎

We can rewrite the definition in equation (0.1.44) in a manner similar to how we rewrote equation (0.1.42) above:

$$\int f \,\mathrm{d}\nu = \sup_{\text{all partitions}} \sum_i \left| \inf_{\omega \in A_i} f(\omega) \right| \nu(A_i). \tag{0.1.45}$$

Note that, just as with the definitions for nonnegative functions above, the integral of a general Borel function may be infinite; in fact, it may be ∞ or $-\infty$.

For what kind of function would the Lebesgue integral not be defined? The Lebesgue integral is not defined for functions for which both the positive part and the negative of the negative part in equation (0.1.44) are ∞. The function $f(x) = \sin(x)/x$ over the positive real line is an example of such a function (but see the section beginning on page 738).

Although the definition allows the integral to be infinite, we use a special term for the case in which the integral is finite. If both $\int f_+ \, d\nu$ and $\int f_- \, d\nu$ are finite, the integral itself is finite, and in that case we say that f is *integrable*. Note that being Borel does not imply that a function is integrable.

We define the *integral over a domain* A as

$$\int_A f \, d\nu = \int \mathrm{I}_A f \, d\nu. \tag{0.1.46}$$

Although we may not explicitly identify the underlying measure space, technically there is one, say $(\Omega, \mathcal{F}, \nu)$, and $A \in \mathcal{F}$ and so $A \subseteq \Omega$.

Measures Defined by Integrals

The integral over a domain together with a nonnegative Borel function leads to an induced measure: If a given measure space $(\Omega, \mathcal{F}, \nu)$ and a given nonnegative Borel function f, let $\lambda(A)$ for $A \subseteq \Omega$ be defined as

$$\lambda(A) = \int_A f \, d\nu. \tag{0.1.47}$$

Then $\lambda(A)$ is a measure over $(\Omega, \mathcal{F})$ (exercise). Furthermore, because

$$\nu(A) = 0 \Rightarrow \lambda(A) = 0,$$

λ is absolutely continuous with respect to ν.

If $f \equiv 1$ the integral with respect to a given measure defines the same measure. This leads to the representation of the probability of an event as an integral. Given a probability space $(\Omega, \mathcal{F}, P)$, $\int_A dP$ is the *probability of* A, written $P(A)$ or $\Pr(A)$.

The properties of a measure defined by an integral depend on the properties of the underlying measure space and the function. For example, in $\mathbb{R}$ with Lebesgue measure ν, the measure for Borel sets of positive reals defined by

$$\lambda(A) = \int_A \frac{1}{x} \, d\nu(x) \tag{0.1.48}$$

is a *Haar measure* (see Definition 0.1.18). More interesting Haar measures are those defined over nonsingular $n \times n$ real matrices,

$$\mu(D) = \int_D \frac{1}{|\det(X)|^n} \, \mathrm{d}X,$$

or over matrices in the orthogonal group. (See Gentle (2007), pages 169–171.)

Properties of the Lebesgue Integral

Lebesgue integrals have several useful and important properties. In this section, we will consider integrals of a finite number of functions (often just one) over a single measure space. (A finite number greater than one is essentially equivalent to two.)

In Section 0.1.6 we will consider countable integrals and integrals of a countable number of functions over a single measure space.

In Section 0.1.6 we will consider integrals over more than one measure space.

The following theorems state several properties of integrals.

Theorem 0.1.17
Let f be a Borel function. Then

$$\nu(\{\omega \,|\, f(\omega) > 0\}) > 0 \implies \int f \mathrm{d}\nu > 0.$$

Proof. Exercise. ∎

Theorem 0.1.18
Let f be a Borel function. Then

$$\int f \mathrm{d}\nu < \infty \implies f < \infty \text{ a.e.}$$

Proof. Exercise. ∎

Theorem 0.1.19
Let f and g be Borel functions. Then
(i). $f = 0$ a.e. $\implies \int f \mathrm{d}\nu = 0$;
(ii). $f \leq g$ a.e. $\implies \int f \mathrm{d}\nu \leq \int g \mathrm{d}\nu$;
(iii). $f = g$ a.e. $\implies \int f \mathrm{d}\nu = \int g \mathrm{d}\nu$.

Proof. Exercise. ∎

One of the most important properties of the integral is the fact that it is a linear operator.

Theorem 0.1.20 (linearity)
For real a and Borel f and g, $\int (af + g) \, \mathrm{d}\nu = a \int f \, \mathrm{d}\nu + \int g \, \mathrm{d}\nu$.

Proof. To see this, we first show it for nonnegative functions f_+ and g_+, using the definition in equation (0.1.41). Then we use the definition in equation (0.1.44) for general Borel functions. ∎

Theorem 0.1.21
(i). $\int |f| \, \mathrm{d}\nu \geq 0$
(ii). $\int |f| \, \mathrm{d}\nu = 0 \Rightarrow f = 0$ *a.e.*

Proof. Follows immediately from the definition. ∎

This fact together with the linearity means that $\int |f| \, \mathrm{d}\nu$ is a *pseudonorm* for functions. A more general pseudonorm based on the integral is $(\int |f|^p \, \mathrm{d}\nu)^{1/p}$ for $1 \leq p$. (Deviating slightly from the usual definition of a norm, it may seem reasonable to allow the implication of $\int |f| \, \mathrm{d}\nu = 0$ to be only almost everywhere. Strictly speaking, however, without this weakened form of equality to 0, $\int |f| \, \mathrm{d}\nu$ is only a pseudonorm. See the discussion on page 638.)

Another important property of the integral is its monotonicity. First, we state this for a finite number of functions and integrals (in fact, for just two because in this case, two is effectively any finite number). Later, in Section 0.1.6, we will consider analogous properties for an infinitely countable number of functions.

Theorem 0.1.22 (finite monotonicity)
For integrable f *and* g, $f \leq g$ a.e. $\Rightarrow \int f \, \mathrm{d}\nu \leq \int g \, \mathrm{d}\nu$.

Proof. Follows immediately from the definition. ∎

Limits of Functions and Limits of Integrals

There are some conditions for interchange of an integration operation and a limit operation that are not so obvious. The following theorems address this issue and are closely related to each other. The fundamental theorems are the monotone convergence theorem, Fatou's lemma, and Lebesgue's dominated convergence theorem. These same three theorems provide important relationships between sequences of expectations and expectations of sequences, as we see on pages 89 and 113.

We begin with a lemma to prove the monotone convergence theorem.

Lemma 0.1.23.1
Assume a measure space $(\Omega, \mathcal{F}, \nu)$*, and Borel measurable functions* f_n *and* f.

$$0 \leq f_n(\omega) \nearrow f(\omega) \; \forall \omega \implies 0 \leq \int f_n \mathrm{d}\nu \nearrow \int f \mathrm{d}\nu.$$

Proof.
First, we observe that $\int f_n \mathrm{d}\nu$ is nondecreasing and is bounded above by $\int f \mathrm{d}\nu$. (This is Theorem 0.1.190.1.19.) So all we need to show is that $\lim_n \int f_n \mathrm{d}\nu \geq \int f \mathrm{d}\nu$. That is, writing the latter integral in the form of equation (0.1.43), we need to show that

$$\lim_n \int f_n \mathrm{d}\nu \geq \sup \sum_i \inf_{\omega \in A_i} f(\omega) \nu(A_i), \tag{0.1.49}$$

where the sup is taken over all finite partitions of Ω.

We consider separately the two cases determined by whether the right-hand side is finite.

First, suppose the right side is finite, and further, that each term in the sum is finite and positive; that is, each $\inf_{\omega\in A_i} f(\omega)$ and each $\nu(A_i)$ are finite and positive. (Recall the convention on page 728 for the terms in this sum that $\infty \cdot 0 = 0 \cdot \infty = 0$.) In this case there exists an ϵ such that for each i, $0 < \epsilon < y_i$, where $y_i = \inf_{\omega\in A_i} f(\omega)$. Now, define the subset of A_i,

$$A_{in} = \{\omega \,|\, \omega \in A_i, f_n(\omega) > y_i - \epsilon\}.$$

We now have a finite partition of Ω (as required in the definition of the integral) consisting of the A_{in} and the complement of their union, and so taking only some of the terms that represent the integral as a sum over that partition, we have

$$\int f_n \mathrm{d}\nu \geq \sum_i (y_i - \epsilon)\nu(A_{in}). \tag{0.1.50}$$

Because $f_n \nearrow f$, we have for each i, $A_{in} \nearrow A_i$ and the complement of their union goes to $\emptyset$. Because of ν is continuous from below, we have for the term on the right above,

$$\sum_i (y_i - \epsilon)\nu(A_{in}) \to \sum_i (y_i - \epsilon)\nu(A_i).$$

Hence, from inequality (0.1.50),

$$\int f_n \mathrm{d}\nu \geq \int f \mathrm{d}\nu - \epsilon \sum_i \nu(A_i),$$

and because all $\nu(A_i)$ are finite and ϵ can be arbitrarily close to 0, we have what we wanted to show, that is, inequality (0.1.49).

Now, still assuming that the right side of (0.1.49) is finite, that is, that $\int f \mathrm{d}\nu$ is finite, we allow some terms to be zero. (They all must still be finite as above, however.) Let us relabel the terms in the finite partition so that for $i \leq m_0$, $y_i\nu(A_i) > 0$ and for $m_0 < i \leq m$, $y_i\nu(A_i) = 0$. If $m_0 < 1$, then all $y_i\nu(A_i) = 0$, and we have inequality (0.1.49) immediately; otherwise for $i \leq m_0$, both y_i and $\nu(A_i)$ are positive and finite. In this case we proceed as before, but only for the positive terms; that is, for $i \leq m_0$, we define A_{in} as above, form the inequality (0.1.50), and by the same steps establish inequality (0.1.49).

Finally, suppose $\int f \mathrm{d}\nu$ is infinite. In this case, for some i_0, both y_{i_0} and $\nu(A_{i_0})$ are positive and at least one is infinite. Choose positive constants δ_y and δ_A and bound them away from 0:

$$0 < \delta_y < y_{i_0} \leq \infty \quad \text{and} \quad 0 < \delta_A < \nu(A_{i_0}) \leq \infty.$$

Now, similarly as before, define a subset of A_{i_0}:

$$A_{i_0 n} = \{\omega \,|\, \omega \in A_{i_0}, f_n(\omega) > \delta_y\}.$$

As before, $f_n \nearrow f \implies A_{i_0 n} \nearrow A_{i_0}$, and so for some n_0, for all $n > n_0$, $\mu(A_{i_0 n}) > \delta_A$. Now, with the partition of Ω consisting of $A_{i_0 n}$ and its complement, we have

$$\int f_n \mathrm{d}\nu \geq \delta_y \delta_A, \quad \text{for } n > n_0,$$

hence, $\lim_n \int f_n \mathrm{d}\nu \geq \delta_y \delta_A$. Now, if $y_{i_0} = \infty$, let $\delta_y \to \infty$ and if $\nu(A_{i_0}) = \infty$, let $\delta_A \to \infty$. Either way, we have $\lim_n \int f_n \mathrm{d}\nu = \infty$ and so we have inequality (0.1.49). ∎

Notice that this lemma applies in the case of pointwise convergence. If convergence is uniform, it would immediately apply in the case of convergence a.e. The next theorem provides the desired generalization.

Theorem 0.1.23 (monotone convergence)
Let $0 \leq f_1 \leq f_2 \leq \cdots$, *and* f *be Borel functions, and let* $\lim_{n\to\infty} f_n = f$ *a.e., then*

$$\int f_n \,\mathrm{d}\nu \nearrow \int f \,\mathrm{d}\nu. \tag{0.1.51}$$

Proof.
Assume the hypothesis: that is, $f_n \nearrow f$ for all $\omega \in A$ where $\nu(A^c) = 0$. Now restrict each function to A, and observe that $f_n \mathrm{I}_A \nearrow f\mathrm{I}_A$ and $\int f_n \mathrm{I}_A \,\mathrm{d}\nu = \int f_n \,\mathrm{d}\nu$ and $\int f \mathrm{I}_A \,\mathrm{d}\nu = \int f \,\mathrm{d}\nu$. Lemma 0.1.23.1 immediately implies $\int f_n \,\mathrm{d}\nu \nearrow \int f \,\mathrm{d}\nu$. ∎

That Theorem 0.1.23 follows so readily from Lemma 0.1.23.1 is another illustration of a principle of real analysis stated by Littlewood that every convergent sequence of functions is "nearly" uniformly convergent (see page 761). In the hypotheses of the lemma, we have only pointwise convergence. Without needing uniform convergence, however, we extend the conclusion to the case of convergence a.e.

Theorem 0.1.24 (Fatou's lemma)
For nonnegative integrable Borel f_n,

$$\int \liminf_n f_n \,\mathrm{d}\nu \leq \liminf_n \int f_n \,\mathrm{d}\nu. \tag{0.1.52}$$

Proof.
Let $g_n = \inf_{k\geq n} f_k$ and $g = \int \lim_n \inf f_n$. As in the monotone convergence theorem, g_n is nonnegative and $g_n \nearrow g$, so $\int g_n \,\mathrm{d}\nu \nearrow \int g \,\mathrm{d}\nu$. Also, for each n, $\int f_n \,\mathrm{d}\nu \geq \int g_n \,\mathrm{d}\nu$; hence, we have the desired conclusion. ∎

The next theorem is the most powerful of the convergence theorems for integrals.

Theorem 0.1.25 (Lebesgue's dominated convergence)
If $\lim_{n\to\infty} f_n = f$ *a.e. and there exists an integrable function* g *such that* $|f_n| \leq g$ *a.e., then*

$$\lim_{n\to\infty} \int f_n \, \mathrm{d}\nu = \int f \, \mathrm{d}\nu. \tag{0.1.53}$$

Proof.
***

Corollary 0.1.25.1 (bounded convergence)
Let $\{f_n\}$ *be a sequence of measurable functions defined on a set* A*, where* $\nu(A) < \infty$*. If for some real number* M*,* $|f_n(\omega)| \leq M$*, and* $\lim_{n\to\infty} f_n(\omega) = f(\omega)$ *for each* $\omega \in A$ *then*

$$\lim_{n\to\infty} \int_A f_n \, \mathrm{d}\nu = \int_A f \, \mathrm{d}\nu. \tag{0.1.54}$$

Integrals over More than One Measure Space

So far the integrals we have discussed have been for functions over a single measure space. We now consider integrals over more than one measure space.

We first consider a space $(\Omega, \mathcal{F}, \nu)$ and a space $(\Lambda, \mathcal{G}, \mu)$ together with a function $f : (\Omega, \mathcal{F}) \mapsto (\Lambda, \mathcal{G})$ that defines the measure μ, that is, $\mu = \nu \circ f^{-1}$. This is change of variables.

We next consider the relation between integration in a product measure space to integration in each of the component measure spaces. Fubini's theorem tells us that the integral over the product space is the same as the iterated integral.

We use Fubini's theorem in a somewhat surprising way to derive a useful formula for integration of products of functions called integration by parts.

Later, in Section 0.1.7, we consider two different measures on the same space and find that if one measure dominates the other, there is a unique function whose integral wrt to the dominating measure defines a measure as in equation (0.1.47) that is the same as the dominated measure. This is the Radon-Nikodym theorem, and leads to a useful function, the Radon-Nikodym derivative.

Change of Variables

Consider two measurable spaces $(\Omega, \mathcal{F})$ and $(\Lambda, \mathcal{G})$, let f be a measurable function from $(\Omega, \mathcal{F})$ to $(\Lambda, \mathcal{G})$, and let ν be a measure on $\mathcal{F}$. As we have seen, $\nu \circ f^{-1}$ is an induced measure on $\mathcal{G}$. Now let g be a Borel function on $(\Lambda, \mathcal{G})$. Then the integral of $g \circ f$ over Ω with respect to ν is the same as the integral of g over Λ with respect to $\nu \circ f^{-1}$:

$$\int_\Omega g \circ f \, \mathrm{d}\nu = \int_\Lambda g \, \mathrm{d}(\nu \circ f^{-1}) \tag{0.1.55}$$

Integration in a Product Space; Fubini's Theorem

Given two measure spaces $(\Omega_1, \mathcal{F}_1, \nu_1)$ and $(\Omega_2, \mathcal{F}_2, \nu_2)$ with σ-finite measures and a Borel function f on $\Omega_1 \times \Omega_2$, the integral over Ω_1, if it exists, is a function of $\omega_2 \in \Omega_2$ a.e., and likewise, the integral over Ω_2, if it exists, is a function of $\omega_2 \in \Omega_1$ a.e. Fubini's theorem shows that if one of these marginal integrals, exists a.e., then the natural extension of an integral to a product space, resulting in the *double integral*, is the same as the *iterated integral*.

Theorem 0.1.26 (Fubini's theorem)
Let $(\Omega_1, \mathcal{F}_1, \nu_1)$ and $(\Omega_2, \mathcal{F}_2, \nu_2)$ be measure spaces where the measures ν_1 and ν_2 are σ-finite. Let f be a Borel function on $\Omega_1 \times \Omega_2$, such that the marginal integral

$$g(\omega_2) = \int_{\Omega_1} f(\omega_1, \omega_2)\, \mathrm{d}\nu_1$$

exists a.e. Then

$$\int_{\Omega_1 \times \Omega_2} f(\omega_1, \omega_2)\, \mathrm{d}\nu_1 \times \mathrm{d}\nu_2 = \int_{\Omega_2} \left(\int_{\Omega_1} f(\omega_1, \omega_2)\, \mathrm{d}\nu_1 \right) \mathrm{d}\nu_2. \tag{0.1.56}$$

Proof. A proof is given in Billingsley (1995), page 233. ∎

Integration by Parts

*** corollary *** If f and g are bounded on the interval $[a, b]$ and have no common points of discontinuity in that interval, then

$$\int_{[a,b]} f(x)\mathrm{d}g(x) = f(b)g(b) - f(a)g(a) - \int_{[a,b]} g(x)\mathrm{d}f(x). \tag{0.1.57}$$

This is proved using Fubini's theorem.

The Riemann Integral

The Riemann integral is one of the simplest integrals. The Riemann integral is defined over intervals of $\mathbb{R}$ (or over rectangles in $\mathbb{R}^k$). The Riemann integral over the interval $[a, b]$ is defined in terms of a partitioning $\{[x_1, x_0[, [x_2, x_1[, \ldots, [x_n - x_{n-1}]\}$. We can define the Riemann integral of a real function f over the interval $[a, b]$ in terms of the Lebesgue measure λ as the real number r such that for any $\epsilon > 0$, there exists a δ such that

$$\left| r - \sum_{i \in P} f(x_i)\lambda(I_i) \right| < \epsilon \tag{0.1.58}$$

where $\{I_i : i \in P\}$ is any finite partition of $]a, b[$ such that for each i, $\lambda(I_i) < \delta$ and $x_i \in I_i$. If the Riemann integral exists, it is the same as the

Lebesgue integral. We use the same notation for the Riemann integral as for the Lebesgue integral, that is, we write r as

$$r = \int_a^b f(x)\mathrm{d}x. \tag{0.1.59}$$

Because of the use of Lebesgue measure in the definition, the integral over $[a, b]$ is the same as over $]a, b[$.

A classic example for which the Lebesgue integral exists, but the Riemann integral does not, is the Dirichlet function (Example 0.1.6) restricted to $]0, 1]$; that is, the function g defined over $]0, 1]$ as $g(x) = 1$ if x is rational, and $g(x) = 0$ otherwise. The Lebesgue integral $\int_0^1 g(x)\,\mathrm{d}x$ exists and equals 0, because $g(x) = 0$ a.e. The Riemann integral, on the other had does not exist because for an arbitrary partition $\{I_i\}$, the integral is 1 if $x_i \in I_i$ is taken as a rational, and the integral is 0 if $x_i \in I_i$ is taken as an irrational number.

The Riemann integral lacks the three convergence properties of the Lebesgue integral given on page 733.

We will not develop the properties of the Riemann integral here. When the Riemann integral exists, it has the same properties as the Lebesgue integral, such as linearity. Hence, the separately important questions involve the existence of the Riemann integral. We list some sufficient conditions for existence below. Proofs of these and other properties of the Riemann integral can be found in texts on advanced calculus, such as Khuri (2003).

- If $f(x)$ is continuous on $[a, b]$, then it is Riemann integrable over $[a, b]$.
- If $f(x)$ is monotone (increasing or decreasing) on $[a, b]$, then it is Riemann integrable over $[a, b]$. (Notice that the function may not be continuous.)
- If $f(x)$ is of bounded variation on $[a, b]$, then it is Riemann integrable over $[a, b]$.

The Riemann-Stieltjes Integral

The Riemann-Stieltjes integral is a generalization of the Riemann integral in which $\mathrm{d}x$ is replaced by $\mathrm{d}g(x)$ and the interval lengths are replaced by changes in $g(x)$. We write it as

$$r_s = \int_a^b f(x)\mathrm{d}g(x). \tag{0.1.60}$$

To define the Riemann-Stieltjes integral we will handle the partitions slightly differently from how they were used in equation (0.1.58) for the Riemann integral. (Either way could be used for either integral, however. This is different from integrals with respect to stochastic differentials, where the endpoints matter; see Section 0.2.2.) Form a partition of $[a, b]$, call it $P = (a = x_0 < x_1 < \cdots < x_n = b)$, and let $\Delta g_i = g(x_i) - g(x_{i-1})$. We now consider the sup and inf of f within each interval of the partition and the inf and sup of sums of over all partitions:

$$\inf_P \sum_{i=1}^{n} \sup_{x\in[x_i,x_{i-1}]} f(x)\Delta g_i$$

and

$$\sup_P \sum_{i=1}^{n} \inf_{x\in[x_i,x_{i-1}]} f(x)\Delta g_i.$$

If these are equal, then the Riemann-Stieltjes integral is defined as their common value:

$$\int_a^b f(x)\mathrm{d}g(x) = \inf_P \sum_{i=1}^{n} \sup_{x\in[x_i,x_{i-1}]} f(x)\Delta g_i = \sup_P \sum_{i=1}^{n} \inf_{x\in[x_i,x_{i-1}]} f(x)\Delta g_i. \tag{0.1.61}$$

There is a simple connection with Riemann-Stieltjes integral and the Riemann integral whenever $g'(x)$ exists and is continuous.

Theorem 0.1.27
Suppose that Riemann-Stieltjes integral $\int_a^b f(x)\mathrm{d}g(x)$ exists and suppose the derivative of g, $g'(x)$ exists and is continuous on $[a,b]$; then

$$\int_a^b f(x)\mathrm{d}g(x) = \int_a^b f(x)g'(x)\mathrm{d}x.$$

Proof. Exercise. (*Hint*: use the mean-value theorem together with the respective definitions.) ∎

The existence of the Riemann-Stieltjes integral depends on f.

Theorem 0.1.28
If $f(x)$ is continuous on $[a,b]$, then Riemann-Stieltjes integrable on $[a,b]$.

Proof. Exercise. (*Hint*: just determine an appropriate $g(x)$.) ∎

The Riemann-Stieltjes integral can exist for discontinuous f (under the same conditions as the Riemann integral), but may fail to exist when f and g are discontinuous at the same point.

The Riemann-Stieltjes integral is often useful when $g(x)$ is a step function. We usually define step functions to be continuous from the right. This allows easy development and interpretation of impulse functions and transfer functions.

Theorem 0.1.29
Let $g(x)$ be a step function on $[a,b]$ such that for the partition

$$P = (a = x_0 < x_1 < \cdots < x_n = b),$$

$g(x)$ is constant over each subinterval in the partition. For $i = 1,\ldots,n$, let $g_i = g(x_{i-1})$ (this means $g(x) = g_i$ on $[x_{i-1},x_i[$), and let $g_{n+1} = g(b)$. Let $\Delta g_i = g_{i+1} - g_i$. If $f(x)$ is bounded on $[a,b]$ and is continuous at $x_1,\cdots,x_n$, then

$$\int_a^b f(x)\mathrm{d}g(x) = \sum_{i=1}^{n} \Delta g_i f(x_i). \tag{0.1.62}$$

Proof. Exercise. ∎

In a special case of this theorem, g is the Heaviside function H, and we have

$$\int_a^b f(x)\mathrm{dH}(x) = \int_a^b f(x)\delta(x)\mathrm{d}x = f(0),$$

where $\delta(x)$ is the Dirac delta function.

Improper Integrals

Because of the restriction on the Lebesgue measure of the subintervals in the definitions of the Riemann and Riemann-Stieltjes integrals, if $a = \infty$ or $b = \infty$, the integral is not defined. We define an "improper" Riemann integral, however, as, for example,

$$\int_a^\infty f(x)\mathrm{d}x = \lim_{b\to\infty} \int_a^b f(x)\mathrm{d}x. \tag{0.1.63}$$

Notice that an analogous definition for such a special case is not necessary for a Lebesgue integral.

Adding the improper Riemann integral to the definition of the integral itself yields an instance where the Riemann integral that exists even though the Lebesgue integral does not. Recall (page 729) that the Lebesgue integral

$$\int_0^\infty \sin(x)/x\,\mathrm{d}x \tag{0.1.64}$$

does not exist because both the positive part and the negative of the negative part are ∞.

Whether the integral is interpreted in the Riemann sense or in the Lebesgue sense, we may be interested in

$$\lim_{t\to\infty} \int_0^t \sin(x)/x\,\mathrm{d}x.$$

(In the Riemann sense, we would just write that as $\int_0^\infty \sin(x)/x\,\mathrm{d}\nu(x)$, but it is not standard to use such notation for Lebesgue integrals unless they exist by Definition 0.1.41.) With some effort (see Billingsley (1995), for example, in which Fubini's theorem is used), we have

$$\lim_{t\to\infty} \int_0^t \frac{\sin(x)}{x}\,\mathrm{d}x = \frac{\pi}{2}. \tag{0.1.65}$$

This is the same value as the Riemann improper integral $\int_0^\infty \sin(x)/x\,\mathrm{d}x$, but we do not write it that way when "$\int$" represents the Lebesgue integral.

0.1.7 The Radon-Nikodym Derivative

Given a measure ν on $(\Omega, \mathcal{F})$ and an integrable function f, we have seen that

$$\lambda(A) = \int_A f \mathrm{d}\nu \quad \forall\, A \in \mathcal{F}$$

is also a measure on $(\Omega, \mathcal{F})$ and that $\lambda \ll \nu$. The Radon-Nikodym theorem says that given two such measures, $\lambda \ll \nu$, then a function f exists.

Theorem 0.1.30 (Radon-Nikodym theorem)
Given two measures ν and λ on the same measurable space, $(\Omega, \mathcal{F})$, such that $\lambda \ll \nu$ and ν is σ-finite. Then there exists a unique a.e. nonnegative Borel function f on Ω such that $\lambda(A) = \int_A f \mathrm{d}\nu \; \forall\, A \in \mathcal{F}$.

Proof. A proof is given in Billingsley (1995), page 422. ∎

Uniqueness a.e. means that if also, for some g, $\lambda(A) = \int_A g \mathrm{d}\nu \; \forall\, A \in \mathcal{F}$ then $f = g$ a.e.

Definition 0.1.42 (Radon-Nikodym derivative)
Let ν and λ be σ-finite measures on the same measurable space and $\lambda \ll \nu$. Let f be the function such that

$$\lambda(A) = \int_A f \mathrm{d}\nu \; \forall\, A \in \mathcal{F}.$$

Then f is called the *Radon-Nikodym derivative* of λ with respect to ν, and we write $f = \mathrm{d}\lambda/\mathrm{d}\nu$. ∎

Notice an important property of the derivative: If $\mathrm{d}\lambda/\mathrm{d}\nu > 0$ over A, but $\lambda(A) = 0$, then $\nu(A) = 0$.

With this definition of a derivative, we have the familiar properties for measures λ, λ_1, λ_2, μ, and ν on the same measurable space, $(\Omega, \mathcal{F})$:

1. If $\lambda \ll \nu$, with ν σ-finite, and $f \geq 0$, then

$$\int f \mathrm{d}\lambda = \int f \frac{\mathrm{d}\lambda}{\mathrm{d}\nu} \mathrm{d}\nu. \tag{0.1.66}$$

2. If $\lambda_1 \ll \nu$ and $\lambda_1 + \lambda_2 \ll \nu$, with ν σ-finite, then

$$\frac{\mathrm{d}(\lambda_1 + \lambda_2)}{\mathrm{d}\nu} = \frac{\mathrm{d}\lambda_1}{\mathrm{d}\nu} + \frac{\mathrm{d}\lambda_2}{\mathrm{d}\nu} \quad \text{a.e. } \nu. \tag{0.1.67}$$

3. If $\lambda \ll \mu \ll \nu$, with μ and ν σ-finite, then

$$\frac{\mathrm{d}\lambda}{\mathrm{d}\nu} = \frac{\mathrm{d}\lambda}{\mathrm{d}\mu} \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \quad \text{a.e. } \nu. \tag{0.1.68}$$

   If $\lambda \equiv \nu$, then

$$\frac{\mathrm{d}\lambda}{\mathrm{d}\nu} = \left(\frac{\mathrm{d}\nu}{\mathrm{d}\lambda} \right)^{-1} \quad \text{a.e. } \nu \text{ and } \mu. \tag{0.1.69}$$

4. If $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$ are measurable spaces, λ_1 and ν_1, with $\lambda_1 \ll \nu_1$, are measures on $(\Omega_1, \mathcal{F}_1)$, λ_2 and ν_2, with $\lambda_2 \ll \nu_2$, are measures on $(\Omega_2, \mathcal{F}_2)$, and ν_1 and ν_2 are σ-finite, then for $\omega_1 \in \Omega_1$ and $\omega_2 \in \Omega_2$

$$\frac{\mathrm{d}(\lambda_1 + \lambda_2)}{\mathrm{d}(\nu_1 + \nu_2)}(\omega_1, \omega_2) = \frac{\mathrm{d}\lambda_1}{\mathrm{d}\nu_1}(\omega_1)\frac{\mathrm{d}\lambda_2}{\mathrm{d}\nu_2}(\omega_2) \quad \text{a.e. } \nu_1 \times \nu_2. \tag{0.1.70}$$

The proofs of all of these are exercises.

An absolutely continuous function has a derivative almost everywhere; and if the derivative is 0 a.e., the function is constant.

0.1.8 Function Spaces

Real-valued linear function spaces are sets of real-valued functions over a common domain that are closed with respect to the standard operations of pointwise addition and scalar multiplication of function values.

In addition to the operations of pointwise addition and scalar multiplication that are basic to a function space, there are other interesting operations on functions. One of the most common is function composition, often denoted by "$\circ$". For the functions f and g, the composition $f \circ g$ is defined by

$$f \circ g(x) = f(g(x)). \tag{0.1.71}$$

Notice that the operation is not commutative and that the range of g must be a subset of the domain of f for the composition to exist.

Two useful types of real function spaces are those that contain smooth functions and those that contain integrable functions. We first describe spaces of smooth functions of various degrees, and then in Section 0.1.9 discuss spaces of integrable functions of various types and the kinds of operations that can be defined on those spaces.

Other useful operations on functions are correlation, convolution, inner product, and other transforms. Whether or not a given type of operation can be defined on a function space may depend on the properties of functions, in particular, the integrability of the functions. We therefore defer discussion of these other operations to Section 0.1.9.

Spaces of Smooth Real Functions

Differentiability is a smoothness property.

Definition 0.1.43 ($\mathcal{C}^k$ space)
For an integer $k \geq 0$, a function f such that all derivatives up to the k^{th} derivative exist and are continuous is said to belong to the *class* $\mathcal{C}^k$. ∎

The notation $\mathcal{C}^k$ does not specify the domain of the functions. Generally, without any further notation, for d-variate functions, the domain is taken to

be $\mathbb{R}^d$. A domain D can be specified by the notation $\mathcal{C}^k(D)$. For example, $\mathcal{C}^0([0,1])$ refers to the class of all continuous functions over the unit interval $[0,1]$.

The class $\mathcal{C}^0$ includes all continuous functions.

If $f \in \mathcal{C}^k$ then $f \in \mathcal{C}^j$ for $j \leq k$.

A $\mathcal{C}^k$ class of functions over the same domain is a linear space (exercise).

The term "smooth" is used in connection with the $\mathcal{C}^k$ classes. In a relative sense, for $j < k$, a function in $\mathcal{C}^k$ is smoother than one in $\mathcal{C}^j$ and not in $\mathcal{C}^k$. In an absolute sense, a function is said to be "smooth" if it is in $\mathcal{C}^\infty$.

Analytic Functions

Not all functions in $\mathcal{C}^\infty$ have a convergent Taylor series at any point (see page 656). The special ones that do are said to be *analytic* over the region in which the Taylor series at the point x_0 converges to the value of the function at x_0. (The Taylor series may not converge, and more remarkably, it may not converge to the value of the function.) We sometimes denote the class of analytic functions as $\mathcal{C}^\omega$. Analytic functions are of course smooth, and $\mathcal{C}^\omega \subseteq \mathcal{C}^\infty$.

********** domain *** real line versus complex plane ********** refer to proof of Theorem 1.17) on page 49 as an example of analytic continuation.

The property of being analytic is quite different for real and complex functions. In the case of a complex function of a complex variable $f(z)$, if the first derivative of f exists at all points within a region D, then the derivatives of all orders exist. Furthermore, the Taylor series converges to the function value within the region over which the function is analytic. (These facts can be shown using the Cauchy integral formula; see Churchill, 1960, Chapters 5 and 6, for example.) The definition of an analytic complex function is usually different from that of a real function. An analytic complex function is defined as one whose (first) derivative exists over a region.

0.1.9 $\mathcal{L}^p$ Real Function Spaces

Definition 0.1.44 ($\mathcal{L}^p$ space)
Given the measure space $(\Omega, \mathcal{F}, \nu)$ and the real number $p \geq 1$. The space of all measurable functions f on Ω for which $\int |f|^p d\nu < \infty$ is called the $\mathcal{L}^p(\nu)$ space, or just the $\mathcal{L}^p$ space. ∎

Although the measure ν is needed to define the integral, we often drop the ν in $\mathcal{L}^p(\nu)$. If the integral is taken only over some $D \in \mathcal{F}$, we may denote the space as $\mathcal{L}^p(D)$, and a more complete notation may be $\mathcal{L}^p(\nu, D)$.

An $\mathcal{L}^p$ space is a linear space (exercise).

An important fact about the $\mathcal{L}^p$ spaces is that they are Banach spaces (that is, among other things, they are complete). This fact is called the Riesz-Fischer theorem and is proved in most texts on real analysis.

There are several types of useful operations on functions in a given function space $\mathcal{F}$. Most binary operations require that the domains of the two functions be the same. Function composition, equation (0.1.71), is a very common operation that only requires that the range of one function be in the domain of the other. Many interesting binary operations on functions involve integration of the functions, and so require that the functions be in some $\mathcal{L}^p$ space.

We now describe some of these operations. Each operation is a mapping. Some binary operations map $\mathcal{L}^p \times \mathcal{L}^p$ to $\mathcal{L}^q$ or map $\mathcal{L}^p \times \mathcal{L}^p$ to $\mathbb{R}$, often for $p = 2$. Some useful unary operations map $\mathcal{L}^p$ to $\mathbb{R}$, to $[-1, 1]$, or to $\bar{\mathbb{R}}_+$. The transforms described in Section 0.1.12 map $\mathcal{L}^1$ to $\mathcal{L}^1$.

Convolutions and Covariances and Correlations

The *convolution* of the functions f and g is

$$(f \star g)(t) = \int_D f(x)g(t-x)\,\mathrm{d}x. \tag{0.1.72}$$

The range of integration is usually either $[0, t]$ or $\mathbb{R}$. The convolution is a function; often we write the convolution without the dummy argument: $f \star g$.

The convolution is a measure of the amount of overlap of one function as it is shifted over another function. The convolution can be thought of as a blending of one function with another.

Several properties follow immediately from the definition:

- commutativity:
$$f \star g = g \star f$$
- associativity:
$$f \star (g \star h) = (f \star g) \star h$$
- distribution over addition:
$$f \star (g + h) = (f \star g) + (f \star h)$$
- distribution of scalar multiplication over convolution:
$$a(f \star g) = (af) \star g.$$

Although because the convolution is commutative the two functions are essentially the same in a convolution, the second function (g in the definition above) is sometimes called the *kernel*.

The convolution of the n-vectors u and v is

$$(u \star v)_t = \sum_{1 \le i, t-i \le n} u_i v_{t-i}. \tag{0.1.73}$$

The indices of vectors in applications involving convolutions are often defined to begin at 0 instead of 1, and in that case, the lower limit above would be 0. The limits for the sum are simpler for infinite-dimensional vectors.

For functions f and g that integrate to zero, that is, if

$$\int_D f(x)\,\mathrm{d}x = \int_D g(x)\,\mathrm{d}x = 0,$$

the *covariance of f and g at lag t* is

$$\mathrm{Cov}(f,g)(t) = \int_D f(x)g(t+x)\,\mathrm{d}x. \tag{0.1.74}$$

The argument of the covariance, t, is called the lag. The covariance of a function with itself is called its *autocovariance*. The autocovariance of a function at zero lag, $\mathrm{Cov}(f,f)(0)$, is called its *variance*.

For functions f and g that integrate to zero, the *correlation of f and g at lag t* is

$$\mathrm{Cor}(f,g)(t) = \frac{\int_D f(x)\,g(t+x)\,\mathrm{d}x}{\sqrt{\mathrm{Cov}(f,f)(0)\ \mathrm{Cov}(g,g)(0)}}. \tag{0.1.75}$$

The argument of the correlation, t, is often called the lag, and the correlation of a function with itself is called its autocorrelation.

The correlation between two functions is a measure of their similarity. If f near the point x has similar values to those of g near the point $x+t$, then $\mathrm{Cor}(f,g)(t)$ will be relatively large (close to 1). In this case, if t is positive, then f *leads* g; if t is negative, then f *lags* g. These terms are symmetric, because

$$\mathrm{Cor}(f,g)(-t) = \mathrm{Cor}(g,f)(t).$$

Inner Products of Functions

Inner products over linear spaces are useful operators. As we saw in Section 0.0.4, they can be used to define norms and metrics. An inner product is also sometimes called a dot product.

Definition 0.1.45 (inner product of functions)
The *inner product* of the real functions f and g over the domain D, denoted by $\langle f,g\rangle_D$ or usually just by $\langle f,g\rangle$, is defined as

$$\langle f,g\rangle_D = \int_D f(x)g(x)\,\mathrm{d}x \tag{0.1.76}$$

if the (Lebesgue) integral exists. ∎

Of course, often $D = \mathbb{R}$ or $D = \mathbb{R}^d$, and we just drop the subscript and write $\langle f,g\rangle$. (For complex functions, we define the inner product as $\int_D f(x)\bar{g}(x)\,\mathrm{d}x$, where $\bar{g}$ is the complex conjugate of g. Our primary interest will be in real-valued functions of real arguments.)

To avoid questions about integrability, we generally restrict attention to functions whose dot products with themselves exist; that is, to functions that

are square Lebesgue integrable over the region of interest. These functions are members of the space $\mathcal{L}^2(D)$.

The standard properties, such as linearity and the Cauchy-Schwarz inequality, obviously hold for the inner products of functions.

We sometimes introduce a weight function, $w(x)$, in the definition of the inner product of two functions. For the functions f and g, we denote this either as $\langle f, g\rangle_{(\mu;D)}$ or as $\langle f, g\rangle_{(w;D)}$, where μ is the measure given by $\mathrm{d}\mu = w(x)\mathrm{d}x$. In any event, the inner product of with respect to f and g with respect to a weight function, $w(x)$, or with respect to the measure μ, where $\mathrm{d}\mu = w(x)\mathrm{d}x$ is defined as

$$\langle f, g\rangle_{(\mu;D)} = \int_D f(x)g(x)w(x)\,\mathrm{d}x, \tag{0.1.77}$$

if the integral exists. Often, both the weight and the range are assumed to be fixed, and the simpler notation $\langle f, g\rangle$ is used.

Norms of Functions

The norm of a function f, denoted generically as $\|f\|$, is a mapping into the nonnegative reals that satisfies the properties of the definition of a norm given on page 637. A norm of a function $\|f\|$ is often defined as some nonnegative, strictly increasing function of the inner product of f with itself, $\langle f, f\rangle$. Not all norms are defined in terms of inner products, however.

The property of a norm of an object x that $\|x\| = 0 \Rightarrow x = 0$ is an awkward property for a function to satisfy. For a function, it is much more convenient to say that if its norm is zero, it must be zero almost everywhere. Modifying the definition of a norm in this slight way yields what is often called a pseudonorm.

The most common type of norm or pseudonorm for a real scalar-valued function is the L_p norm. It is defined similarly to the L_p vector norm (page 641).

Definition 0.1.46 (L_p (pseudo)norm of functions)
For $p \geq 1$, the L_p norm or L_p pseudonorm of the function f, denoted as $\|f\|_p$, is defined as

$$\|f\|_p = \left(\int_D |f(x)|^p w(x)\,\mathrm{d}x\right)^{1/p}, \tag{0.1.78}$$

if the integral exists. ∎

The triangle inequality in this case is another version of Minkowski's inequality (0.0.30) (which, as before, we can prove using a version of Hölder's inequality (0.0.31)). For this reason, L_p (pseudo)norm for functions is also called the Minkowski (pseudo)norm or the Hölder (pseudo)norm.

To emphasize the measure of the weighting function, the notation $\|f\|_w$ or $\|f\|_\mu$ is sometimes used. (The ambiguity of the possible subscripts on $\|\cdot\|$

is usually resolved by the context.) For functions over finite domains, $w(x)$ is often taken as $w(x) \equiv 1$. This is a uniform weighting.

The space of functions for which the integrals in (0.1.78) exist is $\mathcal{L}^p(w, D)$.

It is clear that $\|f\|_p$ satisfies the properties that define a norm except for the requirement that $\|f\|_p = 0 \Rightarrow f = 0$. For this latter condition, we must either substitute $f = 0$ a.e. (and perhaps introduce the concept of equivalence classes of functions), or else settle for $\|f\|_p$ being a pseudonorm. See the discussion on pages 638 and 730.

A common L_p function pseudonorm is the L_2 norm, which is often denoted simply by $\|f\|$. This pseudonorm is related to the inner product:

$$\|f\|_2 = \langle f, f\rangle^{1/2}. \tag{0.1.79}$$

The space consisting of the set of functions whose L_2 pseudonorms over $\mathbb{R}$ exist together with the pseudonorm, that is, $\mathcal{L}^2(\mathbb{R})$, is a Hilbert space.

Another common pseudonorm is the limit of the L_p pseudonorm as $p \to \infty$. Just as with countable sets, as the L_∞ norm for vectors in equation (0.0.34), this may be the supremum of the function.

A related pseudonorm is more useful, however, because it is the limit of equation (0.1.78) as $p \to \infty$ (compare equation (0.0.34)). We define

$$\|f\|_\infty = \text{ess sup}\,|f(x)w(x)|, \tag{0.1.80}$$

where ess sup denotes the *essential supremum* of a function, defined for a given measure μ by

$$\text{ess sup}\, g(x) = \inf\{a : \mu(\{x : x \in D, g(x) > a\}) = 0\}.$$

The *essential infimum* of a function for a given measure μ is defined similarly:

$$\text{ess inf}\, g(x) = \sup\{a : \mu(\{x : x \in D, g(x) < a\}) = 0\}.$$

The pseudonorm defined by equation (0.1.80) is called the L_∞ norm, the *Chebyshev norm*, or the *uniform norm*.

Another type of function norm, called the *total variation*, is an L_∞-type of measure of the amount of variability of the function. For a real-valued scalar function f on the interval $[a, b]$, the total variation of f on $[a, b]$ is

$$\mathrm{V}_a^b(f) = \sup_\pi \sum_i |f(x_{i+1}) - f(x_i)|, \tag{0.1.81}$$

where π is a partition of $[a, b]$, $(a = x_0 < x_1 < \cdots < x_n = b)$.

If f is continuously differentiable over $[a, b]$, then

$$\mathrm{V}_a^b(f) = \int_a^b |f'(x)|\mathrm{d}x. \tag{0.1.82}$$

A *normal function* is a function whose pseudonorm is 1. A normal function is also called a normal function a *normalized function*. Although this term can be used with respect to any pseudonorm, it is generally reserved for the L_2 pseudonorm (that is, the pseudonorm arising from the inner product). A function whose integral (over a relevant range, usually $\mathbb{R}$) is 1 is also called a normal function. (Although this latter definition is similar to the standard one, the latter is broader because it may include functions that are not square-integrable.) Density and weight functions are often normalized (that is, scaled to be normal).

Metrics in Function Spaces

Statistical properties such as bias and consistency are defined in terms of the difference of the estimator and what is being estimated. For an estimator of a function, first we must consider some ways of measuring this difference. These are general measures for functions and are not dependent on the distribution of a random variable. How well one function approximates another function is usually measured by a norm of the difference in the functions over the relevant range.

The most common measure of the difference between two functions, $g(\cdot)$ and $f(\cdot)$, is a metric, $\rho(g, f)$. (See Section 0.0.2 on page 623.) In normed linear spaces, the most useful metric for two elements is the norm of the difference of the two elements (see pages 623, and 638):

$$\rho(g, f) = \|g - f\|,$$

if that norm exists and is finite.

The metric corresponding to the L_p norm is

$$\rho_p(g, f) = \|g - f\|_p.$$

As we mentioned above, the L_p "norm" is not a true norm; hence, the metric induced is only a pseudometric.

When one function is an estimate or approximation of the other function, we may call this difference the "error".

If g is used to approximate f, then $\rho_\infty(g, f)$, that is, $\|g - f\|_\infty$, is likely to be the norm of interest. This is the norm most often used in numerical analysis when the objective is interpolation or quadrature. This norm is also often used in comparing CDFs. If P and Q are CDFs, $\|P - Q\|_\infty$ is called the Kolmogorov distance. For CDFs, this metric always exists and is finite.

In applications with noisy data, or when g may be very different from f, $\|g - f\|_2$ may be the more appropriate (pseudo)norm. This is the norm most often used in estimating probability density functions.

For comparing two functions g and f we can use a metric based on a norm of their difference, $\|g - f\|$. We often prefer to use a pseudometric, which is

the same as a metric except that $\rho(g, f) = 0$ if and only if $g = f$ a.e. (We usually just use this interpretation and call it a metric, however.)

Definition 0.1.47 (Hellinger distance)
Let P be absolutely continuous with respect to Q and $p = \mathrm{d}P$ and $q = \mathrm{d}q$. Then

$$\left(\int_{\mathbb{R}} \left(q^{1/r}(x) - p^{1/r}(x) \right)^r \, \mathrm{d}x \right)^{1/r} \tag{0.1.83}$$

is called the *Hellinger distance* between p and q. ∎

The most common instance has $r = 2$, and in this case the Hellinger distance is also called the *Matusita distance*.

Other Distances between Functions

Sometimes the difference in two functions is defined asymmetrically. A general class of divergence measures for comparing CDFs was introduced independently by Ali and Silvey (1966) and Csiszár (1967) (see also Pardo (2005)). The measure is based on a convex function ϕ of a term similar to the "odds".

Definition 0.1.48 (ϕ-divergence)
Let P be absolutely continuous with respect to Q and ϕ is a convex function,

$$d(P, Q) = \int_{\mathbb{R}} \phi\left(\frac{\mathrm{d}P}{\mathrm{d}Q} \right) \mathrm{d}Q, \tag{0.1.84}$$

if it exists, is called the *ϕ-divergence* from Q to P. ∎

The ϕ-divergence is also called the f-divergence.

The ϕ-divergence is in general not a metric because it is not symmetric. One function is taken as the base from which the other function is measured. The expression often has a more familiar form if both P and Q are dominated by Lebesgue measure and we write $p = \mathrm{d}P$ and $q = \mathrm{d}Q$. The Hellinger distance given in equation (0.1.83) is a ϕ-divergence that is a metric. The Matusita distance is the square root of a ϕ-divergence with $\phi(t) = (\sqrt{t} - 1)^2$.

Another specific instance of ϕ-divergence is the Kullback-Leibler measure.

Definition 0.1.49 (Kullback-Leibler measure)
Let P be absolutely continuous with respect to Q and $p = \mathrm{d}P$ and $q = \mathrm{d}q$. Then

$$\int_{\mathbb{R}} p(x) \log\left(\frac{p(x)}{q(x)} \right) \mathrm{d}x. \tag{0.1.85}$$

is called the *Kullback-Leibler measure* of the difference of p and q. ∎

The Kullback-Leibler measure is not a metric.

Various forms of ϕ-divergence are used in goodness-of-fit analyses. The Pearson chi-squared discrepancy measure, for example, has $\phi(t) = (t-1)^2$:

$$\int_{\mathbb{R}} \frac{(q(x) - p(x))^2}{q(x)} \mathrm{d}x. \tag{0.1.86}$$

See the discussion beginning on page 598 for other applications in which two probability distributions are compared.

Convergence of Functions

We have defined almost everywhere convergence of measurable functions in general measure spaces (see page 726). We will now define two additional types of convergence of measurable functions in normed linear measure spaces. Here we will restrict our attention to real-valued functions over real domains, but the ideas are more general.

The first is convergence in L_p.

Definition 0.1.50 (convergence in L_p)
Let $f_1, f_2, \ldots$ be a sequence of Borel functions in $\mathcal{L}^p$ and let f be another Borel function in $\mathcal{L}^p$. We say that $\{f_n\}$ *converges in* L_p to f if

$$\|f_n - f\|_p \to 0.$$

We write

$$f_n \xrightarrow{\mathrm{L}_p} f.$$

∎

Theorem 0.1.31
Suppose $f, f_1, f_2, \ldots \in \mathcal{L}^p(\nu, D)$ *and* $\nu(D) < \infty$ *Then*

$$f_n \xrightarrow{\mathrm{L}_p} f \;\Rightarrow\; f_n \xrightarrow{\mathrm{L}_r} f \quad \text{for } r \leq p.$$

Proof. Exercise. ∎

Convergence in L_p is different from convergence a.e.; neither implies the other.

give examples:

The second is convergence in measure.

Definition 0.1.51 (convergence in measure)
Let $f_1, f_2, \ldots$ be a sequence of Borel functions on the measure space $(\Omega, \mathcal{F}, \nu)$ and let f be another Borel function on $(\Omega, \mathcal{F}, \nu)$. We say that $\{f_n\}$ *converges in measure* to f if *** We write

$$f_n \xrightarrow{\nu} f.$$

∎

Convergence in measure is weaker than both L_p convergence and a.e. convergence; a.e. implies it.

prove

Basis Sets in Function Spaces

If each function in a linear space can be expressed as a linear combination of the functions in a set G, then G is said to be a *generating set*, a *spanning set*, or a *basis set* for the linear space. (These three terms are synonymous.) The basis sets for finite-dimensional vector spaces are finite; for most function spaces of interest, the basis sets are infinite.

A set of functions $\{q_k\}$ is *orthogonal over the domain D with respect to the nonnegative weight function $w(x)$* if the inner product with respect to $w(x)$ of q_k and q_l, $\langle q_k, q_l\rangle$, is 0 if $k \neq l$; that is,

$$\int_D q_k(x)\bar{q}_l(x)w(x)\mathrm{d}x = 0 \quad k \neq l. \tag{0.1.87}$$

If, in addition,

$$\int_D q_k(x)\bar{q}_k(x)w(x)\mathrm{d}x = 1,$$

the functions are called *orthonormal*.

In the following, we will be concerned with real functions of real arguments, so we can take $\bar{q}_k(x) = q_k(x)$.

The weight function can also be incorporated into the individual functions to form a different set,

$$\tilde{q}_k(x) = q_k(x)w^{1/2}(x).$$

This set of functions also spans the same function space and is orthogonal over D with respect to a constant weight function.

Basis sets consisting of orthonormal functions are generally easier to work with and can be formed from any basis set. Given two nonnull, linearly independent functions, q_1 and q_2, two orthonormal vectors, $\tilde{q}_1$ and $\tilde{q}_2$, that span the same space can be formed as

$$\begin{aligned}
\tilde{q}_1(\cdot) &= \frac{1}{\|q_1\|}\, q_1(\cdot), \\
\tilde{q}_2(\cdot) &= \frac{1}{\|q_2 - \langle \tilde{q}_1, q_2\rangle \tilde{q}_1\|}\, \big(q_2(\cdot) - \langle \tilde{q}_1, q_2\rangle \tilde{q}_1(\cdot)\big).
\end{aligned} \tag{0.1.88}$$

These are the Gram-Schmidt function transformations. They can easily be extended to more than two functions to form a set of orthonormal functions from any set of linearly independent functions.

Series Expansions in Basis Functions

Our objective is to represent a function of interest, $f(x)$, over some domain $D \subseteq \mathbb{R}$, as a linear combination of "simpler" functions, $q_0(x), q_1(x), \ldots$:

$$f(x) = \sum_{k=0}^{\infty} c_k q_k(x). \tag{0.1.89}$$

There are various ways of constructing the q_k functions. If they are developed through a linear operator on a function space, they are called *eigenfunctions*, and the corresponding c_k are called eigenvalues.

We choose a set $\{q_k\}$ that spans some class of functions over the given domain D. A set of orthogonal basis functions is often the best choice because they have nice properties that facilitate computations and a large body of theory about their properties is available.

If the function to be estimated, $f(x)$, is continuous and integrable over a domain D, the orthonormality property allows us to determine the coefficients c_k in the expansion (0.1.89):

$$c_k = \langle f, q_k \rangle. \tag{0.1.90}$$

The coefficients $\{c_k\}$ are called the *Fourier coefficients* of f with respect to the orthonormal functions $\{q_k\}$.

In applications, we *approximate* the function using a truncated orthogonal series. The error due to finite truncation at j terms of the infinite series is the residual function $f - \sum_{k=1}^{j} c_k f_k$. The *mean squared error* over the domain D is the scaled, squared L_2 norm of the residual,

$$\frac{1}{d} \left\| f - \sum_{k=0}^{j} c_k q_k \right\|^2, \tag{0.1.91}$$

where d is some measure of the domain D. (If the domain is the interval $[a, b]$, for example, one choice is $d = b - a$.)

A very important property of Fourier coefficients is that they yield the minimum mean squared error for a given set of basis functions $\{q_i\}$; that is, for any other constants, $\{a_i\}$, and any j,

$$\left\| f - \sum_{k=0}^{j} c_k q_k \right\|^2 \leq \left\| f - \sum_{k=0}^{j} a_k q_k \right\|^2. \tag{0.1.92}$$

In applications of statistical data analysis, after forming the approximation, we then *estimate* the coefficients from equation (0.1.90) by identifying an appropriate probability density that is a factor of the function of interest, f. (Note again the difference in "approximation" and "estimation".) Expected values can be estimated using observed or simulated values of the random variable and the approximation of the probability density function.

The basis functions are generally chosen to be easy to use in computations. Common examples include the Fourier trigonometric functions $\sin(kt)$ and $\cos(kt)$ for $k = 1, 2, \ldots$, orthogonal polynomials such as Legendre, Hermite, and so on, splines, and wavelets.

Orthogonal Polynomials

The most useful type of basis function depends on the nature of the function being estimated. The orthogonal polynomials are useful for a very wide range of functions. Orthogonal polynomials of real variables are their own complex conjugates. It is clear that for the k^{th} polynomial in the orthogonal sequence, we can choose an a_k that does not involve x, such that

$$q_k(x) - a_k x q_{k-1}(x)$$

is a polynomial of degree $k-1$.

Because any polynomial of degree $k-1$ can be represented by a linear combination of the first k members of any sequence of orthogonal polynomials, we can write

$$q_k(x) - a_k x q_{k-1}(x) = \sum_{i=0}^{k-1} c_i q_i(x).$$

Because of orthogonality, all c_i for $i < k-2$ must be 0. Therefore, collecting terms, we have, for some constants a_k, b_k, and c_k, the three-term recursion that applies to any sequence of orthogonal polynomials:

$$q_k(x) = (a_k x + b_k) q_{k-1}(x) - c_k q_{k-2}(x), \quad \text{for } k = 2, 3, \ldots. \tag{0.1.93}$$

This recursion formula is often used in computing orthogonal polynomials. The coefficients in this recursion formula depend on the specific sequence of orthogonal polynomials, of course.

This three-term recursion formula can also be used to develop a formula for the sum of products of orthogonal polynomials $q_i(x)$ and $q_i(y)$:

$$\sum_{i=0}^{k} q_i(x) q_i(y) = \frac{1}{a_{k+1}} \frac{q_{k+1}(x) q_k(y) - q_k(x) q_{k+1}(y)}{x - y}. \tag{0.1.94}$$

This expression, which is called the Christoffel-Darboux formula, is useful in evaluating the product of arbitrary functions that have been approximated by finite series of orthogonal polynomials.

There are several widely used complete systems of univariate orthogonal polynomials. The different systems are characterized by the one-dimensional intervals over which they are defined and by their weight functions. The Legendre, Chebyshev, and Jacobi polynomials are defined over $[-1, 1]$ and hence can be scaled into any finite interval $[a, b]$. The weight function of the Jacobi polynomials is more general, so a finite sequence of them may fit a given function better, but the Legendre and Chebyshev polynomials are simpler and so are often used. The Laguerre polynomials are defined over the half line $[0, \infty[$ and hence can be scaled into any half-finite interval $[a, \infty[$. The Hermite polynomials are defined over the reals, $]-\infty, \infty[$.

Any of these systems of polynomials can be developed easily by beginning with the basis set $1, x, x^2, \ldots$ and orthogonalizing them by use of the Gram-Schmidt equations (0.1.88).

Table 0.2 summarizes the ranges and weight functions for these standard orthogonal polynomials.

Table 0.2. Orthogonal Polynomials

Polynomial Series	Range	Weight Function
Legendre	$[-1,\ 1]$	1 (uniform)
Chebyshev	$[-1,\ 1]$	$(1-x^2)^{1/2}$ (symmetric beta)
Jacobi	$[-1,\ 1]$	$(1-x)^{\alpha}(1+x)^{\beta}$ (beta)
Laguerre	$[0,\ \infty[$	$x^{\alpha-1}\mathrm{e}^{-x}$ (gamma)
Hermite	$]-\infty,\ \infty[$	$\mathrm{e}^{-x^2/2}$ (normal)

The *Legendre polynomials* have a constant weight function and are defined over the interval $[-1, 1]$. Using the Gram-Schmidt transformations on $1, x, x^2, \ldots$, we have

$$\begin{aligned} \widetilde{P}_0(t) &= 1/\sqrt{\textstyle\int_{-1}^{1} 1^2 \mathrm{d}x} &&= 1/\sqrt{2}, \\ \widetilde{P}_1(t) &= (t-0)/\sqrt{\textstyle\int_{-1}^{1} x^2 \mathrm{d}x} &&= \sqrt{3/2}t, \\ &\vdots \end{aligned} \tag{0.1.95}$$

Orthogonal polynomials are often expressed in the simpler, unnormalized form. The first few unnormalized Legendre polynomials are

$$\begin{aligned} P_0(t) &= 1 & P_1(t) &= t \\ P_2(t) &= (3t^2-1)/2 & P_3(t) &= (5t^3-3t)/2 \\ P_4(t) &= (35t^4-30t^2+3)/8 & P_5(t) &= (63t^5-70t^3+15t)/8 \end{aligned} \tag{0.1.96}$$

The normalizing constant that relates the k^{th} unnormalized Legendre polynomial to the normalized form is determined by noting

$$\int_{-1}^{1} (P_k(t))^2 \mathrm{d}t = \frac{2}{2k+1}.$$

The recurrence formula for the Legendre polynomials is

$$P_k(t) = \frac{2k-1}{k} t P_{k-1}(t) - \frac{k-1}{k} P_{k-2}(t). \tag{0.1.97}$$

The *Hermite polynomials* are orthogonal with respect to a Gaussian, or standard normal, weight function. We can form the normalized Hermite polynomials using the Gram-Schmidt transformations on $1, x, x^2, \ldots$, with a weight function of $\mathrm{e}^{x/2}$ similarly to what is done in equations (0.1.95).

The first few unnormalized Hermite polynomials are

$$\begin{array}{ll} H_0^{\mathrm{e}}(t) = 1 & H_1^{\mathrm{e}}(t) = t \\ H_2^{\mathrm{e}}(t) = t^2 - 1 & H_3^{\mathrm{e}}(t) = t^3 - 3t \\ H_4^{\mathrm{e}}(t) = t^4 - 6t^2 + 3 & H_5^{\mathrm{e}}(t) = t^5 - 10t^3 + 15t \end{array} \tag{0.1.98}$$

These are not the standard Hermite polynomials, but they are the ones most commonly used by statisticians because the weight function is proportional to the normal density.

The recurrence formula for the Hermite polynomials is

$$H_k^{\mathrm{e}}(t) = tH_{k-1}^{\mathrm{e}}(t) - (k-1)H_{k-2}^{\mathrm{e}}(t). \tag{0.1.99}$$

These Hermite polynomials are useful in probability and statistics. The Gram-Charlier series and the Edgeworth series for asymptotic approximations are based on these polynomials. See Section 1.2, beginning on page 65.

Multivariate Orthogonal Polynomials

Multivariate orthogonal polynomials can be formed easily as tensor products of univariate orthogonal polynomials. The tensor product of the functions $f(x)$ over D_x and $g(y)$ over D_y is a function of the arguments x and y over $D_x \times D_y$:

$$h(x, y) = f(x)g(y).$$

If $\{q_{1,k}(x_1)\}$ and $\{q_{2,l}(x_2)\}$ are sequences of univariate orthogonal polynomials, a sequence of bivariate orthogonal polynomials can be formed as

$$q_{kl}(x_1, x_2) = q_{1,k}(x_1)q_{2,l}(x_2). \tag{0.1.100}$$

These polynomials are orthogonal in the same sense as in equation (0.1.87), where the integration is over the two-dimensional domain. Similarly as in equation (0.1.89), a bivariate function can be expressed as

$$f(x_1, x_2) = \sum_{k=0}^{\infty} \sum_{l=0}^{\infty} c_{kl} q_{kl}(x_1, x_2), \tag{0.1.101}$$

with the coefficients being determined by integrating over both dimensions.

Although obviously such product polynomials, or radial polynomials, would emphasize features along coordinate axes, they can nevertheless be useful for representing general multivariate functions. Often, it is useful to apply a rotation of the coordinate axes.

The weight functions, such as those for the Jacobi polynomials, that have various shapes controlled by parameters can also often be used in a mixture model of the function of interest. The weight function for the Hermite polynomials can be generalized by a linear transformation (resulting in a normal weight with mean μ and variance σ^2), and the function of interest may be represented as a mixture of general normals.

0.1.10 Distribution Function Spaces

In probability and statistics, one of the most important kinds of function is a cumulative distribution function, or CDF, defined on page 14 both in terms of a probability distribution and in terms of four characterizing properties.

A set of CDFs cannot constitute a linear space, because of the restrictions on the functions. Instead, we will define a *distribution function space* that has similar properties. If $\mathcal{P}$ is a set of CDFs such that for any $w \in [0,1]$ and $P_1, P_2 \in \mathcal{P}$, $(1-w)P_1 + wP_2 \in \mathcal{P}$, then $\mathcal{P}$ is a distribution function space.

The CDFs of the ϵ-mixture distributions defined on page 194 is a simple example of a distribution function space. In that space, one of the CDFs is degenerate.

Important distribution function spaces are those consisting of CDFs P such that for given $p \geq 1$

$$\int \|t\|^p \mathrm{d}P < \infty. \tag{0.1.102}$$

Such a distribution function space is denoted by $\mathcal{P}^p$. (Constrast this with the $\mathcal{L}^p$ space.) It is clear that $\mathcal{P}^{p_1} \subseteq \mathcal{P}^{p_2}$ if $p_1 \geq p_2$.

Spaces of distribution functions are related to divisibility of the distributions. They are useful in robustness studies. Most of the interesting families of probability distributions as discussed in Chapter 2 do not generate distribution function spaces.

0.1.11 Transformation Groups

On page 630 we have an example of a group on a set of bijections. Such transformation groups are important in statistics and are useful in establishing desirable properties of statistical procedures.

Example 0.1.12 (Continuation of Example 0.0.4) Group of linear transformations
A common instance of the group $\mathcal{G}$ of bijections is formed by functions of the form

$$g(x) = bx - c, \quad x, b, c \in \mathbb{R},\ b \neq 0.$$

For given g, we see that $g^{-1}(x) = (x+c)/b \in \mathcal{G}$. ∎

Invariant Functions

Definition 0.1.52 (Invariant function)
Let $\mathcal{G}$ be a transformation group with domain X. A function f with domain X is said to be *invariant* under the transformation group $\mathcal{G}$ if for all $x \in X$ and $g \in \mathcal{G}$,

$$f(g(x)) = f(x). \tag{0.1.103}$$

∎

We also use the phrases "invariant over ..." and "invariant with respect to ..." to denote this kind of invariance.

Example 0.1.13 Invariant function
The function

$$f(x) = \max(d - x^2)$$

is invariant over the group $\mathcal{G} = \{g \ : g(x) = bx - c, \quad x, b, c \in \mathbb{R},\ b \neq 0\}$ and function composition. ∎

A transformation group $\mathcal{G}$ may define an equivalence relation (identity, symmetry, and transitivity) for elements in its domain, X. If $x_1, x_2 \in X$ and there exists a g in $\mathcal{G}$ such that $g(x_1) = x_2$, then we say x_1 and x_2 are equivalent under $\mathcal{G}$, and we write

$$x_1 \equiv x_2 \bmod \mathcal{G}. \tag{0.1.104}$$

Sets of equivalent points are called *orbits* of $\mathcal{G}$. (In other contexts such sets are called "residue classes".) It is clear that a function that is invariant under the transformation group $\mathcal{G}$ must be constant over the orbits of G. A transformation group $\mathcal{G}$ is said to be *transitive* over the set X if for any $x_1, x_2 \in X$, there exists a g in $\mathcal{G}$ such that $g(x_1) = x_2$. (This terminology is not standard. Also note that the equivalence relation between elements is always a transitive relation.) In this case the whole domain is a single orbit. The group in Example 0.1.12 is transitive over $\mathbb{R}$.

Example 0.1.14 Orbits
Consider the group $\mathcal{G}_1 = \{g_1(x) = 1-x, g_e(x) = x \ : \quad x \in [0,1]\}$ and function composition. The group $\mathcal{G}_1$ is not transitive, and the orbits of $\mathcal{G}_1$ are the pairs $(x, 1-x)$. ∎

Definition 0.1.53 (Maximal invariant function)
An invariant function m over $\mathcal{G}$ is called *maximal invariant* over $\mathcal{G}$ if

$$m(x_1) = m(x_2) \quad \Rightarrow \quad \exists\, g \in \mathcal{G} \ni g(x_1) = x_2. \tag{0.1.105}$$

∎

Maximal invariance can be used to characterize invariance. If m is maximal invariant under $\mathcal{G}$, then the function f is invariant under $\mathcal{G}$ if and only if it depends on x only through m; that is, if and only if there exists a function h such that for all x, $f(x) = h(m(x))$.

Any invariant function with respect to a transitive group is maximal invariant.

Equivariant Functions

Definition 0.1.54 (Equivariant function)
A function f is said to be *equivariant* under the transformation group $\mathcal{G}$ with domain X if for all $x \in X$ and $g \in \mathcal{G}$,

$$f(g(x)) = g(f(x)). \tag{0.1.106}$$

∎

We also use the phrases "equivariant over ..." and "equivariant with respect to ..." to denote this kind of equivariance.

0.1.12 Transforms

Many operations on functions can be facilitated by first forming an inner product with the given functions and another specific function that has an additional argument. The inner product with the function having an additional argument, being itself a function, is a *transform* of a given function. Because of the linearity of inner products, these are *linear transforms*. Linear transforms arising from inner products (that is, from integrals) include the familiar Fourier, Laplace, and wavelet integral transforms.

An integral linear transform of the function f is an operator $\mathcal{T}$ of the general form

$$\mathcal{T}f(s) = \int_D \psi(s,x)f(x)\,\mathrm{d}x, \tag{0.1.107}$$

where the integral exists. We will denote a transform of the function f by the operator $\mathcal{T}$ as $f^{\mathcal{T}}$, that is,

$$f^{\mathcal{T}} = \mathcal{T}f.$$

The dummy arguments of the pair of functions f and $f^{\mathcal{T}}$ may range over different domains, which may correspond to different physical entities, such as time and frequency, for example.

The notation for functions and their transforms requires a word of clarification. All three of the symbols f, $\mathcal{T}f$, and $f^{\mathcal{T}}$ represent functions. The corresponding notation in which the dummy arguments appear are the symbols $f(x)$, $\mathcal{T}f(s)$, and $f^{\mathcal{T}}(s)$. We may also write both dummy arguments, as

in $\mathcal{T}(f(x))(s)$, in which x is the argument of the function f to which the transform is being applied, and s is the argument of the transform, the function $\mathcal{T}f$.

The linearity of the transform in equation (0.1.107) is clear:

$$\mathcal{T}(af + g) = a\mathcal{T}f + \mathcal{T}g, \tag{0.1.108}$$

where a is a constant, f and g are functions, and the transform is defined over an appropriate domain. This relation is why it is a linear transform, and of course is a property of any inner product.

There are several useful transforms that correspond to specific functions $\psi(s, x)$ and domains D in equation (0.1.107). The question of the existence of the integral in equation (0.1.107) is of course important, and the choice of $\psi(s, x)$ can determine the class of functions for which the transform is defined. Often $\psi(s, x)$ is chosen so that the integral exists for and $f \in \mathcal{L}^1$.

In the *Fourier transform*, $\psi(s, x) = \mathrm{e}^{2\pi \mathrm{i} s x}$, and the range of integration is the real line:

$$\mathcal{F}f(s) = \int_{-\infty}^{\infty} \mathrm{e}^{2\pi \mathrm{i} s x} f(x)\,\mathrm{d}x.$$

In this expression, i is the imaginary unit, $\sqrt{-1}$. We also write the Fourier transform of the function f as $f^{\mathcal{F}}(s)$.

A linear transform with $\psi(s, x) \propto (\mathrm{e}^{sx})^c$ for some c, such as the Fourier transform, the Laplace transform, and the characteristic function, satisfies the "change of scale property":

$$\mathcal{T}(f(ax))(s) = \frac{1}{|a|}\mathcal{T}(f(x))\left(\frac{s}{a}\right), \tag{0.1.109}$$

where a is a constant. This is easily shown by making a change of variables in the definition (0.1.107). This change of variables is sometimes referred to as "time scaling", because the argument of f often corresponds to a measure of time. A similar scaling applies to the argument of the transform $f^{\mathcal{T}}$, which is sometimes called "frequency scaling".

Transforms in which $\psi(s, x) \propto (\mathrm{e}^{sx})^c$ also have two useful translation properties:

- for a shift in the argument of f,

$$\mathcal{T}(f(x - x_0))(s) = \psi(s, x_0)\mathcal{T}(f(x))(s)\,: \tag{0.1.110}$$

- for a shift in the argument of the transform $\mathcal{T}f$,

$$\mathcal{T}(f(x))(s - s_0) = \mathcal{T}(\psi(-s_0, x)f(x))(s). \tag{0.1.111}$$

These scaling and translation properties are major reasons for the usefulness of the Fourier and Laplace transforms and of the characteristic function in probability theory.

Linear transforms apply to multivariate functions as well as to univariate functions. In the definition of linear transforms (0.1.107), both s and x may be vectors. In most cases s and x are vectors of the same order, and specific transforms have simple extensions. In the characteristic function of multivariate random variable, for example,

$$\psi(s, x) = \mathrm{e}^{\mathrm{i}\langle s,x\rangle}.$$

Fourier Transforms

The Fourier transform of a function $f(x)$ is the function

$$\mathcal{F}f(s) = \int_{-\infty}^{\infty} \mathrm{e}^{2\pi \mathrm{i} s x} f(x)\,\mathrm{d}x, \tag{0.1.112}$$

if the integral exists.

The inverse Fourier transform is

$$f(x) = \int_{-\infty}^{\infty} \mathrm{e}^{-2\pi \mathrm{i} s x} \mathcal{F}f(s)\,\mathrm{d}s. \tag{0.1.113}$$

Instead of $\mathrm{e}^{2\pi \mathrm{i} s x}$ as in equation (0.1.112), the Fourier transform is often defined with the function $\mathrm{e}^{\mathrm{i}\omega x}$, in which ω is called the "angular frequency".

Fourier transforms are linear transforms, and thus enjoy the linearity property (0.1.108). Fourier transforms are inner products with a function of the form $(\mathrm{e}^{sx})^c$, and thus enjoy the change of scale property (0.1.109), and the translation properties (0.1.110) and (0.1.111). Fourier transforms have additional useful properties that derive from the identity

$$\exp(\mathrm{i}\omega s) = \cos(\omega s) + \mathrm{i}\sin(\omega s),$$

in which the real component is an even function and the imaginary component is an odd function. Because of this, we immediately have the following:

- if $f(x)$ is even, then the Fourier transform is even
$$\mathcal{F}f(-s) = \mathcal{F}f(s)$$
- if $f(x)$ is odd, then the Fourier transform is odd
$$\mathcal{F}f(-s) = -\mathcal{F}f(s)$$
- if $f(x)$ is real, then
$$\mathcal{F}f(-s) = \overline{\mathcal{F}f(s)},$$
where the overbar represents the complex conjugate.

Fourier transforms are useful in working with convolutions and correlations because of the following relationships, which follow immediately from the definition of convolutions (0.1.72) and of correlations (0.1.74):

$$\mathcal{F}(f \star g)(s) = \mathcal{F}f(s)\mathcal{F}g(s). \tag{0.1.114}$$

$$\mathcal{F}(\mathrm{Cor}(f,g))(s) = \mathcal{F}f(s)\overline{\mathcal{F}g(s)}. \tag{0.1.115}$$

$$\mathcal{F}(\mathrm{Cor}(f,f))(s) = |\mathcal{F}f(s)|^2. \tag{0.1.116}$$

Equation (0.1.114) is sometimes called the "convolution theorem". Some authors take this as the definition of the convolution of two functions. Equation (0.1.115) is sometimes called the "correlation theorem", and equation (0.1.116), for the autocorrelation is sometimes called the "Wiener-Khinchin theorem",

These relationships are among the reasons that Fourier transforms are so useful in communications engineering. For a signal with amplitude $h(t)$, the *total power* is the integral

$$\int_{-\infty}^{\infty} |h(t)|^2 \mathrm{d}t.$$

From the relations above, we have Parseval's theorem, for the total power:

$$\int_{-\infty}^{\infty} |h(t)|^2 \mathrm{d}t = \int_{-\infty}^{\infty} |\mathcal{F}h(s)|^2 \mathrm{d}s. \tag{0.1.117}$$

0.1.13 Functionals

Functionals are functions whose arguments are functions. The value of a functional may be any kind of object, a real number or another function, for example. The domain of a functional is a set of functions.

If $\mathcal{F}$ is a linear space of functions, that is, if $\mathcal{F}$ is such that $f \in \mathcal{F}$ and $g \in \mathcal{F}$ implies $(af + g) \in \mathcal{F}$ for any real a, then the functional Υ defined on $\mathcal{F}$ is said to be *linear* if $\Upsilon(af + g) = a\Upsilon(f) + \Upsilon(g)$.

A similar expression defines linearity of a functional over a distribution function space $\mathcal{P}$: Υ defined on $\mathcal{P}$ is linear if $\Upsilon((1-w)P_1 + wP_2) = (1-w)\Upsilon(P_1) + w\Upsilon(P_2)$ for $w \in [0,1]$ and $P_1, P_2 \in \mathcal{P}$.

Functionals of CDFs have important uses in statistics as measures of the differences between two distributions or to define distributional measures of interest. A functional applied to a ECDF is a plug-in estimator of the distributional measure defined by the same functional applied to the corresponding CDF.

Derivatives of Functionals

For the case in which the arguments are functions, the cardinality of the possible perturbations is greater than that of the continuum. We can be precise in discussions of continuity and differentiability of a functional Υ at a point (function) F in a domain $\mathcal{F}$ by defining another set $\mathcal{D}$ consisting of difference functions over $\mathcal{F}$; that is the set the functions $D = F_1 - F_2$ for $F_1, F_2 \in \mathcal{F}$.

The concept of differentiability for functionals is necessarily more complicated than for functions over real domains. For a functional Υ over the domain $\mathcal{F}$, we define three levels of differentiability at the function $F \in \mathcal{F}$. All definitions are in terms of a domain $\mathcal{D}$ of difference functions over $\mathcal{F}$, and a linear functional Λ_F defined over $\mathcal{D}$ in a neighborhood of F. The first type of derivative is very general. The other two types depend on a metric ρ on $\mathcal{F} \times \mathcal{F}$ induced by a norm $\|\cdot\|$ on $\mathcal{F}$.

Definition 0.1.55 (Gâteaux differentiable)
Υ is *Gâteaux differentiable at* F iff there exists a linear functional $\Lambda_F(D)$ over $\mathcal{D}$ such that for $t \in \mathbb{R}$ for which $F + tD \in \mathcal{F}$,

$$\lim_{t \to 0} \left(\frac{\Upsilon(F + tD) - \Upsilon(F)}{t} - \Lambda_F(D) \right) = 0. \tag{0.1.118}$$

∎

In this case, the linear functional Λ_F is called the *Gâteaux differential* of Υ at F in the direction of $F + D$.

Definition 0.1.56 (ρ-Hadamard differentiable)
For a metric ρ induced by a norm, Υ is ρ-Hadamard differentiable at F iff there exists a linear functional $\Lambda_F(D)$ over $\mathcal{D}$ such that for any sequence $t_j \to 0 \in \mathbb{R}$ and sequence $D_j \in \mathcal{D}$ such that $\rho(D_j, D) \to 0$ and $F + t_j D_j \in \mathcal{F}$,

$$\lim_{j \to \infty} \left(\frac{\Upsilon(F + t_j D_j) - \Upsilon(F)}{t_j} - \Lambda_F(D_j) \right) = 0. \tag{0.1.119}$$

∎

In this case, the linear functional Λ_F is called the *ρ-Hadamard differential* of Υ at F.

Definition 0.1.57 (ρ-Fréchet differentiable)
Υ is ρ-Fréchet differentiable at F iff there exists a linear functional $\Lambda(D)$ over $\mathcal{D}$ such that for any sequence $F_j \in \mathcal{F}$ for which $\rho(F_j, F) \to 0$,

$$\lim_{j \to \infty} \left(\frac{\Upsilon(F_j) - \Upsilon(F) - \Lambda_F(F_j - F)}{\rho(F_j, F)} \right) = 0. \tag{0.1.120}$$

∎

In this case, the linear functional Λ_F is called the *ρ-Fréchet differential* of Υ at F.

Derivative Expansions of Functionals

****

Notes and References for Section 0.1

After the introductory material in Section 0.0, in Section 0.1 I try to cover the important aspects of real analysis (which means "measure theory") for statistical mathematics.

Measure theory is the most important element of analysis for probability theory and mathematical statistics. In measure theory, we are concerned with collections of subsets, and we identify particular systems of collections of subsets. These systems are called "rings" (Definition 0.1.2) or "fields" (Definition 0.1.3). (The reader should also be aware that these terms are often used differently in algebra. The term "ring" also applies to a mathematical structure consisting of a set and two operations on the set satisfying certain properties. The prototypic ring is the set of integers with ordinary addition and multiplication. The term "field" as in Definition 0.0.3 also applies to a mathematical structure consisting of a set and two operations on the set satisfying certain properties. The prototypic field is the set of real numbers with ordinary addition and multiplication.)

"Littlewood's three principles of real analysis" are heuristics that state that if sets, functions, or series have certain properties, then stronger properties "almost" hold. The third principle, which is illustrated by the results of the Severini-Egorov theorem and the monotone convergence theorem, states that every convergent sequence of functions is "nearly" uniformly convergent. The first principle states that a measurable set is "almost" an open set. In $\mathbb{R}$, this is the statement that for a measurable subset T and any $\epsilon > 0$ there is a sequence of open intervals, O_n such that $\lambda(T \Delta (\cup O_n)) < \epsilon$, where λ is the Lebesgue measure.

Littlewood's second principle states that a measurable function is almost a continuous function. In $\mathbb{R}$, this is the statement that for a measurable real function f and any $\epsilon > 0$ there is an open subset of $\mathbb{R}$, say S, such that f is continuous outside of S, and $\lambda(S) < \epsilon$.

The concept of an integral is one of the most important ones in mathematics. The definition of an integral based on Jordan measure by Bernhard Riemann in the mid-nineteenth century was rigorous and seemed to cover most interesting cases. By 1900, however, a number of examples had been put forth that indicated the inadequacy of the Riemann integral (see Hawkins (1979), for an interesting discussion of the mathematical developments). Lebesgue not only provided generalizations of basic concepts, such as what we now call Lebesgue measure, but took a fundamentally different approach. (It is interesting to read what Lebesgue had to say about generalizations: "It is that a generalization made not for the vain pleasure of generalizing but in

order to solve previously existing problems is always a fruitful generalization" (Lebesgue (1926), page 194 as translated by May, 1966,).

There are many classic and standard texts on real analysis, and it would be difficult to select "best" ones. Many treat measure theory in the context of probability theory, and some of those are listed in the additional references for Chapter 1, beginning on page 145. Below I list a few more that I have found useful. I often refer to Hewitt and Stromberg (1965), from which I first began learning real analysis. Royden (1988) may be more readily available, however. The little book by Boas Jr. (1960) is a delightful read.

My study of complex analysis has been more superficial, and I am not familiar with the standard texts. The text Brown and Churchill (2008) is an updated version of the second edition by Churchill (1960) that I used.

There are a number of useful books on " counterexamples in [X]", such as Gelbaum and Olmsted (1990), Gelbaum and Olmsted (2003), Rajwade and Bhandari (2007), Steen and Seebach Jr. (1995), Stoyanov (1987), Wise and Hall (1993), and Romano and Siegel (1986).

Exercises for Section 0.1

0.1.1. Let Ω be the universal set, and let $\mathcal{F}$ consist of all countable and cocountable subsets of Ω. Show that $\mathcal{F}$ is a σ-field on Ω (Example 0.1.3).

0.1.2. Prove Theorem 0.1.4.

0.1.3. Show that $\mathcal{F}_1 \cup \mathcal{F}_2$ in equation (0.1.2) is not a σ-field.

0.1.4. Show that $\mathcal{F}_B$ in equation (0.1.3) is a σ-field.

0.1.5. Let f be a measurable function from the measurable space $(\Omega, \mathcal{F})$ to $(\Lambda, \mathcal{G})$. Show that $f^{-1}[\mathcal{G}]$ is a sub-σ-field of $\mathcal{F}$.

0.1.6. Show that $\sigma(g \circ f) \subseteq \sigma(f)$ (page 704).

0.1.7. Prove Theorem 0.1.6.

0.1.8. Prove Theorem 0.1.8.

0.1.9. Prove Theorem 0.1.9.

0.1.10. Suppose that $\mu_1, \mu_2, \ldots$ are measures on the measurable space $(\Omega, \mathcal{F})$. Let $\{a_n\}_{n=1}^{\infty}$ be a sequence of positive numbers. Prove that $\mu = \sum_{n=1}^{\infty} a_n \mu_n$ is a measure on $(\Omega, \mathcal{F})$.

0.1.11. Show that the function defined in equation (0.1.12) is a Radon measure.

0.1.12. Let Ω and Λ be arbitrary sets and let $X : \Lambda \mapsto \Omega$ be an arbitrary function.
   a) Show that if $\mathcal{F}$ is a σ-field on Ω then $\mathcal{G} = \{X^{-1}(A) : A \in \mathcal{F}\}$ is a σ-field on Λ
   b) Show that if $\mathcal{G}$ is a σ-field on Λ then $\mathcal{F} = \{A \subseteq \Omega : X^{-1}(A) \in \mathcal{G}\}$ is a σ-field on Ω.

0.1.13. Let λ, μ, and ν be measures on $(\Omega, \mathcal{F})$ and $a \in \mathbb{R}$. Show

$$\lambda \ll \nu \quad \text{and} \quad \mu \ll \nu \Longrightarrow (a\lambda + \mu) \ll \nu$$

and

$$\lambda \perp \nu \quad \text{and} \quad \mu \perp \nu \Longrightarrow (a\lambda + \mu) \perp \nu.$$

0.1.14. Prove parts (ii) and (iii) of Theorem 0.1.12.

0.1.15. Show that the same Borel field $\mathcal{B}(\mathbb{R})$ is generated by the collection of all open sets of $\mathbb{R}$.

0.1.16. Show that the inverse image of a Borel set under a continuous function $f : \mathbb{R} \mapsto \mathbb{R}$ is Borel.

0.1.17. Let $(\Omega, \mathcal{F}, \nu)$ be a measure space, and let $B \in \mathcal{F}$. Now let ν_B be a set function that is the same as ν on $\mathcal{F}_B$ and undefined elsewhere. Show that ν_B is a measure.

0.1.18. Given the measure space $(\Omega, \mathcal{F}, \nu)$ and $\mathcal{F}_c$ and ν_c constructed as on page 712.

a) Show that $(\Omega, \mathcal{F}_c, \nu_c)$ is a complete measure space.

b) Show that for every $A \in \mathcal{F}_c$ there is some $B, C \in \mathcal{F}$ with $\nu(C) = 0$ such that $A = B \cup C$, and

$$\nu_c(A) = \nu(B).$$

0.1.19. Given the measure space $(\Omega, \mathcal{F}, \nu)$ and the measurable space $(\Lambda, \mathcal{G})$. Let f be a function from Ω to Λ that is measurable with respect to $\mathcal{F}$. Show that $\nu \circ f^{-1}$ is a measure and that its domain and range are $\mathcal{G}$. This is the induced measure or the "pushforward" measure.

0.1.20. Let X be a measurable function from the measurable space $(\Omega, \mathcal{F})$ to $(\mathbb{R}, \mathcal{B})$. Prove that $\sigma(X^{-1}[\mathbb{R}]) \subseteq \mathcal{F}$.

0.1.21. Show that $(\mathbb{R}, \mathcal{B})$ is a topological space. What are the open sets of the topology?

0.1.22. Show that Lebesgue measure on $(\mathbb{R}, \mathcal{B})$ is a Radon measure.

Show that Lebesgue measure on $(\mathbb{R}, \mathcal{B})$ is a Haar invariant measure wrt to the group $(\mathbb{R}, +)$.

0.1.23. Show that equation (0.1.22) holds for Lebesgue measure.

0.1.24. Let $\mathcal{C}_{[0,1]}$ be the collection of all open intervals within $[0, 1]$. Show that $\mathcal{B}_{[0,1]} = \sigma(\mathcal{C}_{[0,1]})$.

0.1.25. Under what conditions is the indicator function I_S measurable?

0.1.26. Show that a simple function is Borel measurable.

0.1.27. ****continuity questions

0.1.28. Show that the Weierstrass function is Hölder continuous of order α for any $\alpha < 1$.

0.1.29. Let $(\Omega, \mathcal{F}, \nu)$ be a measure space and f be a nonnegative Borel function. For $A \subseteq \Omega$, show that $\lambda(A) = \int_A f \, \mathrm{d}\nu$ is a measure over $(\Omega, \mathcal{F})$.

0.1.30. Show that the measure defined in equation (0.1.48) is Haar invariant.

0.1.31. Prove Theorem 0.1.19.

0.1.32. In the text, we say that the proofs of Theorems 0.1.20 through 0.1.22 "follow immediately from the definition" of the Lebesgue integral. "Immediately" means that there are one or two reasons that are direct results of the definition. For each of these theorems, state the reason(s).

0.1.33. Prove each of the equations (0.1.66) through (0.1.70) under the conditions given.

0.1.34. Assume f^2 and g^2 are integrable. Show that

$$\left(\int fg\mathrm{d}\nu\right)^2 \leq \int f^2\mathrm{d}\nu \int g^2\mathrm{d}\nu.$$

This is an instance of a famous inequality. What is its name?

0.1.35. Show that the class $\mathcal{C}^k$ of functions over the same domain is a linear space.

0.1.36. Show that $\mathcal{L}^p$ is a linear space.

0.1.37. Show that if $f, f_1, f_2, \ldots \in \mathcal{L}^p(\nu, D)$ and $\nu(D) < \infty$, then

$$f_n \xrightarrow{\mathrm{L}_p} f \Rightarrow f_n \xrightarrow{\mathrm{L}_r} f \quad \text{for } r \leq p.$$

0.1.38. Prove Theorem 0.1.27.

0.1.39. Prove Theorem 0.1.28.

0.1.40. Prove Theorem 0.1.29.

Hint: First prove this for the case that g is the Heaviside function.

0.2 Stochastic Processes and the Stochastic Calculus

Most sections in this chapter generally cover prerequisite material for the rest of the book. This section, on the other hand, depends on some of the material in Chapter 1, and is closely interrelated with the material in Section 1.6.

0.2.1 Stochastic Differential Equations

We consider a stochastic process $\{B_t\}$ in which we generally associate the index t with time. We often write $\{B(t)\}$ in place of $\{B_t\}$, but for all practical purposes, the notation is equivalent. If time is considered continuous, there are two possibilities that we will consider. One, called a *jump process*, is discontinuous in time, and the other, called a *diffusion process*, is not only continuous in time, but it is also differential with respect to time.

We will first briefly discuss a particular kind of jump process and then turn our attention to various kinds of diffusion processes.

The most obvious way of developing a diffusion process is to begin with a differential equation in which some of the terms are random variables. We call such a differential equation a stochastic differential equation or SDE.

In a very important class of stochastic processes, the differences between the values at two time points have normal distributions and the difference between two points is independent of the difference between two nonoverlapping points. The simplest such process is called a Bachelier-Wiener process. We will discuss it first, and then consider some other related processes.

Poisson Jump Process

A jump process is one that is discontinuous in time. The most important jump processes are Poisson processes.

A Poisson process is a sequence of events in which the probability of k events (where $k = 0, 1, \ldots$) in an interval of length Δt, denoted by $g(k, \Delta t)$ satisfies the following conditions:

- $g(1, \Delta t) = \lambda \Delta t + \mathrm{o}(\Delta t)$, where λ is a positive constant and $(\Delta t) > 0$.
- $\sum_{k=2}^{\infty} g(k, \Delta t) \in \mathrm{o}(\Delta t)$.
- The numbers of changes in nonoverlapping intervals are stochastically independent.

This axiomatic characterization of a Poisson process leads to a differential equation whose solution (using mathematical induction) is

$$g(k, \Delta t) = \frac{(\lambda \Delta t)^k \mathrm{e}^{-\lambda \Delta t}}{k!}, \quad \text{for } k = 1, 2, \ldots \tag{0.2.1}$$

which, in turn leads to the familiar probability function for a Poisson distribution

$$p_K(k) = \frac{(\theta)^k \mathrm{e}^{-\theta}}{k!}, \quad \text{for } k = 0, 1, 2, \ldots \tag{0.2.2}$$

Bachelier-Wiener Processes

Suppose in the sequence $B_0, B_1, \ldots$, the distribution of $B_{t+1} - B_t$ is normal with mean 0 and standard deviation 1. In this case, the distribution of $B_{t+2} - B_t$ is normal with mean 0 and standard deviation $\sqrt{2}$, and the distribution of $B_{t+0.5} - B_t$ is normal with mean 0 and standard deviation $\sqrt{0.5}$. More generally, the distribution of the change ΔB in time Δt has a standard deviation of $\sqrt{\Delta t}$

This kind of process with the Markovian property and with a normal distribution of the changes leads to a Brownian motion or a Bachelier-Wiener process.

Consider a process of changes ΔB characterized by two properties:

- The change ΔB during a small period of time Δt is given by

$$\Delta B = Z\sqrt{\Delta t}, \tag{0.2.3}$$

  where Z is a random variable with a $\mathrm{N}(0,1)$ distribution.
- The values of ΔB for any two short intervals of time Δt are independent (with an appropriate definition of "short").

Now, consider N time periods, and let $T = N\Delta t$. We have

$$B(T) - B(0) = \sum_{i=1}^{N} Z_i\sqrt{\Delta t}. \tag{0.2.4}$$

The fact that we have $\sqrt{\Delta t}$ in this equation has important implications.

As in ordinary calculus, we consider $\Delta B/\Delta t$ and take the limit as $\Delta t \to 0$, which we call $\mathrm{d}B/\mathrm{d}t$, and we have the stochastic differential equation

$$\mathrm{d}B = Z\mathrm{d}t. \tag{0.2.5}$$

A random variable formed as $\mathrm{d}B$ above is called a *stochastic differential*.

A stochastic differential arising from a process of changes ΔB with the two properties above is called a *Bachelier-Wiener process* or a *Brownian motion*. In the following, we will generally use the phrase "Bachelier-Wiener process".

We can use the Bachelier-Wiener process to develop a *generalized Bachelier-Wiener process*:

$$\mathrm{d}S = \mu\mathrm{d}t + \sigma\mathrm{d}B, \tag{0.2.6}$$

where μ and σ are constants.

Properties of a Discrete Process Underlying the Bachelier-Wiener Process

With $\Delta B = Z\sqrt{\Delta t}$ and $Z \sim \mathrm{N}(0,1)$, we immediately have

$$\begin{aligned}
\mathrm{E}(\Delta B) &= 0 \\
\mathrm{E}\left((\Delta B)^2\right) &= \mathrm{V}(\Delta B) + (\mathrm{E}(\Delta B))^2 \\
&= \Delta t \\
\mathrm{E}\left((\Delta B)^3\right) &= 0 \\
\mathrm{E}\left((\Delta B)^4\right) &= \mathrm{V}\left((\Delta B)^2\right) + \left(\mathrm{E}\left((\Delta B)^2\right)\right)^2 \\
&= 3(\Delta t)^2
\end{aligned}$$

Because of independence, for $\Delta_i B$ and $\Delta_j B$ representing changes in two nonoverlapping intervals of time,

$$\mathrm{E}((\Delta_i B)(\Delta_j B)) = \mathrm{cov}(\Delta_i B, \Delta_j B) = 0. \tag{0.2.7}$$

The Bachelier-Wiener process is a random variable; that is, it is a real-valued mapping from a sample space Ω. We sometimes use the notation $B(\omega)$ to emphasize this fact.

The Bachelier-Wiener process is a function in continuous time. We sometimes use the notation $B(t, \omega)$ to emphasize the time dependency.

Most of the time we drop the "ω". Also, sometimes we write B_t instead of $B(t)$.

All of these notations are equivalent.

There two additional properties of a Bachelier-Wiener process or Brownian motion that we need in order to have a useful model. We need an initial value, and we need it to be continuous in time.

Because the Bachelier-Wiener process is a random variable, the values it takes are those of a function at some point in the underlying sample space, Ω. Therefore, when we speak of $B(t)$ at some t, we must speak in terms of probabilities of values or ranges of values.

When we speak of a particular value of $B(t)$, unless we specify a specific point $\omega_0 \in \Omega$, the most we can say is that the value occurs almost surely.

- We assume $B(t) = 0$ almost surely at $t = 0$.
- We assume $B(t)$ is almost surely continuous in t.

These two properties together with the limiting forms of the two properties given at the beginning define a Bachelier-Wiener process or Brownian motion.

(There is a theorem due to Kolmogorov that states that given the first three properties, there exists a "version" that is absolutely continuous in t.)

From the definition, we can see immediately that

- the Bachelier-Wiener process is Markovian
- the Bachelier-Wiener process is a martingale.

Generalized Bachelier-Wiener Processes

A Bachelier-Wiener process or Brownian motion is a model for changes. It models diffusion.

If the process drifts over time (in a constant manner), we can add a term for the drift, adt.

More generally, a model for the state of a process that has both a Brownian diffusion and a drift is a generalized Bachelier-Wiener process:

$$\mathrm{d}S = a\mathrm{d}t + b\mathrm{d}B, \tag{0.2.8}$$

where a and b are constants. A generalized Bachelier-Wiener process is a type of a more general "drift-diffusion process".

While the expected value of the Bachelier-Wiener process at any time is 0, the expected value of the state S is not necessarily 0. Likewise, the variance is affected by b. Both the expected value and the variance of S are functions of time.

One of the most interesting properties of a Bachelier-Wiener process is that its first variation is infinite. It is infinitely "wiggly". We can see this by generating normal processes over varying length time intervals, as in Figure 0.2.

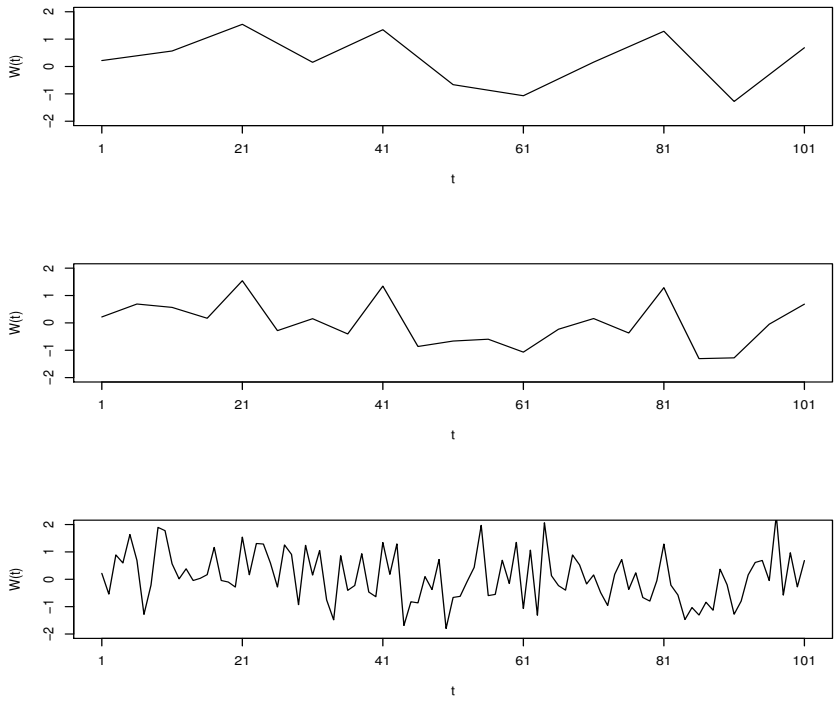


Figure 0.2. A Bachelier-Wiener Process Observed at Varying Length Intervals

Variation of Functionals

The variation of a functional is a measure of its rate of change. It is similar in concept to an integral of a derivative of a function.

For studying variation, we will be interested only in functions from the interval $[0, T]$ to $\mathbb{R}$.

To define the variation of a general function $f : [0, T] \mapsto \mathbb{R}$, we form N intervals $0 = t_0 \leq t_1 \leq \cdots \leq t_N = T$. The intervals are not necessarily of equal length, so we define Δ as the maximum length of any interval; that is,

$$\Delta = \max(t_i - t_{i-1}).$$

Now, we denote the p^{th} variation of f as $\mathrm{V}^p(f)$ and define it as

$$\mathrm{V}^p(f) = \lim_{\Delta \to 0} \sum_{i=1}^{N} |f(t_i) - f(t_{i-1})|^p.$$

(Notice that $\Delta \to 0$ implies $N \to \infty$.)

With equal intervals, Δt, for the first variation, we can write

$$\begin{aligned} \mathrm{V}^1(f) &= \lim_{\Delta t \to 0} \sum_{i=1}^{N} |f(t_i) - f(t_{i-1})| \\ &= \lim_{N \to \infty} \sum_{i=0}^{N-1} \Delta t \frac{|f(t_i + \Delta t) - f(t_i)|}{\Delta t}, \end{aligned}$$

from which we can see that for a differentiable function $f : [0, T] \mapsto \mathbb{R}$,

$$\mathrm{V}^1(f) = \int_0^T \left| \frac{\mathrm{d}f}{\mathrm{d}t} \right| \mathrm{d}t.$$

The notation $FV(f)$, or more properly, $\mathrm{FV}(f)$, is sometimes used instead of $\mathrm{V}^1(f)$.

Again, with equal intervals, Δt, for the second variation, we can write

$$\begin{aligned} \mathrm{V}^2(f) &= \lim_{\Delta t \to 0} \sum_{i=1}^{N} (f(t_i) - f(t_{i-1}))^2 \\ &= \lim_{\Delta t \to 0} \Delta t \lim_{N \to \infty} \sum_{i=0}^{N-1} \Delta t \left(\frac{|f(t_i + \Delta t) - f(t_i)|}{\Delta t} \right)^2. \end{aligned}$$

For a differentiable function $f : [0, T] \mapsto \mathbb{R}$, we have

$$\mathrm{V}^2(f) = \lim_{\Delta t \to 0} \Delta t \int_0^T \left| \frac{\mathrm{d}f}{\mathrm{d}t} \right|^2 \mathrm{d}t.$$

The integrand is bounded, therefore this limit is 0, and we conclude that the second variation of a differentiable function is 0.

If X is a stochastic functional, then $\mathrm{V}^p(X)$ is also stochastic. If it converges to a deterministic quantity, the nature of the convergence must be considered.

First and Second Variation of a Bachelier-Wiener Process

Two important properties of a Bachelier-Wiener process on $[0, T]$ are

- $\mathrm{V}^2(B) = T$ a.s., which as we have seen, implies that $B(t)$ is not differentiable.
- $\mathrm{V}^1(B) = \infty$ a.s.

Notice that because B is a random variable we must temper our statement with a phrase about the probability or expected value.

We now prove these for the quadratic mean instead of a.s. We start with the first one, because it will imply the second one. Let

$$\begin{aligned} X_N &= \sum_{n=0}^{N-1} (B(t_{n+1}) - B(t_n))^2 \\ &= \sum_{n=0}^{N-1} (\Delta_n B)^2 \quad \text{note notation} \end{aligned}$$

We want to show

$$\mathrm{E}\left((X_N - T)^2\right) \to 0 \quad \text{as } |\Delta t| \to 0. \tag{0.2.9}$$

Now,

$$\mathrm{E}\left((X_N - T)^2\right) = \mathrm{E}\left(X_N^2\right) - 2T\mathrm{E}(X_N) + T^2 = \mathrm{E}\left(X_N^2\right) - T^2.$$

So now we want to show

$$\mathrm{E}\left(X_N^2\right) = T^2. \tag{0.2.10}$$

$$\begin{aligned} \mathrm{E}\left(X_N^2\right) &= \mathrm{E}\left(\sum_{i=0}^{N-1} (\Delta_i B)^2 \sum_{j=0}^{N-1} (\Delta_j B)^2\right) \\ &= \mathrm{E}\left(\sum_{i=0}^{N-1} (\Delta_i B)^4\right) + \mathrm{E}\left(\sum_{i \neq j} (\Delta_i B)^2 (\Delta_j B)^2\right) \\ &= \sum_{i=0}^{N-1} (\Delta_i t)^2 + \sum_{i \neq j} (\Delta_i t)(\Delta_j t). \end{aligned}$$

Because $|\Delta t| \to 0$ (or, if we allow different size intervals, $\sup |\Delta_i t| \to 0$), we have

$$\sum_{i=0}^{N-1} (\Delta_i t)^2 \to 0.$$

So the first term goes to 0; now consider $\sum_{i\neq j}(\Delta_i t)(\Delta_j t)$.

$$\begin{aligned}\sum_{i\neq j}(\Delta_i t)(\Delta_j t) &= \sum_{i=0}^{N-1}(\Delta_i t)\left(\sum_{j=0}^{i-1}(\Delta_j t)+\sum_{j=i+1}^{N-1}(\Delta_j t)\right)\\ &= \sum_{i=0}^{N-1}(\Delta_i t)(T-\Delta_i t)\\ &= T\sum_{i=0}^{N-1}(\Delta_i t)-\sum_{i=0}^{N-1}(\Delta_i t)^2\\ &= T^2-0.\end{aligned}$$

So now we have $\mathrm{E}\left((X_N-T)^2\right)\to 0$, or $X_N \overset{\mathrm{L}_2}{\to} T$ as $|\Delta t|\to 0$; that is, $\mathrm{V}^2(B)=T$ in quadratic mean, or in L_2 norm.

(I just realized that I had stated a.s. convergence, and I proved L_2 *convergence. One does not imply the other, but a.s. is also true in this case.)*

Now, although we have already seen that since the second variation is nonzero, B cannot be differentiable.

But also because of the continuity of B in t, it is easy to see that the first variation diverges if the second variation converges to a finite value. This is because

$$\sum_{n=0}^{N-1}(B(t_{n+1})-B(t_n))^2 \leq \sup|B(t_{n+1})-B(t_n)|\sum_{n=0}^{N-1}|B(t_{n+1})-B(t_n)|$$

In the limit the term on the left is $T>0$, and the term on the right is 0 times $\mathrm{V}^1(B)$; therefore $\mathrm{V}^1(B)=\infty$.

Properties of Stochastic Differentials

Although B and $\mathrm{d}B$ are random variables, the product $\mathrm{d}B\mathrm{d}B$ is deterministic.

We can see this by considering the stochastic process $(\Delta B)^2$. We have seen that $\mathrm{V}\left((\Delta B)^2\right)=2(\Delta t)^2$, so the variance of this process is $2(\Delta t)^2$; that is, as $\Delta t\to 0$, the variance of this process goes to 0 faster, as $(\Delta t)^2$.

Also, as we have seen, $\mathrm{E}\left((\Delta B)^2\right)=\Delta t$, and so $(\Delta B)^2$ goes to Δt at the same rate as $\Delta t\to 0$. That is,

$$(\Delta B)(\Delta B)\overset{\text{a.s.}}{\to}\Delta t \quad \text{as } \Delta t\to 0. \tag{0.2.11}$$

The convergence of $(\Delta B)(\Delta B)$ to Δt as $\Delta t\to 0$ yields

$$\mathrm{d}B\mathrm{d}B=\mathrm{d}t. \tag{0.2.12}$$

(This equality is almost sure.) But $\mathrm{d}t$ is a deterministic quantity.

This is one of the most remarkable facts about a Bachelier-Wiener process.

Multidimensional Bachelier-Wiener Processes

If we have two Bachelier-Wiener processes B_1 and B_2, with $\mathrm{V}(\mathrm{d}B_1) = \mathrm{V}(\mathrm{d}B_2) = \mathrm{d}t$ and $\mathrm{cov}(\mathrm{d}B_1, \mathrm{d}B_2) = \rho\mathrm{d}t$ (that is, $\mathrm{Cor}(\mathrm{d}B_1, \mathrm{d}B_2) = \rho$), then by a similar argument as before, we have $\mathrm{d}B_1\mathrm{d}B_2 = \rho\mathrm{d}t$, almost surely.

Again, this is deterministic.

The results of course extend to any vector of Bachelier-Wiener processes $(B_1, \ldots, B_d)$.

If $(B_1, \ldots, B_d)$ arise from

$$\Delta B_i = X_i\sqrt{\Delta t},$$

where the vector of Xs has a multivariate normal distribution with mean 0 and variance-covariance matrix Σ, then the variance-covariance matrix of $(\mathrm{d}B_1, \ldots, \mathrm{d}B_d)$ is $\Sigma\mathrm{d}t$, which is deterministic.

Starting with $(Z_1, \ldots, Z_d) \overset{\text{iid}}{\sim} \mathrm{N}(0, 1)$ and forming the Wiener processes $B = (B_1, \ldots, B_d)$ beginning with

$$\Delta B_i = Z_i\sqrt{\Delta t},$$

we can form a vector of Bachelier-Wiener processes $B = (B_1, \ldots, B_d)$ with variance-covariance matrix $\Sigma\mathrm{d}t$ for $\mathrm{d}B = (\mathrm{d}B_1, \ldots, \mathrm{d}B_d)$ by the transformation

$$B = \Sigma^{1/2}B,$$

or equivalently by

$$B = \Sigma_C B,$$

where Σ_C is a Cholesky factor of Σ, that is, $\Sigma_C^{\mathrm{T}}\Sigma_C = \Sigma$.

Recall, for a fixed matrix A,

$$\mathrm{V}(AY) = A^{\mathrm{T}}\mathrm{V}(Y)A,$$

so from above, for example,

$$\mathrm{V}(\mathrm{d}B) = \Sigma_C^{\mathrm{T}}\mathrm{V}(\mathrm{d}B)\Sigma_C = \Sigma_C^{\mathrm{T}}\mathrm{diag}(\mathrm{d}t)\Sigma_C = \Sigma\mathrm{d}t. \tag{0.2.13}$$

The stochastic differentials such as $\mathrm{d}B$ naturally lead us to consider integration with respect to stochastic differentials, that is, stochastic integrals.

Stochastic Integrals with Respect to Bachelier-Wiener Processes

If B is a Bachelier-Wiener process on $[0, T]$, we may be interested in an integral of the form

$$\int_0^T g(Y(t), t)\mathrm{d}B,$$

where $Y(t)$ is a stochastic process (that is, Y is a random variable) and g is some function. First, however, we must develop a definition of such an integral. We will return to this problem in Section 0.2.2. Before doing that, let us consider some generalizations of the Wiener process.

Ito Processes

An Ito process is a generalized Bachelier-Wiener process $\mathrm{d}X = a\mathrm{d}t + b\mathrm{d}B$, in which the parameters a and b are functions of the underlying variable X and of time t (of course, X is also a function of t).

The functions a and b must be measurable with respect to the filtration generated by $B(t)$ (that is, to the sequence of smallest σ-fields with respect to which $B(t)$ is measurable. (This is expressed more simply by saying $a(X(t), t)$ and $b(X(t), t)$ are adapted to the filtration generated by $B(t)$.)

The Ito process is of the form

$$\mathrm{d}X(t) = a(X(t), t)\mathrm{d}t + b(X(t), t)\mathrm{d}B. \tag{0.2.14}$$

The Ito integral (or any other stochastic integral) gives us a solution to this stochastic differential equation:

$$X(T) = X(0) + \int_0^T a(X(t), t)\mathrm{d}t + \int_0^T b(X(t), t)\mathrm{d}B(t). \tag{0.2.15}$$

(The differential in the first integral is deterministic although the integrand is stochastic. The second integral, however, is a stochastic integral. Other definitions of this integral would require modifications in the interpretation of properties of the Ito process.)

We are often interested in multidimensional Ito processes. Their second-order properties (variances and covariances) behave very similarly to those of Bachelier-Wiener processes, which we discussed earlier.

There are many interesting forms of Ito processes.

Geometric Brownian Motion

The Ito process would be much easier to work with if $\mu(\cdot)$ and $\sigma(\cdot)$ did not depend on the value of the state; that is, if we we use the model

$$\frac{\mathrm{d}X(t)}{X(t)} = \mu(t)\mathrm{d}t + \sigma(t)\mathrm{d}B. \tag{0.2.16}$$

The Ito process would be even easier to work with if $\mu(\cdot)$ and $\sigma(\cdot)$ were constant; that is, if we we just use the model

$$\frac{\mathrm{d}X(t)}{X(t)} = \mu\mathrm{d}t + \sigma\mathrm{d}B. \tag{0.2.17}$$

This model is called a geometric Brownian motion, and is widely used in modeling prices of various financial assets. ("Geometric" refers to series with multiplicative changes, as opposed to "arithmetic series" that have additive changes).

The geometric Brownian motion model is similar to other common statistical models:

$$\frac{\mathrm{d}X(t)}{X(t)} = \mu\mathrm{d}t + \sigma\mathrm{d}B(t)$$

or

$$\text{response} = \text{systematic component} + \text{random error}.$$

Without the stochastic component, the differential equation has the simple solution

$$X(t) = ce^{\mu t},$$

from which we get the formula for continuous compounding for a rate μ.

Ornstein-Ulenbeck Process

Also called Vasicek process

$$\mathrm{d}X(t) = (\theta_1 - \theta_2 X(t))\mathrm{d}t + \theta_3 \mathrm{d}B. \tag{0.2.18}$$

Cox-Ingersoll-Ross Process

Also called Feller process

$$\mathrm{d}X(t) = (\theta_1 - \theta_2 X(t))\mathrm{d}t + \theta_3 \sqrt{X(t)}\mathrm{d}B. \tag{0.2.19}$$

***** move this *** Feller's condition

$$2\theta_1 > \theta_3^2$$

Jump-Diffusion Processes

In financial modeling, we often use a compound process that consists of some smooth process coupled with a jump process. The parameters controlling the frequency of jumps may also be modeled as a stochastic process. The *amount* of the jump is usually modeled as a random variable.

We merely add a pure Poisson jump process $\mathrm{d_j}X(t)$ (see page 765) to the drift-diffusion process,

$$\mathrm{d}X(t) = \mu(X(t), t)\mathrm{d}t + \sigma(X(t), t)\mathrm{d}B(t).$$

After rearranging terms, this yields

$$\begin{aligned} \mathrm{d}X(t) = &\big(\mu(X(t),t) + (\lambda(X(t),t)\textstyle\int_{\mathcal{Z}} z\, p_Z(z;X(t))\mathrm{d}z\big)\,\mathrm{d}t \\ &+\sigma(X(t),t)\mathrm{d}B(t) \\ &+\mathrm{d}_{\mathrm{j}}J_X(t). \end{aligned} \tag{0.2.20}$$

There are two stochastic terms, $\mathrm{d}B(t)$ and $\mathrm{d}_{\mathrm{j}}J_X(t)$.

We will assume that they are independent.

Note that I suppressed the d_{j} on the left hand side, although, clearly, this is a discontinuous process, both because of the compensated process and the discontinuity in the drift.

Multivariate Processes

The multivariate Ito process has the form

$$\mathrm{d}X(t) = a(X,t)\mathrm{d}t + B(X,t)\mathrm{d}B(t), \tag{0.2.21}$$

where $\mathrm{d}X(t)$, $a(X,t)$, and $\mathrm{d}B(t)$ are vectors and $B(X,t)$ is a matrix.

The elements of $\mathrm{d}B(t)$ can come from independent Bachelier-Wiener processes, or from correlated Bachelier-Wiener processes. I think it is easier to work with independent Bachelier-Wiener processes and incorporate any correlations into the $B(X,t)$ matrix. Either way is just as general.

We write the individual terms in a multivariate Ito process in the form

$$\mathrm{d}X_i(t) = a_i(X,t)\mathrm{d}t + b_i(X,t)\mathrm{d}B_i(t), \tag{0.2.22}$$

where the $B_i(t)$ are Bachelier-Wiener processes with

$$\mathrm{Cor}(\mathrm{d}B_i(t), \mathrm{d}B_j(t)) = \rho_{ij}, \tag{0.2.23}$$

for some constants ρ_{ij}. Note that a_i and b_i are functions of all X_j, so the processes are coupled not just through the ρ_{ij}.

Recall that $\mathrm{V}(\mathrm{d}B_i(t)) = \mathrm{V}(\mathrm{d}B_i(t)) = \mathrm{d}t$, and hence $\mathrm{cov}(\mathrm{d}B_i(t), \mathrm{d}B_j(t)) = \rho_{ij}\mathrm{d}t$.

Also recall that $(\mathrm{d}B_i(t))^2 \overset{\text{a.s.}}{=} \mathrm{E}((\mathrm{d}B_i(t))^2) \overset{\mathrm{d}}{=} t$; i.e., $(\mathrm{d}B_i(t))^2$ is non-stochastic. Likewise, $\mathrm{d}B_i(t)\mathrm{d}B_i(t) \overset{\text{a.s.}}{=} \rho_{ij}\mathrm{d}t$.

0.2.2 Integration with Respect to Stochastic Differentials

The problem with developing a definition of an integral of the form

$$\int_0^T g(Y(t),t)\mathrm{d}B \tag{0.2.24}$$

following the same steps as in the definition of a Riemann integral, that is, as a limit of sequences of sums of areas of rectangles, is that because the sides of these rectangles, Y and $\mathrm{d}B$, are random variables, there are different kinds of convergence of a limit.

Also, the convergence of products of $Y(t)$ depend on where $Y(t)$ is evaluated.

The Ito Integral

We begin developing a definition of

$$\int_0^T g(Y(t),t)\mathrm{d}B,$$

by considering how the Riemann integral is defined in terms of the sums

$$I_n(t) = \sum_{i=0}^{n-1} g(Y(\tau_i),\tau_i)(B(t_{i+1}) - B(t_i)),$$

where $0 = t_0 \leq \tau_0 \leq t_1 \leq \tau_1 \leq \cdots \leq \tau_{n-1} \leq t_n = T$.

As in the Riemann case we will define the integral in terms of a limit as the mesh size goes to 0.

First, the existence depends on a finite expectation that is similar to a variance. We assume

$$\mathrm{E}\left(\int_0^T g(Y(t),t)\mathrm{d}t\right) < \infty.$$

The convergence must be qualified because the intervals are random variables; furthermore, (although it is not obvious!) the convergence depends on where τ_i is in the interval $[t_i, t_{i+1}]$.

The first choice in the definition of the Ito stochastic integral is to choose $\tau_i = t_i$. Other choices, such as choosing τ_i to be at the midpoint of the integral, lead to different types of stochastic integrals.

Next is the definition of the type of convergence. In the Ito stochastic integral, the convergence is in mean square, that is L_2 convergence.

With the two choices me have made, we take

$$I_n(t) = \sum_{i=0}^{n-1} g(Y(t_i),t_i)(B(t_{i+1}) - B(t_i)),$$

and the Ito integral is defined as

$$I(t) = \text{ms-lim}_{n\to\infty} I_n(t). \tag{0.2.25}$$

This integral based on a Bachelier-Wiener process is used throughout financial analysis.

Note that this integral is a random variable; in fact, it is a stochastic process. This is because of the fact that the differentials are from a Bachelier-Wiener process.

Also, because the integral is defined by a Bachelier-Wiener process, it is a martingale.

Ito's Lemma

We can formalize the preceding discussion using Ito's lemma.

Suppose X follows an Ito process,

$$\mathrm{d}X(t) = a(X,t)\mathrm{d}t + b(X,t)\mathrm{d}B(t),$$

where $\mathrm{d}B$ is a Bachelier-Wiener process. Let G be an infinitely differentiable function of X and t. Then G follows the process

$$\mathrm{d}G(t) = \left(\frac{\partial G}{\partial X}a(X,t) + \frac{\partial G}{\partial t} + \frac{1}{2}\frac{\partial^2 G}{\partial X^2}b^2\right)\mathrm{d}t + \frac{\partial G}{\partial X}b(X,t)\mathrm{d}B(t). \tag{0.2.26}$$

Thus, Ito's lemma provides a formula that tells us that G also follows an Ito process.

The drift rate is

$$\frac{\partial G}{\partial X}a(X,t) + \frac{\partial G}{\partial t} + \frac{1}{2}\frac{\partial^2 G}{\partial X^2}b^2$$

and the volatility is

$$\frac{\partial G}{\partial X}b(X,t).$$

This allows us to work out expected values and standard deviations of G over time.

First, suppose that G is infinitely of X and an unrelated variable y, and consider a Taylor series expansion for ΔG:

$$\Delta G = \frac{\partial G}{\partial X}\Delta X + \frac{\partial G}{\partial y}\Delta y + \frac{1}{2}\left(\frac{\partial^2 G}{\partial X^2}(\Delta X)^2 + \frac{\partial^2 G}{\partial y^2}(\Delta y)^2 + 2\frac{\partial^2 G}{\partial X \partial y}\Delta X \Delta y\right) + \cdots \tag{0.2.27}$$

In the limit as ΔX and Δy tend to zero, this is the usual "total derivative"

$$\mathrm{d}G = \frac{\partial G}{\partial X}\mathrm{d}X + \frac{\partial G}{\partial y}\mathrm{d}y, \tag{0.2.28}$$

in which the terms in ΔX and Δy have dominated and effectively those in $(\Delta X)^2$ and $(\Delta y)^2$ and higher powers have disappeared.

Now consider an X that follows an Ito process,

$$\mathrm{d}X(t) = a(X,t)\mathrm{d}t + b(X,t)\mathrm{d}B(t),$$

or

$$\Delta X(t) = a(X,t)\Delta t + b(X,t)Z\sqrt{\Delta t}.$$

Now let G be a function of both X and t, and consider the analogue to equation (0.2.27). The factor $(\Delta X)^2$, which could be ignored in moving to equation (0.2.28), now contains a term with the factor Δt, which cannot be ignored. We have

$$(\Delta X(t))^2 = b(X,t)^2 Z^2 \Delta t + \text{ terms of higher degree in } \Delta t.$$

Consider the Taylor series expansion

$$\Delta G = \frac{\partial G}{\partial X}\Delta X + \frac{\partial G}{\partial t}\Delta t + \frac{1}{2}\left(\frac{\partial^2 G}{\partial X^2}(\Delta X)^2 + \frac{\partial^2 G}{\partial t^2}(\Delta t)^2 + 2\frac{\partial^2 G}{\partial X \partial t}\Delta X \Delta t\right) + \cdots \tag{0.2.29}$$

We have seen, under the assumptions of Brownian motion, $(\Delta X(t))^2$ or, equivalently, $Z^2\Delta t$, is nonstochastic; that is, we can treat $Z^2\Delta t$ as equal to its expected value as Δt tends to zero. Therefore, when we substitute for $\Delta X(t)$, and take limits in equation (0.2.29) as ΔX and Δt tend to zero, we get

$$\mathrm{d}G(t) = \frac{\partial G}{\partial X}\mathrm{d}X + \frac{\partial G}{\partial t}\mathrm{d}t + \frac{1}{2}\frac{\partial^2 G}{\partial X^2}b^2\mathrm{d}t \tag{0.2.30}$$

or, after substituting for $\mathrm{d}X$ and rearranging, we have Ito's formula

$$\mathrm{d}G(t) = \left(\frac{\partial G}{\partial X}a(X,t) + \frac{\partial G}{\partial t} + \frac{1}{2}\frac{\partial^2 G}{\partial X^2}b^2\right)\mathrm{d}t + \frac{\partial G}{\partial X}b(X,t)\mathrm{d}B(t).$$

Equation (0.2.30) is also called *Ito's formula*. Compare equation (0.2.30) with equation (0.2.28).

We can think of Ito's formula as a stochastic version of the chain rule.

There is a multivariate version of Ito's formula for a multivariate Ito process. Given an infinitely differential function G of the vector $X = (X_1, \ldots, X_d)$ and the scalar t, Ito's formula in the form of equation (0.2.30), derived in the same way as for the univariate case, is

$$\mathrm{d}G(t) = \sum_{i=1}^{d}\frac{\partial G}{\partial X_i}\mathrm{d}X_i(t) + \frac{\partial G}{\partial t}\mathrm{d}t + \frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{d}\frac{\partial G^2}{\partial X_i \partial X_j}\rho_{ij}b_i(X,t)b_j(X,t)\mathrm{d}t. \tag{0.2.31}$$

The form of equation (0.2.26), for example, is obtained by substituting for $\mathrm{d}X_i(t)$.

Solution of Stochastic Differential Equations

*** existence Feller's condition etc.

The solution of a stochastic differential equation is obtained by integrating both sides and allowing for constant terms. Constant terms are evaluated by satisfying known boundary conditions, or initial values.

In a stochastic differential equation, we must be careful in how the integration is performed, although different interpretations may be equally appropriate.

For example, the SDE that defines an Ito process

$$\mathrm{d}X(t) = a(X,t)\mathrm{d}t + b(X,t)\mathrm{d}B(t), \tag{0.2.32}$$

when integrated from time t_0 to T yields

$$X(T) - X(t_0) = \int_{t_0}^{T} a(X,t)\mathrm{d}t + \int_{t_0}^{T} b(X,t)\mathrm{d}B(t). \tag{0.2.33}$$

The second integral is a stochastic integral. We will interpret it as an Ito integral.

The nature of $a(X,t)$ and $b(X,t)$ determine the complexity of the solution to the SDE.

In the Ito process

$$\mathrm{d}S(t) = \mu(t)S(t)\mathrm{d}t + \sigma(t)S(t)\mathrm{d}B(t),$$

using Ito's formula for the log as before, we get the solution

$$S(T) = S(t_0)\exp\left(\int_{t_0}^{T}\left(\mu(t) - \frac{1}{2}\sigma(t)^2\right)\mathrm{d}t + \int_{t_0}^{T}\sigma(t)\mathrm{d}B(t)\right). \tag{0.2.34}$$

In the simpler version of a geometric Brownian motion model, in which μ and σ are constants, we have

$$S(T) = S(t_0)\exp\left(\left(\mu - \frac{1}{2}\sigma^2\right)\Delta t + \sigma\Delta B\right). \tag{0.2.35}$$

Given a solution of a differential equation we may determine the mean, variance and so on by taking expectations of the random component in the solution.

Sometimes, however, it is easier just to develop an ordinary (nonstochastic) differential equation for the moments. We do this from an Ito process

$$\mathrm{d}X(t) = a(X,t)\mathrm{d}t + b(X,t)\mathrm{d}B(t), \tag{0.2.36}$$

by using Ito's formula on the powers of the variable. So we have

$$\mathrm{d}X^p(t) = \left(pX(t)^{p-1}a(X,t) + \frac{1}{2}p(p-1)X(t)^{p-2}b(X,t)^2\right)\mathrm{d}t + pX(t)^{p-1}b(X,t)\mathrm{d}B(t).$$

** exercise

Taking expectations of both sides, we have an ordinary differential equation in the expected values.

Ito's Formula in Jump-Diffusion Processes

Now suppose we are interested in a process defined by a function g of $S(t)$ and t. This is where Ito's formula is used.

The simple approach is to apply Ito's formula directly to the drift-diffusion part and then consider $\mathrm{d_j}g(t)$ separately. (We have absorbed $S(t)$ into t in the notation $g(t)$.)

As before,we consider the random variable of the magnitude of the change, Δg and write the process as a systematic component plus a random component

$$\begin{aligned}\mathrm{d_j}g(t) &= g(t) - g(t^-) \\ &= \left(\lambda(S(t),t)\int_{\mathcal{D}(\Delta g)} p_{\Delta g}(\Delta g; g(t))\mathrm{d}\Delta g\right)\mathrm{d}t + \mathrm{d_j}J_g(t)\end{aligned}$$

where the random component $\mathrm{d_j}J_g(t)$ is a compensated process as before.

Putting this all together we have

$$\begin{aligned}\mathrm{d}g(t) = \Bigg(&\frac{\partial g}{\partial t} + \mu\frac{\partial g}{\partial S} + \frac{1}{2}\sigma^2\frac{\partial^2 g}{\partial S^2} \\ &+\lambda(t)\int_{\mathcal{D}(\Delta g)} \Delta g\, p_{\Delta g}(\Delta g; g(t))\mathrm{d}\Delta g\Bigg)\mathrm{d}t \\ &+\frac{\partial g}{\partial S}\sigma\mathrm{d}B(t) \\ &+\mathrm{d_j}J_g(t).\end{aligned}$$

We must remember that this is a discontinuous process.

Notes and References for Section 0.2

Study of continuous stochastic processes, such as Brownian motion, requires real analysis using random infinitesimals. This area of statistical mathematics is sometimes called stochastic calculus. Some useful texts on stochastic processes and the stochastic calculus are Bass (2011), Karatzas and Shreve (1991), Øksendal (1998), and Rogers and Williams (2000a), Rogers and Williams (2000b).

Stochastic calculus is widely used in models of prices of financial assets, and many of the developments in the general theory have come from that area of application; see Steele (2001) for example.

0.3 Some Basics of Linear Algebra

In the following we will assume the usual axioms for the reals, $\mathbb{R}$. We will be concerned with two linear structures on $\mathbb{R}$. We denote one as $\mathbb{R}^n$, and call its members *vectors*. We denote another as $\mathbb{R}^{n\times m}$, and call its members *matrices*. For both structures we have scalar multiplication (multiplication of a member of the structure by a member of $\mathbb{R}$), an addition operation, an additive identity, and additive inverses for all elements. The addition operation is denoted by "+" and the additive identity by "0", which are the same two symbols used similarly in $\mathbb{R}$. We also have various types of multiplication operations, all with identities, and some with inverses. In addition, we define various real-valued functions over these structures, the most important of which are inner products and norms.

Both $\mathbb{R}^n$ and $\mathbb{R}^{n\times m}$ with addition and multiplication operations are *linear spaces*.

In this section, we abstract some of the basic material on linear algebra from Gentle (2007).

0.3.1 Inner Products, Norms, and Metrics

Although various inner products could be defined in $\mathbb{R}^n$, "the" inner product or dot product for vectors x and y in $\mathbb{R}^n$ is defined as $\sum_{i=1}^n x_i y_i$, and is often written as $x^{\mathrm{T}}y$. It is easy to see that this satisfies the definition of an inner product (see page 636).

Two elements $x, y \in \mathbb{R}^n$ are said to be *orthogonal* if $\langle x, y\rangle = 0$.

An element $x \in \mathbb{R}^n$ is said to be *normal* or *normalized* if $\langle x, x\rangle = 1$. Any $x \neq 0$ can be normalized, that is, mapped to a normal element, $x/\langle x, x\rangle$. A set of normalized elements that are pairwise orthogonal is called an *orthonormal* set. (On page 685 we discuss a method of forming a set of orthogonal vectors.)

Various inner products could be defined in $\mathbb{R}^{n\times m}$, but "the" inner product or dot product for matrices A and B in $\mathbb{R}^{n\times m}$ is defined as $\sum_{j=1}^m a_j^{\mathrm{T}} b_j$, where a_j is the vector whose elements are those from the j^{th} column of A, and likewise for b_j. Again, it is easy to see that this satisfies the definition of an inner product.

Norms and Metrics

There are various norms that can be defined on $\mathbb{R}^n$. An important class of norms are the L_p norms, defined for $p \geq 1$ by

$$\|x\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}. \tag{0.3.1}$$

It is easy to see that this satisfies the definition of a norm (see page 641).

The norm in $\mathbb{R}^n$ induced by the inner product (that is, "the" inner product) is the Euclidean norm or the L_2 norm:

$$\|x\|_2 = \sqrt{\langle x, x\rangle} = \sqrt{\sum_{i=1}^{n} x_i^2}. \tag{0.3.2}$$

This is the only L_p norm induced by an inner product.

The norm in $\mathbb{R}^{n\times m}$ induced by the inner product exists only for $n = m$. In that case it is $\|A\| = \sum_{j=1}^{n} a_j^{\mathrm{T}} a_j = \sum_{j=1}^{n}\sum_{i=1}^{n} a_{ij}^2$. Note that this is not the L_2 matrix norm; it is the Frobenius norm (see below).

The most common and useful metrics in $\mathbb{R}^n$ and $\mathbb{R}^{n\times m}$ are those induced by the norms. For $\mathbb{R}^n$ the L_2 norm is the most common, and a metric for $x, y \in \mathbb{R}^n$ is defined as

$$\rho(x, y) = \|x - y\|_2. \tag{0.3.3}$$

This metric is called the Euclidean distance.

0.3.2 Matrices and Vectors

Vectors are n-tuples and matrices are n by m rectangular arrays. We will be interested in vectors and matrices whose elements are real numbers. We denote the set of such vectors as $\mathbb{R}^n$ and the set of such matrics as $\mathbb{R}^{n\times m}$.

We generally denote a member of $\mathbb{R}^{n\times m}$ by an upper case letter. A member of $\mathbb{R}^{n\times m}$ consists of nm elements, which we denote by use of two subscripts. We often use a lower-case letter with the two subscripts. For example, for a matrix A, we denote the elements as A_{ij} or a_{ij} with $i = 1, \ldots, n$ and $j = 1, \ldots, m$.

The transpose of a matrix A in $\mathbb{R}^{n\times m}$ is a matrix in $\mathbb{R}^{m\times n}$ denoted by A^{T} such that $(A^{\mathrm{T}})_{ij} = A_{ji}$. Note that this is consistent with the use of $^{\mathrm{T}}$ above for vectors.

If $n = m$ the matrix is *square*.

We define (Cayley) multiplication of the matrix $A \in \mathbb{R}^{n\times m}$ and the matrix $B \in \mathbb{R}^{m\times p}$ as $C = AB \in \mathbb{R}^{n\times p}$, where $c_{ij} = \sum_{k=1}^{m} a_{ik}b_{kj}$.

If x and y are n-vectors, in most cases, we can consider them to be $n \times 1$ matrices. Hence, $x^{\mathrm{T}}y$ is a 1×1 matrix and xy^{T} is an $n \times n$ matrix.

We see from the definition that $x^{\mathrm{T}}y$ is an inner product. This inner product is also called the dot product. The product xy^{T} is called the outer product.

As above, we see that $\sqrt{x^{\mathrm{T}}x}$ is a norm (it is the induced norm). We sometimes denote this norm as $\|x\|_2$, because it is $\left(\sum_{i=1}^{n} |x_i|^2\right)^{1/2}$. We call it the Euclidean norm and also the L_2 norm. More generally, for $p \geq 1$, we define the L_p norm for the n-vector x as $\left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}$.

We denote the L_p norm of x as $\|x\|_p$. We generally denote the Euclidean or L_2 norm simply as $\|x\|$.

Properties, Concepts, and Notation Associated with Matrices and Vectors

Definition 0.3.1 (linear independence)
A set of vectors $x_1, \ldots, x_n \in \mathbb{R}^n$ is said to be *linearly independent* if given $a_i \in \mathbb{R}$, $\sum_{i=1}^n a_i x_i = 0$ implies $a_i = 0$ for $i = 1, \ldots, n$. ∎

Definition 0.3.2 (rank of a matrix)
The rank of a matrix is the maximum number of rows or columns that are linearly independent. (The maximum number of rows that are linearly independent is the same as the maximum number of columns that are linearly independent.) For the matrix A, we write $\mathrm{rank}(A)$. We adopt the convention that $\mathrm{rank}(A) = 0 \Leftrightarrow A = 0$ (the zero matrix). $A \in \mathbb{R}^{n\times m}$ is said to be full rank iff $\mathrm{rank}(A) = \min(n, m)$. ∎

An important fact is

$$\mathrm{rank}(AB) \leq \min(\mathrm{rank}(A), \mathrm{rank}(B)),$$

and a consequence of this is that the rank of an outer product is less than or equal to 1.

To define the determinant of a matrix, we first need to consider permuations of the integers from 1 to n. Let the list of integers $\Pi_j = (j_1, j_2, \ldots, j_n)$ be one of the $n!$ permutations of the integers from 1 to n. Define a permutation to be *even* or *odd* according to the number of times that a smaller element follows a larger one in the permutation. (For example, 1, 3, 2 is an odd permutation, and 3, 1, 2 is an even permutation.) Let $\sigma(\Pi_j) = 1$ if $\Pi_j = (j_1, \ldots, j_n)$ is an even permutation, and let $\sigma(\Pi_j) = -1$ otherwise.

Definition 0.3.3 (determinant of a square matrix)
The *determinant* of an $n \times n$ (square) A, denoted by $|A|$, is defined by

$$|A| = \sum_{\text{all permutations}} \sigma(\Pi_j) a_{1j_1} \cdots a_{nj_n}.$$

∎

The determinant is a real number. We write $|A|$ or $\det(A)$. $|A| \neq 0$ iff A is square and of full rank.

Definition 0.3.4 (identity matrix)
$I \in \mathbb{R}^{n\times n}$ and $I[i, j] = 0$ if $i \neq j$ and $I[i, j] = 1$ if $i \neq i$; that is $I[i, j] = \delta_{ij}$, where δ_{ij} is the Kronecker delta. We write the identity as I_n or just I. ∎

Definition 0.3.5 (inverse of a matrix)
For $A \in \mathbb{R}^{n\times n}$, if a matrix $B \in \mathbb{R}^{n\times n}$ exists, such that $AB = I$, then B is the inverse of A, and is written A^{-1}. ∎

A matrix has an inverse iff it is square and of full rank.

Definition 0.3.6 (generalized inverse of a matrix)
For $A \in \mathbb{R}^{n\times m}$, a matrix $B \in \mathbb{R}^{m\times n}$ such that $ABA = A$ is called a generalized inverse of A, and is written A^-. ∎

If A is nonsingular (square and full rank), then obviously $A^- = A^{-1}$.

Definition 0.3.7 (pseudoinverse or Moore-Penrose inverse of a matrix)

For $A \in \mathbb{R}^{n\times m}$, the matrix $B \in \mathbb{R}^{m\times n}$ such that $ABA = A$, $BAB = B$, $(AB)^{\mathrm{T}} = AB$, and $(BA)^{\mathrm{T}} = BA$ is called the pseudoinverse of A, and is written A^+. ∎

Definition 0.3.8 (orthogonal matrix)
For $A \in \mathbb{R}^{n\times m}$, if $A^{\mathrm{T}}A = I_m$, that is, if the columns are orthonormal and $m \leq n$, or $AA^{\mathrm{T}} = I_n$, that is, if the rows are orthonormal and $n \leq m$, then A is said to be orthogonal. ∎

Definition 0.3.9 (quadratic forms)
For $A \in \mathbb{R}^{n\times n}$ and $x \in \mathbb{R}^n$, the scalar $x^{\mathrm{T}}Ax$ is called a quadratic form. ∎

Definition 0.3.10 (nonnegative definite matrix)
For $A \in \mathbb{R}^{n\times n}$ and any $x \in \mathbb{R}^n$, if $x^{\mathrm{T}}Ax \geq 0$, then is said to be nonnegative definite. We generally restrict the definition to symmetric matrices. This is essentially without loss of generality because if a matrix is nonnegative definite, then there is a similar symmetric matrix. (Two matrices are said to be *similar* if they have exactly the same eigenvalues.) We write $A \succeq 0$ to denote that A is nonnegative definite. ∎

Definition 0.3.11 (positive definite matrix)
For $A \in \mathbb{R}^{n\times n}$ and any $x \in \mathbb{R}^n$, if $x^{\mathrm{T}}Ax \geq 0$ and $x^{\mathrm{T}}Ax = 0$ implies $x = 0$, then is said to be positive definite. As with nonnegative definite matrices, we generally restrict the definition of positive definite matrices to symmetric matrices. We write $A \succ 0$ to denote that A is positive definite. ∎

Definition 0.3.12 (eigenvalues and eigenvectors) *If $A \in \mathbb{R}^{n\times n}$, v is an n-vector (complex), and c is a scalar (complex), and $Av = cv$, then c is an eigenvalue of A and v is an eigenvector of A associated with c.* ∎

The Trace and Some of Its Properties

Definition 0.3.13 (trace of a matrix)
The sum of the diagonal elements of a square matrix is called the *trace* of the matrix. ∎

We use the notation "$\mathrm{tr}(A)$" to denote the trace of the matrix A:

$$\mathrm{tr}(A) = \sum_i a_{ii}.$$

Some properties of the trace that follow immediately from the definition:

$$\mathrm{tr}(A) = \mathrm{tr}(A^{\mathrm{T}}).$$

For a scalar c and an $n \times n$ matrix A,

$$\mathrm{tr}(cA) = c\,\mathrm{tr}(A).$$

If A and B are such that both AB and BA are defined,

$$\mathrm{tr}(AB) = \mathrm{tr}(BA).$$

If x is a vector, we have

$$\|x\|^2 = x^{\mathrm{T}}x = \mathrm{tr}(x^{\mathrm{T}}x) = \mathrm{tr}(xx^{\mathrm{T}}).$$

If x is a vector and A a matrix, we have

$$x^{\mathrm{T}}Ax = \mathrm{tr}(x^{\mathrm{T}}Ax) = \mathrm{tr}(Axx^{\mathrm{T}}).$$

Eigenanalysis of Symmetric Matrices

The eigenvalues and eigenvectors of symmetric matrices have some interesting properties. First of all, for a real symmetric matrix, the eigenvalues are all real. Symmetric matrices are diagonalizable; therefore all of the properties of diagonalizable matrices carry over to symmetric matrices.

Orthogonality of Eigenvectors

In the case of a symmetric matrix A, any eigenvectors corresponding to distinct eigenvalues are orthogonal. This is easily seen by assuming that c_1 and c_2 are unequal eigenvalues with corresponding eigenvectors v_1 and v_2. Now consider $v_1^{\mathrm{T}}v_2$. Multiplying this by c_2, we get

$$c_2v_1^{\mathrm{T}}v_2 = v_1^{\mathrm{T}}Av_2 = v_2^{\mathrm{T}}Av_1 = c_1v_2^{\mathrm{T}}v_1 = c_1v_1^{\mathrm{T}}v_2.$$

Because $c_1 \neq c_2$, we have $v_1^{\mathrm{T}}v_2 = 0$.

Now, consider two eigenvalues $c_i = c_j$, that is, an eigenvalue of multiplicity greater than 1 and distinct associated eigenvectors v_i and v_j. By what we just saw, an eigenvector associated with $c_k \neq c_i$ is orthogonal to the space spanned by v_i and v_j. Assume v_i is normalized and apply a Gram-Schmidt transformation to form

$$\tilde{v}_j = \frac{1}{\|v_j - \langle v_i, v_j\rangle v_i\|}(v_j - \langle v_i, v_j\rangle v_i),$$

yielding a vector orthogonal to v_i. Now, we have

$$\begin{aligned} A\tilde{v}_j &= \frac{1}{\|v_j - \langle v_i, v_j\rangle v_i\|}(Av_j - \langle v_i, v_j\rangle Av_i) \\ &= \frac{1}{\|v_j - \langle v_i, v_j\rangle v_i\|}(c_j v_j - \langle v_i, v_j\rangle c_i v_i) \\ &= c_j \frac{1}{\|v_j - \langle v_i, v_j\rangle v_i\|}(v_j - \langle v_i, v_j\rangle v_i) \\ &= c_j \tilde{v}_j; \end{aligned}$$

hence, $\tilde{v}_j$ is an eigenvector of A associated with c_j. We conclude therefore that the eigenvectors of a symmetric matrix can be chosen to be orthogonal.

A symmetric matrix is orthogonally diagonalizable, and because the eigenvectors can be chosen to be orthogonal, and can be written as

$$A = VCV^{\mathrm{T}}, \tag{0.3.4}$$

where $VV^{\mathrm{T}} = V^{\mathrm{T}}V = I$, and so we also have

$$V^{\mathrm{T}}AV = C. \tag{0.3.5}$$

Such a matrix is orthogonally similar to a diagonal matrix formed from its eigenvalues.

Spectral Decomposition

When A is symmetric and the eigenvectors v_i are chosen to be orthonormal,

$$I = \sum_i v_i v_i^{\mathrm{T}}, \tag{0.3.6}$$

so

$$\begin{aligned} A &= A\sum_i v_i v_i^{\mathrm{T}} \\ &= \sum_i Av_i v_i^{\mathrm{T}} \\ &= \sum_i c_i v_i v_i^{\mathrm{T}}. \end{aligned} \tag{0.3.7}$$

This representation is called the *spectral decomposition* of the symmetric matrix A. It is essentially the same as equation (0.3.4), so $A = VCV^{\mathrm{T}}$ is also called the spectral decomposition.

The representation is unique except for the ordering and the choice of eigenvectors for eigenvalues with multiplicities greater than 1. If the rank of the matrix is r, we have $|c_1| \geq \cdots \geq |c_r| > 0$, and if $r < n$, then $c_{r+1} = \cdots = c_n = 0$.

Note that the matrices in the spectral decomposition are projection matrices that are orthogonal to each other (but they are not orthogonal matrices) and they sum to the identity. Let

$$P_i = v_i v_i^{\mathrm{T}}. \tag{0.3.8}$$

Then we have

$$P_i P_i = P_i, \tag{0.3.9}$$

$$P_i P_j = 0 \text{ for } i \neq j, \tag{0.3.10}$$

$$\sum_i P_i = I, \tag{0.3.11}$$

and the spectral decomposition,

$$A = \sum_i c_i P_i. \tag{0.3.12}$$

The P_i are called *spectral projectors*.

The spectral decomposition also applies to powers of A,

$$A^k = \sum_i c_i^k v_i v_i^{\mathrm{T}}, \tag{0.3.13}$$

where k is an integer. If A is nonsingular, k can be negative in the expression above.

The spectral decomposition is one of the most important tools in working with symmetric matrices.

Although we will not prove it here, all diagonalizable matrices have a spectral decomposition in the form of equation (0.3.12) with projection matrices that satisfy properties (0.3.9) through (0.3.11). These projection matrices cannot necessarily be expressed as outer products of eigenvectors, however. The eigenvalues and eigenvectors of a nonsymmetric matrix might not be real, the left and right eigenvectors might not be the same, and two eigenvectors might not be mutually orthogonal. In the spectral representation $A = \sum_i c_i P_i$, however, if c_j is a simple eigenvalue with associated left and right eigenvectors y_j and x_j, respectively, then the projection matrix P_j is $x_j y_j^{\mathrm{H}} / y_j^{\mathrm{H}} x_j$. (Note that because the eigenvectors may not be real, we take the conjugate transpose.)

Quadratic Forms and the Rayleigh Quotient

Equation (0.3.7) yields important facts about quadratic forms in A. Because V is of full rank, an arbitrary vector x can be written as Vb for some vector b. Therefore, for the quadratic form $x^{\mathrm{T}}Ax$ we have

$$\begin{aligned} x^{\mathrm{T}}Ax &= x^{\mathrm{T}}\sum_i c_i v_i v_i^{\mathrm{T}} x \\ &= \sum_i b^{\mathrm{T}} V^{\mathrm{T}} v_i v_i^{\mathrm{T}} V b c_i \\ &= \sum_i b_i^2 c_i. \end{aligned}$$

This immediately gives the inequality

$$x^{\mathrm{T}}Ax \leq \max\{c_i\} b^{\mathrm{T}} b.$$

(Notice that $\max\{c_i\}$ here is not necessarily c_1; in the important case when all of the eigenvalues are nonnegative, it is, however.) Furthermore, if $x \neq 0$, $b^{\mathrm{T}}b = x^{\mathrm{T}}x$, and we have the important inequality

$$\frac{x^{\mathrm{T}}Ax}{x^{\mathrm{T}}x} \leq \max\{c_i\}. \tag{0.3.14}$$

Equality is achieved if x is the eigenvector corresponding to $\max\{c_i\}$, so we have

$$\max_{x \neq 0} \frac{x^{\mathrm{T}}Ax}{x^{\mathrm{T}}x} = \max\{c_i\}. \tag{0.3.15}$$

If $c_1 > 0$, this is the spectral radius, $\rho(A)$.

*** prove the following add to matrix book

$$\max_x \frac{x^{\mathrm{T}}Ax}{x^{\mathrm{T}}Cx} = \max(\text{e.v.})(C^{-1}A). \tag{0.3.16}$$

or if $A = aa^{\mathrm{T}}$, then

$$\max_x \frac{x^{\mathrm{T}}Ax}{x^{\mathrm{T}}Cx} = a^{\mathrm{T}}C^{-1}a. \tag{0.3.17}$$

The expression on the left-hand side in (0.3.14) as a function of x is called the *Rayleigh quotient* of the symmetric matrix A and is denoted by $R_A(x)$:

$$\begin{aligned} R_A(x) &= \frac{x^{\mathrm{T}}Ax}{x^{\mathrm{T}}x} \\ &= \frac{\langle x, Ax\rangle}{\langle x, x\rangle}. \end{aligned} \tag{0.3.18}$$

Because if $x \neq 0$, $x^{\mathrm{T}}x > 0$, it is clear that the Rayleigh quotient is nonnegative for all x if and only if A is nonnegative definite and is positive for all x if and only if A is positive definite.

The Fourier Expansion

The $v_i v_i^{\mathrm{T}}$ matrices in equation (0.3.7) have the property that $\langle v_i v_i^{\mathrm{T}}, v_j v_j^{\mathrm{T}} \rangle = 0$ for $i \neq j$ and $\langle v_i v_i^{\mathrm{T}}, v_i v_i^{\mathrm{T}} \rangle = 1$, and so the spectral decomposition is a Fourier expansion and the eigenvalues are Fourier coefficients. Because of orthogonality, the eigenvalues can be represented as the dot product

$$c_i = \langle A, v_i v_i^{\mathrm{T}} \rangle. \tag{0.3.19}$$

The eigenvalues c_i have the same properties as the Fourier coefficients in any orthonormal expansion. In particular, the best approximating matrices within the subspace of $n \times n$ symmetric matrices spanned by $\{v_1 v_1^{\mathrm{T}}, \ldots, v_n v_n^{\mathrm{T}}\}$ are partial sums of the form of equation (0.3.7).

Powers of a Symmetric Matrix

If (c, v) is an eigenpair of the symmetric matrix A with $v^{\mathrm{T}} v = 1$, then for any $k = 1, 2, \ldots,$

$$\left(A - cvv^{\mathrm{T}}\right)^k = A^k - c^k vv^{\mathrm{T}}. \tag{0.3.20}$$

This follows from induction on k, for it clearly is true for $k = 1$, and if for a given k it is true that for $k - 1$

$$\left(A - cvv^{\mathrm{T}}\right)^{k-1} = A^{k-1} - c^{k-1} vv^{\mathrm{T}},$$

then by multiplying both sides by $(A - cvv^{\mathrm{T}})$, we see it is true for k:

$$\begin{aligned}
\left(A - cvv^{\mathrm{T}}\right)^k &= \left(A^{k-1} - c^{k-1} vv^{\mathrm{T}}\right)\left(A - cvv^{\mathrm{T}}\right) \\
&= A^k - c^{k-1} vv^{\mathrm{T}} A - cA^{k-1} vv^{\mathrm{T}} + c^k vv^{\mathrm{T}} \\
&= A^k - c^k vv^{\mathrm{T}} - c^k vv^{\mathrm{T}} + c^k vv^{\mathrm{T}} \\
&= A^k - c^k vv^{\mathrm{T}}.
\end{aligned}$$

There is a similar result for nonsymmetric square matrices, where w and v are left and right eigenvectors, respectively, associated with the same eigenvalue c that can be scaled so that $w^{\mathrm{T}} v = 1$. (Recall that an eigenvalue of A is also an eigenvalue of A^{T}, and if w is a left eigenvector associated with the eigenvalue c, then $A^{\mathrm{T}} w = cw$.) The only property of symmetry used above was that we could scale $v^{\mathrm{T}} v$ to be 1; hence, we just need $w^{\mathrm{T}} v \neq 0$. This is clearly true for a diagonalizable matrix (from the definition). It is also true if c is simple (which is somewhat harder to prove).

If w and v are left and right eigenvectors of A associated with the same eigenvalue c and $w^{\mathrm{T}} v = 1$, then for $k = 1, 2, \ldots,$

$$\left(A - cvw^{\mathrm{T}}\right)^k = A^k - c^k vw^{\mathrm{T}}. \tag{0.3.21}$$

We can prove this by induction as above.

The Trace and Sums of Eigenvalues

For a general $n \times n$ matrix A with eigenvalues $c_1, \ldots, c_n$, we have $\mathrm{tr}(A) = \sum_{i=1}^n c_i$. This is particularly easy to see for symmetric matrices because of equation (0.3.4), rewritten as $V^{\mathrm{T}}AV = C$, the diagonal matrix of the eigenvalues. For a symmetric matrix, however, we have a stronger result.

If A is an $n \times n$ symmetric matrix with eigenvalues $c_1 \geq \cdots \geq c_n$, and U is an $n \times k$ orthogonal matrix, with $k \leq n$, then

$$\mathrm{tr}(U^{\mathrm{T}}AU) \leq \sum_{i=1}^{k} c_i. \tag{0.3.22}$$

To see this, we represent U in terms of the columns of V, which span $\mathbb{R}^n$, as $U = VX$. Hence,

$$\begin{aligned} \mathrm{tr}(U^{\mathrm{T}}AU) &= \mathrm{tr}(X^{\mathrm{T}}V^{\mathrm{T}}AVX) \\ &= \mathrm{tr}(X^{\mathrm{T}}CX) \\ &= \sum_{i=1}^{n} x_i^{\mathrm{T}} x_i \, c_i, \end{aligned} \tag{0.3.23}$$

where x_i^{T} is the i^{th} row of X.

Now $X^{\mathrm{T}}X = X^{\mathrm{T}}V^{\mathrm{T}}VX = U^{\mathrm{T}}U = I_k$, so either $x_i^{\mathrm{T}}x_i = 0$ or $x_i^{\mathrm{T}}x_i = 1$, and $\sum_{i=1}^n x_i^{\mathrm{T}}x_i = k$. Because $c_1 \geq \cdots \geq c_n$, therefore $\sum_{i=1}^n x_i^{\mathrm{T}}x_i \, c_i \leq \sum_{i=1}^k c_i$, and so from equation (0.3.23) we have $\mathrm{tr}(U^{\mathrm{T}}AU) \leq \sum_{i=1}^k c_i$.

0.3.2.1 Positive Definite and Nonnegative Definite Matrices

The factorization of symmetric matrices in equation (0.3.4) yields some useful properties of positive definite and nonnegative definite matrices.

Eigenvalues of Positive and Nonnegative Definite Matrices

In this book, we use the terms "nonnegative definite" and "positive definite" only for real symmetric matrices, so the eigenvalues of nonnegative definite or positive definite matrices are real.

Any real symmetric matrix is positive (nonnegative) definite if and only if all of its eigenvalues are positive (nonnegative). We can see this using the factorization (0.3.4) of a symmetric matrix. One factor is the diagonal matrix C of the eigenvalues, and the other factors are orthogonal. Hence, for any x, we have $x^{\mathrm{T}}Ax = x^{\mathrm{T}}VCV^{\mathrm{T}}x = y^{\mathrm{T}}Cy$, where $y = V^{\mathrm{T}}x$, and so

$$x^{\mathrm{T}}Ax > (\geq)\, 0$$

if and only if

$$y^{\mathrm{T}}Cy > (\geq)\, 0.$$

This implies that if P is a nonsingular matrix and D is a diagonal matrix, $P^{\mathrm{T}}DP$ is positive (nonnegative) if and only if the elements of D are positive (nonnegative).

A matrix (whether symmetric or not and whether real or not) all of whose eigenvalues have positive real parts is said to be *positive stable*. Positive stability is an important property in some applications, such as numerical solution of systems of nonlinear differential equations. Clearly, a positive definite matrix is positive stable.

Inverse of Positive Definite Matrices

If A is positive definite and $A = VCV^{\mathrm{T}}$ as in equation (0.3.4), then $A^{-1} = VC^{-1}V^{\mathrm{T}}$ and A^{-1} is positive definite because the elements of C^{-1} are positive.

Diagonalization of Positive Definite Matrices

If A is positive definite, the elements of the diagonal matrix C in equation (0.3.4) are positive, and so their square roots can be absorbed into V to form a nonsingular matrix P. The diagonalization in equation (0.3.5), $V^{\mathrm{T}}AV = C$, can therefore be reexpressed as

$$P^{\mathrm{T}}AP = I. \tag{0.3.24}$$

Square Roots of Positive and Nonnegative Definite Matrices

The factorization (0.3.4) together with the nonnegativity of the eigenvalues of positive and nonnegative definite matrices allows us to define a square root of such a matrix.

Let A be a nonnegative definite matrix and let V and C be as in equation (0.3.4): $A = VCV^{\mathrm{T}}$. Now, let S be a diagonal matrix whose elements are the square roots of the corresponding elements of C. Then $(VSV^{\mathrm{T}})^2 = A$; hence, we write

$$A^{\frac{1}{2}} = VSV^{\mathrm{T}} \tag{0.3.25}$$

and call this matrix the *square root* of A. We also can similarly define $A^{\frac{1}{r}}$ for $r > 0$.

We see immediately that $A^{\frac{1}{2}}$ is symmetric because A is symmetric.

If A is positive definite, A^{-1} exists and is positive definite. It therefore has a square root, which we denote as $A^{-\frac{1}{2}}$.

The square roots are nonnegative, and so $A^{\frac{1}{2}}$ is nonnegative definite. Furthermore, $A^{\frac{1}{2}}$ and $A^{-\frac{1}{2}}$ are positive definite if A is positive definite.

This $A^{\frac{1}{2}}$ is unique, so our reference to it as the square root is appropriate. (There is occasionally some ambiguity in the terms "square root" and "second

root" and the symbols used to denote them. If x is a nonnegative scalar, the usual meaning of its square root, denoted by $\sqrt{x}$, is a nonnegative number, while its second roots, which may be denoted by $x^{\frac{1}{2}}$, are usually considered to be either of the numbers $\pm\sqrt{x}$. In our notation $A^{\frac{1}{2}}$, we mean *the* square root; that is, the nonnegative matrix, if it exists. Otherwise, we say the square root of the matrix does not exist. For example, $I_2^{\frac{1}{2}} = I_2$, and while if $J = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$, $J^2 = I_2$, we do not consider J to be a square root of I_2.)

Forming a Vector from the Elements of a Matrix: vec(·), vecsy(·), and vech(·)

It is sometimes useful to consider the elements of a matrix to be elements of a single vector. The most common way this is done is to string the columns of the matrix end-to-end into a vector. The vec(·) function does this:

$$\text{vec}(A) = (a_1^{\mathrm{T}}, a_2^{\mathrm{T}}, \ldots, a_m^{\mathrm{T}}), \tag{0.3.26}$$

where $a_1, a_2, \ldots, a_m$ are the column vectors of the matrix A. The vec function is also sometimes called the "pack" function. The vec function is a mapping $\mathbb{R}^{n\times m} \mapsto \mathbb{R}^{nm}$.

For a symmetric matrix A with elements a_{ij}, the "vecsy" function stacks the unique elements into a vector:

$$\text{vecsy}(A) = (a_{11}, a_{21}, a_{22}, a_{31}, a_{32}, a_{33}, \ldots, a_{n1}, a_{n2}, \ldots, a_{nn}). \tag{0.3.27}$$

The vecsy function is called the V^2 function by Kollo and von Rosen (2005). It is the "symmetric storage mode" used by numerical analysts since the 1950s.

There are other ways that the unique elements could be stacked. The "vech" function is

$$\text{vech}(A) = (a_{11}, a_{21}, \ldots, a_{n1}, a_{22}, \ldots, a_{n2}, \ldots, a_{nn}). \tag{0.3.28}$$

The vecsy and vech functions are mappings $\mathbb{R}^{n\times n} \mapsto \mathbb{R}^{n(n+1)/2}$.

The Kronecker Product

Kronecker multiplication, denoted by $\otimes$, is defined for any two matrices $A_{n\times m}$ and $B_{p\times q}$ as

$$A \otimes B = \begin{bmatrix} a_{11}B & \ldots & a_{1m}B \\ \vdots & \ldots & \vdots \\ a_{n1}B & \ldots & a_{nm}B \end{bmatrix}.$$

The Kronecker product of A and B is $np \times mq$; that is, Kronecker matrix multiplication is a mapping

$$\mathbb{R}^{n\times m} \times \mathbb{R}^{p\times q} \mapsto \mathbb{R}^{np\times mq}.$$

The Kronecker product is also called the "right direct product" or just *direct product*. (A left direct product is a Kronecker product with the factors reversed.)

Kronecker multiplication is not commutative, but it is associative and it is distributive over addition, as we will see below.

The identity for Kronecker multiplication is the 1×1 matrix with the element 1; that is, it is the same as the scalar 1.

The determinant of the Kronecker product of two square matrices $A_{n\times n}$ and $B_{m\times m}$ has a simple relationship to the determinants of the individual matrices:

$$|A \otimes B| = |A|^m |B|^n. \tag{0.3.29}$$

The proof of this, like many facts about determinants, is straightforward but involves tedious manipulation of cofactors. The manipulations in this case can be facilitated by using the vec-permutation matrix.

We can understand the properties of the Kronecker product by expressing the (i, j) element of $A \otimes B$ in terms of the elements of A and B,

$$(A \otimes B)_{i,j} = A_{[i/p]+1,\,[j/q]+1} B_{i-p[i/p],\,j-q[i/q]}, \tag{0.3.30}$$

where $[\cdot]$ is the greatest integer function.

Some additional properties of Kronecker products that are immediate results of the definition are, assuming the matrices are conformable for the indicated operations,

$$\begin{aligned}(aA) \otimes (bB) &= ab(A \otimes B)\\ &= (abA) \otimes B\\ &= A \otimes (abB), \text{ for scalars } a, b,\end{aligned} \tag{0.3.31}$$

$$(A + B) \otimes (C) = A \otimes C + B \otimes C, \tag{0.3.32}$$

$$(A \otimes B) \otimes C = A \otimes (B \otimes C), \tag{0.3.33}$$

$$(A \otimes B)^{\mathrm{T}} = A^{\mathrm{T}} \otimes B^{\mathrm{T}}, \tag{0.3.34}$$

$$(A \otimes B)(C \otimes D) = AC \otimes BD. \tag{0.3.35}$$

These properties are all easy to see by using equation (0.3.30) to express the (i, j) element of the matrix on either side of the equation, taking into account the size of the matrices involved. For example, in the first equation, if A is $n \times m$ and B is $p \times q$, the (i, j) element on the left-hand side is

$$aA_{[i/p]+1,\,[j/q]+1} bB_{i-p[i/p],\,j-q[i/q]}$$

and that on the right-hand side is

$$abA_{[i/p]+1,\,[j/q]+1}B_{i-p[i/p],\,j-q[i/q]}.$$

Another property of the Kronecker product of square matrices is

$$\mathrm{tr}(A \otimes B) = \mathrm{tr}(A)\mathrm{tr}(B). \tag{0.3.36}$$

This is true because the trace of the product is merely the sum of all possible products of the diagonal elements of the individual matrices.

The Kronecker product and the vec function often find uses in the same application. For example, an $n \times m$ normal random matrix X with parameters M, Σ, and Ψ can be expressed in terms of an ordinary np-variate normal random variable $Y = \mathrm{vec}(X)$ with parameters $\mathrm{vec}(M)$ and $\Sigma \otimes \Psi$.

A relationship between the vec function and Kronecker multiplication is

$$\mathrm{vec}(ABC) = (C^{\mathrm{T}} \otimes A)\mathrm{vec}(B) \tag{0.3.37}$$

for matrices A, B, and C that are conformable for the multiplication indicated.

Matrix Factorizations

There are a number of useful ways of factorizing a matrix.

- the LU (and LR and LDU) factorization of a general matrix:
- the QR factorization of a general matrix,
- the similar canonical factorization or "diagonal factorization" of a diagonalizable matrix (which is necessarily square):

  $$A = VCV^{-1},$$

  where V is a matrix whose columns correspond to the eigenvectors of A and is nonsingular, and C is a diagonal matrix whose entries are the eigenvalues corresponding to the columns of V.
- the singular value factorization of a general $n \times m$ matrix A:

  $$A = UDV^{\mathrm{T}},$$

  where U is an $n \times n$ orthogonal matrix, V is an $m \times m$ orthogonal matrix, and D is an $n \times m$ diagonal matrix with nonnegative entries. (An $n \times m$ diagonal matrix has $\min(n, m)$ elements on the diagonal, and all other entries are zero.)
- the square root of a nonnegative definite matrix A (which is necessarily symmetric):

  $$A = A^{1/2}A^{1/2}$$
- the Cholesky factorization of a nonnegative definite matrix:

  $$A = A_{\mathrm{c}}^{\mathrm{T}}A_{\mathrm{c}},$$

  where A_{c} is an upper triangular matrix with nonnegative diagonal elements.

Spectral Decomposition

For a symmetric matrix A, we can always write $A = VCV^{\mathrm{T}}$, as above. This is called the spectral decomposition, and is unique except for the ordering and the choice of eigenvectors for eigenvalues with multiplicities greater than 1. We can also write

$$A = \sum_i c_i P_i,$$

where the P_i are the outer products of the eigenvectors,

$$P_i = v_i v_i^{\mathrm{T}},$$

and are called spectral projectors.

Matrix Norms

A matrix norm is generally required to satisfy one more property in addition to those listed above for the definition of a norm. It is the consistency property: $\|AB\| \leq \|A\| \, \|B\|$. The L_p matrix norm for the $n \times m$ matrix A is defined as

$$\|A\|_p = \max_{\|x\|_p=1} \|Ax\|_p.$$

The L_2 matrix norm has the interesting relationship

$$\|A\|_2 = \sqrt{\rho(A^{\mathrm{T}}A)},$$

where $\rho(\cdot)$ is the spectral radius (the modulus of the eigenvalue with the maximum modulus).

The "usual" matrix norm is the Frobenius norm:

$$\|A\|_{\mathrm{F}} = \sqrt{\sum_{i,j} a_{ij}^2}.$$

Idempotent and Projection Matrices

A matrix A such that $AA = A$ is called an *idempotent matrix*. An idempotent matrix is square, and it is either singular or it is the identity matrix. (It must be square in order to be conformable for the indicated multiplication. If it is not singular, we have $A = (A^{-1}A)A = A^{-1}(AA) = A^{-1}A = I$; hence, an idempotent matrix is either singular or it is the identity matrix.)

If A is idempotent and $n \times n$, then $(I - A)$ is also idempotent, as we see by multiplication.

In this case, we also have

$$\text{rank}(I - A) = n - \text{rank}(A).$$

Because the eigenvalues of A^2 are the squares of the eigenvalues of A, all eigenvalues of an idempotent matrix must be either 0 or 1. The number of eigenvalues that are 1 is the rank of the matrix. We therefore have for an idempotent matrix A,

$$\text{tr}(A) = \text{rank}(A).$$

Because $AA = A$, any vector in the column space of A is an eigenvector of A.

For a given vector space $\mathcal{V}$, a symmetric idempotent matrix A whose columns span $\mathcal{V}$ is said to be a *projection matrix* onto $\mathcal{V}$; in other words, a matrix A is a projection matrix onto $\text{span}(A)$ if and only if A is symmetric and idempotent.

It is easy to see that for any vector x, if A is a projection matrix onto $\mathcal{V}$, the vector Ax is in $\mathcal{V}$, and the vector $x - Ax$ is in $\mathcal{V}^\perp$ (the vectors Ax and $x - Ax$ are orthogonal). For this reason, a projection matrix is sometimes called an "orthogonal projection matrix". Note that an orthogonal projection matrix is not an orthogonal matrix, however, unless it is the identity matrix. Stating this in alternate notation, if A is a projection matrix and $A \in \mathbb{R}^{n\times n}$, then A maps $\mathbb{R}^n$ onto $\mathcal{V}(A)$, and $I - A$ is also a projection matrix (called the *complementary projection matrix* of A), and it maps $\mathbb{R}^n$ onto the orthogonal complement, $\mathcal{N}(A)$. These spaces are such that $\mathcal{V}(A) \oplus \mathcal{N}(A) = \mathbb{R}^n$.

Useful projection matrices often encountered in statistical linear models are A^+A and AA^+.

If x is a general vector in $\mathbb{R}^n$, that is, if x has order n and belongs to an n-dimensional space, and A is a projection matrix of rank $r \leq n$, then Ax has order n and belongs to $\text{span}(A)$, which is an r-dimensional space.

Because a projection matrix is idempotent, the matrix projects any of its columns onto itself, and of course it projects the full matrix onto itself: $AA = A$. More generally, if x and y are vectors in $\text{span}(A)$ and a is a scalar, then

$$A(ax + y) = ax + y.$$

(To see this, we merely represent x and y as linear combinations of columns (or rows) of A and substitute in the equation.)

The projection of a vector y onto a vector x is

$$\frac{x^{\mathrm{T}}y}{x^{\mathrm{T}}x}x.$$

The projection matrix to accomplish this is the "outer/inner products matrix",

$$\frac{1}{x^{\mathrm{T}}x}xx^{\mathrm{T}}.$$

The outer/inner products matrix has rank 1. It is useful in a variety of matrix transformations. If x is normalized, the projection matrix for projecting a vector on x is just xx^{T}. The projection matrix for projecting a vector onto a unit vector e_i is $e_ie_i^{\mathrm{T}}$, and $e_ie_i^{\mathrm{T}}y = (0, \ldots, y_i, \ldots, 0)$.

Inverses of Matrices

Often in applications we need inverses of various sums of matrices. A simple general result, which we can verify by multiplication, is that if A is a full-rank $n \times n$ matrix, B is a full-rank $m \times m$ matrix, C is any $n \times m$ matrix, and D is any $m \times n$ matrix such that $A + CBD$ is full rank, then

$$(A + CBD)^{-1} = A^{-1} - A^{-1}C(B^{-1} + DA^{-1}C)^{-1}DA^{-1}.$$

From this it follows that if A is a full-rank $n \times n$ matrix and b and c are n-vectors such that $(A + bc^{\mathrm{T}})$ is full rank, then

$$(A + bc^{\mathrm{T}})^{-1} = A^{-1} - \frac{A^{-1}bc^{\mathrm{T}}A^{-1}}{1 + c^{\mathrm{T}}A^{-1}b}.$$

If A and B are full rank matrices of the same size, the following relationships are easy to show directly.

$$\begin{aligned}
(I + A^{-1})^{-1} &= A(A + I)^{-1} \\
(A + BB^{\mathrm{T}})^{-1}B &= A^{-1}B(I + B^{\mathrm{T}}A^{-1}B)^{-1} \\
(A^{-1} + B^{-1})^{-1} &= A(A + B)^{-1}B \\
A - A(A + B)^{-1}A &= B - B(A + B)^{-1}B \\
A^{-1} + B^{-1} &= A^{-1}(A + B)B^{-1} \\
(I + AB)^{-1} &= I - A(I + BA)^{-1}B \\
(I + AB)^{-1}A &= A(I + BA)^{-1}
\end{aligned}$$

From the relationship $\det(AB) = \det(A)\det(B)$ for square matrices mentioned earlier, it is easy to see that for nonsingular A,

$$\det(A) = 1/\det(A^{-1}).$$

For a square matrix A, $\det(A) = 0$ if and only if A is singular.

Partitioned Matrices

We often find it useful to partition a matrix into submatrices, and we usually denote those submatrices with capital letters with subscripts indicating the relative positions of the submatrices. Hence, we may write

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},$$

where the matrices A_{11} and A_{12} have the same number of rows, A_{21} and A_{22} have the same number of rows, A_{11} and A_{21} have the same number of columns, and A_{12} and A_{22} have the same number of columns.

The term "submatrix" is also sometimes used to refer to a matrix formed from another one by deleting various rows and columns of the given matrix. In this terminology, B is a submatrix of A if for each element b_{ij} there is an a_{kl} with $k \geq i$ and $l \geq j$, such that $b_{ij} = a_{kl}$; that is, the rows and/or columns of the submatrix are not contiguous in the original matrix.

A submatrix whose principal diagonal elements are elements of the principal diagonal of the given matrix is called a *principal submatrix*; A_{11} is a principal submatrix in the example above, and if A_{22} is square it is also a principal submatrix. Sometimes the term "principal submatrix" is restricted to square submatrices.

A principal submatrix that contains the $(1, 1)$ and whose rows and columns are contiguous in the original matrix is called a *leading principal submatrix*. A_{11} is a principal submatrix in the example above.

Multiplication and other operations with matrices, such as transposition, are carried out with their submatrices in the obvious way. Thus,

$$\begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \end{bmatrix}^{\mathrm{T}} = \begin{bmatrix} A_{11}^{\mathrm{T}} & A_{21}^{\mathrm{T}} \\ A_{12}^{\mathrm{T}} & A_{22}^{\mathrm{T}} \\ A_{13}^{\mathrm{T}} & A_{23}^{\mathrm{T}} \end{bmatrix},$$

and, assuming the submatrices are conformable for multiplication,

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} = \begin{bmatrix} A_{11}B_{11} + A_{12}B_{21} & A_{11}B_{12} + A_{12}B_{22} \\ A_{21}B_{11} + A_{22}B_{21} & A_{21}B_{12} + A_{22}B_{22} \end{bmatrix}.$$

Sometimes a matrix may be partitioned such that one partition is just a single column or row, that is, a vector or the transpose of a vector. In that case, we may use a notation such as

$$[X\ y]$$

or

$$[X \mid y],$$

where X is a matrix and y is a vector. We develop the notation in the obvious fashion; for example,

$$[X\ y]^{\mathrm{T}}\,[X\ y] = \begin{bmatrix} X^{\mathrm{T}}X & X^{\mathrm{T}}y \\ y^{\mathrm{T}}X & y^{\mathrm{T}}y \end{bmatrix}.$$

Partitioned matrices may also have useful patterns. A "block diagonal" matrix is one of the form

$$\begin{bmatrix} \mathtt{X} & \mathtt{0} & \cdots & \mathtt{0} \\ \mathtt{0} & \mathtt{X} & \cdots & \mathtt{0} \\ & & \ddots & \\ \mathtt{0} & \mathtt{0} & \cdots & \mathtt{X} \end{bmatrix},$$

where `0` represents a submatrix with all zeros, and `X` represents a general submatrix, with at least some nonzeros. The diag(·) function previously introduced for a vector is also defined for a list of matrices:

$$\text{diag}(A_1, A_2, \ldots, A_k)$$

denotes the block diagonal matrix with submatrices $A_1, A_2, \ldots, A_k$ along the diagonal and zeros elsewhere.

Inverses of Partitioned Matrices

If A is nonsingular, and can be partitioned as

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},$$

where both A_{11} and A_{22} are nonsingular, it is easy to see that the inverse of A is given by

$$A^{-1} = \begin{bmatrix} A_{11}^{-1} + A_{11}^{-1}A_{12}Z^{-1}A_{21}A_{11}^{-1} & -A_{11}^{-1}A_{12}Z^{-1} \\ -Z^{-1}A_{21}A_{11}^{-1} & Z^{-1} \end{bmatrix},$$

where $Z = A_{22} - A_{21}A_{11}^{-1}A_{12}$. In this partitioning Z is called the *Schur complement* of A_{11} in A.

If

$$A = [Xy]^{\mathrm{T}}[Xy]$$

and is partitioned as into $X^{\mathrm{T}}X$ and $y^{\mathrm{T}}y$ on the diagonal, and X is of full column rank, then the Schur complement of $X^{\mathrm{T}}X$ in $[Xy]^{\mathrm{T}}[Xy]$ is

$$y^{\mathrm{T}}y - y^{\mathrm{T}}X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y.$$

This particular partitioning is useful in linear regression analysis, where this Schur complement is the residual sum of squares.

Gramian Matrices and Generalized Inverses

A matrix of the form $Z^{\mathrm{T}}Z$ is called a *Gramian matrix*. Such matrices arise often in statistical applications.

Some interesting properties of a Gramian matrix $Z^{\mathrm{T}}Z$ are

- $Z^{\mathrm{T}}Z$ is symmetric;
- $Z^{\mathrm{T}}Z$ is of full rank if and only if Z is of full column rank, or, more generally,

$$\text{rank}(Z^{\mathrm{T}}Z) = \text{rank}(Z);$$

- $Z^{\mathrm{T}}Z$ is nonnegative definite, and positive definite if and only if Z is of full column rank;
- $Z^{\mathrm{T}}Z = 0 \implies Z = 0$.

The generalized inverses of $Z^{\mathrm{T}}Z$ have useful properties. First, we see from the definition, for any generalized inverse, $(Z^{\mathrm{T}}Z)^-$ that $((Z^{\mathrm{T}}Z)^-)^{\mathrm{T}}$ is also a generalized inverse of $Z^{\mathrm{T}}Z$. (Note that $(Z^{\mathrm{T}}Z)^-$ is not necessarily symmetric.)

Another useful property of a Gramian matrix is that for any matrices B and C (that are conformable for the operations indicated),

$$BZ^{\mathrm{T}}Z = CZ^{\mathrm{T}}Z \iff BZ^{\mathrm{T}} = CZ^{\mathrm{T}}.$$

The implication from right to left is obvious, and we can see the left to right implication by writing

$$(BZ^{\mathrm{T}}Z - CZ^{\mathrm{T}}Z)(B^{\mathrm{T}} - C^{\mathrm{T}}) = (BZ^{\mathrm{T}} - CZ^{\mathrm{T}})(BZ^{\mathrm{T}} - CZ^{\mathrm{T}})^{\mathrm{T}},$$

and then observing that if the left side is null, then so is the right side, and if the right side is null, then $BZ^{\mathrm{T}} - CZ^{\mathrm{T}} = 0$. Similarly, we have

$$Z^{\mathrm{T}}ZB = Z^{\mathrm{T}}ZC \iff Z^{\mathrm{T}}B = Z^{\mathrm{T}}C.$$

Also,

$$Z(Z^{\mathrm{T}}Z)^-Z^{\mathrm{T}}Z = Z.$$

This means that $(Z^{\mathrm{T}}Z)^-Z^{\mathrm{T}}$ is a generalized inverse of Z

An important property of $Z(Z^{\mathrm{T}}Z)^-Z^{\mathrm{T}}$ is its invariance to the choice of the generalized inverse of $Z^{\mathrm{T}}Z$. Suppose G is any generalized inverse of $Z^{\mathrm{T}}Z$. Then we have

$$ZGZ^{\mathrm{T}} = Z(Z^{\mathrm{T}}Z)^-Z^{\mathrm{T}};$$

that is, $Z(Z^{\mathrm{T}}Z)^-Z^{\mathrm{T}}$ is invariant to choice of generalized inverse.

The squared norm of the residual vector obtained from any generalized inverse of $Z^{\mathrm{T}}Z$ has some interesting properties. First, just by direct multiplication, we get the "Pythagorean property" of the norm of the predicted values and the residuals:

$$\|X - Z\beta\|^2 = \|X - Z\widehat{\beta}\|^2 + \|Z\widehat{\beta} - Z\beta\|^2$$

where $\widehat{\beta} = (Z^{\mathrm{T}}Z)^-Z^{\mathrm{T}}X$ for any generalized inverse. We also have

$$\mathrm{E}(Z\widehat{\beta}) = Z\beta,$$

and

$$\mathrm{E}((Z\widehat{\beta} - Z\beta)^{\mathrm{T}}(Z\widehat{\beta} - Z\beta)) = \mathrm{V}(Z\widehat{\beta}).$$

Because for any vector y, we have

$$\|y\|^2 = y^{\mathrm{T}}y = \mathrm{tr}(y^{\mathrm{T}}y),$$

we can derive an interesting expression for $\mathrm{E}(\|X - Z\beta\|^2)$:

$$\begin{aligned}
\mathrm{E}(\|X - Z\widehat{\beta}\|^2) &= \mathrm{tr}(\mathrm{E}(\|X - Z\widehat{\beta}\|^2)) \\
&= \mathrm{tr}(\mathrm{E}((X - Z\beta)^{\mathrm{T}}(X - Z\beta)) - \mathrm{E}((Z\widehat{\beta} - Z\beta)^{\mathrm{T}}(Z\widehat{\beta} - Z\beta))) \\
&= \mathrm{tr}(\mathrm{V}(X) - \mathrm{V}(Z\widehat{\beta}) \\
&= n\sigma^2 - \mathrm{tr}((Z(Z^{\mathrm{T}}Z)^{-}Z^{\mathrm{T}})\sigma^2 I(Z(Z^{\mathrm{T}}Z)^{-}Z^{\mathrm{T}})) \\
&= \sigma^2(n - \mathrm{tr}((Z^{\mathrm{T}}Z)^{-}Z^{\mathrm{T}}Z)).
\end{aligned}$$

The trace in the latter expression is the "regression degrees of freedom".

The Moore-Penrose Inverse

The Moore-Penrose inverse, or the pseudoinverse, of Z has an interesting relationship with a generalized inverse of $Z^{\mathrm{T}}Z$:

$$ZZ^{+} = Z(Z^{\mathrm{T}}Z)^{-}Z^{\mathrm{T}}.$$

This can be established directly from the definition of the Moore-Penrose inverse.

0.3.3 Vector/Matrix Derivatives and Integrals

The operations of differentiation and integration of vectors and matrices are logical extensions of the corresponding operations on scalars. There are three objects involved in this operation:

- the variable of the operation;
- the operand (the function being differentiated or integrated); and
- the result of the operation.

In the simplest case, all three of these objects are of the same type, and they are scalars. If either the variable or the operand is a vector or a matrix, however, the structure of the result may be more complicated. This statement will become clearer as we proceed to consider specific cases.

In this section, we state or show the form that the derivative takes in terms of simpler derivatives. We state high-level rules for the nature of the differentiation in terms of simple partial differentiation of a scalar with respect to a scalar. We do not consider whether or not the derivatives exist. In general, if the simpler derivatives we write that comprise the more complicated object exist, then the derivative of that more complicated object exists. Once a shape of the derivative is determined, definitions or derivations in ϵ-δ terms could be given, but we will refrain from that kind of formal exercise. The purpose of this section is not to develop a calculus for vectors and matrices but rather to consider some cases that find wide applications in statistics. For a more careful treatment of differentiation of vectors and matrices see Gentle (2007).

Basics of Differentiation

It is useful to recall the heuristic interpretation of a derivative. A derivative of a function is the infinitesimal rate of change of the function with respect to the variable with which the differentiation is taken. If both the function and the variable are scalars, this interpretation is unambiguous. If, however, the operand of the differentiation, Φ, is a more complicated function, say a vector or a matrix, and/or the variable of the differentiation, Ξ, is a more complicated object, the changes are more difficult to measure. Change in the value both of the function,

$$\delta\Phi = \Phi_{\text{new}} - \Phi_{\text{old}},$$

and of the variable,

$$\delta\Xi = \Xi_{\text{new}} - \Xi_{\text{old}},$$

could be measured in various ways, by using various norms, for example. (Note that the subtraction is not necessarily ordinary scalar subtraction.)

Furthermore, we cannot just divide the function values by $\delta\Xi$. We do not have a definition for division by that kind of object. We need a mapping, possibly a norm, that assigns a positive real number to $\delta\Xi$. We can define the change in the function value as just the simple difference of the function evaluated at the two points. This yields

$$\lim_{\|\delta\Xi\|\to 0} \frac{\Phi(\Xi+\delta\Xi)-\Phi(\Xi)}{\|\delta\Xi\|}. \tag{0.3.38}$$

So long as we remember the complexity of $\delta\Xi$, however, we can adopt a simpler approach. Since for both vectors and matrices, we have definitions of multiplication by a scalar and of addition, we can simplify the limit in the usual definition of a derivative, $\delta\Xi \to 0$. Instead of using $\delta\Xi$ as the element of change, we will use $t\Upsilon$, where t is a scalar and Υ is an element to be added to Ξ. The limit then will be taken in terms of $t \to 0$. This leads to

$$\lim_{t\to 0} \frac{\Phi(\Xi+t\Upsilon)-\Phi(\Xi)}{t} \tag{0.3.39}$$

as a formula for the derivative of Φ with respect to Ξ.

The expression (0.3.39) may be a useful formula for evaluating a derivative, but we must remember that it is not the derivative. The type of object of this formula is the same as the type of object of the function, Φ; it does not accommodate the type of object of the argument, Ξ, unless Ξ is a scalar. As we will see below, for example, if Ξ is a vector and Φ is a scalar, the derivative must be a vector, yet in that case the expression (0.3.39) is a scalar.

The expression (0.3.38) is rarely directly useful in evaluating a derivative, but it serves to remind us of both the generality and the complexity of the concept. Both Φ and its arguments could be functions, for example. In functional analysis, various kinds of functional derivatives are defined, such as a Gâteaux

derivative. These derivatives find applications in developing robust statistical methods. Here we are just interested in the combinations of three possibilities for Φ, namely scalar, vector, and matrix, and the same three possibilities for Ξ and Υ.

Continuity

It is clear from the definition of continuity that for the derivative of a function to exist at a point, the function must be continuous at that point. A function of a vector or a matrix is continuous if it is continuous for each element of the vector or matrix. Just as scalar sums and products are continuous, vector/matrix sums and all of the types of vector/matrix products we have discussed are continuous. A continuous function of a continuous function is continuous.

Many of the vector/matrix functions we have discussed are clearly continuous. For example, the L_p vector norms are continuous over the nonnegative reals but not over the reals unless p is an even (positive) integer. The determinant of a matrix is continuous, as we see from the definition of the determinant and the fact that sums and scalar products are continuous. The fact that the determinant is a continuous function immediately yields the result that cofactors and hence the adjugate are continuous. From the relationship between an inverse and the adjugate, we see that the inverse is a continuous function.

Notation and Properties

We write the differential operator with respect to the dummy variable x as $\partial/\partial x$ or $\partial/\partial x^{\mathrm{T}}$. We usually denote differentiation using the symbol for "partial" differentiation, ∂, whether the operator is written ∂x_i for differentiation with respect to a specific scalar variable or ∂x for differentiation with respect to the array x that contains all of the individual elements. Sometimes, however, if the differentiation is being taken with respect to the whole array (the vector or the matrix), we use the notation $\mathrm{d}/\mathrm{d}x$.

The operand of the differential operator $\partial/\partial x$ is a function of x. (If it is not a function of x—that is, if it is a constant function with respect to x—then the operator evaluates to 0.) The result of the operation, written $\partial f/\partial x$, is also a function of x, with the same domain as f, and we sometimes write $\partial f(x)/\partial x$ to emphasize this fact. The value of this function at the fixed point x_0 is written as $\partial f(x_0)/\partial x$. (The derivative of the constant $f(x_0)$ is identically 0, but it is not necessary to write $\partial f(x)/\partial x|_{x_0}$ because $\partial f(x_0)/\partial x$ is interpreted as the value of the function $\partial f(x)/\partial x$ at the fixed point x_0.)

If $\partial/\partial x$ operates on f, and $f : S \to T$, then $\partial/\partial x : S \to U$. The nature of S, or more directly the nature of x, whether it is a scalar, a vector, or a matrix, and the nature of T determine the structure of the result U. For example, if x is an n-vector and $f(x) = x^{\mathrm{T}}x$, then

$$f : \mathbb{R}^n \to \mathbb{R}$$

and

$$\partial f/\partial x : \mathbb{R}^n \to \mathbb{R}^n,$$

as we will see. The outer product, $h(x) = xx^{\mathrm{T}}$, is a mapping to a higher rank array, but the derivative of the outer product is a mapping to an array of the same rank; that is,

$$h : \mathbb{R}^n \to \mathbb{R}^{n\times n}$$

and

$$\partial h/\partial x : \mathbb{R}^n \to \mathbb{R}^n.$$

(Note that "rank" here means the number of dimensions. This term is often used in this way in numerical software. See Gentle (2007), page 5.)

As another example, consider $g(\cdot) = \det(\cdot)$, so

$$g : \mathbb{R}^{n\times n} \mapsto \mathbb{R}.$$

In this case,

$$\partial g/\partial X : \mathbb{R}^{n\times n} \mapsto \mathbb{R}^{n\times n};$$

that is, the derivative of the determinant of a square matrix is a square matrix, as we will see later.

Higher-order differentiation is a composition of the $\partial/\partial x$ operator with itself or of the $\partial/\partial x$ operator and the $\partial/\partial x^{\mathrm{T}}$ operator. For example, consider the familiar function in linear least squares

$$f(b) = (y - Xb)^{\mathrm{T}}(y - Xb).$$

This is a mapping from $\mathbb{R}^m$ to $\mathbb{R}$. The first derivative with respect to the m-vector b is a mapping from $\mathbb{R}^m$ to $\mathbb{R}^m$, namely $2X^{\mathrm{T}}Xb - 2X^{\mathrm{T}}y$. The second derivative with respect to b^{T} is a mapping from $\mathbb{R}^m$ to $\mathbb{R}^{m\times m}$, namely, $2X^{\mathrm{T}}X$.

We see from expression (0.3.38) that differentiation is a linear operator; that is, if $\mathcal{D}(\Phi)$ represents the operation defined in expression (0.3.38), Ψ is another function in the class of functions over which $\mathcal{D}$ is defined, and a is a scalar that does not depend on the variable Ξ, then $\mathcal{D}(a\Phi + \Psi) = a\mathcal{D}(\Phi) + \mathcal{D}(\Psi)$. This yields the familiar rules of differential calculus for derivatives of sums or constant scalar products. Other usual rules of differential calculus apply, such as for differentiation of products and composition (the chain rule). We can use expression (0.3.39) to work these out. For example, for the derivative of the product $\Phi\Psi$, after some rewriting of terms, we have the numerator

$$\begin{aligned}
&\Phi(\Xi)\big(\Psi(\Xi + t\Upsilon) - \Psi(\Xi)\big) \\
&+\Psi(\Xi)\big(\Phi(\Xi + t\Upsilon) - \Phi(\Xi)\big) \\
&+\big(\Phi(\Xi + t\Upsilon) - \Phi(\Xi)\big)\big(\Psi(\Xi + t\Upsilon) - \Psi(\Xi)\big).
\end{aligned}$$

Now, dividing by t and taking the limit, assuming that as

$$t \to 0,$$
$$(\Phi(\Xi + t\Upsilon) - \Phi(\Xi)) \to 0,$$

we have

$$\mathcal{D}(\Phi\Psi) = \mathcal{D}(\Phi)\Psi + \Phi\mathcal{D}(\Psi), \tag{0.3.40}$$

where again $\mathcal{D}$ represents the differentiation operation.

Differentials

For a differentiable scalar function of a scalar variable, $f(x)$, the *differential of f at c with increment u* is $u\mathrm{d}f/\mathrm{d}x|_c$. This is the linear term in a truncated Taylor series expansion:

$$f(c+u) = f(c) + u\frac{\mathrm{d}}{\mathrm{d}x}f(c) + r(c,u). \tag{0.3.41}$$

Technically, the differential is a function of both x and u, but the notation $\mathrm{d}f$ is used in a generic sense to mean the differential of f. For vector/matrix functions of vector/matrix variables, the differential is defined in a similar way. The structure of the differential is the same as that of the function; that is, for example, the differential of a matrix-valued function is a matrix.

Types of Differentiation

In the following sections we consider differentiation with respect to different types of objects first, and we consider differentiation of different types of objects.

Differentiation with Respect to a Scalar

Differentiation of a structure (vector or matrix, for example) with respect to a scalar is quite simple; it just yields the ordinary derivative of each element of the structure in the same structure. Thus, the derivative of a vector or a matrix with respect to a scalar variable is a vector or a matrix, respectively, of the derivatives of the individual elements.

Differentiation with respect to a vector or matrix, which we will consider below, is often best approached by considering differentiation with respect to the individual elements of the vector or matrix, that is, with respect to scalars.

Derivatives of Vectors with Respect to Scalars

The derivative of the vector $y(x) = (y_1, \ldots, y_n)$ with respect to the scalar x is the vector

$$\partial y/\partial x = (\partial y_1/\partial x, \ldots, \partial y_n/\partial x). \tag{0.3.42}$$

The second or higher derivative of a vector with respect to a scalar is likewise a vector of the derivatives of the individual elements; that is, it is an array of higher rank.

Derivatives of Matrices with Respect to Scalars

The derivative of the matrix $Y(x) = (y_{ij})$ with respect to the scalar x is the matrix

$$\partial Y(x)/\partial x = (\partial y_{ij}/\partial x). \tag{0.3.43}$$

The second or higher derivative of a matrix with respect to a scalar is likewise a matrix of the derivatives of the individual elements.

Derivatives of Functions with Respect to Scalars

Differentiation of a function of a vector or matrix that is linear in the elements of the vector or matrix involves just the differentiation of the elements, followed by application of the function. For example, the derivative of a trace of a matrix is just the trace of the derivative of the matrix. On the other hand, the derivative of the determinant of a matrix is not the determinant of the derivative of the matrix (see below).

Higher-Order Derivatives with Respect to Scalars

Because differentiation with respect to a scalar does not change the rank of the object ("rank" here means rank of an array or "shape"), higher-order derivatives $\partial^k/\partial x^k$ with respect to scalars are merely objects of the same rank whose elements are the higher-order derivatives of the individual elements.

Differentiation with Respect to a Vector

Differentiation of a given object with respect to an n-vector yields a vector for each element of the given object. The basic expression for the derivative, from formula (0.3.39), is

$$\lim_{t\to 0} \frac{\Phi(x+ty) - \Phi(x)}{t} \tag{0.3.44}$$

for an arbitrary conformable vector y. The arbitrary y indicates that the derivative is omnidirectional; it is the rate of change of a function of the vector in any direction.

Derivatives of Scalars with Respect to Vectors; The Gradient

The derivative of a scalar-valued function with respect to a vector is a vector of the partial derivatives of the function with respect to the elements of the vector. If $f(x)$ is a scalar function of the vector $x = (x_1, \ldots, x_n)$,

$$\frac{\partial f}{\partial x} = \left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n}\right), \tag{0.3.45}$$

if those derivatives exist. This vector is called the *gradient* of the scalar-valued function, and is sometimes denoted by $\mathrm{g}_f(x)$ or $\nabla f(x)$, or sometimes just g_f or ∇f:

$$\mathrm{g}_f = \nabla f = \frac{\partial f}{\partial x}. \tag{0.3.46}$$

The notation g_f or ∇f implies differentiation with respect to "all" arguments of f, hence, if f is a scalar-valued function of a vector argument, they represent a vector.

This derivative is useful in finding the maximum or minimum of a function. Such applications arise throughout statistical and numerical analysis.

Inner products, bilinear forms, norms, and variances are interesting scalar-valued functions of vectors. In these cases, the function Φ in equation (0.3.44) is scalar-valued and the numerator is merely $\Phi(x+ty) - \Phi(x)$. Consider, for example, the quadratic form $x^{\mathrm{T}}Ax$. Using equation (0.3.44) to evaluate $\partial x^{\mathrm{T}}Ax/\partial x$, we have

$$\begin{aligned}
&\lim_{t\to 0} \frac{(x+ty)^{\mathrm{T}}A(x+ty) - x^{\mathrm{T}}Ax}{t} \\
&= \lim_{t\to 0} \frac{x^{\mathrm{T}}Ax + ty^{\mathrm{T}}Ax + ty^{\mathrm{T}}A^{\mathrm{T}}x + t^2y^{\mathrm{T}}Ay - x^{\mathrm{T}}Ax}{t} \\
&= y^{\mathrm{T}}(A + A^{\mathrm{T}})x,
\end{aligned} \tag{0.3.47}$$

for an arbitrary y (that is, "in any direction"), and so $\partial x^{\mathrm{T}}Ax/\partial x = (A+A^{\mathrm{T}})x$.

This immediately yields the derivative of the square of the Euclidean norm of a vector, $\|x\|_2^2$, and the derivative of the Euclidean norm itself by using the chain rule. Other L_p vector norms may not be differentiable everywhere because of the presence of the absolute value in their definitions. The fact that the Euclidean norm is differentiable everywhere is one of its most important properties.

The derivative of the quadratic form also immediately yields the derivative of the variance. The derivative of the correlation, however, is slightly more difficult because it is a ratio.

The operator $\partial/\partial x^{\mathrm{T}}$ applied to the scalar function f results in $\mathrm{g}_f^{\mathrm{T}}$.

The second derivative of a scalar-valued function with respect to a vector is a derivative of the first derivative, which is a vector. We will now consider derivatives of vectors with respect to vectors.

Derivatives of Vectors with Respect to Vectors; The Jacobian

The derivative of an m-vector-valued function of an n-vector argument consists of nm scalar derivatives. These derivatives could be put into various structures. Two obvious structures are an $n \times m$ matrix and an $m \times n$ matrix. For a function $f : S \subseteq \mathbb{R}^n \to \mathbb{R}^m$, we define $\partial f^{\mathrm{T}}/\partial x$ to be the $n \times m$ matrix, which is the natural extension of $\partial/\partial x$ applied to a scalar function, and

$\partial f/\partial x^{\mathrm{T}}$ to be its transpose, the $m \times n$ matrix. Although the notation $\partial f^{\mathrm{T}}/\partial x$ is more precise because it indicates that the elements of f correspond to the columns of the result, we often drop the transpose in the notation. We have

$$\begin{aligned}\frac{\partial f}{\partial x} &= \frac{\partial f^{\mathrm{T}}}{\partial x} \quad \text{by convention} \\ &= \left[\frac{\partial f_1}{\partial x} \cdots \frac{\partial f_m}{\partial x}\right] \\ &= \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_2}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_1} \\ \frac{\partial f_1}{\partial x_2} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_2} \\ & & \cdots & \\ \frac{\partial f_1}{\partial x_n} & \frac{\partial f_2}{\partial x_n} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}\end{aligned} \tag{0.3.48}$$

if those derivatives exist. This derivative is called the *matrix gradient* and is denoted by G_f or ∇f for the vector-valued function f. (Note that the ∇ symbol can denote either a vector or a matrix, depending on whether the function being differentiated is scalar-valued or vector-valued.)

The $m \times n$ matrix $\partial f/\partial x^{\mathrm{T}} = (\nabla f)^{\mathrm{T}}$ is called the *Jacobian* of f and is denoted by J_f:

$$\mathrm{J}_f = \mathrm{G}_f^{\mathrm{T}} = (\nabla f)^{\mathrm{T}}. \tag{0.3.49}$$

The absolute value of the determinant of the Jacobian appears in integrals involving a change of variables. (Occasionally, the term "Jacobian" is used to refer to the absolute value of the determinant rather than to the matrix itself.)

To emphasize that the quantities are functions of x, we sometimes write $\partial f(x)/\partial x$, $\mathrm{J}_f(x)$, $\mathrm{G}_f(x)$, or $\nabla f(x)$.

Derivatives of Matrices with Respect to Vectors

The derivative of a matrix with respect to a vector is a three-dimensional object that results from applying equation (0.3.45) to each of the elements of the matrix. For this reason, it is simpler to consider only the partial derivatives of the matrix Y with respect to the individual elements of the vector x; that is, $\partial Y/\partial x_i$. The expressions involving the partial derivatives can be thought of as defining one two-dimensional layer of a three-dimensional object.

Using the rules for differentiation of powers that result directly from the definitions, we can write the partial derivatives of the inverse of the matrix Y as

$$\frac{\partial}{\partial x} Y^{-1} = -Y^{-1}\left(\frac{\partial}{\partial x} Y\right) Y^{-1}. \tag{0.3.50}$$

Beyond the basics of differentiation of constant multiples or powers of a variable, the two most important properties of derivatives of expressions are

the linearity of the operation and the chaining of the operation. These yield rules that correspond to the familiar rules of the differential calculus. A simple result of the linearity of the operation is the rule for differentiation of the trace:

$$\frac{\partial}{\partial x}\mathrm{tr}(Y) = \mathrm{tr}\left(\frac{\partial}{\partial x}Y\right).$$

Higher-Order Derivatives with Respect to Vectors; The Hessian

Higher-order derivatives are derivatives of lower-order derivatives. As we have seen, a derivative of a given function with respect to a vector is a more complicated object than the original function. The simplest higher-order derivative with respect to a vector is the second-order derivative of a scalar-valued function. Higher-order derivatives may become uselessly complicated.

In accordance with the meaning of derivatives of vectors with respect to vectors, the second derivative of a scalar-valued function with respect to a vector is a matrix of the partial derivatives of the function with respect to the elements of the vector. This matrix is called the *Hessian*, and is denoted by H_f or sometimes by $\nabla\nabla f$ or $\nabla^2 f$:

$$\mathrm{H}_f = \frac{\partial^2 f}{\partial x \partial x^{\mathrm{T}}} = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_m} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_m} \\ & \cdots & & \\ \frac{\partial^2 f}{\partial x_m \partial x_1} & \frac{\partial^2 f}{\partial x_m \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_m^2} \end{bmatrix}. \tag{0.3.51}$$

The Hessian is a function of x. We write $\mathrm{H}_f(x)$ or $\nabla\nabla f(x)$ or $\nabla^2 f(x)$ for the value of the Hessian at x.

Summary of Derivatives with Respect to Vectors

As we have seen, the derivatives of functions are complicated by the problem of measuring the change in the function, but often the derivatives of functions with respect to a vector can be determined by using familiar scalar differentiation. In general, we see that

- the derivative of a scalar (a quadratic form) with respect to a vector is a vector and
- the derivative of a vector with respect to a vector is a matrix.

Table 0.3 lists formulas for the vector derivatives of some common expressions. The derivative $\partial f/\partial x^{\mathrm{T}}$ is the transpose of $\partial f/\partial x$.

Table 0.3. Formulas for Some Vector Derivatives

$f(x)$	$\partial f/\partial x$
ax	a
$b^{\mathrm{T}}x$	b
$x^{\mathrm{T}}b$	b^{T}
$x^{\mathrm{T}}x$	$2x$
xx^{T}	$2x^{\mathrm{T}}$
$b^{\mathrm{T}}Ax$	$A^{\mathrm{T}}b$
$x^{\mathrm{T}}Ab$	$b^{\mathrm{T}}A$
$x^{\mathrm{T}}Ax$	$(A+A^{\mathrm{T}})x$
	$2Ax$, if A is symmetric
$\exp(-\frac{1}{2}x^{\mathrm{T}}Ax)$	$-\exp(-\frac{1}{2}x^{\mathrm{T}}Ax)Ax$, if A is symmetric
$\|x\|_2^2$	$2x$
$\mathrm{V}(x)$	$2x/(n-1)$

In this table, x is an n-vector, a is a constant scalar, b is a constant conformable vector, and A is a constant conformable matrix.

Differentiation with Respect to a Matrix

The derivative of a function with respect to a matrix is a matrix with the same shape consisting of the partial derivatives of the function with respect to the elements of the matrix. This rule defines what we mean by differentiation with respect to a matrix.

By the definition of differentiation with respect to a matrix X, we see that the derivative $\partial f/\partial X^{\mathrm{T}}$ is the transpose of $\partial f/\partial X$. For scalar-valued functions, this rule is fairly simple. For example, consider the trace. If X is a square matrix and we apply this rule to evaluate $\partial\,\mathrm{tr}(X)/\partial X$, we get the identity matrix, where the nonzero elements arise only when $j=i$ in $\partial(\sum x_{ii})/\partial x_{ij}$. If AX is a square matrix, we have for the (i,j) term in $\partial\,\mathrm{tr}(AX)/\partial X$, $\partial\sum_i\sum_k a_{ik}x_{ki}/\partial x_{ij} = a_{ji}$, and so $\partial\,\mathrm{tr}(AX)/\partial X = A^{\mathrm{T}}$, and likewise, inspecting $\partial\sum_i\sum_k x_{ik}x_{ki}/\partial x_{ij}$, we get $\partial\,\mathrm{tr}(X^{\mathrm{T}}X)/\partial X = 2X^{\mathrm{T}}$. Likewise for the scalar-valued $a^{\mathrm{T}}Xb$, where a and b are conformable constant vectors, for $\partial\sum_m(\sum_k a_k x_{km})b_m/\partial x_{ij} = a_i b_j$, so $\partial a^{\mathrm{T}}Xb/\partial X = ab^{\mathrm{T}}$.

Now consider $\partial|X|/\partial X$. Using an expansion in cofactors, the only term in $|X|$ that involves x_{ij} is $x_{ij}(-1)^{i+j}|X_{-(i)(j)}|$, and the cofactor $(x_{(ij)}) = (-1)^{i+j}|X_{-(i)(j)}|$ does not involve x_{ij}. Hence, $\partial|X|/\partial x_{ij} = (x_{(ij)})$, and so $\partial|X|/\partial X = (\mathrm{adj}(X))^{\mathrm{T}}$. We can write this as $\partial|X|/\partial X = |X|X^{-\mathrm{T}}$.

The chain rule can be used to evaluate $\partial\log|X|/\partial X$.

Applying the rule stated at the beginning of this section, we see that the derivative of a matrix Y with respect to the matrix X is

$$\frac{\mathrm{d}Y}{\mathrm{d}X} = Y\otimes\frac{\mathrm{d}}{\mathrm{d}X}. \tag{0.3.52}$$

Table 0.4 lists some formulas for the matrix derivatives of some common expressions. The derivatives shown in Table 0.4 can be obtained by evaluating expression (0.3.52), possibly also using the chain rule.

Table 0.4. Formulas for Some Matrix Derivatives

General X

$f(X)$	$\partial f/\partial X$
$a^{\mathrm{T}}Xb$	ab^{T}
$\mathrm{tr}(AX)$	A^{T}
$\mathrm{tr}(X^{\mathrm{T}}X)$	$2X^{\mathrm{T}}$
BX	$I_n \otimes B$
XC	$C^{\mathrm{T}} \otimes I_m$
BXC	$C^{\mathrm{T}} \otimes B$

Square and Possibly Invertible X

$f(X)$	$\partial f/\partial X$
$\mathrm{tr}(X)$	I_n
$\mathrm{tr}(X^k)$	kX^{k-1}
$\mathrm{tr}(BX^{-1}C)$	$-(X^{-1}CBX^{-1})^{\mathrm{T}}$
$\lvert X\rvert$	$\lvert X\rvert X^{-\mathrm{T}}$
$\log\lvert X\rvert$	$X^{-\mathrm{T}}$
$\lvert X\rvert^k$	$k\lvert X\rvert^k X^{-\mathrm{T}}$
$BX^{-1}C$	$-(X^{-1}C)^{\mathrm{T}} \otimes BX^{-1}$

In this table, X is an $n \times m$ matrix, a is a constant n-vector, b is a constant m-vector, A is a constant $m \times n$ matrix, B is a constant $p \times n$ matrix, and C is a constant $m \times q$ matrix.

There are some interesting applications of differentiation with respect to a matrix in maximum likelihood estimation. Depending on the structure of the parameters in the distribution, derivatives of various types of objects may be required. For example, the determinant of a variance-covariance matrix, in the sense that it is a measure of a volume, often occurs as a normalizing factor in a probability density function; therefore, we often encounter the need to differentiate a determinant with respect to a matrix.

0.3.4 Optimization of Functions

*** move this to Appendix on Optimization ***

Because a derivative measures the rate of change of a function, a point at which the derivative is equal to 0 is a stationary point, which may be a maximum or a minimum of the function. Differentiation is therefore a very useful tool for finding the optima of functions, and so, for a given function $f(x)$, the gradient vector function, $\mathrm{g}_f(x)$, and the Hessian matrix function, $\mathrm{H}_f(x)$, play important roles in optimization methods.

We may seek either a maximum or a minimum of a function. Since maximizing the scalar function $f(x)$ is equivalent to minimizing $-f(x)$, we can always consider optimization of a function to be minimization of a function. Thus, we generally use terminology for the problem of finding a minimum of a function. Because the function may have many ups and downs, we often use the phrase *local minimum* (or local maximum or local optimum).

Except in the very simplest of cases, the optimization method must be iterative, moving through a sequence of points, $x^{(0)}, x^{(1)}, x^{(2)}, \ldots$, that approaches the optimum point arbitrarily closely. At the point $x^{(k)}$, the direction of *steepest descent* is clearly $-\mathrm{g}_f(x^{(k)})$, but because this direction may be continuously changing, the steepest descent direction may not be the best direction in which to seek the next point, $x^{(k+1)}$.

In the following subsection we describe some specific methods of optimization in the context of vector/matrix differentiation. We will discuss optimization in somewhat more detail in Section 0.4.

Stationary Points of Functions

The first derivative helps only in finding a stationary point. The matrix of second derivatives, the Hessian, provides information about the nature of the stationary point, which may be a local minimum or maximum, a saddlepoint, or only an inflection point.

The so-called second-order optimality conditions are the following (see a general text on optimization for their proofs).

- If (but not only if) the stationary point is a local minimum, then the Hessian is nonnegative definite.
- If the Hessian is positive definite, then the stationary point is a local minimum.
- Likewise, if the stationary point is a local maximum, then the Hessian is nonpositive definite, and if the Hessian is negative definite, then the stationary point is a local maximum.
- If the Hessian has both positive and negative eigenvalues, then the stationary point is a saddlepoint.

Newton's Method

We consider a differentiable scalar-valued function of a vector argument, $f(x)$. By a Taylor series about a stationary point x_*, truncated after the second-order term

$$f(x) \approx f(x_*) + (x - x_*)^{\mathrm{T}} \mathrm{g}_f(x_*) + \frac{1}{2}(x - x_*)^{\mathrm{T}} \mathrm{H}_f(x_*)(x - x_*), \tag{0.3.53}$$

because $\mathrm{g}_f(x_*) = 0$, we have a general method of finding a stationary point for the function $f(\cdot)$, called Newton's method. If x is an m-vector, $\mathrm{g}_f(x)$ is an m-vector and $\mathrm{H}_f(x)$ is an $m \times m$ matrix.

Newton's method is to choose a starting point $x^{(0)}$, then, for $k = 0, 1, \ldots$, to solve the linear systems

$$\mathrm{H}_f(x^{(k)})p^{(k+1)} = -\mathrm{g}_f(x^{(k)}) \tag{0.3.54}$$

for $p^{(k+1)}$, and then to update the point in the domain of $f(\cdot)$ by

$$x^{(k+1)} = x^{(k)} + p^{(k+1)}. \tag{0.3.55}$$

The two steps are repeated until there is essentially no change from one iteration to the next. If $f(\cdot)$ is a quadratic function, the solution is obtained in one iteration because equation (0.3.53) is exact. These two steps have a very simple form for a function of one variable.

Linear Least Squares

In a least squares fit of a linear model

$$y = X\beta + \epsilon, \tag{0.3.56}$$

where y is an n-vector, X is an $n \times m$ matrix, and β is an m-vector, we replace β by a variable b, define the residual vector

$$r = y - Xb, \tag{0.3.57}$$

and minimize its Euclidean norm,

$$f(b) = r^{\mathrm{T}} r, \tag{0.3.58}$$

with respect to the variable b. We can solve this optimization problem by taking the derivative of this sum of squares and equating it to zero. Doing this, we get

$$\begin{aligned} \frac{\mathrm{d}(y - Xb)^{\mathrm{T}}(y - Xb)}{\mathrm{d}b} &= \frac{\mathrm{d}(y^{\mathrm{T}}y - 2b^{\mathrm{T}}X^{\mathrm{T}}y + b^{\mathrm{T}}X^{\mathrm{T}}Xb)}{\mathrm{d}b} \\ &= -2X^{\mathrm{T}}y + 2X^{\mathrm{T}}Xb \\ &= 0, \end{aligned}$$

which yields the normal equations

$$X^{\mathrm{T}}Xb = X^{\mathrm{T}}y. \tag{0.3.59}$$

The solution to the normal equations is a stationary point of the function (0.3.58). The Hessian of $(y - Xb)^{\mathrm{T}}(y - Xb)$ with respect to b is $2X^{\mathrm{T}}X$ and

$$X^{\mathrm{T}}X \succeq 0.$$

Because the matrix of second derivatives is nonnegative definite, the value of b that solves the system of equations arising from the first derivatives is a local minimum of equation (0.3.58).

Quasi-Newton Methods

All gradient-descent methods determine the path $p^{(k)}$ to take in the k^{th} step by a system of equations of the form

$$R^{(k)}p^{(k)} = -\mathrm{g}_f\big(x^{(k-1)}\big).$$

In the steepest-descent method, $R^{(k)}$ is the identity, I, in these equations. For functions with eccentric contours, the steepest-descent method traverses a zigzag path to the minimum. In Newton's method, $R^{(k)}$ is the Hessian evaluated at the previous point, $\mathrm{H}_f\big(x^{(k-1)}\big)$, which results in a more direct path to the minimum. Aside from the issues of consistency of the resulting equation and the general problems of reliability, a major disadvantage of Newton's method is the computational burden of computing the Hessian, which requires $\mathrm{O}(m^2)$ function evaluations, and solving the system, which requires $\mathrm{O}(m^3)$ arithmetic operations, at each iteration.

Instead of using the Hessian at each iteration, we may use an approximation, $B^{(k)}$. We may choose approximations that are simpler to update and/or that allow the equations for the step to be solved more easily. Methods using such approximations are called *quasi-Newton* methods or *variable metric* methods.

Because

$$\mathrm{H}_f\big(x^{(k)}\big)\big(x^{(k)} - x^{(k-1)}\big) \approx \mathrm{g}_f\big(x^{(k)}\big) - \mathrm{g}_f\big(x^{(k-1)}\big),$$

we choose $B^{(k)}$ so that

$$B^{(k)}\big(x^{(k)} - x^{(k-1)}\big) = \mathrm{g}_f\big(x^{(k)}\big) - \mathrm{g}_f\big(x^{(k-1)}\big). \tag{0.3.60}$$

This is called the *secant condition*.

We express the secant condition as

$$B^{(k)}s^{(k)} = y^{(k)}, \tag{0.3.61}$$

where

$$s^{(k)} = x^{(k)} - x^{(k-1)}$$

and

$$y^{(k)} = \mathrm{g}_f(x^{(k)}) - \mathrm{g}_f(x^{(k-1)}),$$

as above.

The system of equations in (0.3.61) does not fully determine $B^{(k)}$ of course. Because $B^{(k)}$ should approximate the Hessian, we may require that it be symmetric and positive definite.

The most common approach in quasi-Newton methods is first to choose a reasonable starting matrix $B^{(0)}$ and then to choose subsequent matrices by additive updates,

$$B^{(k+1)} = B^{(k)} + B_a^{(k)}, \tag{0.3.62}$$

subject to preservation of symmetry and positive definiteness. An approximate Hessian $B^{(k)}$ may be used for several iterations before it is updated; that is, $B_a^{(k)}$ may be taken as 0 for several successive iterations.

Multiparameter Likelihood Functions

For a sample $y = (y_1, \ldots, y_n)$ from a probability distribution with probability density function $p(\cdot; \theta)$, the *likelihood function* is

$$L(\theta;\, y) = \prod_{i=1}^{n} p(y_i; \theta), \tag{0.3.63}$$

and the *log-likelihood function* is $l(\theta;\, y) = \log(L(\theta;\, y))$. It is often easier to work with the log-likelihood function.

The log-likelihood is an important quantity in information theory and in unbiased estimation. If Y is a random variable with the given probability density function with the r-vector parameter θ, the *Fisher information* matrix that Y contains about θ is the $r \times r$ matrix

$$I(\theta) = \mathrm{Cov}_\theta \left(\frac{\partial l(t, Y)}{\partial t_i}, \frac{\partial l(t, Y)}{\partial t_j} \right), \tag{0.3.64}$$

where Cov_θ represents the variance-covariance matrix of the functions of Y formed by taking expectations for the given θ. (I use different symbols here because the derivatives are taken with respect to a *variable*, but the θ in Cov_θ cannot be the variable of the differentiation. This distinction is somewhat pedantic, and sometimes I follow the more common practice of using the same symbol in an expression that involves both Cov_θ and $\partial l(\theta, Y)/\partial \theta_i$.)

For example, if the distribution is the d-variate normal distribution with mean d-vector μ and $d \times d$ positive definite variance-covariance matrix Σ, the likelihood, equation (0.3.63), is

$$L(\mu, \Sigma;\, y) = \frac{1}{\left((2\pi)^{d/2} |\Sigma|^{1/2}\right)^n} \exp\left(-\frac{1}{2} \sum_{i=1}^{n} (y_i - \mu)^{\mathrm{T}} \Sigma^{-1} (y_i - \mu) \right).$$

(Note that $|\Sigma|^{1/2} = |\Sigma^{\frac{1}{2}}|$. The square root matrix $\Sigma^{\frac{1}{2}}$ is often useful in transformations of variables.)

Anytime we have a quadratic form that we need to simplify, we should recall the useful fact: $x^{\mathrm{T}}Ax = \mathrm{tr}(Axx^{\mathrm{T}})$. Using this, and because, as is often the case, the log-likelihood is easier to work with, we write

$$l(\mu, \Sigma; y) = c - \frac{n}{2}\log|\Sigma| - \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\sum_{i=1}^{n}(y_i - \mu)(y_i - \mu)^{\mathrm{T}}\right), \tag{0.3.65}$$

where we have used c to represent the constant portion. Next, we use the "Pythagorean equation" on the outer product to get

$$l(\mu, \Sigma; y) = c - \frac{n}{2}\log|\Sigma| - \frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}\sum_{i=1}^{n}(y_i - \bar{y})(y_i - \bar{y})^{\mathrm{T}}\right) - \frac{n}{2}\mathrm{tr}\left(\Sigma^{-1}(\bar{y} - \mu)(\bar{y} - \mu)^{\mathrm{T}}\right). \tag{0.3.66}$$

In maximum likelihood estimation, we seek the maximum of the likelihood function (0.3.63) with respect to θ while we consider y to be fixed. If the maximum occurs within an open set and if the likelihood is differentiable, we might be able to find the maximum likelihood estimates by differentiation. In the log-likelihood for the d-variate normal distribution, we consider the parameters μ and Σ to be variables. To emphasize that perspective, we replace the parameters μ and Σ by the variables $\hat{\mu}$ and $\widehat{\Sigma}$. Now, to determine the maximum, we could take derivatives with respect to $\hat{\mu}$ and $\widehat{\Sigma}$, set them equal to 0, and solve for the maximum likelihood estimates. Some subtle problems arise that depend on the fact that for any constant vector a and scalar b, $\Pr(a^{\mathrm{T}}X = b) = 0$, but we do not interpret the likelihood as a probability.

Often in working out maximum likelihood estimates, students immediately think of differentiating, setting to 0, and solving. As noted above, this requires that the likelihood function be differentiable, that it be concave, and that the maximum occur at an interior point of the parameter space. Keeping in mind exactly what the problem is — one of finding a maximum — often leads to the correct solution more quickly.

0.3.5 Vector Random Variables

The simplest kind of vector random variable is one whose elements are independent. Such random vectors are easy to work with because the elements can be dealt with individually, but they have limited applications. More interesting random vectors have a multivariate structure that depends on the relationships of the distributions of the individual elements. The simplest nondegenerate multivariate structure is of second degree; that is, a covariance or correlation structure. The probability density of a random vector with a multivariate structure generally is best represented by using matrices. In the case

of the multivariate normal distribution, the variances and covariances together with the means completely characterize the distribution. For example, the fundamental integral that is associated with the d-variate normal distribution, sometimes called Aitken's integral, equation (0.0.88) on page 682, provides that constant. The rank of the integral is the same as the rank of the integrand. ("Rank" is used here in the sense of "number of dimensions".) In this case, the integrand and the integral are scalars.

Equation (0.0.88) is a simple result that follows from the evaluation of the individual single integrals after making the change of variables $y_i = x_i - \mu_i$. If Σ^{-1} is positive definite, Aitken's integral can also be evaluated by writing $P^{\mathrm{T}}\Sigma^{-1}P = I$ for some nonsingular matrix P. Now, after the translation $y = x - \mu$, which leaves the integral unchanged, we make the linear change of variables $z = P^{-1}y$, with the associated Jacobian $|\det(P)|$. From $P^{\mathrm{T}}\Sigma^{-1}P = I$, we have $|\det(P)| = (\det(\Sigma))^{1/2} = |\Sigma|^{1/2}$ because the determinant is positive. Aitken's integral therefore is

$$\begin{aligned}\int_{\mathbb{R}^d} \mathrm{e}^{-y^{\mathrm{T}}\Sigma^{-1}y/2}\,\mathrm{d}y &= \int_{\mathbb{R}^d} \mathrm{e}^{-(Pz)^{\mathrm{T}}\Sigma^{-1}Pz/2}\,(\det(\Sigma))^{1/2}\mathrm{d}z \\ &= \int_{\mathbb{R}^d} \mathrm{e}^{-z^{\mathrm{T}}z/2}\,\mathrm{d}z\,(\det(\Sigma))^{1/2} \\ &= (2\pi)^{d/2}(\det(\Sigma))^{1/2}.\end{aligned}$$

The expected value of a function f of the vector-valued random variable X is

$$\mathrm{E}(f(X)) = \int_{D(X)} f(x)p_X(x)\,\mathrm{d}x, \tag{0.3.67}$$

where $D(X)$ is the support of the distribution, $p_X(x)$ is the probability density function evaluated at x, and $x\ \mathrm{d}x$ are dummy vectors whose elements correspond to those of X. Interpreting $\int_{D(X)} \mathrm{d}x$ as a nest of univariate integrals, the result of the integration of the vector $f(x)p_X(x)$ is clearly of the same type as $f(x)$. For example, if $f(x) = x$, the expectation is the mean, which is a vector. For the normal distribution, we have

$$\begin{aligned}\mathrm{E}(X) &= (2\pi)^{-d/2}|\Sigma|^{-1/2}\int_{\mathbb{R}^d} x\mathrm{e}^{-(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)/2}\,\mathrm{d}x \\ &= \mu.\end{aligned}$$

For the variance of the vector-valued random variable X,

$$\mathrm{V}(X),$$

the function f in expression (0.3.67) above is the matrix $(X - \mathrm{E}(X))(X - \mathrm{E}(X))^{\mathrm{T}}$, and the result is a matrix. An example is the normal variance:

$$\begin{aligned}\mathrm{V}(X) &= \mathrm{E}\left((X - \mathrm{E}(X))(X - \mathrm{E}(X))^{\mathrm{T}}\right) \\ &= (2\pi)^{-d/2}|\Sigma|^{-1/2}\int_{\mathbb{R}^d} \left((x-\mu)(x-\mu)^{\mathrm{T}}\right)\mathrm{e}^{-(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)/2}\,\mathrm{d}x \\ &= \Sigma.\end{aligned}$$

0.3.6 Transition Matrices

An important use of matrices in statistics is in models of transitions of a stochastic process from one state to another. In a discrete-state Markov chain, for example, the probability of going from state j to state i may be represented as elements of a *transition matrix*, which can any square matrix with nonnegative elements and such that the sum of the elements in any column is 1. Any square matrix with nonnegative elements whose columns each sum to 1 is called a *right stochastic matrix*.

(Note that many people who work with Markov chains define the transition matrix as the transpose of K above. This is not a good idea, because in applications with state vectors, the state vectors would naturally have to be row vectors. Until about the middle of the twentieth century, many mathematicians thought of vectors as row vectors; that is, a system of linear equations would be written as $xA = b$. Nowadays, almost all mathematicians think of vectors as column vectors in matrix algebra. Even in some of my previous writings, e.g., Gentle (2007), I have called the transpose of K the transition matrix, and I defined a stochastic matrix in terms of the transpose. I think that it is time to adopt a notation that is more consistent with current matrix/vector notation. This is merely a change in notation; no concepts require any change.)

There are various properties of transition matrices that are important for studying Markov chains.

Irreducible Matrices

Any nonnegative square matrix that can be permuted into the form

$$\begin{bmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{bmatrix}$$

with square diagonal submatrices is said to be *reducible*; a matrix that cannot be put into that form is *irreducible*. An alternate term for reducible is *decomposable*, with the associated term *indecomposable*.

We see from the definition that a positive matrix is irreducible.

We now consider irreducible square nonnegative matrices. This class includes positive matrices.

Irreducible matrices have several interesting properties. An $n \times n$ nonnegative matrix A is irreducible if and only if $(I + A)^{n-1}$ is a positive matrix; that is,

$$A \text{ is irreducible} \iff (I + A)^{n-1} > 0. \tag{0.3.68}$$

To see this, first assume $(I+A)^{n-1} > 0$; thus, $(I+A)^{n-1}$ clearly is irreducible. If A is reducible, then there exists a permutation matrix E_π such that

$$E_\pi^{\mathrm{T}} A E_\pi = \begin{bmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{bmatrix},$$

and so

$$\begin{aligned} E_\pi^{\mathrm{T}}(I+A)^{n-1}E_\pi &= \left(E_\pi^{\mathrm{T}}(I+A)E_\pi\right)^{n-1} \\ &= \left(I+E_\pi^{\mathrm{T}}AE_\pi\right)^{n-1} \\ &= \begin{bmatrix} I_{n_1}+B_{11} & B_{12} \\ 0 & I_{n_2}+B_{22} \end{bmatrix}. \end{aligned}$$

This decomposition of $(I+A)^{n-1}$ cannot exist because it is irreducible; hence we conclude A is irreducible if $(I+A)^{n-1} > 0$. We can see that $(I+A)^{n-1}$ must be a positive matrix by first observing that the $(i,j)^{\text{th}}$ element of $(I+A)^{n-1}$ can be expressed as

$$\left((I+A)^{n-1}\right)_{ij} = \left(\sum_{k=0}^{n-1}\binom{n-1}{k}A^k\right)_{ij}. \tag{0.3.69}$$

Hence, for $k = 1,\ldots,n-1$, we consider the $(i,j)^{\text{th}}$ entry of A^k. Let $a_{ij}^{(k)}$ represent this quantity.

Given any pair (i,j), for some $l_1, l_2, \ldots, l_{k-1}$, we have

$$a_{ij}^{(k)} = \sum_{l_1,l_2,\ldots,l_{k-1}} a_{1l_1}a_{l_1l_2}\cdots a_{l_{k-1}j}.$$

Now $a_{ij}^{(k)} > 0$ if and only if $a_{1l_1}, a_{l_1l_2}, \ldots, a_{l_{k-1}j}$ are all positive; that is, if there is a path $v_1, v_{l_1}, \ldots, v_{l_{k-1}}, v_j$ in $\mathcal{G}(A)$. If A is irreducible, then $\mathcal{G}(A)$ is strongly connected, and hence the path exists. So, for any pair (i,j), we have from equation (0.3.69) $\left((I+A)^{n-1}\right)_{ij} > 0$; that is, $(I+A)^{n-1} > 0$.

The positivity of $(I+A)^{n-1}$ for an irreducible nonnegative matrix A is a very useful property because it allows us to extend some conclusions of the Perron theorem to irreducible nonnegative matrices.

Properties of Square Irreducible Nonnegative Matrices; the Perron-Frobenius Theorem

If A is a square irreducible nonnegative matrix, then we have the following properties. These following properties are the conclusions of the *Perron-Frobenius theorem*.

1. $\rho(A)$ is an eigenvalue of A. This eigenvalue is called the *Perron root*, as before.
2. The Perron root $\rho(A)$ is simple. (That is, the algebraic multiplicity of the Perron root is 1.)
3. The dimension of the eigenspace of the Perron root is 1. (That is, the geometric multiplicity of $\rho(A)$ is 1.)
4. The eigenvector associated with $\rho(A)$ is positive. This eigenvector is called the *Perron vector*, as before.

The relationship (0.3.68) allows us to prove properties 1 and 4.

The one property of square positive matrices that does not carry over to square irreducible nonnegative matrices is that $r = \rho(A)$ is the only eigenvalue on the spectral circle of A. For example, the small irreducible nonnegative matrix

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

has eigenvalues 1 and -1, and so both are on the spectral circle.

It turns out, however, that square irreducible nonnegative matrices that have only one eigenvalue on the spectral circle also have other interesting properties that are important, for example, in Markov chains. We therefore give a name to the property:

> A square irreducible nonnegative matrix is said to be *primitive* if it has only one eigenvalue on the spectral circle.

In modeling with Markov chains and other applications, the limiting behavior of A^k is an important property.

If A is a primitive matrix, then we have the useful result

$$\lim_{k\to\infty} \left(\frac{A}{\rho(A)}\right)^k = vw^{\mathrm{T}}, \tag{0.3.70}$$

where v is an eigenvector of A associated with $\rho(A)$ and w is an eigenvector of A^{T} associated with $\rho(A)$, and w and v are scaled so that $w^{\mathrm{T}}v = 1$. (Such eigenvectors exist because $\rho(A)$ is a simple eigenvalue. They also exist because they are both positive. Note that A is not necessarily symmetric, and so its eigenvectors may include imaginary components; however, the eigenvectors associated with $\rho(A)$ are real, and so we can write w^{T} instead of w^{H}.)

To see equation (0.3.70), we consider $\left(A - \rho(A)vw^{\mathrm{T}}\right)$. First, if (c_i, v_i) is an eigenpair of $\left(A - \rho(A)vw^{\mathrm{T}}\right)$ and $c_i \neq 0$, then (c_i, v_i) is an eigenpair of A. We can see this by multiplying both sides of the eigen-equation by vw^{T}:

$$\begin{aligned} c_i vw^{\mathrm{T}}v_i &= vw^{\mathrm{T}}\left(A - \rho(A)vw^{\mathrm{T}}\right)v_i \\ &= \left(vw^{\mathrm{T}}A - \rho(A)vw^{\mathrm{T}}vw^{\mathrm{T}}\right)v_i \\ &= \left(\rho(A)vw^{\mathrm{T}} - \rho(A)vw^{\mathrm{T}}\right)v_i \\ &= 0; \end{aligned}$$

hence,

$$\begin{aligned} Av_i &= \left(A - \rho(A)vw^{\mathrm{T}}\right)v_i \\ &= c_i v_i. \end{aligned}$$

Next, we show that

$$\rho\left(A - \rho(A)vw^{\mathrm{T}}\right) < \rho(A). \tag{0.3.71}$$

If $\rho(A)$ were an eigenvalue of $\left(A - \rho(A)vw^{\mathrm{T}}\right)$, then its associated eigenvector, say w, would also have to be an eigenvector of A, as we saw above. But since as an eigenvalue of A the geometric multiplicity of $\rho(A)$ is 1, for some scalar s, $w = sv$. But this is impossible because that would yield

$$\begin{aligned}\rho(A)sv &= \left(A - \rho(A)vw^{\mathrm{T}}\right)sv \\ &= sAv - s\rho(A)v \\ &= 0,\end{aligned}$$

and neither $\rho(A)$ nor sv is zero. But as we saw above, any eigenvalue of $\left(A - \rho(A)vw^{\mathrm{T}}\right)$ is an eigenvalue of A and no eigenvalue of $\left(A - \rho(A)vw^{\mathrm{T}}\right)$ can be as large as $\rho(A)$ in modulus; therefore we have inequality (0.3.71).

Finally, with w and v as defined above, and with the eigenvalue $\rho(A)$,

$$\left(A - \rho(A)vw^{\mathrm{T}}\right)^k = A^k - (\rho(A))^k vw^{\mathrm{T}}, \tag{0.3.72}$$

for $k = 1, 2, \ldots$.

Dividing both sides of equation (0.3.72) by $(\rho(A))^k$ and rearranging terms, we have

$$\left(\frac{A}{\rho(A)}\right)^k = vw^{\mathrm{T}} + \frac{\left(A - \rho(A)vw^{\mathrm{T}}\right)}{\rho(A)}. \tag{0.3.73}$$

Now

$$\rho\left(\frac{\left(A - \rho(A)vw^{\mathrm{T}}\right)}{\rho(A)}\right) = \frac{\rho\left(A - \rho(A)vw^{\mathrm{T}}\right)}{\rho(A)},$$

which is less than 1; hence, we have

$$\lim_{k\to\infty}\left(\frac{\left(A - \rho(A)vw^{\mathrm{T}}\right)}{\rho(A)}\right)^k = 0;$$

so, taking the limit in equation (0.3.73), we have equation (0.3.70).

Applications of the Perron-Frobenius theorem are far-ranging.

Notes and References for Section 0.3

Matrix algebra arises in many areas of statistics. The study and application of linear models is inseparable from matrix/vector operations. In other areas of statistics, such as stochastic processes, matrices play an important role. In these areas, the matrices are often of a type different from the important ones in linear models.

There are many texts on matrix algebra, some with an orientation toward applications in statistics. I have referred often to Gentle (2007) just because I am most familiar with it. Harville (1997) is a large theorem-proof compendium of facts about matrices that are especially useful in linear models. Almost half of Kollo and von Rosen (2005) is devoted to matrix theory. The rest of the book covers various multivariate distributions.

0.4 Optimization

Optimization problems — maximization or minimization — arise in many areas of statistics. Statistical estimation and modeling both are usually special types of optimization problems. In a common method of statistical estimation, we *maximize* a likelihood, which is a function proportional to a probability density at the point of the observed data. In another method of estimation and in standard modeling techniques, we *minimize* a norm of the residuals. The best fit of a model is often defined in terms of a minimum of a norm, such as least squares. Other uses of optimization in statistical applications occur prior to collection of data, for example, when we design an experiment or a survey so as to minimize experimental or sampling errors.

When a statistical method is based on the solution of an optimization problem, to formulate that problem unambiguously helps us both to understand the method and to decide whether the method is appropriate to the purposes for which it is applied.

Some of the simpler and more common optimization problems in statistics can be solved easily, often by solving a system of linear equations. Many other problems, however, do not have closed-form solutions, and the solutions must be approximated by iterative methods.

0.4.1 Overview of Optimization

Optimization means to find a **maximum** or a **maximum** of an **objective function**, $f : D \subseteq \mathbb{R}^d \mapsto \mathbb{R}$.

Local optimization means optimization within a some subset of the domain of the objective function **Global** optimization results in the optimum of all local optima.

In **unconstrained** optimization, we take all points in D to be feasible.

Important Properties of the Objective Function

- domain dense or not
- differentiable or not
  - to what order
  - easy or hard to compute
- concave (or convex) or neither
  - if neither, there may be **local** optima

In the following, let $f(x)$ be the objective function, and assume we want to maximize it.
(To minimize, $f(x) \leftarrow -f(x)$ and convex $\leftarrow$ concave.)

Methods

- **Analytic**: yields closed form for all local maxima.
- **Iterative**: for $k = 1, 2, \ldots$, given $x^{(k-1)}$ choose $x^{(k)}$ so that $f(x^{(k)}) \rightarrow$local maximum of f.
  We need
  - a **starting point:** $x^{(0)}$;
  - a method to **choose** $\tilde{x}$ with good prospects of being $x^{(k)}$;
  - a method to **decide** whether $\tilde{x}$ should be $x^{(k)}$.

  How we **choose** and **decide** determines the differences between optimization algorithms.

How to **choose** may be based on derivative information or on some systematic method of exploring the domain.

How to **decide** may be based on a deterministic criterion, such as requiring $f(x^{(k)}) > f(x^{(k-1)})$,
or the decision may be randomized.

Metamethods: General Tools

- Transformations (for either analytic or iterative methods).
- Any trick you can think of (for either analytic or iterative methods), e.g., alternating conditional optimization.
- Conditional bounding functions (for iterative methods).

Convergence of Iterative Algorithms

In an iterative algorithm, we have a sequence $\left\{\left(f\left(x^{(k)}\right), x^{(k)}\right)\right\}$.

The first question is *whether* the sequence converges to the correct solution.

If there is a unique maximum, and if x_* is the point at which the maximum occurs, the first question can be posed more precisely as, given ϵ_1 does there exist M_1 such that for $k > M_1$,

$$\left|f(x^{(k)}) - f(x_*)\right| < \epsilon_1;$$

or, alternatively, for a given ϵ_2 does there exist M_2 such that for $k > M_2$,

$$\left\|x^{(k)} - x_*\right\| < \epsilon_2.$$

Recall that $f : \mathbb{R}^d \mapsto \mathbb{R}$, so $|\cdot|$ above is the absolute value, while $\|\cdot\|$ is some kind of vector norm.

There are complications if x_* is not unique as the point at which the maximum occurs.

Similarly, there are comlications if x_* is merely a point of local maximum.

Assessing Convergence

In practice, we must *decide* when convergence has occurred; that is, whether the iterations have become close enough to the solution. Since we don't know the solution, we cannot base this decision on the convergence criteria above.

We put faith in our algorithm, and decide convergence has occurred if, for some $e_1, e_2 > 0$, either

$$|f(x^{(k)}) - f(x^{(k-1)})| \leq e_1$$

or

$$\|x^{(k)} - x^{(k-1)}\| \leq e_2,$$

or both.

Notice, that lacking any known value, we trust the algorithm to do the right thing; both $x^{(k)}$ and $x^{(k-1)}$ are just values in the algorithmic sequence. The fact that this particular sequence — or any sequence, even ones yielding nonincreasing function values — converges does not really get at the question of whether $x^{(k)} \to x_*$.

Note also the subtle change from "$<$" to "$\leq$".

For some special class of functions $\nabla f(x)$ may exist, and we may know that at the solution, $\nabla f(x_*) = 0$. In these cases, we may have another criterion for deciding convergence has occurred:

$$\|\nabla f(x_*)\| \leq e_3.$$

Rate of Convergence of Iterative Algorithms

If the answer to the first question is "yes", that is, if the algorithmic sequence converges, the next question is *how fast* the sequence converges. (We address this question assuming it converges to the correct solution.)

The rate of convergence is a measure of how fast the "error" decreases. Any of three quantities we mentioned in discussing convergence, $f(x^{(k)}) - f(x_*)$, $x^{(k)} - x_*$, or $\nabla f(x_*)$, could be taken to be the error. If we take

$$e_k = x^{(k)} - x_*$$

to be the error at step k, we might define the magnitude of the error as $\|e_k\|$ (for some norm $\|\cdot\|$). If the algorithm converges to the correct solution,

$$\lim_{k\to\infty} \|e_k\| = 0.$$

Our interest is in how fast $\|e_k\|$ decreases.

Sometimes there is no reasonable way of quantifying the rate at which this quantity decreases.

In the happy case (and a common case for simple algorithms), if there exist $r > 0$ and $c > 0$ such that

$$\lim_{k\to\infty} \frac{\|e_k\|}{\|e_{k-1}\|^r} = c,$$

we say the **rate of convergence** is r and the **rate constant** is c.

The Steps in Iterative Algorithms For a Special Class of Functions

The steps in iterative algorithms are often based on some analytic relationship between $f(x)$ and $f(x^{(k-1)})$. For a continuously differentiable function, the most common relationship is the Taylor series expansion:

$$\begin{aligned} f(x) = f(x^{(k-1)}) \; + \\ (x - x^{(k-1)})^{\mathrm{T}} \nabla f(x^{(k-1)}) \; + \\ \frac{1}{2}(x - x^{(k-1)})^{\mathrm{T}} \nabla^2 f(x^{(k-1)})(x - x^{(k-1)}) \; + \\ \cdots \end{aligned}$$

Note this limitation: "For a continuously differentiable function, ...".
We cannot use this method on just any old function.

In the following, we will consider only this restricted class of functions.

Steepest Ascent (Descent)

The steps are defined by truncating the Taylor series. A truncation to two terms yields the steepest ascent direction. For a steepest ascent step, we find $x^{(k)}$ along the path $\nabla f(x^{(k-1)})$ from $x^{(k-1)}$.

If $\nabla f(x^{(k-1)}) \geq 0$, then moving along the path $\nabla f(x^{(k-1)})$ can increase the function value. If $f(x)$ is bounded above (i.e., if the maximization problem makes sense), then at some point along this path, the function begins to decrease.

This point is not necessarily the maximum of the function, of course. Finding the maximum along the path, is a one-dimensional "line search".

After moving to the point $x^{(k)}$ in the direction of $\nabla f(x^{(k-1)})$, if $\nabla f(x^{(k)}) = 0$, we are at a stationary point. This may not be a maximum, but it is as good as we can do using steepest ascent from $x^{(k-1)}$. (In practice, we check $\|\nabla f(x^{(k)})\| \leq \epsilon$, for some norm $\|\cdot\|$ and some positive ϵ.)

If $\nabla f(x^{(k)}) < 0$ (remember we're maximizing the function), we change directions and move in the direction of $\nabla f(x^{(k)}) < 0$.

Knowing that we will probably be changing direction anyway, we often truncate the line search before we find the best $x^{(k)}$ in the direction of $\nabla f(x^{(k-1)})$.

Newton's Method

At the maximum x_*, $\nabla f(x_*) = 0$.
"Newton's method" for optimization is based on this fact.

Newton's method for optimization just solves the system of equations $\nabla f(x) = 0$ using

Newton's iterative method for solving equations:
to solve the system of n equations in n unknowns, $g(x) = 0$, we move from point $x^{(k-1)}$ to point $x^{(k)}$ by

$$x^{(k)} = x^{(k-1)} - \left(\nabla g(x^{(k-1)})^{\mathrm{T}}\right)^{-1} g(x^{(k-1)}).$$

Hence, applying this to solving $\nabla f(x^{(k-1)}) = 0$, we have the k^{th} step in Newton's method for optimization:

$$x^{(k)} = x^{(k-1)} - \nabla^2 f(x^{(k-1)})^{-1} \nabla f(x^{(k-1)}).$$

The direction of the step is $d_k = x^{(k)} - x^{(k-1)}$.

For numerical reasons, it is best to think of this as the problem of solving the equations

$$\nabla^2 f(x^{(k-1)}) d_k = -\nabla f(x^{(k-1)}),$$

and then taking $x^{(k)} = x^{(k-1)} + d_k$.

The Hessian

The Hessian $H(x) = \nabla^2 f(x)$ clearly plays an important role in Newton's method; if it is singular, the Newton step based on the solution to

$$\nabla^2 f(x^{(k-1)}) d_k = -\nabla f(x^{(k-1)}),$$

is undetermined.

The relevance of the Hessian goes far beyond this, however. The Hessian reveals important properties of the shape of the surface $f(x)$ at $x^{(k-1)}$.

The shape is especially interesting at a stationary point; that is a point x_* at which $\nabla f(x) = 0$.

If the Hessian is negative definite at x_*, $f(x_*)$ is a local maximum.

If the Hessian is positive definite at x_*, $f(x_*)$ is a local maximum.

If the Hessian is nonsingular, but neither negative definite nor positive definite at x_*, it is a saddlepoint.

If the Hessian is singular, the stationary point is none of the above.

In minimization problems, such as least squares, we hope the Hessian is positive definite, in which case the function is concave. In least squares fitting of the standard linear regression model, the Hessian is the famous $X^{\mathrm{T}}X$ matrix.

In maximization problems, such as MLE, it is particularly interesting to know whether $H(x)$ is negative definite everywhere (or -$H(x)$ is positive definite everywhere). In this case, the function is convex.

When $H(x)$ (in minimization problems or -$H(x)$ in maximization problems) is positive definite but nearly singular, it may be helpful to regularize the problem by adding a diagonal matrix with positive elements: $H(x) + D$.

One kind of regularization is ridge regression, in which the Hessian is replaced by $X^{\mathrm{T}}X + dI$.

Modifications of Newton's Method

In the basic Newton step, the direction d_k from $x^{(k-1)}$ is the best direction, but the point $d_k + x^{(k-1)}$ may not be the best point. In fact, the algorithm can often be speeded up by not going quite that far; that is, by "damping" the Newton step and taking $x^{(k)} = \alpha_k d_k + x^{(k-1)}$. This is a line search, and there are several ways of doing this. In the context of least squares, a common way of damping is the **Levenberg-Marquardt** method.

Rather than finding $\nabla^2 f(x^{(k-1)})$, we might find an approximate Hessian at $x^{(k-1)}$, $\widetilde{H}_k$, and then solve

$$\widetilde{H}_k d_k = -\nabla f(x^{(k-1)}).$$

This is called a **quasi-Newton** method.

In MLE, we may take the objective function to be the log likelihood, with the variable θ. In this case, the Hessian, $H(\theta)$, is $\partial^2 \log L(\theta;\ x)/\partial\theta(\partial\theta)^{\mathrm{T}}$. Under very general regularity conditions, the expected value of $H(\theta)$ is the negative of the expected value of $(\partial \log L(\theta;\ x)/\partial\theta)(\partial \log L(\theta;\ x)\partial\theta)^{\mathrm{T}}$, which is the Fisher information matrix, $I(\theta)$. This quantity plays an important role in statistical estimation. In MLE it is often possible to compute $I(\theta)$, and take the Newton step as

$$I(\theta^{(k)})d_k = \nabla \log L(\theta^{(k-1)};\ x).$$

This quasi-Newton method is called **Fisher scoring**.

More Modifications of Newton's Method

The method of solving the Newton or quasi-Newton equations may itself be iterative, such as a conjugate gradient or Gauss-Seidel method. (These are "inner loop iterations".) Instead of continuing the inner loop iterations to the solution, we may stop early. This is called a **truncated Newton method**.

The best gains in iterative algorithms often occur in the first steps. When the optimization is itself part of an iterative method, we may get an acceptable approximate solution to the optimization problem by stopping the optimization iterations early. Sometimes we may stop the optimization after just one iteration. If Newton's method is used, this is called a **one-step Newton method**.

0.4.2 Alternating Conditional Optimization

The computational burden in a single iteration for solving the optimization problem can sometimes be reduced by more than a linear amount by separating x into two subvectors. The optimum is then computed by alternating between computations involving the two subvectors, and the iterations proceed in a zigzag path to the solution.

Each of the individual sequences of iterations is simpler than the sequence of iterations on the full x.

For the problem

$$\min_x f(x)$$

if $x = (x_1, x_2)$ that is, x is a vector with at least two elements, and x_1 and x_2 may be vectors), an iterative alternating conditional optimization algorithm may start with $x_2^{(0)}$, and then for $k = 0, 1, \ldots,$

1. $x_1^{(k)} = \arg\min_{x_1} f\left(x_1, x_2^{(k-1)}\right)$
2. $x_2^{(k)} = \arg\min_{x_2} f\left(x_1^{(k)}, x_2\right)$

Use of Conditional Bounding Functions: MM Methods

In an iterative method to maximize $f(x)$, the idea, given $x^{(k-1)}$ at step k, is to try to find a function $g\left(x;\ x^{(k-1)}\right)$ with these properties:

- is easy to work with (that is, is easy to maximize)
- $g\left(x;\ x^{(k-1)}\right) \leq f(x) \quad \forall x$
- $g\left(x^{(k-1)};\ x^{(k-1)}\right) = f\left(x^{(k-1)}\right)$

If we can find $x^{(k)} \ni g\left(x^{(k)};\ x^{(k-1)}\right) > g\left(x^{(k-1)};\ x^{(k-1)}\right)$, we have the "sandwich inequality":

$$f\left(x^{(k)}\right) \geq g\left(x^{(k)};\ x^{(k-1)}\right) > g\left(x^{(k-1)};\ x^{(k-1)}\right) = f\left(x^{(k-1)}\right).$$

An equivalent (but more complicated) method for seeing this inequality uses the fact that

$$f\left(x^{(k)}\right) - g\left(x^{(k)};\ x^{(k-1)}\right) \geq f\left(x^{(k-1)}\right) - g\left(x^{(k)};\ x^{(k-1)}\right).$$

(From the properties above,

$$g\left(x;\ x^{(k-1)}\right) - f\left(x\right)$$

attains its maximum at $x^{(k-1)}$.)

Hence,

$$\begin{aligned} f\left(x^{(k)}\right) &= g\left(x^{(k)};\ x^{(k-1)}\right) + f\left(x^{(k)}\right) - g\left(x^{(k)};\ x^{(k-1)}\right) \\ &> g\left(x^{(k-1)};\ x^{(k-1)}\right) + f\left(x^{(k-1)}\right) - g\left(x^{(k-1)};\ x^{(k-1)}\right) \\ &= f\left(x^{(k-1)}\right). \end{aligned}$$

The relationship between $f\left(x^{(k)}\right)$ and $f\left(x^{(k-1)}\right)$, that is, whether we have "$>$" or "$\geq$" in the inequalities, depends on the relationship between $g\left(x^{(k)};\ x^{(k-1)}\right)$ and $g\left(x^{(k-1)};\ x^{(k-1)}\right)$.

We generally require $g\left(x^{(k)};\ x^{(k-1)}\right) > g\left(x^{(k-1)};\ x^{(k-1)}\right)$.
Clearly, the best step would be

$$x^{(k)} = \arg\min_x g\left(x;\ x^{(k-1)}\right),$$

but the overall efficiency of the method may be better if we don't work too hard to find the maximum, but just accept some $x^{(k)}$ that satisfies $g\left(x^{(k)};\ x^{(k-1)}\right) \leq g\left(x^{(k-1)};\ x^{(k-1)}\right)$.

After moving to $x^{(k)}$, we must find a new $g\left(x;\ x^{(k)}\right)$.

Equivalent notations:

$$g\left(x;\ x^{(k-1)}\right) \leftrightarrow g^{(k)}(x) \leftrightarrow g_k(x)$$

Note the logical difference in k and $k-1$, although both determine the same g.

The g that we maximize is a "minorizing" function.
Thus, we Minorize then Maximize: MM.
EM methods.
Alternatively, we Majorize then Minimize: MM.
Reference: Lange et al. (2000).

Maximization in Alternating Algorithms

In alternating multiple step methods such as alternating conditional maximization methods and methods that use a conditional bounding function, at least one of the alternating steps involves maximization of some function.

As we indicated in discussing conditional bounding functions, instead of finding a point that actually maximizes the function, which may be a difficult task, we may just find a point that increases the value of the function. Under this weaker condition, the methods still work.

We may relax the requirement even further, so that for some steps we only require that the function not be decreased. So long as we maintain the requirement that the function actually be increased in a sufficient number of steps, the methods still work.

The most basic requirement is that $g\left(x^{(k)};\ x^{(k)}\right) \geq g\left(x^{(k-1)};\ x^{(k-1)}\right)$. (Even this requirement is relaxed in the class of optimization algorithms based on annealing. A reason for relaxing this requirement may be to avoid getting trapped in local optima.)

0.4.3 Simulated Annealing

Stochastic optimization ******* Spall (2012)

Simulated annealing is a method that simulates the thermodynamic process in which a metal is heated to its melting temperature and then is allowed

to cool slowly so that its structure is frozen at the crystal configuration of lowest energy. In this process the atoms go through continuous rearrangements, moving toward a lower energy level as they gradually lose mobility due to the cooling. The rearrangements do not result in a monotonic decrease in energy, however. The density of energy levels at a given temperature ideally is exponential, the so-called Boltzmann distribution, with a mean proportional to the absolute temperature. (The constant of proportionality is called "Boltzmann's constant"). This is analogous to a sequence of optimization iterations that occasionally go uphill. If the function has local minima, going uphill occasionally is desirable.

Metropolis et al. (1953) developed a stochastic relaxation technique that simulates the behavior of a system of particles approaching thermal equilibrium. (This is the same paper that described the Metropolis sampling algorithm.) The energy associated with a given configuration of particles is compared to the energy of a different configuration. If the energy of the new configuration is lower than that of the previous one, the new configuration is immediately accepted. If the new configuration has a larger energy, it is accepted with a nonzero probability. This probability is larger for small increases than for large increases in the energy level. One of the main advantages of simulated annealing is that the process is allowed to move away from a local optimum.

Although the technique is heuristically related to the cooling of a metal, as in the application of Metropolis et al. (1953), it can be successfully applied to a wide range of optimization problems.

The Basic Algorithm

In simulated annealing, a "temperature" parameter controls the probability of moving uphill; when the temperature is high, the probability of acceptance of any given point is high, and the process corresponds to a pure random walk. When the temperature is low, however, the probability of accepting any given point is low; and in fact, only downhill points are accepted. The behavior at low temperatures corresponds to a gradient search.

As the iterations proceed and the points move lower on the surface (it is hoped), the temperature is successively lowered. An "annealing schedule" determines how the temperature is adjusted.

In the description of simulated annealing in Algorithm 0.1, recognizing the common applications in combinatorial optimization, we refer to the argument of the objective function as a "state", rather than as a "point".

Algorithm 0.1 Simulated Annealing

0. Set $k = 1$ and initialize state s.
1. Compute $T(k)$.
2. Set $i = 0$ and $j = 0$.

3. Generate state r and compute $\delta f = f(r) - f(s)$.
4. Based on δf, decide whether to move from state s to state r.
   If $\delta f \leq 0$,
       accept;
   otherwise,
       accept with a probability $P(\delta f, T(k))$.
   If state r is accepted, set $i = i + 1$.
5. If i is equal to the limit for the number of successes at a given temperature, go to step 1.
6. Set $j = j + 1$. If j is less than the limit for the number of iterations at given temperature, go to step 3.
7. If $i = 0$,
       deliver s as the optimum; otherwise,
       if $k < k_{\max}$,
           set $k = k + 1$ and go to step 1;
       otherwise,
           issue message that
           'algorithm did not converge in $k_{\max}$ iterations'. ∎

For optimization of a continuous function over a region, the state is a point in that region. A new state or point may be selected by choosing a radius r and point on the d dimensional sphere of radius r centered at the previous point. For a continuous objective function, the movement in step 3 of Algorithm 0.1 may be a random direction to step in the domain of the objective function. In combinatorial optimization, the selection of a new state in step 3 may be a random rearrangement of a given configuration.

Parameters of the Algorithm: The Probability Function

There are a number of tuning parameters to choose in the simulated annealing algorithm. These include such relatively simple things as the number of repetitions or when to adjust the temperature. The probability of acceptance and the type of temperature adjustments present more complicated choices.

One approach is to assume that at a given temperature, T, the states have a known probability density (or set of probabilities, if the set of states is countable), $p_S(s, T)$, and then to define an acceptance probability to move from state s_k to s_{k+1} in terms of the relative change in the probability density from $p_S(s_k, T)$ to $p_S(s_{k+1}, T)$. In the original application of Metropolis et al., the objective function was the energy of a given configuration, and the probability of an energy change of δf at temperature T is proportional to $\exp(-\delta f/T)$.

Even when there is no underlying probability model, the probability in step 4 of Algorithm 0.1 is often taken as

$$P(\delta f, T(k)) = e^{-\delta f/T(k)}, \tag{0.4.1}$$

although a completely different form could be used. The exponential distribution models energy changes in ensembles of molecules, but otherwise it has no intrinsic relationship to a given optimization problem.

The probability can be tuned in the early stages of the computations so that some reasonable proportion of uphill steps are taken.

Parameters of the Algorithm: The Cooling Schedule

There are various ways the temperature can be updated in step 1.

The probability of the method converging to the global optimum depends on a slow decrease in the temperature. In practice, the temperature is generally decreased by some proportion of its current value:

$$T(k+1) = b(k)T(k). \tag{0.4.2}$$

We would like to decrease T as rapidly as possible, yet have a high probability of determining the global optimum. Under the assumptions that the energy distribution is Gaussian and the acceptance probability is of the form (0.4.1), the probability of convergence goes to 1 if the temperature decreases as the inverse of the logarithm of the time, that is, if $b(k) = (\log(k))^{-1}$ in equation (0.4.2). Under the assumption that the energy distribution is Cauchy, a similar argument allows $b(k) = k^{-1}$, and a uniform distribution over bounded regions allows $b(k) = \exp(-c_k k^{1/d})$, where c_k is some constant, and d is the number of dimensions.

A constant temperature is often used in simulated annealing for optimization of continuous functions and the additive and multiplicative adjustments, $c(k)$ and $b(k)$ are usually taken as constants, rather than varying with k.

Notes and References for Section 0.4

There is an extensive literature on optimization, much of it concerned with practical numerical algorithms. Software for optimization is widely available, both in special-purpose programs and in general-purpose packages such as R and Matlab.

Appendices

A

Important Probability Distributions

Development of stochastic models is facilitated by identifying a few probability distributions that seem to correspond to a variety of data-generating processes, and then studying the properties of these distributions. In the following tables, I list some of the more useful distributions, both discrete distributions and continuous ones. The names listed are the most common names, although some distributions go by different names, especially for specific values of the parameters. In the first column, following the name of the distribution, the parameter space is specified.

There are two very special continuous distributions, for which I use special symbols: the uniform over the interval $[a, b]$, designated $\mathrm{U}(a, b)$, and the normal (or Gaussian), denoted by $\mathrm{N}(\mu, \sigma^2)$. Notice that the second parameter in the notation for the normal is the variance. Sometimes, such as in the functions in R, the second parameter of the normal distribution is the standard deviation instead of the variance. A normal distribution with $\mu = 0$ and $\sigma^2 = 1$ is called the standard normal. I also often use the notation $\phi(x)$ for the PDF of a standard normal and $\Phi(x)$ for the CDF of a standard normal, and these are generalized in the obvious way as $\phi(x|\mu, \sigma^2)$ and $\Phi(x|\mu, \sigma^2)$.

Except for the uniform and the normal, I designate distributions by a name followed by symbols for the parameters, for example, binomial(n, π) or gamma(α, β). Some families of distributions are subfamilies of larger families. For example, the usual gamma family of distributions is a the two-parameter subfamily of the three-parameter gamma.

There are other general families of probability distributions that are defined in terms of a differential equation or of a form for the CDF. These include the Pearson, Johnson, Burr, and Tukey's lambda distributions.

Most of the common distributions fall naturally into one of two classes. They have either a countable support with positive probability at each point in the support, or a continuous (dense, uncountable) support with zero probability for any subset with zero Lebesgue measure. The distributions listed in the following tables are divided into these two natural classes.

There are situations for which these two distinct classes are not appropriate. For many such situations, however, a mixture distribution provides an appropriate model. We can express a PDF of a mixture distribution as

$$p_M(y) = \sum_{j=1}^{m} \omega_j p_j(y \,|\, \theta_j),$$

where the m distributions with PDFs p_j can be either discrete or continuous. A simple example is a probability model for the amount of rainfall in a given period, say a day. It is likely that a nonzero probability should be associated with zero rainfall, but with no other amount of rainfall. In the model above, m is 2, ω_1 is the probability of no rain, p_1 is a degenerate PDF with a value of 1 at 0, $\omega_2 = 1 - \omega_1$, and p_2 is some continuous PDF over $\mathbb{R}_+$, possibly similar to a distribution in the exponential family.

A mixture family that is useful in robustness studies is the ϵ-mixture distribution family, which is characterized by a given family with CDF P that is referred to as the reference distribution, together with a point x_c and a weight ϵ. The CDF of a ϵ-mixture distribution family is

$$P_{x_c,\epsilon}(x) = (1-\epsilon)P(x) + \epsilon \mathrm{I}_{[x_c,\infty[}(x),$$

where $0 \leq \epsilon \leq 1$.

Another example of a mixture distribution is a binomial with constant parameter n, but with a nonconstant parameter π. In many applications, if an identical binomial distribution is assumed (that is, a constant π), it is often the case that "over-dispersion" will be observed; that is, the sample variance exceeds what would be expected given an estimate of some other parameter that determines the population variance. This situation can occur in a model, such as the binomial, in which a single parameter determines both the first and second moments. The mixture model above in which each p_j is a binomial PDF with parameters n and π_j may be a better model.

Of course, we can extend this kind of mixing even further. Instead of $\omega_j p_j(y \,|\, \theta_j)$ with $\omega_j \geq 0$ and $\sum_{j=1}^{m} \omega_j = 1$, we can take $\omega(\theta)p(y \,|\, \theta)$ with $\omega(\theta) \geq 0$ and $\int \omega(\theta)\, \mathrm{d}\theta = 1$, from which we recognize that $\omega(\theta)$ is a PDF and θ can be considered to be the realization of a random variable.

Extending the example of the mixture of binomial distributions, we may choose some reasonable PDF $\omega(\pi)$. An obvious choice is a beta PDF. This yields the *beta-binomial distribution*, with PDF

$$p_{X,\Pi}(x,\pi) = \binom{n}{x} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \pi^{x+\alpha-1}(1-\pi)^{n-x+\beta-1} \mathrm{I}_{\{0,1,\ldots,n\}\times]0,1[}(x,\pi).$$

This is a standard distribution but I did not include it in the tables below.

This distribution may be useful in situations in which a binomial model is appropriate, but the probability parameter is changing more-or-less continuously.

We recognize a basic property of any mixture distribution: It is a joint distribution factored as a marginal (prior) for a random variable, which is often not observable, and a conditional distribution for another random variable, which is usually the observable variable of interest.

In Bayesian analyses, the first two assumptions (a prior distribution for the parameters and a conditional distribution for the observable) lead immediately to a mixture distribution. The beta-binomial above arises in a canonical example of Bayesian analysis.

Some families of distributions are recognized because of their relationship to sampling distributions. These include the t, the chi-squared, and the Wishart. Other families are recognized because of their use as conjugate priors. These include the inverted chi-squared and the inverted Wishart.

General References

Evans et al. (2000) give general descriptions of 40 probability distributions. Balakrishnan and Nevzorov (2003) provide an overview of the important characteristics that distinguish different distributions and then describe the important characteristics of many common distributions. Leemis and McQueston (2008) present an interesting compact graph of the relationships among a large number of probability distributions. Likewise, Morris and Lock (2009) give a graph that illustrates various interrelationships among natural exponential families.

The six books by Johnson, Kotz et al. (Johnson et al. (1995a), Kotz et al. (2000), Johnson et al. (1997), Johnson et al. (1994), Johnson et al. (1995b), and Johnson et al. (2005)) contain a wealth of information above many families of distributions.

Currently, the most readily accessible summary of common probability distributions is Wikipedia: `http://wikipedia.org/` Search under the name of the distribution.

Table A.1. Discrete Distributions (PDFs are wrt counting measure)

discrete uniform	PDF	$\frac{1}{m}, \quad y = a_1, \ldots, a_m$
$a_1, \ldots, a_m \in \mathbb{R}$	mean	$\sum a_i/m$
	variance	$\sum(a_i - \bar{a})^2/m$, where $\bar{a} = \sum a_i/m$
Bernoulli	PDF	$\pi^y(1-\pi)^{1-y}, \quad y = 0, 1$
$\pi \in]0, 1[$	mean	π
	variance	$\pi(1-\pi)$
binomial (n Bernoullis)	PDF	$\binom{n}{y}\pi^y(1-\pi)^{n-y}, \quad y = 0, 1, \ldots, n$
$n = 1, 2, \ldots; \quad \pi \in]0, 1[$	CF	$(1 - \pi + \pi e^{it})^n$
	mean	$n\pi$
	variance	$n\pi(1-\pi)$
geometric	PDF	$\pi(1-\pi)^y, \quad$ y=0,1,2,...
$\pi \in]0, 1[$	mean	$(1-\pi)/\pi$
	variance	$(1-\pi)/\pi^2$
negative binomial (n geometrics)	PDF	$\binom{y+n-1}{n-1}\pi^n(1-\pi)^y, \quad y = 0, 1, 2, \ldots$
$n = 1, 2, \ldots; \quad \pi \in]0, 1[$	CF	$\left(\frac{\pi}{1-(1-\pi)e^{it}}\right)^n$
	mean	$n(1-\pi)/\pi$
	variance	$n(1-\pi)/\pi^2$
multinomial	PDF	$\frac{n!}{\prod y_i!}\prod_{i=1}^{d}\pi_i^{y_i}, \; y_i = 0, 1, \ldots, n, \; \sum y_i = n$
$n = 1, 2, \ldots,$	CF	$\left(\sum_{i=1}^{d}\pi_i e^{it_i}\right)^n$
for $i = 1, \ldots, d$, $\pi_i \in]0, 1[$, $\sum \pi_i = 1$	means	$n\pi_i$
	variances	$n\pi_i(1-\pi_i)$
	covariances	$-n\pi_i\pi_j$
hypergeometric	PDF	$\frac{\binom{M}{y}\binom{N-M}{n-y}}{\binom{N}{n}}$, $y = \max(0, n-N+M), \ldots, \min(n, M)$
$N = 2, 3, \ldots;$	mean	nM/N
$M = 1, \ldots, N$; $n = 1, \ldots, N$	variance	$(nM/N)(1-M/N)(N-n)/(N-1)$

continued ...

Table A.1. Discrete Distributions (continued)

Poisson	PDF	$\theta^y \mathrm{e}^{-\theta}/y!, \quad y = 0, 1, 2, \ldots$
$\theta \in \mathbb{R}_+$	CF	$\mathrm{e}^{\theta(\mathrm{e}^{\mathrm{i}t}-1)}$
	mean	θ
	variance	θ
power series	PDF	$\frac{h_y}{c(\theta)}\theta^y, \quad y = 0, 1, 2, \ldots$
$\theta \in \mathbb{R}_+$	CF	$\sum_y h_y(\theta \mathrm{e}^{\mathrm{i}t})^y / c(\theta)$
$\{h_y\}$ positive constants	mean	$\theta \frac{\mathrm{d}}{\mathrm{d}\theta}(\log(c(\theta))$
$c(\theta) = \sum_y h_y \theta^y$	variance	$\theta \frac{\mathrm{d}}{\mathrm{d}\theta}(\log(c(\theta)) + \theta^2 \frac{\mathrm{d}^2}{\mathrm{d}\theta^2}(\log(c(\theta))$
logarithmic	PDF	$-\frac{\pi^y}{y \log(1-\pi)}, \quad y = 1, 2, 3, \ldots$
$\pi \in]0, 1[$	mean	$-\pi/((1-\pi)\log(1-\pi))$
	variance	$-\pi(\pi + \log(1-\pi))/((1-\pi)^2(\log(1-\pi))^2)$
Benford's	PDF	$\log_b(y+1) - \log_b(y), \quad y = 1, \ldots, b-1$
b integer ≥ 3	mean	$b - 1 - \log_b((b-1)!)$

Table A.2. The Normal Distributions

normal; $\mathrm{N}(\mu, \sigma^2)$	PDF	$\phi(y\|\mu,\sigma^2) \stackrel{\text{def}}{=} \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-(y-\mu)^2/2\sigma^2}$
$\mu \in \mathbb{R}$; $\sigma \in \mathbb{R}_+$	CF	$\mathrm{e}^{\mathrm{i}\mu t - \sigma^2 t^2/2}$
	mean	μ
	variance	σ^2
multivariate normal; $\mathrm{N}_d(\mu, \Sigma)$	PDF	$\frac{1}{(2\pi)^{d/2}\lvert\Sigma\rvert^{1/2}} \mathrm{e}^{-(y-\mu)^{\mathrm{T}}\Sigma^{-1}(y-\mu)/2}$
$\mu \in \mathbb{R}^d$; $\Sigma \succ 0 \in \mathbb{R}^{d\times d}$	CF	$\mathrm{e}^{\mathrm{i}\mu^{\mathrm{T}}t - t^{\mathrm{T}}\Sigma t/2}$
	mean	μ
	covariance	Σ
matrix normal	PDF	$\frac{1}{(2\pi)^{nm/2}\lvert\Psi\rvert^{n/2}\lvert\Sigma\rvert^{m/2}} \mathrm{e}^{-\mathrm{tr}(\Psi^{-1}(Y-M)^{\mathrm{T}}\Sigma^{-1}(Y-M))/2}$
$M \in \mathbb{R}^{n\times m}$, $\Psi \succ 0 \in \mathbb{R}^{m\times m}$,	mean	M
$\Sigma \succ 0 \in \mathbb{R}^{n\times n}$	covariance	$\Psi \otimes \Sigma$
complex multivariate normal	PDF	$\frac{1}{(2\pi)^{d/2}\lvert\Sigma\rvert^{1/2}} \mathrm{e}^{-(z-\mu)^{\mathrm{H}}\Sigma^{-1}(z-\mu)/2}$
$\mu \in \mathbb{C}^d$, $\Sigma \succ 0 \in \mathbb{C}^{d\times d}$	mean	μ
	covariance	Σ

Table A.3. Sampling Distributions from the Normal Distribution

chi-squared; χ^2_ν	PDF	$\frac{1}{\Gamma(\nu/2)2^{\nu/2}} y^{\nu/2-1} \mathrm{e}^{-y/2} \, \mathrm{I}_{\mathbb{R}_+}(y)$
$\nu \in \mathbb{R}_+$	mean	ν
if $\nu \in \mathbb{Z}_+$,	variance	2ν
t	PDF	$\frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\nu\pi}} (1+y^2/\nu)^{-(\nu+1)/2}$
$\nu \in \mathbb{R}_+$	mean	0
	variance	$\nu/(\nu-2)$, for $\nu > 2$
F	PDF	$\frac{\nu_1^{\nu_1/2}\nu_2^{\nu_2/2}\Gamma(\nu_1+\nu_2)}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)} \frac{y^{\nu_1/2-1}}{(\nu_2+\nu_1 y)^{(\nu_1+\nu_2)/2}} \, \mathrm{I}_{\mathbb{R}_+}(y)$
$\nu_1, \nu_2 \in \mathbb{R}_+$	mean	$\nu_2/(\nu_2-2)$, for $\nu_2 > 2$
	variance	$2\nu_2^2(\nu_1+\nu_2-2)/(\nu_1(\nu_2-2)^2(\nu_2-4))$, for $\nu_2 > 4$
Wishart	PDF	$\frac{\lvert W\rvert^{(\nu-d-1)/2}}{2^{\nu d/2}\lvert\Sigma\rvert^{\nu/2}\Gamma_d(\nu/2)} \exp\left(-\mathrm{trace}(\Sigma^{-1}W)\right) \mathrm{I}_{\{M \mid M \succ 0 \in \mathbb{R}^{d\times d}\}}(W)$
$d = 1, 2, \ldots;$	mean	$\nu\Sigma$
$\nu > d-1 \in \mathbb{R}$;	covariance	$\mathrm{Cov}(W_{ij}, W_{kl}) = \nu(\sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk})$, where $\Sigma = (\sigma_{ij})$
$\Sigma \succ 0 \in \mathbb{R}^{d\times d}$		
noncentral chi-squared	PDF	$\frac{\mathrm{e}^{-\lambda/2}}{2^{\nu/2}} y^{\nu/2-1} \mathrm{e}^{-y/2} \sum_{k=0}^{\infty} \frac{(\lambda/2)^k}{k!} \frac{1}{\Gamma(\nu/2+k)2^k} y^k \, \mathrm{I}_{\mathbb{R}_+}(y)$
$\nu, \lambda \in \mathbb{R}_+$	mean	$\nu + \lambda$
	variance	$2(\nu + 2\lambda)$
noncentral t	PDF	$\frac{\nu^{\nu/2}\mathrm{e}^{-\lambda^2/2}}{\Gamma(\nu/2)\pi^{1/2}} (\nu+y^2)^{-(\nu+1)/2} \times$ $\sum_{k=0}^{\infty} \Gamma\left(\frac{\nu+k+1}{2}\right) \frac{(\lambda y)^k}{k!} \left(\frac{2}{\nu+y^2}\right)^{k/2}$
$\nu \in \mathbb{R}_+, \lambda \in \mathbb{R}$	mean	$\frac{\lambda(\nu/2)^{1/2}\Gamma((\nu-1)/2)}{\Gamma(\nu/2)}$, for $\nu > 1$
	variance	$\frac{\nu}{\nu-2}(1+\lambda^2) - \frac{\nu}{2}\lambda^2 \left(\frac{\Gamma((\nu-1)/2)}{\Gamma(\nu/2)}\right)^2$, for $\nu > 2$
noncentral F	PDF	$\left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2} \mathrm{e}^{-\lambda/2} y^{\nu_1/2-1} \left(\frac{\nu_2}{\nu_2+\nu_1 y}\right)^{\nu_1/2+\nu_2/2} \times$ $\sum_{k=0}^{\infty} \frac{(\lambda/2)^k \Gamma(\nu_2+\nu_1+k)}{\Gamma(\nu_2)\Gamma(\nu_1+k)k!} \left(\frac{\nu_1}{\nu_2}\right)^k y^k \left(\frac{\nu_2}{\nu_2+\nu_1 y}\right)^k \, \mathrm{I}_{\mathbb{R}_+}(y)$
$\nu_1, \nu_2, \lambda \in \mathbb{R}_+$	mean	$\nu_2(\nu_1+\lambda)/(\nu_1(\nu_2-2))$, for $\nu_2 > 2$
	variance	$2\left(\frac{\nu_2}{\nu_1}\right)^2 \left(\frac{(\nu_1+\lambda)^2 + (\nu_1+2\lambda)(\nu_2-2)}{(\nu_2-2)^2(\nu_2-4)}\right)$, for $\nu_2 > 4$

Table A.4. Distributions Useful as Priors for the Normal Parameters

inverted gamma	PDF	$\frac{1}{\Gamma(\alpha)\beta^\alpha}\left(\frac{1}{y}\right)^{\alpha+1} \mathrm{e}^{-1/\beta y}\, \mathrm{I}_{\overline{\mathbb{R}}_+}(y)$
$\alpha, \beta \in \mathbb{R}_+$	mean	$1/\beta(\alpha-1)$ for $\alpha > 1$
	variance	$1/(\beta^2(\alpha-1)^2(\alpha-2))$ for $\alpha > 2$
inverted chi-squared	PDF	$\frac{1}{\Gamma(\nu/2)2^{\nu/2}}\left(\frac{1}{y}\right)^{\nu/2+1} \mathrm{e}^{-1/2y}\, \mathrm{I}_{\overline{\mathbb{R}}_+}(y)$
$\nu \in \mathbb{R}_+$	mean	$1/(\nu-2)$ for $\nu > 2$
	variance	$2/((\nu-2)^2(\nu-4))$ for $\nu > 4$
inverted Wishart	PDF	**********
**********	mean	**********
	variance	**********

Table A.5. Distributions Derived from the Univariate Normal

lognormal	PDF	$\frac{1}{\sqrt{2\pi}\sigma} y^{-1} \mathrm{e}^{-(\log(y)-\mu)^2/2\sigma^2}\, \mathrm{I}_{\overline{\mathbb{R}}_+}(y)$
$\mu \in \mathbb{R};\ \sigma \in \mathbb{R}_+$	mean	$\mathrm{e}^{\mu+\sigma^2/2}$
	variance	$\mathrm{e}^{2\mu+\sigma^2}(\mathrm{e}^{\sigma^2}-1)$
inverse Gaussian	PDF	$\sqrt{\frac{\lambda}{2\pi y^3}} \mathrm{e}^{-\lambda(y-\mu)^2/2\mu^2 y}\, \mathrm{I}_{\overline{\mathbb{R}}_+}(y)$
$\mu, \lambda \in \mathbb{R}_+$	mean	μ
	variance	μ^3/λ
skew normal	PDF	$\frac{1}{\pi\sigma} \mathrm{e}^{-(y-\mu)^2/2\sigma^2} \int_{-\infty}^{\lambda(y-\mu)/\sigma} \mathrm{e}^{-t^2/2}\, \mathrm{d}t$
$\mu, \lambda \in \mathbb{R};\ \sigma \in \mathbb{R}_+$	mean	$\mu + \sigma\sqrt{\frac{2\lambda}{\pi(1+\lambda^2)}}$
	variance	$\sigma^2(1 - 2\lambda^2/\pi)$

Table A.6. Other Continuous Distributions (PDFs are wrt Lebesgue measure)

beta	PDF	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}\ \mathrm{I}_{[0,1]}(y)$
$\alpha, \beta \in \mathrm{IR}_+$	mean	$\alpha/(\alpha+\beta)$
	variance	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
Dirichlet	PDF	$\frac{\Gamma(\sum_{i=1}^{d+1}\alpha_i)}{\prod_{i=1}^{d+1}\Gamma(\alpha_i)} \prod_{i=1}^{d} y_i^{\alpha_i-1} \left(1-\sum_{i=1}^{d} y_i\right)^{\alpha_{d+1}-1} \mathrm{I}_{[0,1]^d}(y)$
$\alpha \in \mathrm{IR}_+^{d+1}$	mean	$\alpha/\lVert\alpha\rVert_1$ ($\alpha_{d+1}/\lVert\alpha\rVert_1$ is the "mean of Y_{d+1}".)
	variance	$\frac{\alpha\,(\lVert\alpha\rVert_1-\alpha)}{\lVert\alpha\rVert_1^2(\lVert\alpha\rVert_1+1)}$
uniform; $\mathrm{U}(\theta_1,\theta_2)$	PDF	$\frac{1}{\theta_2-\theta_1}\ \mathrm{I}_{[\theta_1,\theta_2]}(y)$
$\theta_1 < \theta_2 \in \mathrm{IR}$	mean	$(\theta_2+\theta_1)/2$
	variance	$(\theta_2^2-2\theta_1\theta_2+\theta_1^2)/12$
Cauchy	PDF	$\frac{1}{\pi\beta\left(1+\left(\frac{y-\gamma}{\beta}\right)^2\right)}$
$\gamma \in \mathrm{IR};\ \beta \in \mathrm{IR}_+$	mean	does not exist
	variance	does not exist
logistic	PDF	$\frac{\mathrm{e}^{-(y-\mu)/\beta}}{\beta(1+\mathrm{e}^{-(y-\mu)/\beta})^2}$
$\mu \in \mathrm{IR};\ \beta \in \mathrm{IR}_+$	mean	μ
	variance	$\beta^2\pi^2/3$
Pareto	PDF	$\frac{\alpha\gamma^\alpha}{y^{\alpha+1}}\ \mathrm{I}_{[\gamma,\infty[}(y)$
$\alpha, \gamma \in \mathrm{IR}_+$	mean	$\alpha\gamma/(\alpha-1)$ for $\alpha > 1$
	variance	$\alpha\gamma^2/((\alpha-1)^2(\alpha-2))$ for $\alpha > 2$
power function	PDF	$(y/\beta)^\alpha\ \mathrm{I}_{[0,\beta[}(y)$
$\alpha, \beta \in \mathrm{IR}_+$	mean	$\alpha\beta/(\alpha+1)$
	variance	$\alpha\beta^2/((\alpha+2)(\alpha+1)^2)$
von Mises	PDF	$\frac{1}{2\pi\mathrm{I}_0(\kappa)}\mathrm{e}^{\kappa\cos(x-\mu)}\ \mathrm{I}_{[\mu-\pi,\mu+\pi]}(y)$
$\mu \in \mathrm{IR};\ \kappa \in \mathrm{IR}_+$	mean	μ
	variance	$1-(\mathrm{I}_1(\kappa)/\mathrm{I}_0(\kappa))^2$

continued ...

Table A.6. Other Continuous Distributions (continued)

gamma	PDF	$\frac{1}{\Gamma(\alpha)\beta^\alpha} y^{\alpha-1} \mathrm{e}^{-y/\beta}\, \mathrm{I}_{\overline{\mathbb{R}}_+}(y)$
$\alpha, \beta \in \mathbb{R}_+$	mean	$\alpha\beta$
	variance	$\alpha\beta^2$
three-parameter gamma	PDF	$\frac{1}{\Gamma(\alpha)\beta^\alpha} (y-\gamma)^{\alpha-1} \mathrm{e}^{-(y-\gamma)/\beta}\, \mathrm{I}_{]\gamma,\infty[}(y)$
$\alpha, \beta \in \mathbb{R}_+; \gamma \in \mathbb{R}$	mean	$\alpha\beta + \gamma$
	variance	$\alpha\beta^2$
exponential	PDF	$\theta^{-1} \mathrm{e}^{-y/\theta}\, \mathrm{I}_{\overline{\mathbb{R}}_+}(y)$
$\theta \in \mathbb{R}_+$	mean	θ
	variance	θ^2
double exponential	PDF	$\frac{1}{2\theta} \mathrm{e}^{-\lvert y-\mu\rvert/\theta}$
$\mu \in \mathbb{R};\ \theta \in \mathbb{R}_+$	mean	μ
(folded exponential)	variance	$2\theta^2$
Weibull	PDF	$\frac{\alpha}{\beta} y^{\alpha-1} \mathrm{e}^{-y^\alpha/\beta}\, \mathrm{I}_{\overline{\mathbb{R}}_+}(y)$
$\alpha, \beta \in \mathbb{R}_+$	mean	$\beta^{1/\alpha}\Gamma(\alpha^{-1}+1)$
	variance	$\beta^{2/\alpha}\left(\Gamma(2\alpha^{-1}+1) - (\Gamma(\alpha^{-1}+1))^2\right)$
extreme value (Type I)	PDF	$\frac{1}{\beta} \mathrm{e}^{-(y-\alpha)/\beta} \exp(\mathrm{e}^{-(y-\alpha)/\beta})$
$\alpha \in \mathbb{R};\ \beta \in \mathbb{R}_+$	mean	$\alpha - \beta\Gamma'(1)$
	variance	$\beta^2\pi^2/6$

B

Useful Inequalities in Probability

Inequalities involving functions of events and random variables are important throughout the field of probability and statistics. Two important uses are for showing that one procedure is better than another and for showing that some sequence converges to a given object (a constant, a function, or a set).

In the following, for simplicity, we will assume $X \in \mathbb{R}$.

B.1 Preliminaries

A simple, but surprisingly useful inequality states that if $\mathrm{E}(X^2) < \infty$, then the variance is the minimum expectation of the form $\mathrm{E}((X-c)^2)$ for any constant c. In other words, the minimum of $\mathrm{E}((X-c)^2)$ occurs at $c = \mathrm{E}(X)$). We see this by writing

$$\begin{aligned}
\mathrm{E}((X-c)^2) &= \mathrm{E}((X-\mathrm{E}(X)+\mathrm{E}(X)-c)^2) \\
&= \mathrm{E}((X-\mathrm{E}(X))^2) + \mathrm{E}((\mathrm{E}(X)-c)^2) + \\
&\quad\; 2\mathrm{E}((X-\mathrm{E}(X))(\mathrm{E}(X)-c)) \\
&= \mathrm{E}((X-\mathrm{E}(X))^2) + \mathrm{E}((\mathrm{E}(X)-c)^2) \\
&\geq \mathrm{E}((X-\mathrm{E}(X))^2). \qquad \text{(B.1)}
\end{aligned}$$

We will use various inequalities often, so we collect a number of them in this appendix, where we have categorized them into five types depending of the kinds of expressions in the inequalities. These five types involve relations between

- $\Pr(X \in A_i)$ and $\Pr(X \in \cup A_i)$ or $\Pr(X \in \cap A_i)$, e.g., Bonferroni's
- $\Pr(X \in A)$ and $\mathrm{E}(f(X))$, e.g., Chebyshev
- $\mathrm{E}(f(X))$ and $f(\mathrm{E}(X))$, e.g., Jensen's
- $\mathrm{E}(f_1(X,Y))$ and $\mathrm{E}(f_2(X))$ and $\mathrm{E}(f_3(Y))$, e.g., covariance, Cauchy-Schwarz, information

- $\mathrm{V}(Y)$ and $\mathrm{V}\left(\mathrm{E}(Y|X)\right)$, e.g., Rao-Blackwell

Any special case of these involves an appropriate definition of A or f (e.g., nonnegative, convex, etc.)

A more general case of the inequalities is to replace distributions, and hence expected values, by conditioning on a sub-σ-field, $\mathcal{A}$.

For each type of inequality there is an essentially straightforward method of proof.

Some of these inequalities involve absolute values of the random variable. To work with these inequalities, it is useful to recall the triangle inequality for the absolute value of real numbers:

$$|x+y| \leq |x| + |y|. \tag{B.2}$$

We can prove this merely by considering all four cases for the signs of x and y.

This inequality generalizes immediately to $|\sum x_i| \leq \sum |x_i|$.

Expectations of absolute values of functions of random variables are functions of norms. (A *norm* is a function of x that (1) is positive unless $x = 0$ a.e., that (2) is equivariant to scalar multiplication, and that (3) satisfies the triangle inequality.) The important form $(\mathrm{E}(|X|^p))^{1/p}$ for $1 \leq p$ is an L_p norm, $\|X\|_p$. Some of the inequalities given below involving expectations of absolute values of random variables are essentially triangle inequalities and their truth establishes the expectation as a norm.

Some of the expectations discussed below are recognizable as familiar norms over vector spaces. For example, the expectation in Minkowski's inequality is essentially the L_p norm of a vector, which is defined for an n-vector x in a finite-dimensional vector space as $\|x\|_p \equiv (\sum |x_i|^p)^{1/p}$. Minkowski's inequality in this case is $\|x+y\|_p \leq \|x\|_p + \|y\|_p$. For $p = 1$, this is the triangle inequality for absolute values given above.

B.2 $\Pr(X \in A_i)$ and $\Pr(X \in \cup A_i)$ or $\Pr(X \in \cap A_i)$

These inequalities are often useful in working with order statistics and in tests of multiple hypotheses. Instead of $\Pr(X \in A_i)$, we may write $\Pr(A_i)$.

Theorem B.1 (Bonferroni's inequality)
Given events $A_1, \ldots, A_n$, we have

$$\Pr(\cap A_i) \geq \sum \Pr(A_i) - n + 1. \tag{B.3}$$

Proof. We will use induction. For $n = 1$, we have $\Pr(A_1) \geq \Pr(A_1)$, and for $n = 2$, we have $\Pr(A_1 \cap A_2) = \Pr(A_1) + \Pr(A_2) - \Pr(A_1 \cup A_2) \geq \Pr(A_1) + \Pr(A_2) - 1$.

Now assume $\Pr(\cap_{i=1}^k A_i) \geq \sum_{i=1}^k A_i \Pr(A_i) - k + 1$, and consider $\tilde{k} = k + 1$. We have (from the $n = 2$ case) $\Pr(\cap_{i=1}^k A_i \cap A_{k+1}) \geq \Pr(\cap_{i=1}^k A_i) +$

$\Pr(A_{k+1}) - 1$, and now simplifying and substituting the induction hypothesis for $\Pr(\cap_{i=1}^{k} A_i)$, we have

$$\Pr(\cap_{i=1}^{k+1} A_i) \geq \sum_{i=1}^{k+1} \Pr(A_i) - k.$$

∎

Corollary B.2.0.1
Given a random sample $X_1, \ldots, X_n$ and fixed constants $a_1, \ldots, a_n$. For the order statistics, $X_{(1)}, \ldots, X_{(n)}$, we have

$$\Pr\left(\cap\{X_{(i)} \leq a_i\}\right) \geq \prod \Pr(X_{(i)} \leq a_i). \tag{B.4}$$

Proof. Same. ∎

B.3 $\Pr(X \in A)$ and $\mathrm{E}(f(X))$

An important class of inequalities bound tail probabilities of a random variable, that is, limit the probability that the random variable will take on a value beyond some distance from the expected value of the random variable.

The important general form involving $\Pr(X \in A)$ and $\mathrm{E}(f(X))$ is Markov's inequality involving absolute moments. Chebyshev's inequality, of which there are two forms, is a special case of it. Useful generalizations of Markov's inequality involve sums of sequences of random variables. The Bernstein inequalities, including the special case of Hoeffding's inequality, and the Hájek-Rènyi inequality, including the special case of Kolmogorov's inequality, apply to sums of sequences of independent random variables. There are extensions of these inequalities for sums of sequences with weak dependence, such as martingales. Following the basic Markov's and Chebyshev's inequalities, we state without proof some inequalities for sums of independent but not necessarily identically distributed random variables. In Section 1.6 we consider an extension of Hoeffding's inequality to martingales (Azuma's inequality) and an extension of Kolmogorov's inequality to submartingales (Doob's submartingale inequality).

Theorem B.3.1 (Markov's inequality)
For $\epsilon > 0$, $k > 0$, and r.v. $X \ni \mathrm{E}(|X|^k)$ exists,

$$\Pr(|X| \geq \epsilon) \leq \frac{1}{\epsilon^k} \mathrm{E}\left(|X|^k\right). \tag{B.5}$$

Proof. For a nonnegative random variable Y,

$$\mathrm{E}(Y) \geq \int_{y \geq \epsilon} y \, \mathrm{d}P(y) \geq \epsilon \int_{y \geq \epsilon} \mathrm{d}P(y) = \epsilon \Pr(Y \geq \epsilon).$$

Now let $Y = |X|^k$. ∎

Corollary B.3.1.1 (Chebyshev's inequality)
For $\epsilon > 0$,

$$\Pr(|X - \mathrm{E}(X)| \geq \epsilon) \leq \frac{1}{\epsilon^2}\mathrm{V}(X) \tag{B.6}$$

Proof. In Markov's inequality, let $k = 2$, and replace X by $X - \mathrm{E}(X)$. ∎

Corollary B.3.1.2 (Chebyshev's inequality (another form))
For $f \ni f(x) \geq 0$ *and* $\epsilon > 0$,

$$\Pr(f(X) \geq \epsilon) \leq \frac{1}{\epsilon}\mathrm{E}(f(X)) \tag{B.7}$$

Proof. Same as Markov's inequality; start with $\mathrm{E}(f(X))$. ∎

Chebyshev's inequality is often useful for $\epsilon = \sqrt{\mathrm{V}(X)}$. There are also versions of Chebyshev's inequality for specific families of distributions.

- 3σ **rule for a unimodal random variable**
  If X is a random variable with a unimodal absolutely continuous distribution, and $\sigma = \sqrt{\mathrm{V}(X)}$, then

  $$\Pr(|X - \mathrm{E}(X)| \geq 3\sigma) \leq \frac{4}{81}. \tag{B.8}$$

  See Dharmadhikari and Joag-Dev (1988).
- **Normal tail probability**
  If $X \sim \mathrm{N}(\mu, \sigma^2)$, then

  $$\Pr(|X - \mu| \geq k\sigma) \leq \frac{1}{3k^2}. \tag{B.9}$$

  See DasGupta (2000).

There are a number of inequalities that generalize Chebyshev's inequality to finite partial sums of a sequence of independent random variables $X_1, X_2, \ldots$ over a common probability space such that for each, $\mathrm{E}(X_i) = 0$ and $\mathrm{E}(X_i^2) < \infty$. (The common probability space means that $\mathrm{E}(\cdot)$ has the same meaning for each i, but the X_i do not necessarily have the same distribution.) These inequalities are often called the *The Bernstein inequalities*, but some of them have other names.

Theorem B.3.2 (the Hoeffding inequality)
Let $X_1, \ldots, X_n$ *be independent,* $\mathrm{E}(X_i) = 0$, *and* $\Pr(|X_i| \leq c) = 1$. *Then for any* $t > 0$,

$$\Pr\left(\sum_{i=1}^{n} X_i > t\right) \leq \exp\left(-\frac{t^2/2}{\sum_{i=1}^{n}\mathrm{E}(X_i^2) + ct/3}\right). \tag{B.10}$$

The Hoeffding inequality is a special case of the Azuma inequality for martingales (see Section 1.6).

Theorem B.3.3 (Kolmogorov's inequality)
For a sequence of independent random variables $X_1, X_2, \ldots$ over a common probability space such that for each, $\mathrm{E}(X_i) = 0$ and $\mathrm{E}(X_i^2) < \infty$, let $S_k = \sum_{i=1}^{k} X_i$. Then for any positive integer n and any $\epsilon > 0$,

$$\Pr\left(\max_{1\le k\le n} |S_k| \ge \epsilon\right) \le \frac{1}{\epsilon^2}\mathrm{V}(S_n). \tag{B.11}$$

This is a special case of Doob's submartingale inequality (see page 133 in Section 1.6). It is also a special case of the Hájek-Rènyi inequality:

Theorem B.3.4 (the Hájek-Rènyi inequality)
Let $X_1, X_2, \ldots$ be a sequence of independent random variables over a common probability space such that for each $\mathrm{E}(X_i^2) < \infty$. Then for any positive integer n and any $\epsilon > 0$,

$$\Pr\left(\max_{1\le k\le n} c_k \left|\sum_{i=1}^{k}(X_i - \mathrm{E}(X_i))\right| \ge \epsilon\right) \le \frac{1}{\epsilon^2}\sum_{i=1}^{n} c_i^2\mathrm{V}(X_i), \tag{B.12}$$

where $c_1 \ge \cdots \ge c_n > 0$ are constants.

B.4 E($f(X)$) and f(E(X))

Theorem B.4.1 (Jensen's inequality)
For f a convex function over the support of the r.v. X (and all expectations shown exist),

$$f(\mathrm{E}(X)) \le \mathrm{E}(f(X)). \tag{B.13}$$

Proof. By the definition of convexity, f convex over $D \Rightarrow \exists\, c \ni \forall\, t \in D \ni c(x-t) + f(t) \le f(x)$. (Notice that $L(x) = c(x-t) + f(t)$ is a straight line through the point $(t, f(t))$. By the definition of convexity, f is convex if its value at the weighted average of two points does not exceed the weighted average of the function at those two points.) Now, given this, let $t = \mathrm{E}(X)$ and take expectations of both sides of the inequality. ∎

If f is strictly convex, it is clear

$$f(\mathrm{E}(X)) < \mathrm{E}(f(X)) \tag{B.14}$$

unless $f(X) = \mathrm{E}(f(X))$ with probability 1.

For a concave function, the inequality is reversed. (The negative of a concave function is convex.)

Some simple examples for a nonconstant positive random variable X:

- Monomials of even power: for $k = 2, 4, 6, \ldots$,

$$\mathrm{E}(X)^k \le \mathrm{E}(X^k).$$

This inequality implies the familiar fact that $\mathrm{E}(X) \ge 0$.

- Reciprocals:
$$\frac{1}{\mathrm{E}(X)} \leq \mathrm{E}\left(\frac{1}{X}\right)$$
- Logs:
$$\mathrm{E}(\log(X)) \leq \log(\mathrm{E}(X)).$$

The canonical picture is that of a quadratic function of a uniform random variable:

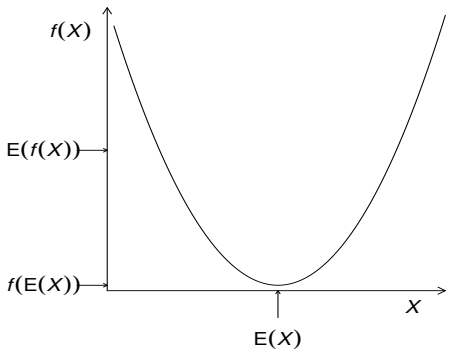


Figure B.1. Jensen's Inequality

There are several other consequences of Jensen's inequality. The first one we state applies to Kullback-Leibler information; that is, the entropy distance.

Corollary B.4.1.1 (Nonnegativity of the entropy distance)
If f and g are probability densities, $\mathrm{E}_f(\log(f(X)/g(X)))$, is the entropy distance between f and g with respect to g. It is also called the Kullback-Leibler information or Kullback-Leibler distance. It is nonnegative:

$$\mathrm{E}_f(\log(f(X)/g(X))) \geq 0. \tag{B.15}$$

Proof.

$$\begin{aligned}\mathrm{E}_f(\log(f(X)/g(X))) &= -\mathrm{E}_f(\log(g(X)/f(X))) \\ &\geq -\log(\mathrm{E}_f(g(X)/f(X))) \\ &= 0.\end{aligned}$$

∎

A related fact applies to any nonnegative integrable functions f and g on a measure space with a σ-finite measure ν, for which $\int f \mathrm{d}\nu \geq \int g \mathrm{d}\nu > 0$:

$$\int f(\log(f/g))\mathrm{d}\nu \geq 0. \tag{B.16}$$

This can be proved as above by normalizing the functions, thus forming densities.

Applying Jensen's inequality to the definition of the entropy distance, we have

Corollary B.4.1.2
For the PDFs f and g.

$$\mathrm{E}_f(\log(f(X))) \geq \mathrm{E}_f(\log(g(X))). \tag{B.17}$$

This inequality, which is important for showing the convergence of the EM algorithm, is also sometimes called the "information inequality" (but see (B.25)).

The strict form of Jensen's inequality (B.14) also applies to the consequent inequalities, and hence, in the case of equality of the expectations, we can get equality of the functions. For example,

Corollary B.4.1.3
For the PDFs f and g,

$$\mathrm{E}_f(\log(f(X))) = \mathrm{E}_f(\log(g(X))) \Leftrightarrow f(X) = g(X) \text{ a.s.} \tag{B.18}$$

Proof.:
$\Rightarrow$
By the equality of $\mathrm{E}_f(\log(f(X)))$ and $\mathrm{E}_f(\log(g(X)))$ we have

$$\int_{\{f>0\}} g(x)\mathrm{d}x = 1,$$

and so for any A,

$$\begin{aligned}\int_A g(x)\mathrm{d}x &= \int_{A\cap\{f>0\}} g(x)\mathrm{d}x \\ &= \mathrm{E}_f\left(g(X)/f(X)|X \in A \cap \{f > 0\}\right) \\ &= \Pr\left(X \in A \cap \{f > 0\}\right) \\ &= \int_A f(x)\mathrm{d}x,\end{aligned}$$

hence $f(X) = g(X)$ a.s.

The proof of $\Leftarrow$ is similar. ∎

B.5 E($f(X,Y)$) and E($g(X)$) and E($h(Y)$)

In many of the inequalities in this section, the functions f, g, and h are norms. The inequalities hold for general L_p norms, and although we will consider the inequality relationship between expected values, similar inequalities often for real numbers, vectors, or random variables.

The inequalities are basically of two types:

- Hölder: $\mathrm{E}(|XY|) \leq \left(\mathrm{E}(|X|^p)\right)^{1/p}\left(\mathrm{E}(|Y|^q)\right)^{1/q}$
- Minkowski: $(\mathrm{E}(|X+Y|^p))^{1/p} \leq (\mathrm{E}(|X|^p))^{1/p} + (\mathrm{E}(|Y|^p))^{1/p}$

Hölder inequality is somewhat more basic; it is used in the proof of Minkowski's inequality. Compare inequalities (0.0.30) and (0.0.31) for vectors in $\mathbb{R}^d$, and see the discussion on page 642.

Note that Minkowski's inequality has an interesting consequence: it means that $(\mathrm{E}(|\cdot|^p))^{1/p}$ is a norm.

Several other inequalities are special cases of these two.

In some inequalities in this section, the functions are second-degree monomials. The basic special inequality of this form is the Cauchy-Schwartz inequality, which then leads to one of the most important inequalities in applications in statistics, the covariance inequality. The covariance inequality, in turn, leads to fundamental bounds on the variances of estimators.

Theorem B.5.1 (Hölder's inequality)
For $p, q > 1$ and $\frac{1}{p} + \frac{1}{q} = 1$ (and if all expectations shown exist),

$$\mathrm{E}(|XY|) \leq \left(\mathrm{E}(|X|^p)\right)^{1/p}\left(\mathrm{E}(|Y|^q)\right)^{1/q}. \tag{B.19}$$

Note that $q = p/(p-1)$; p and q as in this inequality are called *dual* indices.

Proof. If $\mathrm{E}(|X|^p) = 0$ or $\mathrm{E}(|Y|^q) = 0$, then true because both sides $= 0$ wp1. Hence, assume both > 0.

For p and q as in hypothesis, $\forall a, b > 0$, $\exists\, s, t \ni a = \mathrm{e}^{s/p}$ and $b = \mathrm{e}^{t/q}$. Now e^x is convex, so $\mathrm{e}^{s/p+t/q} \leq \frac{1}{p}\mathrm{e}^s + \frac{1}{q}\mathrm{e}^t$, or $ab \leq a^p/p + b^q/q$.

Let

$$a = \left|\frac{X(\omega)}{\left(\mathrm{E}(|X|^p)\right)^{1/p}}\right| \quad \text{and} \quad b = \left|\frac{Y(\omega)}{\left(\mathrm{E}(|Y|^q)\right)^{1/q}}\right|$$

and so

$$|X(\omega)Y(\omega)| \leq \left(\mathrm{E}(|X|^p)\right)^{1/p}\left(\mathrm{E}(|Y|^q)\right)^{1/q}\left(\frac{|X(\omega)|^p}{\mathrm{E}(|X|^p)}\frac{1}{p} + \frac{|Y(\omega)|^q}{\mathrm{E}(|Y|^q)}\frac{1}{q}\right).$$

Now take expectations. (The notation $X(\omega)$ and $Y(\omega)$ is meant to emphasize how to take expectation of XY.) ∎

There are several inequalities that derive from Hölder's inequality. Some of these inequalities are given in the following corollaries to Theorem B.5.1.First is a special case of Jensen's inequality.

Corollary B.5.1.1 (special case of Jensen's inequality)

$$\mathrm{E}(|X|) \leq \Big(\mathrm{E}(|X|^p)\Big)^{1/p},$$

Proof. Set $Y \equiv 1$ in Hölder's inequality. ∎

Note that with $p = 2$, Corollary B.5.1.1 is a special case of the Cauchy-Schwarz inequality,

$$\mathrm{E}(|X|) \leq \Big(\mathrm{E}(X^2)\Big)^{1/2},$$

in Corollary B.5.1.3 below.

Corollary B.5.1.2 (Lyapunov's inequality)
For $1 \leq r \leq s$ (and if all expectations shown exist),

$$(\mathrm{E}(|X|^r))^{1/r} \leq (\mathrm{E}(|X|^s))^{1/s}. \tag{B.20}$$

Proof. First, we observe this is true for $r = s$, and for $r = 1$ (in which it is a form of Jensen's inequality). If $1 < r < s$, replace $|X|$ in the special case of Hölder's inequality above with $|X|^r$, and let $s = pr$ for $1 < p$. This yields $(\mathrm{E}(|X|^r))^{1/r} \leq (\mathrm{E}(|X|^s))^{1/s}$. ∎

Corollary B.5.1.3 (Schwarz inequality, or Cauchy-Schwarz inequality)

$$\mathrm{E}(|XY|) \leq \Big(\mathrm{E}(X^2)\mathrm{E}(Y^2)\Big)^{1/2}. \tag{B.21}$$

Proof. Let $p = q = 2$ in Hölder's inequality. ∎
Another proof: For nonnegative r.v. X and Y and all t (real), $\mathrm{E}((tX+Y)^2) = t^2\mathrm{E}(X^2) + 2t\mathrm{E}(XY) + \mathrm{E}(Y^2) \geq 0$. Hence the discriminant of the quadratic formula ≤ 0. Now, for any r.v., take absolute value. ∎

Corollary B.5.1.4 (covariance inequality) *(see page 36)*
If the second moments of X and Y are finite, then

$$\Big(\mathrm{E}\big((X - \mathrm{E}(X))(Y - \mathrm{E}(Y))\big)\Big)^2 \leq \mathrm{E}\big((X - \mathrm{E}(X))^2\big)\,\mathrm{E}\big((Y - \mathrm{E}(Y))^2\big) \tag{B.22}$$

or

$$\big(\mathrm{Cov}(X,Y)\big)^2 \leq \mathrm{V}(X)\,\mathrm{V}(Y). \tag{B.23}$$

Proof. The covariance inequality is essentially the same as the Cauchy-Schwarz inequality. ∎

The covariance inequality leads to useful lower bounds on the variances of estimators. These are of two types. One type includes the Hammersley-Chapman-Robbins inequality and its extension, the Kshirsagar inequality. The other type, which is based on Fisher information, requires some "regularity conditions".

Corollary B.5.1.5 (Hammersley-Chapman-Robbins inequality)
Let X be a random variable in $\mathbb{R}^d$ with PDF $p(x;\theta)$ and let $\mathrm{E}_\theta(T(X) = g(\theta)$. Let μ be a fixed measure on $\mathcal{X} \subseteq \mathbb{R}^d$ such that $p(x;\theta) \ll \mu$. Now define $S(\theta)$ such that

$$p(x;\theta) > 0 \text{ a.e. } x \in S(\theta)$$
$$p(x;\theta) = 0 \text{ a.e. } x \notin S(\theta).$$

Then

$$\mathrm{V}(T(X)) \geq \sup_{t \ni S(\theta) \supseteq S(\theta+t)} \frac{(g(\theta+t) - g(\theta))^2}{\mathrm{E}_\theta\left(\left(\frac{p(X;\theta+t)}{p(X;\theta)}\right)^2\right)}. \tag{B.24}$$

Proof. This inequality follows from the covariance inequality, by first considering the case for an arbitrary t such that $g(\theta+t) \neq g(\theta)$. ∎

Corollary B.5.1.6 (Kshirsagar inequality)

Corollary B.5.1.7 (information inequality)
Subject to some "regularity conditions" (see Section 2.3), if X has PDF $p(x;\theta)$, then

$$\mathrm{V}(T(X)) \geq \frac{\left(\frac{\partial \mathrm{E}(T(X))}{\partial \theta}\right)^2}{\mathrm{E}_\theta\left(\left(\frac{\partial \log p(X;\theta)}{\partial \theta}\right)^2\right)} \tag{B.25}$$

The denominator of the quantity on the right side of the inequality is called the Fisher information, or just the information. Notice the similarity of this inequality to the Hammersley-Chapman-Robbins inequality, although the information inequality requires more conditions.

Under the regularity conditions, which basically allow the interchange of integration and differentiation, the information inequality follows immediately from the covariance inequality.

We consider the multivariate form of this inequality to Section 3.1.3. Our main interest will be in its application in unbiased estimation, in Section 5.1. If $T(X)$ is an unbiased estimator of a differentiable function $g(\theta)$, the right side of the inequality together with derivatives of $g(\theta)$ forms the Cramér-Rao lower bound, inequality (3.39), and the Bhattacharyya lower bound, inequality (5.29).

Theorem B.5.2 (Minkowski's inequality)
For $1 \leq p$,

$$(\mathrm{E}(|X+Y|^p))^{1/p} \leq (\mathrm{E}(|X|^p))^{1/p} + (\mathrm{E}(|Y|^p))^{1/p} \tag{B.26}$$

This is a triangle inequality for L_p norms and related functions.

Proof. First, observe the truth for $p = 1$ using the triangle inequality for the absolute value, $|x+y| \leq |x| + |y|$, giving $\mathrm{E}(|X+Y|) \leq \mathrm{E}(|X|) + \mathrm{E}(|Y|)$.

Now assume $p > 1$. Now,

$$\begin{aligned}\mathrm{E}(|X+Y|^p) &= \mathrm{E}(|X+Y||X+Y|^{p-1}) \\ &\leq \mathrm{E}(|X||X+Y|^{p-1}) + \mathrm{E}(|Y||X+Y|^{p-1}),\end{aligned}$$

where the inequality comes from the triangle inequality for absolute values. From Hölder's inequality on the terms above with $q = p/(p-1)$, we have

$$\mathrm{E}(|X+Y|^p) \leq (\mathrm{E}(|X|^p))^{1/p}(\mathrm{E}(|X+Y|^p))^{1/q} + (\mathrm{E}(|Y|^p))^{1/p}(\mathrm{E}(|X+Y|^p))^{1/q}.$$

Now, if $\mathrm{E}(|X+Y|^p) = 0$, Minkowski's inequality holds. On the other hand, if $\mathrm{E}(|X+Y|^p) \neq 0$, it is positive, and so divide through by $(\mathrm{E}(|X+Y|^p))^{1/q}$, recalling again that $q = p/(p-1)$. ∎

Minkowski's inequality is a special case of two slightly tighter inequalities; one for $p \in [1, 2]$ due to Esseen and von Bahr (1965), and one for $p \geq 2$ due to Marcinkiewicz and Zygmund (1937).

An inequality that derives from Minkowski's inequality, but which applies directly to real numbers or random variables, is the following.

- For $0 \leq p$,
$$|X+Y|^p \leq 2^p(|X|^p + |Y|^p) \tag{B.27}$$
This is true because $\forall \omega \in \Omega$, $\|X(\omega) + Y(\omega)\| \leq 2\max\{\|X(\omega)\|, \|Y(\omega)\|\}$, and so
$$\begin{aligned}\|X(\omega) + Y(\omega)\|^p &\leq \max\{2^p\|X(\omega)\|^p, 2^p\|Y(\omega)\|^p\} \\ &\leq 2^p\|X(\omega)\|^p + 2^p\|Y(\omega)\|^p.\end{aligned}$$

B.6 $\mathrm{V}(Y)$ and $\mathrm{V}(\mathrm{E}(Y|X))$

- **Rao-Blackwell inequality**
$$\mathrm{V}\big(\mathrm{E}(Y|X)\big) \leq \mathrm{V}(Y) \tag{B.28}$$
This follows from the equality $\mathrm{V}(Y) = \mathrm{V}\big(\mathrm{E}(Y|X)\big) + \mathrm{E}\big(\mathrm{V}(Y|X)\big)$.

B.7 Multivariate Extensions

There are multivariate extensions of most of these inequalities. In some cases, the multivariate extensions apply to the minimum or maximum element of a vector.

Some inequalities involving simple inequalities are extended by conditions on vector norms, and the ones involving variances are usually extended by positive (or semi-positive) definiteness of the difference of two variance-covariance matrices.

Notes and Further Reading

Chebyshev's inequality, Corollary B.3.1.1, in its various forms is one of the most useful inequalities in probability theory. DasGupta (2000) discusses various forms of this basic inequality.

Another useful general-purpose relationship is the Cauchy-Schwarz inequality, Corollary B.5.1.3. Steele (2004) discusses origins, various forms, and various proofs of this inequality, and in so doing illustrates interesting relationships among diverse mathematical inequalities.

Inequalities are very useful in developing asymptotic results and in proving limit theorems. DasGupta (2008) on asymptotic theory contains a very extensive compendium of inequalities on pages 633 to 687. None are proved there, but each is accompanied by a reference to a proof. Petrov (1995) begins with a survey of inequalities in probability theory, including proofs, and then gives a number of limit theorem theorems the proofs of which often rely on the inequalities.

C

Notation and Definitions

All notation used in this work is "standard". I have opted for simple notation, which, of course, results in a one-to-many map of notation to object classes. Within a given context, however, the overloaded notation is generally unambiguous. I have endeavored to use notation consistently.

This appendix is not intended to be a comprehensive listing of definitions.

C.1 General Notation

There are some standard phrases widely used in mathematical statistics, and so for these we adopt special symbols. These phrases are sometime omitted because the property or condition is implicitly assumed. I try to be explicit about my assumptions. If we agree on simple character strings to represent these properties and conditions, I am more likely to state them explicitly when they are relevant.

wrt	"with respect to".
wlog	"without loss of generality".
iid or $\overset{\text{iid}}{\sim}$	"independent and identically distributed (as)".

Uppercase italic Latin and Greek letters, such as A, B, E, Λ, etc., are generally used to represent sets, random variables, and matrices. Realizations of random variables and placeholders in functions associated with random variables are usually represented by lowercase letters corresponding to the uppercase letters; thus, ϵ may represent a realization of the random variable E.

Parameters in models (that is, unobservables in the models) are generally represented by Greek letters. Uppercase Latin and Greek letters are also used

to represent cumulative distribution functions. Symbols whose meaning is context-independent are usually written in an upright font, whereas symbols representing variables are written in a slant or italic font; for example, Γ is used to represent the gamma function, while $\mathit{\Gamma}$ may be used to represent a variable or a parameter. An upright font is also used to represent a special object, such as a sample space or a parameter space.

The Greek Alphabet

alpha	A	α	nu	N	ν
beta	B	β	xi	Ξ	ξ
gamma	$\mathit{\Gamma}$	γ	omicron	O	o
delta	$\mathit{\Delta}$	δ	pi	$\mathit{\Pi}$	π, ϖ
epsilon	E	ϵ, ε	rho	P	ρ, ϱ
zeta	Z	ζ	sigma	Σ	σ, ς
eta	H	η	tau	T	τ
theta	Θ	θ, ϑ	upsilon	Υ	υ
iota	I	ι	phi	Φ	ϕ, φ
kappa	K	κ, $\varkappa$	chi	X	χ
lambda	Λ	λ	psi	Ψ	ψ
mu	M	μ	omega	Ω	ω

Symbols for Structures or Elements within a Structure

Lowercase Latin and Greek letters are used to represent ordinary scalar or vector variables and functions. **No distinction in the notation is made between scalars and vectors**; thus, β may represent a vector and β_i may represent the i^{th} element of the vector β. In another context, however, β may represent a scalar. All vectors are considered to be column vectors, although we may write a vector as $x = (x_1, x_2, \ldots, x_n)$. Transposition of a vector or a matrix is denoted by the superscript "T".

Uppercase calligraphic Latin letters, such as $\mathcal{D}$, $\mathcal{V}$, and $\mathcal{W}$, are generally used to represent special collections of sets, vector spaces, or transforms (functionals).

A single symbol in an italic font is used to represent a single variable. A Roman font or a special font is often used to represent a standard operator or a standard mathematical structure. Sometimes a string of symbols in a Roman font is used to represent an operator (or a standard function); for example, $\exp(\cdot)$ represents the exponential function. But a string of symbols in an italic font on the same baseline should be interpreted as representing a composition (probably by multiplication) of separate objects; for example, *exp* represents the product of e, x, and p. Likewise a string of symbols in a Roman font (usually a single symbol) is used to represent a fundamental

constant; for example, e represents the base of the natural logarithm, while e represents a variable.

Subscripts generally represent indexes to a larger structure; for example, x_{ij} may represent the $(i,j)^{\text{th}}$ element of a matrix, X. A subscript in parentheses represents an order statistic. A superscript in parentheses represents an iteration; for example, $x_i^{(k)}$ may represent the value of x_i at the k^{th} step of an iterative process.

x_i	The i^{th} element of a structure (including a sample, which, if the labels are ignored, is a multiset).
$x_{(i)}$ or $x_{(i:n)}$	The i^{th} order statistic in a sample of size n.
$x^{(i)}$	The value of x at the i^{th} iteration.

Symbols for Fixed Mathematical Structures

Some important mathematical structures and other objects are:

$\mathbb{R}$	The field of reals or the set over which that field is defined.
$\mathbb{R}_+$	The set of positive reals.
$\overline{\mathbb{R}}$	The "extended reals"; $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$.
$\bar{\mathbb{R}}_+$	The nonnegative reals; $\bar{\mathbb{R}}_+ = \mathbb{R}_+ \cup \{0\}$.
$\overline{\mathbb{R}}_+$	The extended nonnegative reals; $\overline{\mathbb{R}}_+ = \mathbb{R}_+ \cup \{0, \infty\}$.
$\mathbb{R}^d$	The usual d-dimensional vector space over the reals or the set of all d-tuples with elements in $\mathbb{R}$.
$\mathbb{R}^{n\times m}$	The vector space of real $n \times m$ matrices.
$\mathbb{Z}$	The ring of integers or the set over which that ring is defined.
$\mathbb{Z}_+$	The set of positive integers.
$\mathbb{C}$	The field of complex numbers or the set over which that field is defined. The notation $\mathbb{C}^d$ and $\mathbb{C}^{n\times m}$ have meanings analogous to the corresponding real spaces.

e	The base of the natural logarithm. This is a constant; the symbol "e" may be used to represent a variable. (Note the difference in the font.)
i	The imaginary unit, $\sqrt{-1}$. This is a constant; the symbol "i" may be used to represent a variable. (Note the difference in the font.)

C.2 General Mathematical Functions and Operators

Functions such as sin, max, span, and so on that are commonly associated with strings of Latin letters are generally represented by those letters in a Roman font.

Operators such as d (the differential operator) that are commonly associated with a Latin letter are generally represented by that letter in a Roman font.

Note that some symbols, such as $|\cdot|$, are overloaded.

$\lvert x\rvert$	The modulus of the real or complex number x; if x is real, $\lvert x\rvert$ is the absolute value of x.
$\lceil x\rceil$	The ceiling function evaluated at the real number x: $\lceil x\rceil$ is the smallest integer greater than or equal to x. For any x, $\lfloor x\rfloor \leq x \leq \lceil x\rceil$.
$\lfloor x\rfloor$	The floor function evaluated at the real number x: $\lfloor x\rfloor$ is the largest integer less than or equal to x.
$x!$	The factorial of x. If $x = 0$ $$x! = 0! = 1;$$ if x is a positive integer, $$x! = x(x-1)\cdots 2\cdot 1;$$ otherwise, for all x except nonpositive integers, $$x! = \Gamma(x+1).$$ If $x = -1, -2, \ldots$, $x!$ is undefined.
0^0	For convenience, we define 0^0 as $\lim_{x\to 0} x^0 = 1$.

$x^{[r]}$ — The r^{th} factorial of x. If x is a positive integer,

$$x^{[r]} = x(x-1)\cdots(x-(r-1)).$$

C_k^n or $\binom{n}{k}$ — The binomial coefficient, $n!/(k!(n-k)!)$. If n is a positive integer, and k is a nonnegative integer no greater than n, then this is the number of ways k items can be chosen from a set of n items.

$\Pi(A)$ or $\Pi(n)$ — For the set A with finite cardinality n, $\Pi(A)$ is an n-tuple consisting of the elements of A, each occurring once. For the positive integer n, $\Pi(n)$ an n-tuple consisting of the elements $1,\ldots,n$. There are $n!$ possible values of either $\Pi(A)$ or $\Pi(n)$.

d — The differential operator.

Δ — A perturbation operator; Δx represents a perturbation of x and not a multiplication of x by Δ, even if x is a type of object for which a multiplication is defined.

$\Delta(\cdot,\cdot)$ — A real-valued difference function; $\Delta(x,y)$ is a measure of the difference of x and y. For simple objects, $\Delta(x,y) = |x-y|$. For more complicated objects, a subtraction operator may not be defined, and Δ is a generalized difference.

$\tilde{x}$ — A perturbation of the object x; $\Delta(x,\tilde{x}) = \Delta x$.

$\tilde{x}$ — An average of a sample of objects generically denoted by x.

$\bar{x}$ — The mean of a sample of objects generically denoted by x.

$\mathrm{O}(f(n))$ The order class big O with respect to $f(n)$.

$$g(n) \in \mathrm{O}(f(n))$$

means there exists some fixed c such that $\|g(n)\| \leq c\|f(n)\|$ $\forall n$. In particular, $g(n) \in \mathrm{O}(1)$ means $g(n)$ is bounded.
In one special case, we will use $\mathrm{O}(f(n))$ to represent some unspecified scalar or vector $x \in \mathrm{O}(f(n))$. This is the case of a convergent series. An example is

$$s = f_1(n) + \cdots + f_k(n) + \mathrm{O}(f(n)),$$

where $f_1(n), \ldots, f_k(n)$ are finite constants.
We may also express the order class defined by convergence as $x \to a$ as $\mathrm{O}(f(x))_{x\to a}$ (where a may be infinite). Hence, $g \in \mathrm{O}(f(x))_{x\to a}$ iff

$$\limsup_{x\to a} \|g(n)\|/\|f(n)\| < \infty.$$

$\mathrm{o}(f(n))$ Little o; $g(n) \in \mathrm{o}(f(n))$ means for all $c > 0$ there exists some fixed N such that $0 \leq g(n) < cf(n)$ $\forall n \geq N$. (The functions f and g and the constant c could all also be negative, with a reversal of the inequalities.) Hence, $g(n) = \mathrm{o}(f(n))$ means $\|g(n)\|/\|f(n)\| \to 0$ as $n \to \infty$.
In particular, $g(n) \in \mathrm{o}(1)$ means $g(n) \to 0$.
We also use $\mathrm{o}(f(n))$ to represent some unspecified scalar or vector $x \in \mathrm{o}(f(n))$ in special case of a convergent series, as above:

$$s = f_1(n) + \cdots + f_k(n) + \mathrm{o}(f(n)).$$

We may also express this kind of convergence in the form $g \in \mathrm{o}(f(x))_{x\to a}$ as $x \to a$ (where a may be infinite).

Spaces of Functions

$\mathcal{C}^k$ For an integer $k \geq 0$, the class of functions whose derivatives up to the k^{th} derivative exist and are continuous.

$\mathcal{L}^p$ For a real number $p \geq 1$, the class of functions f on a measure space $(\Omega, \mathcal{F}, \nu)$ with a metric $\|\cdot\|$ such that $\int \|f\|^p \mathrm{d}\nu < \infty$.

Functions of Convenience

$\mathrm{I}_S(\cdot)$

The indicator function:

$$\begin{aligned} \mathrm{I}_S(x) &= 1, \text{ if } x \in S, \\ &= 0, \text{ otherwise.} \end{aligned} \tag{C.1}$$

If x is a scalar, the set S is often taken as the interval $]-\infty, y[$, and in this case, the indicator function is the Heaviside function, H, evaluated at the difference of the argument and the upper bound on the interval:

$$\mathrm{I}_{]-\infty,y[}(x) = \mathrm{H}(y-x).$$

(An alternative definition of the Heaviside function is the same as this except that $\mathrm{H}(0) = \frac{1}{2}$.) It is interesting to note that

$$\mathrm{I}_{]-\infty,y[}(x) = \mathrm{I}_{]x,\infty[}(y).$$

In higher dimensions, the set S is often taken as the product set,

$$\begin{aligned} A^d &=]-\infty, y_1[\times]-\infty, y_2[\times\cdots\times]-\infty, y_d[\\ &= A_1 \times A_2 \times \cdots \times A_d, \end{aligned}$$

and in this case,

$$\mathrm{I}_{A^d}(x) = \mathrm{I}_{A_1}(x_1)\mathrm{I}_{A_2}(x_2)\cdots\mathrm{I}_{A_d}(x_d),$$

where $x = (x_1, x_2, \ldots, x_d)$.
The derivative of the indicator function is the Dirac delta function, $\delta(\cdot)$.

$\delta(\cdot)$ The Dirac delta "function", defined by

$$\delta(x) = 0, \quad \text{for } x \neq 0,$$

and

$$\int_{-\infty}^{\infty} \delta(t)\,\mathrm{d}t = 1.$$

The Dirac delta function is not a function in the usual sense. We do, however, refer to it as a function, and treat it in many ways as a function. For any continuous function f, we have the useful fact

$$\begin{aligned}\int_{-\infty}^{\infty} f(y)\,\mathrm{dI}_{]-\infty,y[}(x) &= \int_{-\infty}^{\infty} f(y)\,\delta(y-x)\,\mathrm{d}y \\ &= f(x).\end{aligned}$$

Special Functions

Various common mathematical functions are referred to collectively as "special functions". These include the trigonometric functions, both circular and hyperbolic, the various orthogonal polynomial systems, and solutions to special differential equations, such as Bessel functions.

I list only a few below. The functions are often written without parentheses enclosing the arguments, for example $\log x$, but I usually enclose the arguments in parentheses.

Good general references on special functions in mathematics are Olver et al. (2010) and Thompson (1997).

$\log(x)$ The natural logarithm evaluated at x.

$\sin(x)$ The sine evaluated at x (in radians) and similarly for other trigonometric functions.

$\Gamma(\alpha)$ The complete gamma function:

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}\mathrm{e}^{-t}\mathrm{d}t. \tag{C.2}$$

(This is called Euler's integral.) Integration by parts immediately gives the replication formula

$$\Gamma(\alpha+1) = \alpha\Gamma(\alpha),$$

and so if α is a positive integer, $\Gamma(\alpha+1) = \alpha!$. More generally, $\Gamma(\alpha+1)$ can be taken as the definition of $\alpha!$. This does not exist for negative integers, but does for all other real α.
Direct evaluation of the integral yields $\Gamma(1/2) = \sqrt{\pi}$. Using this and the replication formula, with some manipulation we get for the positive integer j

$$\Gamma(j+1/2) = \frac{1\cdot 2\cdots(2j-1)}{2^j}\sqrt{\pi}.$$

The notation $\Gamma_d(\alpha)$ denotes the multivariate gamma function, where α is a d-vector. (In other literature this notation denotes the incomplete univariate gamma function, for which I use $\gamma(\alpha, d)$; see below.)

Associated with the gamma function are some other useful functions:

$\psi(\alpha)$ The psi function or the digamma function:

$$\psi(\alpha) = \mathrm{d}\log(\Gamma(\alpha))/\mathrm{d}\alpha. \tag{C.3}$$

$\psi'(\alpha)$ The trigamma function,

$$\psi'(\alpha) = \mathrm{d}\psi(\alpha)/\mathrm{d}\alpha. \tag{C.4}$$

More general are the polygamma functions, for $n = 1, 2, \ldots$, $\psi^{(n)}(\alpha) = \mathrm{d}^{(n)}\psi(\alpha)/(\mathrm{d}\alpha)^{(n)}$, which for a fixed n, is called the $(n+2)$-gamma function.

$\gamma(\alpha, x)$ The incomplete gamma function,

$$\gamma(\alpha, x) = \int_0^x t^{\alpha-1}\mathrm{e}^{-t}\mathrm{d}t. \tag{C.5}$$

This is also often denoted as $\Gamma_x(\alpha)$.

$P(\alpha, x)$ The regularized incomplete gamma function, which is the CDF of the standard gamma distribution,

$$P(\alpha, x) = \frac{\gamma(\alpha, x)}{\Gamma(\alpha)}. \tag{C.6}$$

$\mathrm{B}(\alpha, \beta)$ The beta function,

$$\begin{aligned} \mathrm{B}(\alpha, \beta) &= \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} && \text{(C.7)} \\ &= \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\mathrm{d}t && \text{(C.8)} \\ &= \int_0^\infty \frac{t^{\alpha-1}}{(1+t)^{\alpha+\beta}}. && \text{(C.9)} \end{aligned}$$

The integral in equation (C.8) is called Euler's beta integral.

$I_x(\alpha, \beta)$ The regularized incomplete beta function, which is the CDF of the beta distribution,

$$I_x(\alpha, \beta) = \frac{1}{\mathrm{B}(\alpha, \beta)} \int_0^x t^{\alpha-1}(1-t)^{\beta-1}\mathrm{d}t. \tag{C.10}$$

C.3 Sets, Measure, and Probability

The notation listed below does not always represent the things associated with it here, but for these objects, I generally use either this notation or other symbols in the same font.

Ω Sample space; the universal set in a given space or probability distribution.

$\#A$ The cardinality of the set A.

A^{c} The complement of the set A; $A^{\mathrm{c}} = \Omega - A$.

$A_1 \cup A_2$ The union of the sets A_1 and A_2; $x \in A_1 \cup A_2$ iff $x \in A_1$ or $x \in A_2$.

$A_1 \cap A_2$	The intersection of the sets A_1 and A_2; $x \in A_1 \cap A_2$ iff $x \in A_1$ and $x \in A_2$.
$A_1 - A_2$	The set A_1 minus the set A_2; $x \in A_1 - A_2$ iff $x \in A_1$ and $x \notin A_2$.
$A_1 \Delta A_2$	The symmetric difference of the sets A_1 and A_2; $A_1 \Delta A_2 = (A_1 - A_2) \cup (A_2 - A_1)$.
$A_1 \times A_2$	Cartesian (or cross) product of the sets A_1 and A_2; $(a_1, a_2) \in A_1 \times A_2$ iff $a_1 \in A_1$ and $a_2 \in A_2$.

The following objects require a notion of "open sets", either as the collection of sets that define a topology (Section 0.0.2) or as defined in a metric space (Sections 0.0.2 and 0.0.5).

A°	The set of interior points of the set A.
$\overline{A}$	The set of closure points of the set A.
∂A	The set of boundary points of the set A: $\partial A = \overline{A} - A^\circ$.
$\mathcal{F}$	A σ-field.
$\mathcal{B}(\Omega)$	The Borel σ-field generated by a collection of open sets defining a topology in Ω. This requires definition of a collection, so we may also use the notation $\mathcal{B}(\Omega, \mathcal{T})$, where $\mathcal{T}$ is a collection of sets defining a topology.
$\mathcal{B}$	The Borel σ-field $\mathcal{B}(\mathbb{R})$.
$\mathcal{B}^d$	The Borel σ-field $\mathcal{B}(\mathbb{R}^d)$.
$\mathcal{B}_I$	The Borel σ-field restricted to the interval I; that is, the σ-field generated by all open intervals contained in I and $\Omega = I$.
$(\Omega, \mathcal{F})$	A measurable space: the sample space Ω and the σ-field $\mathcal{F}$.
$(\Omega, \mathcal{F}, \nu)$	A measure space: the sample space Ω, the σ-field $\mathcal{F}$, and the measure ν defined over the sets in $\mathcal{F}$.

$\lambda \ll \nu$	The measure ν dominates the measure λ; that is, λ is *absolutely continuous with respect to* ν*:* $$\nu(A) = 0 \quad \Rightarrow \quad \lambda(A) = 0,$$ *for any set* A *in the domain of both* λ *and* ν*.*
$\lambda \perp \nu$	The measures ν and λ on a common measurable space $(\Omega, \mathcal{F})$ are singular with respect to each other; that is, there exists two disjoint sets A and B in $\mathcal{F}$ such that $A \cup B = \Omega$ and for any measurable set $A_1 \subseteq A$, $\nu(A_1) = 0$, while for any measurable set $B_1 \subseteq B$, $\mu(B_1) = 0$.
$(\Omega, \mathcal{F}, P)$	The "probability triple": the sample space Ω, the σ-field $\mathcal{F}$, and the probability measure P.
$\mathcal{P}$	A family of probability distributions.
Θ	Parameter space.
$\mathcal{X}$	The range of a random variable.
$\mathrm{O}_P(f(n))$	Bounded convergence in probability; $X(n) \in \mathrm{O}_P(f(n))$ means that for any positive ϵ, there is a constant C_ϵ such that $\sup_n \Pr(\|X(n)\| \geq C_\epsilon \|f(n)\|) < \epsilon$.
$\mathrm{o}_P(f(n))$	Convergent in probability; $X(n) \in \mathrm{o}_P(f(n))$ means that for any positive ϵ, $\Pr(\|X(n) - f(n)\| > \epsilon) \to 0$ as $n \to \infty$.

C.4 Linear Spaces and Matrices

$\mathcal{V}(G)$	For the set of vectors (all of the same order) G, the vector space generated by that set.
$\mathcal{V}(X)$	For the matrix X, the vector space generated by the columns of X.
$\dim(\mathcal{V})$	The dimension of the vector space $\mathcal{V}$; that is, the maximum number of linearly independent vectors in the vector space.
$\mathrm{span}(Y)$	For Y either a set of vectors or a matrix, the vector. space $\mathcal{V}(Y)$

$\mathrm{tr}(A)$	The trace of the square matrix A, that is, the sum of the diagonal elements.
$\mathrm{rank}(A)$	The rank of the matrix A, that is, the maximum number of independent rows (or columns) of A.
$\rho(A)$	The spectral radius of the matrix A (the maximum absolute value of its eigenvalues).
$A > 0$ $A \geq 0$	If A is a matrix, this notation means, respectively, that each element of A is positive or nonnegative. These may also be written as $0 < A$ or $0 \leq A$.
$A \succ 0$ $A \succeq 0$	This notation means that A is a symmetric matrix and that it is, respectively, positive definite or nonnegative definite. These may also be written as $0 \prec A$ or $0 \preceq A$.
$A \succ B$ $A \succeq B$	This notation means that A and B are symmetric matrices and that $A - B$ is, respectively, positive definite or nonnegative definite. These may also be written as $B \prec A$ or $B \preceq A$.
A^{T}	For the matrix A, its transpose (also used for a vector to represent the corresponding row vector).
A^{H}	The conjugate transpose, also called the adjoint, of the matrix A; $A^{\mathrm{H}} = \bar{A}^{\mathrm{T}} = \overline{A^{\mathrm{T}}}$.
A^{-1}	The inverse of the square, nonsingular matrix A.
$A^{-\mathrm{T}}$	The inverse of the transpose of the square, nonsingular matrix A.
A^{+}	The $m \times n$ g_4 inverse, the Moore-Penrose inverse, or the pseudoinverse of the $n \times m$ matrix A; that is, A^{+} a matrix such that $AA^{+}A = A$; $A^{+}AA^{+} = A^{+}$; $A^{+}A$ is symmetric; and AA^{+} is symmetric.
A^{-}	An $m \times n$ g_1 inverse, or generalized inverse of the matrix $n \times m$ A; that is, A^{-} a matrix such that $AA^{-}A = A$.
$A^{\frac{1}{2}}$	The square root of a nonnegative definite or positive definite matrix A; $(A^{\frac{1}{2}})^2 = A$.
$A^{-\frac{1}{2}}$	The square root of the inverse of a positive definite matrix A; $(A^{-\frac{1}{2}})^2 = A^{-1}$.

Norms and Inner Products

L_p	For real $p \geq 1$, a norm formed by accumulating the p^{th} powers of the moduli of individual elements in an object and then taking the $(1/p)^{\text{th}}$ power of the result.
$\|\cdot\|$	In general, the norm of the object $\cdot$.
$\|\cdot\|_p$	In general, the L_p norm of the object $\cdot$.
$\|x\|_p$	For the vector x, the L_p norm $$\|x\|_p = \left(\sum |x_i|^p\right)^{\frac{1}{p}}.$$
$\|X\|_p$	For the matrix X, the L_p norm $$\|X\|_p = \max_{\|v\|_p=1} \|Xv\|_p.$$
$\|X\|_F$	For the matrix X, the Frobenius norm $$\|X\|_F = \sqrt{\sum_{i,j} x_{ij}^2}.$$
$\langle x, y\rangle$	The inner product or dot product of x and y.
$\kappa_p(A)$	The L_p condition number of the nonsingular square matrix A with respect to inversion.

Notation Relating to Matrix Determinants

$\|A\|$	The determinant of the square matrix A, $\|A\| = \det(A)$.
$\det(A)$	The determinant of the square matrix A, $\det(A) = \|A\|$.
$\|A_{(i_1,\ldots,i_k)}\|$	A principal minor of a square matrix A; in this case, it is the minor corresponding to the matrix formed from rows $i_1, \ldots, i_k$ and columns $i_1, \ldots, i_k$ from a given matrix A.
$\|A_{-(i)(j)}\|$	The minor associated with the $(i,j)^{\text{th}}$ element of a square matrix A.

$a_{(ij)}$ — The cofactor associated with the $(i,j)^{\text{th}}$ element of a square matrix A; that is, $a_{(ij)} = (-1)^{i+j}|A_{-(i)(j)}|$.

$\text{adj}(A)$ — The adjugate, also called the classical adjoint, of the square matrix A: $\text{adj}(A) = (a_{(ji)})$; that is, the matrix of the same size as A formed from the cofactors of the elements of A^{T}.

Matrix-Vector Differentiation

$\text{d}t$ — The differential operator on the scalar, vector, or matrix t. This is an operator; d may be used to represent a variable. (Note the difference in the font.)

g_f or ∇f — For the scalar-valued function f of a vector variable, the vector whose i^{th} element is $\partial f/\partial x_i$. This is the gradient, also often denoted as g_f.

∇f — For the vector-valued function f of a vector variable, the matrix whose element in position (i,j) is

$$\frac{\partial f_j(x)}{\partial x_i}.$$

This is also written as $\partial f^{\text{T}}/\partial x$ or just as $\partial f/\partial x$. This is the transpose of the Jacobian of f.

J_f — For the vector-valued function f of a vector variable, the Jacobian of f denoted as J_f. The element in position (i,j) is

$$\frac{\partial f_i(x)}{\partial x_j}.$$

This is the transpose of (∇f): $\text{J}_f = (\nabla f)^{\text{T}}$.

H_f or $\nabla\nabla f$ — The Hessian of the scalar-valued function f of a vector variable. The Hessian is the transpose of the Jacobian of the gradient. Except in pathological cases, it is symmetric. The element in position (i,j) is

$$\frac{\partial^2 f(x)}{\partial x_i \partial x_j}.$$

The symbol $\nabla^2 f$ is also sometimes used to denote the Hessian of f, but I prefer not to use that notation.

$\nabla^2 f$	For the vector-valued function f of a vector variable, the trace of the Hessian. This is also called the Laplacian.

References

The number(s) following a reference indicate the page(s) on which the reference is cited. A few of these references are general ones that are not cited in the text.

Martin Aigner and Günter M. Ziegler. *Proofs from THE BOOK*. Springer-Verlag, Berlin, fourth edition, 2010. 674, 688

S. M. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society, Series B*, 28:131–142, 1966. 747

Robert B. Ash and Catherine A. Doleans-Dade. *Probability & Measure Theory*. Academic Press, New York, second edition, 1999. 145

K. B. Athreya and P. E. Ney. *Branching Processes*. Springer, Berlin, 1972. (Reprinted by Dover Publications, New York, 2004.). 144, 201

Krishna B. Athreya and Soumen N. Lahiri. *Measure Theory and Probability Theory*. Springer, New York, 2006. 145

George Bachman and Lawrence Narici. *Functional Analysis*. Dover Publications, Inc., Mineola, New York, 2000. (reprint with list of errata of the book published by Academic Press, New York, 1966). 638

R. R. Bahadur. A note on quantiles in large samples. *Annals of Mathematical Statistics*, 37:577–580, 1966. 97

N. Balakrishnan and Chin-Diew Lai. *Continuous Bivariate Distributions*. Springer, New York, second edition, 2009. 141

N. Balakrishnan and V. B. Nevzorov. *A Primer on Statistical Distributions*. Wiley–Interscience, New York, 2003. 837

O. E. Barndorff-Nielson. *Information and Exponential Families in Statistical Theory*. Wiley, Chichester, 1978. 200

O. E. Barndorff-Nielson and D. R. Cox. *Inference and Asymptotics*. Chapman and Hall, New York, 1994. 145, 321

Richard F. Bass. *Stochastic Processes*. Cambridge University Press, Cambridge, United Kingdom, 2011. 780

Martin Baxter and Andrew Rennie. *Financial Calculus: An Introduction to Derivative Pricing*. Cambridge University Press, Cambridge, United Kingdom, 1996. 144

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57:289–300, 1995. 538

James 0. Berger. Could Fisher, Jeffreys and Neyman have agreed on testing? *Statistical Science*, 18:1–31, 2003. 558

James O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, New York, second edition, 1985. 328, 381

James O. Berger and Thomas Sellke. Testing a point null hypothesis: The irreconcilability of p values and evidence (with discussion). *Journal of the American Statistical Association*, 82:112–139, 1987. 381

James O. Berger and Robert L. Wolpert. *The Likelihood Principle*. The Institute of Mathematical Statistics, Hayward, California, second edition, 1988. 318

R. N. Bhattacharya and R. Ranga Rao. *Normal Approximations and Asymptotic Expansions*. John Wiley & Sons, New York, 1976. 143

P. J. Bickel and E. L. Lehmann. Unbiased estimation in convex families. *Annals of Mathematical Statistics*, 40:1523–1535, 1969. 51, 404

Peter J. Bickel and David A. Freedman. Some asymptotic theory for the bootstrap. *Annals of Statistics*, 9:1196–1217, 1981. 600

Christophe Biernacki. Testing for a global maximum of the likelihood. *Journal of Computational and Graphical Statistics*, 14:657–674, 2005. 502

Patrick Billingsley. *Probability and Measure*. John Wiley & Sons, New York, third edition, 1995. 48, 86, 88, 104, 119, 145, 691, 710, 713, 735, 738, 739

Allan Birnbaum. On the foundations of statistical inference (with discussion by L. J. Savage, George Barnard, Jerome Cornfield, Irwin Bross, George E. P. Box, I. J. Good, D. V. Lindley, C. W. Clunies-Ross, John W. Pratt, Howard Levene, Thomas Goldman, A. P. Dempster, and Oscar Kempthorne). *Journal of the American Statistical Association*, 57:269–326, 1962. 318

David Blackwell and M. A. Girshick. *Theory of Games and Statistical Decisions*. John Wiley & Sons, New York, 1954. (Reprinted by Dover Publications, New York, 1979.). 319

Colin R. Blyth. Some probability paradoxes in choice from among random alternatives. *Journal of the American Statistical Association*, 67:366–373, 1972. 220

Ralph Boas Jr. *A Primer on Real Functions*. The Mathematical Association of America, Washington, DC, 1960. 723, 762

Leo Breiman. Statistical modeling: The two cultures (with discussion). *Statistical Science*, 16:199–231, 2001. 322

Leo Breiman. *Probability*. Addison-Wesley, New York, 1968. (Reprinted by the Society for Industrial and Applied Mathematics, Philadelphia, 1992.). 145, 200

Lyle D. Broemeling. *Bayesian Analysis of Linear Models*. Chapman & Hall/CRC, Boca Raton, 1984. 382

James Ward Brown and Ruel V. Churchill. *Complex Variables and Applications*. McGraw-Hill Book Company, New York, eighth edition, 2008. 762

Lawrence D. Brown. *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory*. Institute of Mathematical Statistics, Hayward, California, 1986. 173, 200

H. D. Brunk. On an extension of the concept of conditional expectation. *Proceedings of the American Mathematical Society*, 14:298–304, 1963. 141

H. D. Brunk. Conditional expectation given a σ-lattice and applications. *Annals of Mathematical Statistics*, 36:1339–1350, 1965. 141

A. Buse. The likelihood ratio, Wald, and Lagrange multiplier tests: An expository note. *The American Statistician*, 36:153–157, 1982. 558

George Casella and Roger L. Berger. Reconciling Bayesian and frequentist evidence in the one-sided testing problem, (with discussion). *Journal of the American Statistical Association*, 82:106–111 and 123–139, 1987. 381

Kai Lai Chung. *A Course in Probability Theory*. Academic Press, New York, second revised edition, 2000. 48, 145

Daren B. H. Cline and Sidney I. Resnick. Multivariate subexponential distributions. *Stochastic Processes and their Applications*, 42:49–72, 1992. 200

Noel Cressie and Timothy R. C. Read. Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society, Series B*, 46:440–464, 1984. 253

I. Csiszár. Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica*, 2:299–318, 1967. 747

Anirban DasGupta. Best constants in chebyshev inequalities with various applications. *Metrika*, 51:185–200, 2000. 848, 856

Anirban DasGupta. *Asymptotic Theory of Statistics and Probability*. Springer, New York, 2008. 321, 856

H. A. David. *The Method of Paired Comparisons*. Griffith/Oxford University Press, London, second edition, 1988. 319

H. T. David and Shawki A. Salem. Three shrinkage constructions for Pitman-closeness in the one-dimensional location case. *Communications in Statistics — Theory and Methods*, 20:3605–3627, 1991. 319

Herbert A. David and H. N. Nagaraja. *Order Statistics*. John Wiley & Sons, New York, third edition, 2003. 97, 109, 143, 149

Laurie Davies and Ursula Gather. Robust statistics. In James E. Gentle, Wolfgang Härdle, and Yuichi Mori, editors, *Handbook of Computational Statistics; Concepts and Methods*, pages 711–750, Berlin, 2012. Springer. 610

Bruno de Finetti. Probability and exchangeability from a subjective point of view. *International Statistical Review*, 47:139–135, 1979. 138

Laurens de Haan and Ana Ferreira. *Extreme Value Theory. An Introduction*. Springer, New York, 2006. 109, 143

Morris. H. DeGroot. *Optimal Statistical Decisions*. John Wiley & Sons, New York, 1970. (A paperback edition with some additional material was published in the Wiley Classics Series in 2004.). 140

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, 45:51–59, 1977. 470, 479, 502

A. P. Dempster, N. M. Laird, and D. B. Rubin. Iteratively reweighted least squares for linear regression when errors are normal/independent distributed. In Paruchuri R. Krishnaiah, editor, *Multivariate Analysis V*, pages 35–57, Amsterdam, 1980. North-Holland. 480

Dipak Dey, Sujit K. Ghosh, and Bani K. Mallick, editors. *Generalized Linear Models: A Bayesian Perspective*. Marcel Dekker, Inc, New York, 2000. 382

Sudhakar Dharmadhikari and Kumar Joag-Dev. *Unimodality, Convexity, and Applications*. Academic Press, New York, 1988. 848

Persi Diaconis. Finite forms of de Finetti's theorem on exchangeability. *Synthese*, 36:271–281, 1977. 114

J. L. Doob. *Stochastic Processes*. John Wiley & Sons, New York, 1953. (A paperback edition with some additional material was published in the Wiley Classics Series in 1990.). 144

J. L. Doob. *Measure Theory*. Springer-Verlag, New York, 1994.

R. M. Dudley. *Real Analysis and Probability*. Cambridge University Press, Cambridge, United Kingdom, second edition, 2002. 145

R. M. Dudley. *Uniform Central Limit Theorems*. Cambridge University Press, Cambridge, United Kingdom, 1999. 143, 144

J. Durbin. Estimation of parameters in time series regression models. *Journal of the Royal Statistical Society, Series B*, 22:139–153, 1960. 320, 503

A. Dvoretzky, J. Kiefer, and J. Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Annals of Mathematical Statistics*, 27:642–669, 1956. 135

Morris L. Eaton. *Group Invariance Applications in Statistics*. Institute of Mathematical Statistics, Hayward, California, 1989. 320

A. W. F. Edwards. *Likelihood*. Johns Hopkins University Press, Baltimore, expanded edition, 1992. 502, 558

B. Efron. Biased versus unbiased estimation. *Advances in Mathematics*, 16: 259–277, 1975. 220

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, New York, 1993. 321

F. Eiker. Asymptotic normality and consistency of the least squares estimators for families of linear regressions. *Annals of Mathematical Statistics*, 34:447–456, 1963. 321

Merran Evans, Nicholas Hastings, and Brian Peacock. *Statistical Distributions*. John Wiley & Sons, New York, third edition, 2000. 837

Eugene F. Fama and Richard Roll. Parameter estimates for symmetric stable distributions. *Journal of the American Statistical Association*, 66:331–338, 1971. 252

William Feller. *An Introduction to Probability Theory and Its Applications, Volume I*. John Wiley & Sons, New York, 1957. 145

William Feller. *An Introduction to Probability Theory and Its Applications, Volume II*. John Wiley & Sons, New York, 1971. 62, 145, 200

Thomas S. Ferguson. *Mathematical Statistics. A Decision Theoretic Approach*. Academic Press, New York, 1967. 319, 361

Thomas S. Ferguson. A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1:209–230, 1973. 382

R. A. Fisher and L. H. C. Tippett. Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Proceedings of the Cambridge Philosophical Society*, 24:180–190, 1928. 109, 143

Bernard Flury and Alice Zoppè. Exercises in EM. *The American Statistician*, 54:207–209, 2000. 471

David A. Freedman. How can the score test be inconsistent? *The American Statistician*, 61:291–295, 2007. 558

David A. Freedman. Bernard Friedman's urn. *Annals of Mathematical Statistics*, 36:965–970, 1965. 131

Janos Galambos. *The Asymptotic Theory of Extreme Order Statistics*. John Wiley & Sons, New York, 1978. 143

R. C. Geary. The distribution of "Student's" ratio for non-normal samples. *Journal of the Royal Statistical Society, Supplement 3*, pages 178–184, 1936. 142

Seymour Geisser. *Predictive Inference: An Introduction*. Chapman & Hall, New York, 1993. 321

Bernard R. Gelbaum and John M. H. Olmsted. *Counterexamples in Analysis*. Dover Publications, Inc., Mineola, New York, 2003. (corrected reprint of the second printing published by Holden-Day, Inc., San Francisco, 1965). 688, 762

Bernard R. Gelbaum and John M. H. Olmsted. *Theorems and Counterexamples in Mathematics*. Springer, New York, 1990. 688, 762

Alan E. Gelfand and Adrian F. M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85:398–409, 1990. (Reprinted in Samuel Kotz and Norman L. Johnson (Editors) (1997), *Breakthroughs in Statistics, Volume III*, Springer-Verlag, New York, 526–550.). 378

Christopher Genovese and Larry Wasserman. Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society, Series B*, 64:499–517, 2002. 538

James E. Gentle. *Matrix Algebra: Theory, Computations, and Applications in Statistics*. Springer, New York, 2007. 127, 144, 176, 432, 433, 685, 730, 781, 801, 804, 818, 821

James E. Gentle. *Computational Statistics*. Springer, New York, 2009. 321, 644

J. K. Ghosh. A new proof of the Bahadur representation of quantiles and an application. *Annals of Mathematical Statistics*, 42:1957–1961, 1971. 97

Malay Ghosh. Objective priors: An introduction for frequentists (with discussion). *Statistical Science*, 26:187–211, 2011. 381

Malay Ghosh and Glen Meeden. *Bayesian Methods for Finite Population Sampling*. Chapman & Hall/CRC, Boca Raton, 1998. 382

Malay Ghosh and Pranab Kumar Sen. Bayesian Pitman closeness. *Communications in Statistics — Theory and Methods*, 20:3423–3437, 1991. 381

B. Gnedenko. Sur la distribution limite du terme maximum d'une serie aleatoire. *Annals of Mathematics, Second Series*, 44:423–453, 1943. 109, 143

B. V. Gnedenko and A. N. Kolmogorov. *Limit Distributions for Sums of Independent Random Variables*. Addison-Wesley Publishing Company, Inc., Reading, Massachusetts, 1954. (Translated from the 1949 Russian edition by K. L. Chung, with an appendix by J. L. Doob.). 1, 142, 143, 145

Boris V. Gnedenko. *Theory of Probability*. Gordon and Breach Science Publishers, Amsterdam, sixth edition, 1997. (Translated from the 1988 Russian edition by Igor A. Ushakov, after the death of Gnedenko in 1995.). 145

V. P. Godambe. An optimum property of regular maximum likelihood estimation. *Annals of Mathematical Statistics*, 31:1208–1211, 1960. 320, 503

Irving J. Good. *Good Thinking. The Foundations of Probability and Its Applications*. University of Minnesota Press, Minneapolis, 1983. (Reprinted by Dover Publications, New York, 2009.). 380

Ronald L. Graham, Donald E. Knuth, and Oren Patashnik. *Concrete Mathematics*. Addison-Wesley, Upper Saddle River, NJ, second revised edition, 1994. viii, 689

Igor Griva, Stephen G. Nash, and Ariela Sofer. *Linear and Nonlinear Optimization*. Society for Industrial and Applied Mathematics, Philadelphia, second edition, 2009. 658

Allan Gut. *Probability: A Graduate Course*. Springer, New York, 2005. 48, 145

Ian Hacking. *The Emergence of Probability: A Philosophical Study of Early Ideas about Probability, Induction and Statistical Inference*. Cambridge University Press, Cambridge, United Kingdom, 1975. 140

P. Hall and C. C. Heyde. *Martingale Limit Theory and Its Application*. Academic Press, New York, 1980. 144

Peter Hall. *The Bootstrap and Edgeworth Expansion*. Springer, New York, 1992. 143

Paul R. Halmos. The theory of unbiased estimation. *Annals of Mathematical Statistics*, 17:34–43, 1946. 320, 442

Hugo C. Hamaker. Probability and exchangeability from an objective point of view. *International Statistical Review*, 45:223–231, 1977. 138

Theodore E. Harris. *The Theory of Branching Processes*. Dover Publications, Inc., Mineola, New York, 1989. (reprint with corrections and additional material of the book published simultaneously by Springer-Verlag, Berlin, and Prentice-Hall, Englewood Cliffs, New Jersey, 1963). 144

H. O. Hartley. In Dr. Bayes' consulting room. *The American Statistician*, 17(1):22–24, 1963. 381

David A. Harville. Quadratic unbiased estimation of variance components for the one-way classification. *Biometrika*, 56:313–326, 1969. 503

David A. Harville. *Matrix Algebra from a Statistician's Point of View*. Springer, New York, 1997. 821

Thomas Hawkins. *Lebesgue's Theory of Integration: Its Origins and Development*. Chelsea Publishing Company, New York, second edition, 1979. (Reprinted by the American Mathematical Society, 2001). 761

Leon H. Herbach. Properties of model II — Type analysis of variance tests, A: Optimum nature of the F-test for model II in the balanced case. *Annals of Mathematical Statistics*, 30:939–959, 1959. 503

Thomas P. Hettmansperger and Lawrence A. Klimko. A note on the strong convergence of distributions. *Annals of Statistics*, 2:597–598, 1974. 83

Edwin Hewitt and Karl Stromberg. *Real and Abstract Analysis*. Springer-Verlag, Berlin, 1965. A corrected second printing was published in 1969. viii, 75, 633, 645, 648, 762

C. C. Heyde. On a property of the lognormal distribution. *Journal of the Royal Statistical Society, Series B*, 29:392–393, 1963. 142

Christopher C. Heyde. *Quasi-Likelihood and Its Application: A General Approach to Optimal Parameter Estimation*. Springer, New York, 1997. 320, 503

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics*, 19:293–325, 1948. 443

Robert V. Hogg. Adaptive robust procedures: A partial review and some suggestions for future applications and theory (with discussion). *Journal of the American Statistical Association*, 69:909–927, 1974. 610

Robert V. Hogg and Russell V. Lenth. A review of some adaptive statistical techniques. *Communications in Statistics — Theory and Methods*, 13:1551–1579, 1984. 610

Peter J. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, I:221–233, 1967. 321

Peter J. Huber and Elvezio M. Ronchetti. *Robust Statistics*. John Wiley & Sons, New York, second edition, 2009. 610

Aleksander Janicki and Aleksander Weron. *Simulation and Chaotic Behavior of α-Stable Stochastic Processes*. Marcel Dekker, Inc., New York, 1994. 200

E. T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge, United Kingdom, 2003. 319

E. T. Jaynes. Information theory and statistical mechanics. *Physical Review Series II*, 106:620–630, 1957a. 382

E. T. Jaynes. Information theory and statistical mechanics ii. *Physical Review Series II*, 108:171–190, 1957b. 382

Harold Jeffreys. *Theory of Probability*. Oxford University Press, Oxford, third edition, 1961. 368, 380

Jiming Jiang. *Large Sample Techniques for Statistics*. Springer, New York, 2010. 311, 321

Norman L. Johnson, Samuel Kotz, and N. Balakrishnan. *Continuous Univariate Distributions. Volume 1*. John Wiley & Sons, New York, 1994. 837

Norman L. Johnson, Samuel Kotz, and N. Balakrishnan. *Continuous Multivariate Distributions. Volume 1: Models and Applications*. John Wiley & Sons, New York, second edition, 1995a. 837

Norman L. Johnson, Samuel Kotz, and N. Balakrishnan. *Continuous Univariate Distributions. Volume 2*. John Wiley & Sons, New York, 1995b. 837

Norman L. Johnson, Samuel Kotz, and N. Balakrishnan. *Discrete Multivariate Distributions*. John Wiley & Sons, New York, 1997. 837

Norman L. Johnson, Adrienne W. Kemp, and Samuel Kotz. *Univariate Discrete Distributions*. John Wiley & Sons, New York, third edition, 2005. 837

Valen E. Johnson and David Rossell. On the use of non-local densities in Bayesian hypothesis tests. *Journal of the Royal Statistical Society, Series B*, 72:143–170, 2010. 372

L. B. W. Jolley. *Summation of Series*. Dover Publications, Inc., New York, second revised edition, 1961. 689

Joseph B. Kadane, Mark J. Schervish, and Teddy Seidenfeld, editors. *Rethinking the Foundations of Statistics*. Cambridge University Press, Cambridge, United Kingdom, 1999. 380

John D. Kalbfleisch and Ross L. Prentice. *The Statistical Analysis of Failure Time Data*. John Wiley & Sons, New York, second edition, 2002. 609

Ioannis Karatzas and Steven E. Shreve. *Brownian Motion and Stochastic Calculus*. Springer-Verlag, New York, second edition, 1991. 780

Robert E. Kass and Larry Wasserman. The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91:1343–1370, 1996. 382

M. G. Kendall. Regression, structure and functional relationship. Part I. *Biometrika*, 38:11–25, 1951. 320

J. M. Keynes. *A Treatise on Probability*. MacMillan & Co., London, 1921. 138

André I. Khuri. *Advanced Calculus with Applications in Statistics*. John Wiley & Sons, New York, second edition, 2003. 688, 736

J. Kiefer. On Wald's complete class theorems. *Annals of Mathematical Statistics*, 24:70–75, 1953. 320

J. Kiefer. On Bahadur's representation of sample quantiles. *Annals of Mathematical Statistics*, 38:1323–1342, 1967. 97

Dong K. Kim and Jeremy M. G. Taylor. The restricted EM algorithm for maximum likelihood estimation under linear restrictions on the parameters. *Journal of the American Statistical Association*, 90:708–716, 1995. 480

Stephen M. Kogan and Douglas B. Williams. Characteristic function based estimation of stable distribution parameters. In Robert J. Adler, Raisa E.

Feldman, and Murad S. Taqqu, editors, *Statistics and Related Topics*, pages 311–335, Boston, 1998. Birkhäuser. 252

Tõnu Kollo and Dietrich von Rosen. *Advanced Multivariate Statistics with Matrices*. Springer, Dordrecht, The Netherlands, 2005. 141, 792, 821

A. N. Kolmogorov. *Foundations of the Theory of Probability*. Chelsea Publishing Company, New York, 1956. translated from the German. 139

A. N. Kolmogorov and S. V. Fomin. *Elements of the Theory of Functions and Functional Analysis, in two volumes*. Gaylock Press, Rochester, NY, 1954, 1960. (translated from the Russian) Reprinted in one volume (1999) by Dover Publications, Inc., Mineola, NY.

Samuel Kotz, N. Balakrishnan, and Norman L. Johnson. *Continuous Multivariate Distributions. Volume 2*. John Wiley & Sons, New York, second edition, 2000. 837

Ioannis A. Koutrouvelis. Regression-type estimation of the parameters of stable laws. *Journal of the American Statistical Association*, 75:918–928, 1980. 252

Jeanne Kowalski and Xin M. Tu. *Modern Applied U-Statistics*. John Wiley & Sons, New York, 2008. 443

Charles H. Kraft, John W. Pratt, and A. Seidenberg. Intuitive probability on finite sets. *Annals of Mathematical Statistics*, 30:408–419, 1959. 139

H. O. Lancaster. Zero correlation and independence. *Australian Journal of Statistics*, 1:53–56, 1959. 185

Kenneth Lange, David R. Hunter, and Ilsoon Yang. Optimization transfer using surrogate objective functions (with discussion. *Journal of Computational and Graphical Statistics*, 9:1–59, 2000. 829

L. Le Cam. On some asymptotic properties of maximum likelihood estimates and related Bayes estimates. *University of California Publications in Statistics*, 1:277–330, 1953. Though this is a widely-cited paper and a review, MR0054913, appeared in *Mathematical Reviews*, no such serial as *University of California Publications in Statistics* seems to be widely available. 421, 422

L. Le Cam. An extension of Wald's theory of statistical decision functions. *Annals of Mathematical Statistics*, 26:69–81, 1955. 320

Lucien Le Cam and Grace Lo Yang. *Asymptotics in Statistics: Some Basic Concepts*. Springer, New York, second edition, 2000. 200

Henri Lebesgue. Sur le développement de la notion d'intégrale. *Revue de métaphysique et de morale*, 34:149–167, 1926. Translated by Kenneth O. May and reprinted in *Measure and the Integral*, edited by Kenneth O. May (1966), Holden-Day, Inc., San Francisco, 178–194. 762

Hugues Leblanc. *Statistical and Inductive Probabilities*. Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1962. Reprinted by Dover Publications, New York, 2006. 138

Lawrence M. Leemis and Jacquelyn T. McQueston. Univariate distribution relationships. *The American Statistician*, 62:45–53, 2008. 837

E. L. Lehmann. A general concept of unbiasedness. *Annals of Mathematical Statistics*, 22:587–592, 1951. 320

E. L. Lehmann. *Elements of Large-Sample Theory*. Springer, New York, 1999. 321

E. L. Lehmann and George Casella. *Theory of Point Estimation*. Springer, New York, second edition, 1998. vii

E. L. Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses*. Springer, New York, third edition, 2005. vii

D. V. Lindley. A statistical paradox. *Biometrika*, 44:187–192, 1957. 371

D. V. Lindley and L. D. Phillips. Inference for a Bernoulli process (A Bayesian view). *The American Statistician*, 30:112–119, 1976. 1, 381, 558

J. E. Littlewood. *Lectures on the Theory of Functions*. Oxford University Press, Oxford, UK, 1944. 726

Thomas A. Louis. Finding the observed information matrix when using the EM algorithm. *Journal of the the Royal Statistical Society, Series B*, 44: 226–233, 1982. 479

Eugene Lukacs. *Characteristic Functions*. Hafner Publishing Company, New York, second edition, 1970. 142

Charles F. Manski. *Analog Estimation Methods in Econometrics*. Chapman & Hall, New York, 1988. 319

P. Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Annals of Probability*, 18:1269–1283, 1990. 135, 144

Sharon Bertsch McGrayne. *The Theory that Would Not Die*. Yale University Press, New Haven, 2011. 381

X.-L. Meng and D. B. Rubin. Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80:267–278, 1993. 479

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equations of state calculation by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092, 1953. (Reprinted in Samuel Kotz and Norman L. Johnson (Editors) (1997), *Breakthroughs in Statistics, Volume III*, Springer-Verlag, New York, 127–139.). 830

Sean Meyn and Richard L. Tweedie. *Markov Chains and Stochastic Stability*. Cambridge University Press, Cambridge, United Kingdom, second edition, 2009. 144

Thomas Mikosch. Copulas: Tales and facts (with discussion). *Extremes*, 9: 3–66, 2006. 141

B. J. T. Morgan, K. J. Palmer, and M. S. Ridout. Negative score test statistic. *The American Statistician*, 61:285–288, 2007. 533, 535

Carl N. Morris. Natural exponential families with quadratic variance functions. *Annals of Statistics*, 10:65–80, 1982. 200

Carl N. Morris and Kari F. Lock. Unifying the named natural exponential families and their relatives. *The American Statistician*, 63:247–253, 2009. 173, 200, 837

MS2. *Shao (2003)*. 232, 271, 275, 287, 316, 397, 398, 413, 418, 419, 420, 440, 442, 486, 525, 529, 530, 531, 546, 547, 551, 558, 609

Roger B. Nelsen. *An Introduction to Copulas*. Springer, New York, second edition, 2006. 40

J. Neyman and E. L. Scott. Consistent estimates based on partially consistent observations. *Econometrica*, 16:1–32, 1948. 503

Shu Kay Ng, Thriyambakam Krishnan, and Geoffrey J. McLachlan. The EM algorithm. In James E. Gentle, Wolfgang Härdle, and Yuichi Mori, editors, *Handbook of Computational Statistics; Concepts and Methods*, pages 139–172, Berlin, 2012. Springer. 502

Bernt Øksendal. *Stochastic Differential Equations. An Introduction with Applications*. Springer, Heidelberg, fifth edition, 1998. 780

Frank W. J. Olver, Daniel W. Lozier, Ronald F. Boisvert, and Charles W. Clark, editors. *NIST Handbook of Mathematical Functions*. National Institute of Standards and Technology and Cambridge University Press, Gaithersburg, Maryland and Cambridge, United Kingdom, 2010. Available at `http://dlmf.nist.gov/`. 466, 864

Art B. Owen. *Empirical Likelihood*. Chapman & Hall/CRC, Boca Raton, 2001. 503

Leandro Pardo. *Statistical Inference Based on Divergence Measures*. Chapman & Hall/CRC, Boca Raton, 2005. 319, 747

Valentin V. Petrov. *Limit Theorems of Probability Theory*. Oxford University Press, Oxford, United Kingdom, 1995. 102, 142, 856

E. J. G. Pitman. Subexponential distribution functions. *Journal of the Australian Mathematical Society (Series A)*, 29:337–347, 1980. 201

E. J. G. Pitman. The "closest" estimates of statistical parameters. *Communications in Statistics — Theory and Methods*, 20:3423–3437, 1991. 319

David Pollard. *A User's Guide to Mearure Theoretic Probability*. Cambridge University Press, Cambridge, United Kingdom, 2003. second printing with corrections. 145

Raquel Prado and Mike West. *Time Series: Modelling, Computation and Inference*. Chapman & Hall/CRC Press, Boca Raton, 2010. 382

S. James Press. Estimation in univariate and multivariate stable distributions. *Journal of the American Statistical Association*, 67:842–846, 1972. 252

S. James Press and Judith M. Tanur. *The Subjectivity of Scientists and the Bayesian Approach*. Wiley-Interscience, New York, 2001. 317, 318

A. R. Rajwade and A. K. Bhandari. *Surprises and Counterexamples in Real Function Theory*. Hindustan Book Agency, New Delhi, 2007. 762

C. R. Rao. Some comments on the minimum mean square as a criterion of estimation. In M. Csörgö, D. A. Dawson, J. N. K. Rao, and A. K. Md. E. Saleh, editors, *Statistics and Related Topics*, pages 123–143, Amsterdam, 1981. North-Holland. 218, 319

J. N. K. Rao. Impact of frequentist and Bayesian methods on survey sampling practice: A selective appraisal (with discussion). *Statistical Science*, 26:240–270, 2011. 382

J. N. K. Rao and J. T. Webster. On two methods of bias reduction in the estimation of ratios. *Biometrika*, 53:571–577, 1966. 303

Herbert Robbins. Statistical methods related to the law of the iterated logarithm. *Annals of Mathematical Statistics*, 41:1397–1409, 1970. 132

Christian P. Robert. *The Bayesian Choice*. Springer, New York, second edition, 2001. 381

L. C. G. Rogers and David Williams. *Diffusions, Markov Processes, and Martingales: Volume 1, Foundations*. Cambridge University Press, Cambridge, United Kingdom, second edition, 2000a. 780

L. C. G. Rogers and David Williams. *Diffusions, Markov Processes, and Martingales: Volume 2, Itô Calculus*. Cambridge University Press, Cambridge, United Kingdom, second edition, 2000b. 780

Joseph P. Romano and Andrew F. Siegel. *Counterexamples in probability and statistics*. Wadsworth & Brooks/Cole, Monterey, California, 1986. 79, 80, 82, 391, 401, 420, 421, 461, 484, 688, 762

Jutta Roosen and David A. Hennessy. Testing for the monotone likelihood ratio assumption. *Journal of Business & Economic Statistics*, 22:358–366, 2004. 558

Richard M. Royall. *Statistical Evidence: A Likelihood Paradigm*. Chapman & Hall/CRC, Boca Raton, 1997. 318, 541, 558

H. L. Royden. *Real Analysis*. MacMillan, New York, third edition, 1988. 762

Francisco J. Samaniego. *A Comparison of the Bayesian and Frequentist Approaches to Estimation*. Springer-Verlag, New York, 2010. 318

Gennady Samorodnitsky and Murad S. Taqqu. *Stable Non-Gaussian Random Processes*. Chapman & Hall/CRC, Boca Raton, 1994. 200

Carl-Erik Särndal, Bengt Swensson, and Jan Wretman. *Model Assisted Survey Sampling*. Springer-Verlag, New York, 1997. 439

Leonard J. Savage. *The Foundations of Statistics*. Dover Publications, New York, second revised edition, 1972. (First edition by John Wiley & Sons, New York, 1954.). 380

Henry Scheffé. A useful convergence theorem for probability distributions. *Annals of Mathematical Statistics*, 18:434–438, 1947. 83

Mark J. Schervish. *Theory of Statistics*. Springer, New York, 1995. 75, 114

Robert Schlaifer. *Probability and Statistics for Business Decisions*. McGraw-Hill Book Company, New York, 1959. 380

David W. Scott. Multivariate density estimation and visualization. In James E. Gentle, Wolfgang Härdle, and Yuichi Mori, editors, *Handbook of Computational Statistics; Concepts and Methods*, pages 549–570, Berlin, 2012. Springer. 579

David W. Scott. *Multivariate Density Estimation*. John Wiley & Sons, New York, 1992. 579

Shayle R. Searle, George Casella, and Charles E. McCulloch. *Variance Components*. Wiley-Interscience, New York, 1992. 503

Robert J. Serfling. *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, New York, 1980. 143, 321, 412, 443

J. Sethuraman. A constructive definition of dirichlet priors. *Statistica Sinica*, 4:639–650, 1994. 382

Glenn Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, New Jersey, 1976. 381

Glenn Shafer. Lindley's paradox (with discussion by D. V. Lindley, Morris H. DeGroot, I. J. Good, Bruce M. Hill, and Robert E. Kass). *Journal of the American Statistical Association*, 77:325–351, 1982. 372

Jun Shao. *Mathematical Statistics*. Springer, New York, second edition, 2003. vii

Jun Shao. *Mathematical Statistics: Exercises and Solutions*. Springer, New York, 2005.

Jun Shao and Dongsheng Tu. *The Jackknife and Bootstrap*. Springer-Verlag, New York, 1995. 321

J. A. Shohat and J. D. Tamarkin. *The Problem of Moments*. American Mathematical Society, New York, 1943. 142

Galen R. Shorack and Jon A. Wellner. *Empirical Processes With Applications to Statistics (Classics in Applied Mathematics)*. Society for Industrial and Applied Mathematics, Philadelphia, 2009. (Originally published in 1986 by John Wiley & Sons. This "Classic Edition" includes errata as well as some updated material.). 144

Christopher G. Small. *Expansions and Asymptotics for Statistics*. Chapman & Hall/CRC, Boca Raton, 2010. 609

Christopher G. Small and Jinfang Wang. *Numerical Methods for Nonlinear Estimating Functions*. Oxford University Press, Oxford, 2003. 320

Daniel Solow. *How to Read and Do Proofs*. John Wiley & Sons, New York, third edition, 2003. 688

Ehsan S. Soofi. Capturing the intangible concept of information. *Journal of the American Statistical Association*, 89:1243–1254, 1994. 318

James C. Spall. Stochastic optimization. In James E. Gentle, Wolfgang Härdle, and Yuichi Mori, editors, *Handbook of Computational Statistics; Concepts and Methods*, pages 173–202, Berlin, 2012. Springer. 829

Robert G. Staudte and Simon J. Sheather. *Robust Estimation and Testing*. John Wiley & Sons, New York, 1990. 610

J. Michael Steele. *Stochastic Calculus and Financial Applications*. Springer-Verlag, New York, 2001. 780

J. Michael Steele. *The Cauchy-Schwarz Master Class: An Introduction to the Art of Mathematical Inequalities*. Cambridge University Press, Cambridge, United Kingdom, 2004. 856

Lynn Arthur Steen and J. Arthur Seebach Jr. *Counterexamples in Topology*. Dover Publications, Inc., Mineola, New York, 1995. (reprint of the second edition published by Springer-Verlag, New York, 1978). 762

Frederick F. Stephan. The expected value and variance of the reciprocal and other negative powers of a positive bernoullian variate. *Annals of Mathematical Statistics*, 16:50–61, 1945. 460

Fred W. Steutel. *Preservation of Infinite Divisibility under Mixing and Related Topics*. Mathematisch Centrum Amsterdam, Amsterdam, The Netherlands, 1970. 200

Fred W. Steutel and Klaas van Harn. *Infinite Divisibility of Probability Distributions on the Real Line*. Marcel Dekker, Inc., New York, 2004. 200

Stephen M. Stigler. *A History of Statistics: The Measurement of Uncertainty before 1900*. Belknap Press of Harvard University Press, Cambridge, 1986. 380

John D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B*, 64:479–498, 2002. 559

John D. Storey. The positive false discovery rate: A Bayesian interpretation and the q-value. *Annals of Statistics*, 31:2013–2035, 2003. 559

Jordan M. Stoyanov. *Counterexamples in Probability*. John Wiley & Sons, Ltd., Chichester, United Kingdom, 1987. 762

William E. Strawderman. Proper Bayes minimax estimators of the multivariate normal mean. *Annals of Mathematical Statistics*, 42:385–388, 1971. 273

Robert Lee Taylor, Peter Z. Daffer, and Ronald F. Patterson. *Limit Theorems for Sums of Exchangeable Random Variables*. Rowman & Allanheld, Totowa, NJ, 1985. 143

Jozef L. Teugels. The class of subexponential distributions. *Annals of Probability*, 3:1000–1011, 1975. 200

William J. Thompson. *Atlas for Computing Mathematical Functions: An Illustrated Guide for Practitioners with Programs in Fortran 90 and Mathematica*. John Wiley & Sons, New York, 1997. 864

W. A. Thompson Jr. The problem of negative estimates of variance components. *Annals of Mathematical Statistics*, 33:273–289, 1962. 503

TPE2. *Lehmann and Casella (1998)*. 233, 311, 442

TSH3. *Lehmann and Romano (2005)*. 200, 528, 558

Richard Valliant, Alan H. Dorfman, and Richard M. Royall. *Finite Population Sampling and Inference: A Prediction Approach*. John Wiley & Sons, New York, 2000. 439

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, Cambridge, United Kingdom, 1998. 321

Various Authors. Chapter 4, Theory and Methods of Statistics. In Adrian E. Raftery, Martin A. Tanner, and Martin T. Wells, editors, *Statistics in the 21st Century*, New York, 2002. Chapman and Hall.

Geert Verbeke and Geert Molenberghs. What can go wrong with the score test? *The American Statistician*, 61:289–290, 2007. 558

R. von Mises (de Misès). La distribution de la plus grande de n valeurs. *Revue Mathématique de l'Union Interbalkanique*, pages 141–160, 1939. 143

R. von Mises (v. Mises). Fundamentalsätze der wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 4:1–97, 1919a. 138

R. von Mises (v. Mises). Grundlagen der wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 5:52–99, 1919b. 138

Abraham Wald. Contributions to the theory of statistical estimation and testing hypotheses. *Annals of Mathematical Statistics*, 10:299–326, 1939. 320

Abraham Wald. Sequential tests of statistical hypotheses. *Annals of Mathematical Statistics*, 16:117–186, 1945. 559

Abraham Wald. An essentially complete class of admissible decision functions. *Annals of Mathematical Statistics*, 18:549–555, 1947a. 320

Abraham Wald. Foundations of a general theory of sequential decision functions. *Econometrica*, 15:279–313, 1947b. 559

Abraham Wald. Note on the consistency of the maximum likelihood estimate. *Annals of Mathematical Statistics*, 20:595–601, 1949. 242, 449

Abraham Wald. *Statistical Decision Functions*. John Wiley & Sons, New York, 1950. Reprinted by Chelsea Publishing Company, New York, 1971. 319, 320, 559

A. M. Walker. On the asymptotic behaviour of posterior distributions. *Journal of the Royal Statistical Society, Series B*, 31:80–88, 1969. 334

R.W.M. Wedderburn. Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, 61:439–447, 1974. 499, 503

Mike West and Jeff Harrison. *Bayesian Forecasting and Dynamic Models*. Springer-Verlag, New York, second edition, 1997. 382

Halbert White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48:817–838, 1980. 321

Peter Whittle. *Probability via Expectation*. Springer, New York, fourth edition, 2000. 140

R. A. Wijsman. On the attainment of the Cramér-Rao lower bound. *Annals of Statistics*, 1:538–542, 1973. 443

Gary L. Wise and Eric B. Hall. *Counterexamples in Probability and Real Analysis*. The Clarendon Press, Oxford University Press, New York, 1993. 762

J. Wolfowitz. On ϵ-complete classes of decision functions. *Annals of Mathematical Statistics*, 22:461–465, 1951. 265

C.F.J. Wu. Jackknife, bootstrap and other resampling methods in regression analysis (with discussion). *Annals of Statistics*, 14:1261–1350, 1986. 321

Nailong Wu. *The Maximum Entropy Method*. Springer, New York, 1997. 319

Ronald R. Yager and Liping Liu, editors. *Classic Works of the Dempster-Shafer Theory of Belief Functions*. Springer-Verlag, New York, 2008. 381

Ryszard Zieliński. Kernel estimators and the Dvoretzky-Kiefer-Wolfowitz inequality. *Applicationes Mathematicae*, 34:401–404, 2007. 594

A. Zygmund. A remark on characteristic funtions. *Annals of Mathematical Statistics*, 18:272–276, 1947. 148

Index