Lecture Notes on *Statistical Theory*[1]

Ryan Martin
Department of Mathematics, Statistics, and Computer Science
University of Illinois at Chicago
www.math.uic.edu/~rgmartin

January 8, 2015

[1]These notes are meant to supplement the lectures for Stat 411 at UIC given by the author. The course roughly follows the text by Hogg, McKean, and Craig, *Introduction to Mathematical Statistics*, 7th edition, 2012, henceforth referred to as HMC. The author makes no guarantees that these notes are free of typos or other, more serious errors.

Contents

Chapter 1

Statistics and Sampling Distributions

1.1 Introduction

Statistics is closely related to probability theory, but the two fields have entirely different goals. Recall, from Stat 401, that a typical probability problem starts with some assumptions about the distribution of a random variable (e.g., that it's binomial), and the objective is to derive some properties (probabilities, expected values, etc) of said random variable based on the stated assumptions. The statistics problem goes almost completely the other way around. Indeed, in statistics, a sample from a given population is observed, and the goal is to learn something about that population based on the sample. In other words, the goal in statistics is to reason from sample to population, rather than from population to sample as in the case of probability. So while the two things—probability and statistics—are closely related, there is clearly a sharp difference. One could even make the case that a statistics problem is actually more challenging than a probability problem because it requires more than just mathematics to solve. (The point here is that, in a statistics problem, there's simply too much information missing about the population to be able to derive *the* answer via the deductive reasoning of mathematics.) The goal of this course is to develop the mathematical theory of statistics, mostly building on calculus and probability.

To understand the goal a bit better, let's start with some notation. Let $X_1, \ldots, X_n$ be a random sample (independent and identically distributed, iid) from a distribution with cumulative distribution function (CDF) $F(x)$. The CDF admits a probability mass function (PMF) in the discrete case and a probability density function (PDF) in the continuous case; in either case, write this function as $f(x)$. One can imagine that $f(x)$ characterizes the population from which $X_1, \ldots, X_n$ is sampled from. Typically, there is something about this population that is unknown; otherwise, there's not much point in sampling from it. For example, if the population in question is of registered voters in Cook county, then one might be interested in the *unknown* proportion that would vote democrat in the upcoming election. The goal would be to "estimate" this proportion from a sample. But the point here is that the population/distribution of interest is not completely known. Mathematically, we handle this by introducing a quantity θ, taking values in some $\Theta \subseteq \mathbb{R}^d$, $d \geq 1$, and weakening the

initial assumption by saying that the distribution in question has PMF or PDF of the form $f_\theta(x)$ for some $\theta \in \Theta$. That is, the statistician believes that the data was produced by a distribution in a class indexed by Θ, and the problem boils down to picking a "good" value of θ to characterize the data-generating distribution.

Example 1.1. Suppose the population of registered voters in Cook county is divided into two groups: those who will vote democrat in the upcoming election, and those that will vote republican. To each individual in the population is associated a number, either 0 or 1, depending on whether he/she votes republican or democrat. If a sample of n individuals is taken completely at random, then the number X of democrat voters is a binomial random variable, written $X \sim \mathsf{Bin}(n, \theta)$, where $\theta \in \Theta = [0, 1]$ is the unknown proportion of democrat voters. The statistician wants to use the data $X = x$ to learn about θ.[1]

Throughout we will refer to θ as the *parameter* and Θ the *parameter space*. The typical problem we will encounter will begin with something like "Suppose $X_1, \ldots, X_n$ is an independent sample from a distribution with PDF $f_\theta(x) \ldots$" For the most part, we shall omit the (important) step of choosing the functional form of the PMF/PDF; Section 1.2 discusses this topic briefly. So we shall mostly take the functional form of $f_\theta(x)$ as fixed and focus on finding good ways to use the data to learn, or *make inference* about the value of θ. The two main statistical inference problems are summarized in Section 1.3. In Stat 411, we will focus mostly on the simplest of these problems, namely *point estimation*, since this is the easiest to understand and most of the fundamental concepts can be formulated in this context. But regardless of the statistical inference problem at hand, the first step of a statistical analysis is to produce some summary of the information in the data about the unknown parameter.[2] Such summaries are called *statistics*, and Section 1.4 gives an introduction. Once a summary statistic has been chosen, the sampling distribution of this statistic is required to construct a statistical inference procedure. Various characteristics of this sampling distribution will help not only for developing the procedure itself but for comparing procedures.

1.2 Model specification

The starting point for the problems in this course is that data $X_1, \ldots, X_n$ are an observed sample from a population characterized by a PMF or PDF $f_\theta(x)$, where the parameter θ is unknown. But it's not immediately clear where the knowledge about the functional form of $f_\theta(x)$ comes from. All the statistician has is a list of numbers representing the observed values of the random variables $X_1, \ldots, X_n$—how can he/she pick a reasonable model for these data? The most common way this is done is via *exploratory data analysis*, discussed briefly next.

[1] This is essentially what's being done when the results of exit polls are reported on the news; the only difference is that their sampling schemes are more sophisticated, so that they can report their results with a desired level of accuracy.

[2] There are other approaches, besides the so-called "frequentist" approach emphasized in Stat 411, which do not start with this kind of summary. Later on we shall briefly discuss the "Bayesian" approach, which is fundamentally and operationally very different.

In some select cases, however, the statistical model can be obtained by other means. For example, in some applied problems in physics, chemistry, etc, there may be a physical model coming from existing theory that determines the functional form of the statistical model. In other cases, the definition of experiment can determine the statistical model. Example 1.1 is one such case. There, because the population is dichotomous (democrat or republican), the number of democrats in an independent sample of size n is, by definition, a binomial random variable, so the PMF $f_\theta(x)$ is determined, modulo θ. But these situations are atypical.

When little or no information is available concerning the population in question, the statistician must rely on exploratory data analysis methods and some imagination to cook up the functional form of the PMF/PDF. These data analytic methods include drawing plots and calculating summary statistics, etc. Such methods are discussed in more detail in applied statistics courses. In addition to basic summaries like the mean and standard deviation, histograms five-number summaries, and boxplots are particularly useful.

1.3 Two kinds of inference problems

1.3.1 Point estimation

Suppose $X_1, \ldots, X_n$ are iid with PDF/PMF $f_\theta(x)$. The point estimation problem seeks to find a quantity $\hat{\theta}$, called an *estimator*, depending on the values of $X_1, \ldots, X_n$, which is a "good" guess, or estimate, of the unknown θ. The choice of $\hat{\theta}$ depends, not only on the data, but on the assumed model and the definition of "good." Initially, it seems obvious what should be meant by "good"—that $\hat{\theta}$ is close to θ—but as soon as one remembers that θ itself is unknown, the question of what it means even for $\hat{\theta}$ to be close to θ becomes uncertain. We'll discuss this more throughout the course.

It would be quite unreasonable to believe that one's point estimate $\hat{\theta}$ hits the unknown θ on the nose. Indeed, one should expect that $\hat{\theta}$ will miss θ by some positive amount. Therefore, in addition to a point estimate $\hat{\theta}$, it would be helpful to have some estimate of the amount by which $\hat{\theta}$ will miss θ. The necessary information is encoded in the sampling distribution (see Section 1.5) of $\hat{\theta}$, and we can summarize this by say, reporting the standard error or variance of $\hat{\theta}$, or with a confidence interval.[3]

1.3.2 Hypothesis testing

Unlike the point estimation problem which starts with a vague question like "what is θ?," the hypothesis testing problem starts with a specific question like "is θ equal to θ_0?," where θ_0 is some specified value. The popular t-test problem is one in this general class. In this case, notice that the goal is somewhat different than that of estimating an unknown parameter. The general setup is to construct a *decision rule*, depending on data, by which one can decide if $\theta = \theta_0$ or $\theta \neq \theta_0$. Often times this rule consists of taking a point estimate $\hat{\theta}$

[3]Confidence intervals are widely used in statistics and would be covered in detail in some applied courses. We'll discuss this a little bit later.

and comparing it with the hypothesized value θ_0: if $\hat{\theta}$ is too far from θ_0, conclude $\theta \neq \theta_0$, otherwise, conclude $\theta = \theta_0$. Later in the course we shall discuss how to define "too far" and what sorts of properties this decision rule should have.

1.4 Statistics

As discussed previously, the approach taken in Stat 411 starts with a summary of the information about θ in the data $X_1, \ldots, X_n$. This summary is called a *statistic*.

Definition 1.1. Let $X_1, \ldots, X_n$ be a sample whose distribution may or may not depend on an unknown parameter θ. Then any ("measurable") function $T = T(X_1, \ldots, X_n)$ that does not depend on θ is a *statistic*.

Examples of statistics $T = T(X_1, \ldots, X_n)$ include:

$$
\begin{aligned}
&T = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \text{the sample mean;} \\
&T = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2, \quad \text{the sample variance;} \\
&T = M_n, \quad \text{the sample median;} \\
&T \equiv 7; \\
&T = I_{(-\infty,7]}(X_1); \quad \text{etc, etc, etc.}
\end{aligned}
$$

(Here $I_A(x)$ denotes the *indicator function*, which takes value 1 if $x \in A$ and value 0 if $x \notin A$.) Apparently, a statistic T can be basically anything, some choices being more reasonable than others. The choice of statistic will depend on the problem at hand. Note that T need not be a continuous function. And by "depend on θ," we mean that θ cannot appear in the formula for T; it is OK (and actually preferable) if the distribution of T depends on θ.

An interesting class of statistics, of which the median is a special case, are the *order statistics*; Chapter 4.4 of HMC.

Definition 1.2. Let $X_1, \ldots, X_n$ be a random sample. Then the order statistics are the sorted values, denoted by $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$.

For example, if n is odd, then $X_{(\frac{n+1}{2})}$ is the sample median. The first and third quantiles (25th and 75th percentiles) can be defined similarly. The point is that even if the distribution of $X_1, \ldots, X_n$ depends on θ, the act of sorting them in ascending order has nothing to do with the particular value of θ; therefore, the order statistics are, indeed, statistics.

Example 1.2. Let $X_1, \ldots, X_n$ be an independent sample from a distribution with CDF $F(x)$ and PDF $f(x) = dF(x)/dx$. Fix an integer $k \leq n$. The the CDF of the kth order

statistic $X_{(k)}$ can be derived as follows.

$$\begin{aligned} G_k(x) &= \mathsf{P}(X_{(k)} \le x) = \mathsf{P}(\text{at least } k \text{ of } X_1, \ldots, X_n \text{ are} \le x) \\ &= \sum_{j=k}^{n} \binom{n}{j} [F(x)]^j [1 - F(x)]^{n-j}. \end{aligned}$$

Therefore, the PDF of $X_{(k)}$ is obtained by differentiating[4] with respect to x:

$$g_k(x) = \frac{dG_k(x)}{dx} = k\binom{n}{k} f(x)[F(x)]^{k-1}[1 - F(x)]^{n-k}.$$

In particular, if $F(x)$ is a $\mathsf{Unif}(0, 1)$ CDF, then

$$g_k(x) = k\binom{n}{k} x^{k-1}(1 - x)^{n-k}, \quad x \in (0, 1),$$

which is the PDF of a $\mathsf{Beta}(k, n - k + 1)$ distribution.

Exercise 1.1. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Unif}(0, 1)$. Find the expected value $\mathsf{E}(X_{(k)})$ and variance $\mathsf{V}(X_{(k)})$ of the kth order statistic. (Hint: Look up the beta distribution in HMC, p. 163.)

1.5 Sampling distributions

1.5.1 Basics

A sampling distribution is just the name given to the distribution of a statistic $T_n = T(X_1, \ldots, X_n)$; here T_n is a random variable because it is a function of the random variables $X_1, \ldots, X_n$. Similar ideas come up in discussions of "convergence in probability" (Chapter 5.1 of HMC), "convergence in distribution" (Chapter 5.2 of HMC), and especially the "central limit theorem" (Chapter 5.3 of HMC).

Here's one way to visualize the sampling distribution of T_n.

$$\begin{array}{rcl} X_1^{(1)}, \ldots, X_n^{(1)} & \to & T_n^{(1)} = T(X_1^{(1)}, \ldots, X_n^{(1)}) \\ X_1^{(2)}, \ldots, X_n^{(2)} & \to & T_n^{(2)} = T(X_1^{(2)}, \ldots, X_n^{(2)}) \\ & \vdots & \\ X_1^{(s)}, \ldots, X_n^{(s)} & \to & T_n^{(s)} = T(X_1^{(s)}, \ldots, X_n^{(s)}) \\ & \vdots & \end{array}$$

That is, for each sample $X_1^{(s)}, \ldots, X_n^{(s)}$ of size n, there is a corresponding $T_n^{(s)}$ obtained by applying the function $T(\cdot)$ to that particular sample. Then the sampling distribution of T_n is what would be approximated by a histogram of $T_n^{(1)}, T_n^{(2)}, \ldots, T_n^{(s)}, \ldots$; see Section 1.5.3

[4] This is a somewhat tedious calculation...

below. Mathematically, the sampling distribution of T_n is characterized by the CDF $F_{T_n}(t) = \mathsf{P}(T_n \leq t)$, where the probability $\mathsf{P}(\cdot)$ is with respect to the joint distribution of $X_1, \ldots, X_n$. In all but the strangest of cases, the sampling distribution of T_n will depend on n (and also any parameter θ involved in the joint distribution of $X_1, \ldots, X_n$). What follows is a simple but important example from Stat 401.

Example 1.3. Suppose $X_1, \ldots, X_n$ are iid $\mathsf{N}(\mu, \sigma^2)$. Let $T_n = \bar{X}$ be the sample mean, i.e., $T_n = n^{-1} \sum_{i=1}^n X_i$. Since T_n is a linear function of $X_1, \ldots, X_n$ and linear functions of normals are also normal (see Theorem 3.4.2 in HMC), it follows that T_n is, too, a normal random variable. In particular, $T_n \sim \mathsf{N}(\mu, \sigma^2/n)$.

Example 1.4. Consider again the setup in Example 1.3. Suppose that μ is a known number, and let $T_n = \sqrt{n}(\bar{X} - \mu)/S$, where $S^2 = (n-1)^{-1} \sum_{i=1}^n (X_i - \bar{X})^2$ is the sample variance and $S = \sqrt{S^2}$ is the sample standard deviation. It was shown in Stat 401 (See Theorem 3.6.1d in HMC) that the sampling distribution of this T_n is a Student-t distribution with $n-1$ degrees of freedom, written as $T_n \sim \mathsf{t}_{n-1}$.

Exercise 1.2. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Exp}(1)$, i.e., exponential distribution with mean 1. Let $T_n = \sum_{i=1}^n X_i$. Use moment-generating functions to show that $T_n \sim \mathsf{Gamma}(n, 1)$, i.e., a gamma distribution with shape parameter n and scale parameter 1. (Hint: See the proof of Theorem 3.3.2 in HMC, using the fact that $\mathsf{Exp}(1) \equiv \mathsf{Gamma}(1, 1)$.)

Except for special cases (like those in Examples 1.3–1.4), the exact sampling distribution of T_n will not be available. In such cases, we'll have to rely on some kind of approximation. There are asymptotic (large-n) approximations, and numerical approximations. These are discussed next in turn.

1.5.2 Asymptotic results

In most cases, the exact sampling distribution of T_n not available, probably because it's too hard to derive. One way to handle this is to approximate the sampling distribution by assuming n is "close to infinity." In other words, approximate the sampling distribution of T_n by its limiting distribution (if the latter exists). See Chapter 5.2 of HMC for details about convergence in distribution. The most important of such results is the famous *central limit theorem*, or CLT for short. This result gives the limiting distribution for sums/averages of almost any kind of random variables.

Theorem 1.1 (Central Limit Theorem, CLT)**.** *Let $X_1, \ldots, X_n$ be an iid sample from a distribution with mean μ and variance $\sigma^2 < \infty$. For $\bar{X}_n$ the sample mean, let $Z_n = \sqrt{n}(\bar{X}_n - \mu)/\sigma$. Then Z_n converges in distribution to $\mathsf{N}(0, 1)$ as $n \to \infty$.*

In other words, if n is large, then the distribution of $\bar{X}_n$ is approximately $\mathsf{N}(\mu, \sigma^2/n)$, which matches up with the (exact) conclusion for $\bar{X}$ in Example 1.3 in the special case where the underlying distribution was normal.

Here is a rough sketch of one proof of the CLT; for details, see p. 301–309 in HMC. Modify the notation of the statement a bit by writing $Z_i = (X_i - \mu)/\sigma$ and $\bar{Z}_n = n^{-1}\sum_{i=1}^n Z_i$. Then the moment-generating function for $\bar{Z}_n$ is given by

$$M_{\bar{Z}_n}(t) = \mathsf{E}(e^{t\bar{Z}_n}) = [\text{missing details}] = M_Z(t/\sqrt{n})^n, \tag{1.1}$$

where M_Z is the (common) moment-generating function for the Z_i's. Write a two-term Taylor approximation for $M_Z(t)$ at $t = 0$; the choice of $t = 0$ is reasonable since $t/\sqrt{n} \to 0$ as $n \to \infty$. We get

$$M_{\bar{Z}_n}(t) \approx \Big[1 + M_Z'(0) \cdot \frac{t}{\sqrt{n}} + \frac{M_Z''(0)}{2} \cdot \frac{t^2}{n}\Big]^n.$$

Since $\mathsf{E}(Z) = 0$ and $\mathsf{V}(Z) = \mathsf{E}(Z^2) = 1$, this simplifies to

$$M_{\bar{Z}_n}(t) \approx \Big[1 + \frac{t^2}{2n}\Big]^n.$$

Using tools from calculus, it can be verified that the right-hand side above converges to $e^{t^2/2}$ as $n \to \infty$. This is the moment-generating function of $\mathsf{N}(0,1)$, so the CLT follows by the uniqueness of moment-generating functions.

Exercise 1.3. Fill in the "missing details" in (1.1).

Exercise 1.4. Look up the definition of *convergence in probability*; see p. 289 in HMC. Using the notation in Theorem 1.1, argue heuristically that the CLT implies $\bar{X}_n \to \mu$ in probability. This argument demonstrates that the law of large numbers (LLN; Theorem 5.5.1 in HMC) is a consequence of the CLT.

Example 1.5. It's important to note that the CLT applies to other things besides the original data sequence. Here's an example. Let $U_1, \ldots, U_n$ be iid $\mathsf{Unif}(0,1)$, and define $X_i = g(U_i)$ for some specified function g such that $\int_0^1 g(u)^2\,du < \infty$; e.g., $g(u) = u^2$, $g(u) = u^{-1/2}$, etc. Set $\mu = \int_0^1 g(u)\,du$ and $\sigma^2 = \int_0^1 [g(u) - \mu]^2\,du$. Then the CLT says that $\sqrt{n}(\bar{X} - \mu)/\sigma$ converges in distribution to $\mathsf{N}(0,1)$ or, alternatively, $\bar{X}$ is approximately $\mathsf{N}(\mu, \sigma^2/n)$ provided n is large. This type of calculation is frequently used in conjunction with the Monte Carlo method described in Section 1.5.3.

In Stat 411 we will encounter a number of CLT-like results for statistics T_n used to estimate an unknown parameter θ. There will be occasions where we need to transform T_n by a function g and we'll be interested in the limiting distribution of $g(T_n)$. When a limiting distribution of T_n is available, there is s slick way to handle smooth transformations via the *Delta Theorem*; see Section 5.2.2 in HMC.

Theorem 1.2 (Delta Theorem)**.** *For a sequence $\{T_n\}$, suppose there are numbers θ and $v_\theta > 0$ such that $\sqrt{n}(T_n - \theta)$ converges in distribution to $\mathsf{N}(0, v_\theta)$. Let $g(t)$ be a function, differentiable at $t = \theta$, with $g'(\theta) \neq 0$. Then $\sqrt{n}[g(T_n) - g(\theta)]$ converges in distribution to $\mathsf{N}(0, [g'(\theta)]^2 v_\theta)$, as $n \to \infty$.*

Proof. Take a first-order Taylor expansion of $g(T_n)$ around θ:

$$g(T_n) = g(\theta) + g'(\theta)\,(T_n - \theta) + R_n,$$

where R_n is the remainder. I'll leave off the details, but it can be shown that $\sqrt{n}R_n$ converges in probability to 0. Rearranging terms above gives

$$\sqrt{n}[g(T_n) - g(\theta)] = g'(\theta) \cdot \sqrt{n}(T_n - \theta) + \sqrt{n}R_n.$$

Since $g'(\theta) \cdot \sqrt{n}(T_n - \theta) \to \mathsf{N}(0, [g'(\theta)]^2 v_\theta)$ in distribution, and $\sqrt{n}R_n \to 0$ in probability, the result follows from Slutsky's theorem (Theorem 5.2.5 in HMC). □

Example 1.6. Let $X_1, \ldots, X_n$ be an independent sample from $\mathsf{Exp}(\theta)$, the exponential distribution with PDF $f_\theta(x) = e^{-x/\theta}/\theta$, with $x > 0$ and $\theta > 0$. One can check that the mean and variance of this distribution are θ and θ^2, respectively. It follows from the CLT that $\sqrt{n}(\bar{X} - \theta) \to \mathsf{N}(0, \theta^2)$ in distribution. Now consider $\bar{X}^{1/2}$. In the Delta Theorem context, $T_n = \bar{X}$ and $g(t) = t^{1/2}$. By simple calculus, $g'(\theta) = 1/2\theta^{1/2}$. Therefore, by the Delta Theorem, $\sqrt{n}(\bar{X}^{1/2} - \theta^{1/2}) \to \mathsf{N}(0, \theta/4)$ in distribution.

1.5.3 Two numerical approximations

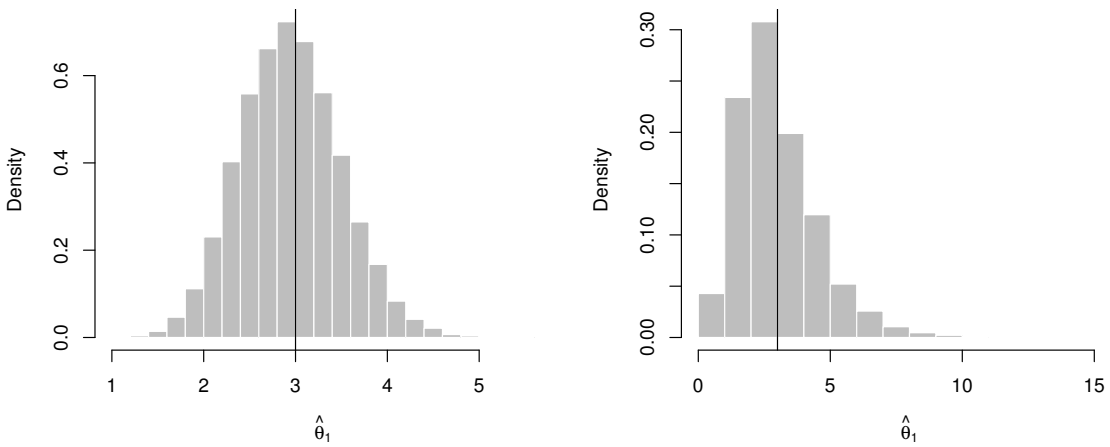
In cases where the exact sampling distribution of the statistic T_n is not known, and n it too small to consider asymptotic theory, there are numerical approximations that one can rely on. High-powered computing is now readily available, so these numerical methods are very popular in modern applied statistics.

Monte Carlo

(Chapter 4.8 in HMC provides some details about Monte Carlo; Examples 4.8.1 and 4.8.4 are good.) The basic principle of the Monte Carlo method is contained in the "illustration" in Section 1.5.1. That is, information about the sampling distribution of T_n can be obtained by actually performing the sample of $X_1^{(s)}, \ldots, X_n^{(s)}$ for lots of s values and looking at the histogram of the corresponding $T_n^{(s)}$.

Mathematically, the LLN/CLT guarantees that the method will work. For example, let g be some function and suppose we want to know $\mathsf{E}[g(T_n)]$. If the sampling distribution of T_n were available to us, then this would boil down to, say, a calculus (integration) problem. But if we don't know the PDF of T_n, then calculus is useless. According to the law of large numbers, if we take lots of samples of T_n, say $\{T_n^{(s)} : s = 1, \ldots, S\}$, then the expectation $\mathsf{E}[g(T_n)]$ can be approximated by the average $S^{-1} \sum_{s=1}^{S} g(T_n^{(s)})$, in the sense that the latter converges in probability to the former. The CLT makes this convergence even more precise.

Example 1.7. Let $X_1, \ldots, X_n$ be an independent sample from $\mathsf{Pois}(\theta)$, the Poisson distribution with mean θ. Suppose that we'd like to use data $X_1, \ldots, X_n$ to try to estimate an unknown θ. Since $\mathsf{E}(X_1) = \mathsf{V}(X_1) = \theta$, one could consider either $\hat{\theta}_1 = \bar{X}$, the sample



(a) Histogram of $\hat{\theta}_1$ values

(b) Histogram of $\hat{\theta}_2$ values

Figure 1.1: Monte Carlo results from Example 1.7.

mean, or $\hat{\theta}_2 = S^2$, the sample variance. Both are "reasonable" estimators in the sense that $\mathsf{E}(\hat{\theta}_1) = \mathsf{E}(\hat{\theta}_2) = \theta$ and both converge to θ in probability as $n \to \infty$. However, suppose that n is too small to consider asymptotics. Which of $\hat{\theta}_1$ and $\hat{\theta}_2$ is better for estimating θ? One way to compare is via the variance; that is, the one with smaller variance is the better of the two. But how do we calculate these variances? The Monte Carlo method is one approach.

Suppose $n = 10$ and the value of the unknown θ is 3. The Monte Carlo method will simulate S samples of size $n = 10$ from the $\mathsf{Pois}(\theta = 3)$ distribution. For each sample, the two estimates will be evaluated; resulting in the list of S values for each of the two estimates. Then the respective variances will be estimated by taking the sample variance for each of these two lists. R code for the simulation is given in Section 1.6.1. The table below shows the estimated variances of $\hat{\theta}_1$ and $\hat{\theta}_2$ for several values of S, larger S means more precise approximations.

S	1,000	5,000	10,000	50,000
$\mathsf{V}(\hat{\theta}_1)$	0.298	0.303	0.293	0.300
$\mathsf{V}(\hat{\theta}_2)$	2.477	2.306	2.361	2.268

Histograms approximating the sampling distributions of $\hat{\theta}_1$ and $\hat{\theta}_2$ are shown in Figure 1.1. There we see that $\hat{\theta}_1$ has a much more concentrated distribution around $\theta = 3$. Therefore, we would conclude that $\hat{\theta}_1$ is better for estimating θ than $\hat{\theta}_2$. Later we will see how this fact can be proved mathematically, without numerical approximations.

The problem of simulating the values of $X_1, \ldots, X_n$ to perform the Monte Carlo approximation is, itself, often a difficult one. The R software has convenient functions for sampling from many known distributions. For distributions that have a CDF for which an inverse can be written down, Theorem 4.8.1 in HMC and a $\mathsf{Unif}(0, 1)$ generator (e.g., `runif()` in

R) is all that's needed. In other cases, more sophisticated methods are needed, such as the accept-reject algorithm in Section 5.8.1 of HMC, or importance sampling. Simulation of random variables and the full power of the Monte Carlo method would be covered in a formal statistical computing course.

Bootstrap

Let $X_1, \ldots, X_n$ be a random sample from some distribution. Suppose we'd like to estimate $\mathsf{E}[g(X_1)]$, where g is some specified function. Clearly this quantity depends on the distribution in question. A natural estimate would be something like $T_n = n^{-1} \sum_{i=1}^{n} g(X_i)$, the sample mean of the $g(X_i)$'s. But if we don't know anything about the distribution which generated the $X_1, \ldots, X_n$, can we say anything about the sampling distribution of T_n? The immediate answer would seem to be "no" but the bootstrap method provides a clever little trick.

The bootstrap method is similar to the Monte Carlo method in that they are both based on sampling. However, in the Monte Carlo context, the distribution of $X_1, \ldots, X_n$ must be completely known, whereas, here, we know little or nothing about this distribution. So in order to do the sampling, we must somehow insert some proxy for that underlying distribution. The trick behind the bootstrap is to use the observed sample $\{X_1, \ldots, X_n\}$ as an approximation for this underlying distribution. Now the same Monte Carlo strategy can be applied, but instead of sampling $X_1^{(b)}, \ldots, X_n^{(b)}$ from the known distribution, we sample them (*with replacement*) from the observed sample $\{X_1, \ldots, X_n\}$. That is, e.g., the sampling distribution of T_n can be approximated by sampling

$$X_1^{(b)}, \ldots, X_n^{(b)} \overset{\text{iid}}{\sim} \mathsf{Unif}(\{X_1, \ldots, X_n\}), \quad b = 1, \ldots, B$$

and looking at the distribution of the bootstrap sample $T_n^{(b)} = T(X_1^{(b)}, \ldots, X_n^{(b)})$, for $b = 1, \ldots, B$. If B is sufficiently large, then the sampling distribution of T_n can be approximated by, say, the histogram of the bootstrap sample.

Example 1.8. (Data here is taken from Example 4.9.1 in HMC.) Suppose the following sample of size $n = 20$ is observed:

```
131.7  183.7  73.3  10.7  150.4  42.3  22.2   17.9  264.0  154.4
  4.3  256.6  61.9  10.8   48.8  22.5   8.8  150.6  103.0   85.9
```

To estimate the variance of the population from which this sample originated, we propose to use the sample variance S^2. Since nothing is known about the underlying population, we shall use the bootstrap method to to approximate the sampling distribution of S^2; R code is given in Section 1.6.2. Figure 1.2 shows a histogram of a bootstrap sample of $B = 5000$ S^2_{boot} values. The 5th and 95th percentiles of this bootstrap distribution are, respectively, 3296 and 9390, which determines a (90% bootstrap confidence) interval for the unknown variance of the underlying population. Other things, like $\mathsf{V}(S^2)$ can also be estimated.

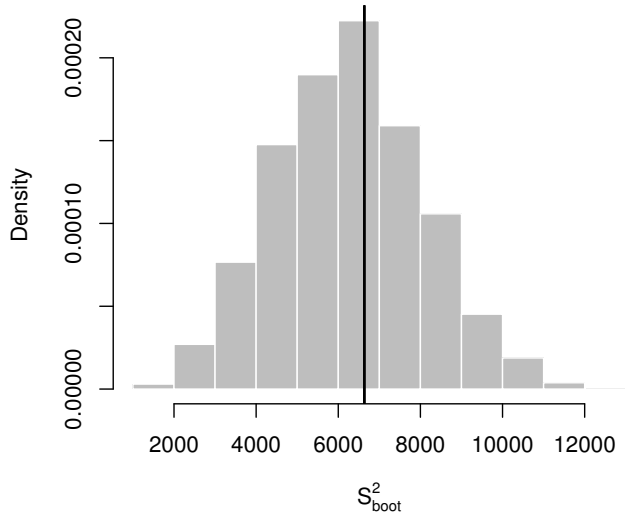


Figure 1.2: Bootstrap distribution of S^2 in Example 1.8.

The bootstrap method is not as ad hoc as it may seem on this first presentation. In fact, there is a considerable amount of elegant mathematical theory to show that the bootstrap method works[5] in a wide range of problems. There are actually lots of variations on the bootstrap, tailored for specific situations, and an upper-level course on computational statistics or resampling methods would delve more deeply into these methods. But it should be pointed out that there are well-documented classes of problems for which the bootstrap method fails, so one must use caution in applications.

1.6 Appendix

1.6.1 R code for Monte Carlo simulation in Example 1.7

```
poisson.var <- function(n, theta, S) {

  theta.1 <- theta.2 <- numeric(S)
  for(s in 1:S) {

    X <- rpois(n, theta)
    theta.1[s] <- mean(X)
    theta.2[s] <- var(X)

  }
  print(cbind(var.1=var(theta.1), var.2=var(theta.2)))
  return(cbind(theta1=theta.1, theta2=theta.2))
```

[5]Even the definition of what it means for the bootstrap to "work" is too technical to present here.

```
}
theta <- 3
o <- poisson.var(n=10, theta=theta, S=50000)
hist(o[,1], freq=FALSE, xlab=expression(hat(theta)[1]), col="gray", border="white")
abline(v=theta)
hist(o[,2], freq=FALSE, xlab=expression(hat(theta)[1]), col="gray", border="white")
abline(v=theta)
```

1.6.2 R code for bootstrap calculation in Example 1.8

```
boot.approx <- function(X, f, S) {

  n <- length(X)
  out <- numeric(S)
  for(s in 1:S) out[s] <- f(sample(X, n, replace=TRUE))
  return(out)

}
X <- c(131.7, 183.7, 73.3, 10.7, 150.4, 42.3, 22.2, 17.9, 264.0, 154.4, 4.3,
       256.6, 61.9, 10.8, 48.8, 22.5, 8.8, 150.6, 103.0, 85.9)
out <- boot.approx(X, var, 5000)
hist(out, freq=FALSE, xlab=expression(S[boot]^2), col="gray",
     border="white", main="")
abline(v=var(X), lwd=2)
print(quantile(out, c(0.05, 0.95)))
```

Chapter 2

Point Estimation Basics

2.1 Introduction

The statistical inference problem starts with the identification of a population of interest, about which something is unknown. For example, before introducing a law that homes be equipped with radon detectors, government officials should first ascertain whether radon levels in local homes are, indeed, too high. The most efficient (and, surprisingly, often the most accurate) way to gather this information is to take a sample of local homes and record the radon levels in each.[1] Now that the sample is obtained, how should this information be used to answer the question of interest? Suppose that officials are interested in the mean radon level for all homes in their community—this quantity is unknown, otherwise, there'd be no reason to take the sample in the first place. After some careful exploratory data analysis, the statistician working the project determines a statistical model, i.e., the functional form of the PDF that characterizes radon levels in homes in the community. Now the statistician has a model, which depends on the unknown mean radon level (and possibly other unknown population characteristics), and a sample from that distribution. His/her charge is to use these two pieces of information to make inference about the unknown mean radon level. The simplest of such inferences is simply to *estimate* this mean. In this section we will discuss some of the basic principles of statistical estimation. This will be an important theme throughout the course.

2.2 Notation and terminology

The starting point is a statement of the model. Let $X_1, \ldots, X_n$ be a sample from a distribution with CDF F_θ, depending on a parameter θ which is unknown. In some cases, it will

[1] In this course, we will take the sample as given, that is, we will not consider the question of *how* the sample is obtained. In general it is not an easy task to obtain a bona fide completely random sample; carefully planning of experimental/survey designs is necessary.

be important to also know the parameter space Θ, the set of possible values of θ.[2] Point estimation is the problem of find a function of the data that provides a "good" estimate of the unknown parameter θ.

Definition 2.1. Let $X_1, \ldots, X_n$ be a sample from a distribution F_θ with $\theta \in \Theta$. A point estimate of θ is a function $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ taking values in Θ.

This (point) estimator $\hat{\theta}$ (read as "theta-hat") is a special case of a statistic discussed previously. What distinguishes an estimator from a general statistic is that it is required to take values in the parameter space Θ, so that it makes sense to compare θ and $\hat{\theta}$. But besides this, $\hat{\theta}$ can be anything, although some choices are better than others.

Example 2.1. Suppose that $X_1, \ldots, X_n$ are iid $\mathsf{N}(\mu, \sigma^2)$. Then the following are all point estimates of the mean μ:

$$\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \hat{\mu}_2 = M_n \text{ (sample median)}, \quad \hat{\mu}_3 = \frac{X_{(1)} + X_{(n)}}{2}.$$

The sampling distribution of $\hat{\mu}_1$ is known (what is it?), but not for the others. However, some asymptotic theory is available that may be helpful for comparing these as estimators of μ; more on this later.

Exercise 2.1. Modify the R code in Section 1.6.1 to get Monte Carlo approximations of the sampling distributions of $\hat{\mu}_1$, $\hat{\mu}_2$, and $\hat{\mu}_3$ in Example 2.1. Start with $n = 10$, $\mu = 0$, and $\sigma = 1$, and draw histograms to compare. What happens if you change n, μ, or σ?

Example 2.2. Let θ denote the proportion of individuals in a population that favor a particular piece of legislation. To estimate θ, sample $X_1, \ldots, X_n$ iid $\mathsf{Ber}(\theta)$; that is, $X_i = 1$ if sampled individual i favors the legislation, and $X_i = 0$ otherwise. Then an estimate of θ is the sample mean, $\hat{\theta} = n^{-1}\sum_{i=1}^{n} X_i$. Since the summation $\sum_{i=1}^{n} X_i$ has a known sampling distribution (what is it?), many properties of $\hat{\theta}$ can be derived without too much trouble.

We will focus mostly on problems where there is only one unknown parameter. However, there are also important problems where θ is actually a vector and Θ is a subset of $\mathbb{R}^d$, $d > 1$. One of the most important examples is the normal model where both the mean and variance are unknown. In this case $\theta = (\mu, \sigma^2)$ and $\Theta = \{(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+\} \subset \mathbb{R}^2$.

The properties of estimators $\hat{\theta}$ will depend on its sampling distribution. Here I need to elaborate a bit on notation. Since the distribution of $X_1, \ldots, X_n$ depends on θ, so does the sampling distribution of $\hat{\theta}$. So when we calculate probabilities, expected values, etc it is often important to make clear under what parameter value these are being take. Therefore, we will highlight this dependence by adding a subscript to the familiar probability and expected value operators P and E. That is, P_θ and E_θ will mean probability and expected value with respect to the joint distribution of $(X_1, \ldots, X_n)$ under F_θ.

[2] HMC uses "Ω" (Omega) for the parameter space, instead of Θ; however, I find it more convenient to use the same Greek letter with the lower- and upper-case to distinguish the meaning.

In general (see Example 2.1) there can be more than one "reasonable" estimator of an unknown parameter. One of the goals of mathematical statistics is to provide a theoretical framework by which an "optimal" estimator can be identified in a given problem. But before we can say anything about which estimator is best, we need to know something about the important properties estimator should have.

2.3 Properties of estimators

Properties of estimators are all consequences of their sampling distributions. Most of the time, the full sampling distribution of $\hat{\theta}$ is not available; therefore, we focus on properties that do not require complete knowledge of the sampling distribution.

2.3.1 Unbiasedness

The first, and probably the simplest, property is called *unbiasedness*. In words, an estimator $\hat{\theta}$ is unbiased if, when applied to many different samples from F_θ, $\hat{\theta}$ equals the true parameter θ, *on average*. Equivalently, unbiasedness means the sampling distribution of $\hat{\theta}$ is, in some sense, centered around θ.

Definition 2.2. The *bias* of an estimator is $b_\theta(\hat{\theta}) = \mathsf{E}_\theta(\hat{\theta}) - \theta$. Then $\hat{\theta}$ is an *unbiased* estimator of θ if $b_\theta(\hat{\theta}) = 0$ for all θ.

That is, no matter the actual value of θ, if we apply $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ to many data sets $X_1, \ldots, X_n$ sampled from F_θ, then the average of these $\hat{\theta}$ values will equal θ—in other words, $\mathsf{E}_\theta(\hat{\theta}) = \theta$ for all θ. This is clearly not an unreasonable property, and a lot of work in mathematical statistics has focused on unbiased estimation.

Example 2.3. Let $X_1, \ldots, X_n$ be iid from some distribution having mean μ and variance σ^2. This distribution could be normal, but it need not be. Consider $\hat{\mu} = \bar{X}$, the sample mean, and $\hat{\sigma}^2 = S^2$, the sample variance. Then $\hat{\mu}$ and $\hat{\sigma}^2$ are unbiased estimators of μ and σ^2, respectively. The proof for $\hat{\mu}$ is straightforward—try it yourself! For $\hat{\sigma}^2$, recall the following decomposition of the sample variance:

$$\hat{\sigma}^2 = S^2 = \frac{1}{n-1}\Big\{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\Big\}.$$

Drop the subscript (μ, σ^2) on $\mathsf{E}_{\mu,\sigma^2}$ for simplicity. Recall the following two general facts:

$$\mathsf{E}(X^2) = \mathsf{V}(X) + \mathsf{E}(X)^2 \quad \text{and} \quad \mathsf{V}(\bar{X}) = n^{-1}\mathsf{V}(X_1).$$

Then using linearity of expectation,

$$\begin{aligned}
\mathsf{E}(\hat{\sigma}^2) &= \frac{1}{n-1}\Big\{\mathsf{E}\Big(\sum_{i=1}^{n} X_i^2\Big) - n\mathsf{E}(\bar{X}^2)\Big\} \\
&= \frac{1}{n-1}\Big\{n\big(\mathsf{V}(X_1) + \mathsf{E}(X_1)^2\big) - n\mathsf{V}(\bar{X}) - n\mathsf{E}(\bar{X})\Big\} \\
&= \frac{1}{n-1}\big\{n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2\big\} \\
&= \sigma^2.
\end{aligned}$$

Therefore, the sample variance is an unbiased estimator of the population variance, regardless of the model.

While unbiasedness is a nice property for an estimator to have, it doesn't carry too much weight. Specifically, an estimator can be unbiased but otherwise very poor. For an extreme example, suppose that $\mathsf{P}_\theta\{\hat{\theta} = \theta + 10^5\} = 1/2 = \mathsf{P}_\theta\{\hat{\theta} = \theta - 10^5\}$. In this case, $\mathsf{E}_\theta(\hat{\theta}) = \theta$, but $\hat{\theta}$ is always very far away from θ. There is also a well-known phenomenon (bias–variance trade-off) which says that often allowing the bias to be non-zero will improve on estimation accuracy; more on this below. The following example highlights some of the problems of focusing primarily on the unbiasedness property.

Example 2.4. (See Remark 7.6.1 in HMC.) Let X be a sample from a $\mathsf{Pois}(\theta)$ distribution. Suppose the goal is to estimate $\eta = e^{-2\theta}$, not θ itself. We know that $\hat{\theta} = X$ is an unbiased estimator of θ. However, the natural estimator e^{-2X} is *not* an unbiased estimator of $e^{-2\theta}$. Consider instead $\hat{\eta} = (-1)^X$. This estimator is unbiased:

$$\mathsf{E}_\theta[(-1)^X] = \sum_{x=0}^{\infty} e^{-\theta}\frac{(-1)^x\theta^x}{x!} = e^{-\theta}\sum_{x=0}^{\infty}\frac{(-\theta)^x}{x!} = e^{-\theta}e^{-\theta} = e^{-2\theta}.$$

In fact, it can even be shown that $(-1)^X$ is the "best" of all unbiased estimators; cf. the Lehmann–Scheffe theorem. But even though it's unbiased, it can only take values ± 1 so, depending on θ, $(-1)^X$ may never be close to $e^{-2\theta}$.

Exercise 2.2. Prove the claim in the previous example that e^{-2X} is *not* an unbiased estimator of $e^{-2\theta}$. (Hint: use the Poisson moment-generating function, p. 152 in HMC.)

In general, for a given function g, if $\hat{\theta}$ is an unbiased estimator of θ, then $g(\hat{\theta})$ is not an unbiased estimator of $g(\theta)$. But there is a nice method by which an unbiased estimator of $g(\theta)$ can often be constructed; see *method of moments* in Section 2.4. It is also possible that certain (functions of) parameters may not be unbiasedly estimable.

Example 2.5. Let $X_1, \ldots, X_n$ be iid $\mathsf{Ber}(\theta)$ and suppose we want to estimate $\eta = \theta/(1-\theta)$, the so-called odds ratio. Suppose $\hat{\eta}$ is an unbiased estimator of η, so that $\mathsf{E}_\theta(\hat{\eta}) = \eta = \theta/(1-\theta)$ for all θ or, equivalently,

$$(1-\theta)\mathsf{E}_\theta(\hat{\eta}) - \theta = 0 \quad \text{for all } \theta.$$

Here the joint PMF of $(X_1, \ldots, X_n)$ is $f_\theta(x_1, \ldots, x_n) = \theta^{x_1+\cdots+x_n}(1-\theta)^{n-(x_1+\cdots+x_n)}$. Writing out $\mathsf{E}_\theta(\hat{\eta})$ as a weighted average with weights given by $f_\theta(x_1, \ldots, x_n)$, we get

$$(1-\theta) \sum_{\text{all } (x_1,\ldots,x_n)} \hat{\eta}(x_1, \ldots, x_n)\theta^{x_1+\cdots+x_n}(1-\theta)^{n-(x_1+\cdots+x_n)} - \theta = 0 \quad \text{for all } \theta.$$

The quantity on the left-hand side is a polynomial in θ of degree $n+1$. From the Fundamental Theorem of Algebra, there can be at most $n+1$ real roots of the above equation. However, unbiasedness requires that there be infinitely many roots. This contradicts the fundamental theorem, so we must conclude that there are no unbiased estimators of η.

2.3.2 Consistency

Another reasonable property is that the estimator $\hat{\theta} = \hat{\theta}_n$, which depends on the sample size n through the dependence on $X_1, \ldots, X_n$, should get close to the true θ as n gets larger and larger. To make this precise, recall the following definition (see Definition 5.1.1 in HMC).[3]

Definition 2.3. Let T and $\{T_n : n \geq 1\}$ be random variables in a common sample space. Then T_n converges to T in probability if, for any $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathsf{P}\{|T_n - T| > \varepsilon\} = 0.$$

The *law of large numbers* (LLN, Theorem 5.1.1 in HMC) is an important result on convergence in probability.

Theorem 2.1 (Law of Large Numbers, LLN)**.** *If $X_1, \ldots, X_n$ are iid with mean μ and finite variance σ^2, then $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$ converges in probability to μ.*[4]

The LLN is a powerful result and will be used throughout the course. Two useful tools for proving convergence in probability are the inequalities of Markov and Chebyshev. (These are presented in HMC, Theorems 1.10.2–1.10.3, but with different notation.)

- *Markov's inequality.* Let X be a positive random variable, i.e., $\mathsf{P}(X > 0) = 1$. Then, for any $\varepsilon > 0$, $\mathsf{P}(X > \varepsilon) \leq \varepsilon^{-1}\mathsf{E}(X)$.
- *Chebyshev's inequality.* Let X be a random variable with mean μ and variance σ^2. Then, for any $\varepsilon > 0$, $\mathsf{P}\{|X - \mu| > \varepsilon\} \leq \varepsilon^{-2}\sigma^2$.

It is through convergence in probability that we can say that an estimator $\hat{\theta} = \hat{\theta}_n$ gets close to the estimand θ as n gets large.

[3] To make things simple, here we shall focus on the real case with distance measured by absolute difference. When $\hat{\theta}$ is vector-valued, we'll need to replace the absolute difference by a normed difference. More generally, the definition of convergence in probability can handle sequences of random elements in any space equipped with a metric.

[4] We will not need this in Stat 411, but note that the assumption of finite variance can be removed and, simultaneously, the mode of convergence can be strengthened.

Definition 2.4. An estimator $\hat{\theta}_n$ of θ is consistent if $\hat{\theta}_n \to \theta$ in probability.

A rough way to understand consistency of an estimator $\hat{\theta}_n$ of θ is that the sampling distribution of $\hat{\theta}_n$ gets more and more concentrated as $n \to \infty$. The following example demonstrates both a theoretical verification of consistency and a visual confirmation via Monte Carlo.

Example 2.6. Recall the setup of Example 2.3. It follows immediately from the LLN that $\hat{\mu}_n = \bar{X}$ is a consistent estimator of the mean μ. Moreover, the sample variance $\hat{\sigma}_n^2 = S^2$ is also a consistent estimator of the variance σ^2. To see this, recall that

$$\hat{\sigma}_n^2 = \frac{n}{n-1}\Big\{\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2\Big\}.$$

The factor $n/(n-1)$ converges to 1; the first term in the braces convergence in probability to $\sigma^2 - \mu^2$ by the LLN applied to the X_i^2's; the second term in the braces converges in probability to μ^2 by the LLN and Theorem 5.1.4 in HMC (see, also, the Continuous Mapping Theorem below). Putting everything together, we find that $\hat{\sigma}_n^2 \to \sigma^2$ in probability, making it a consistent estimator. To see this property visually, suppose that the sample originates from a Poisson distribution with mean $\theta = 3$. We can modify the R code in Example 7 in Notes 01 to simulate the sampling distribution of $\hat{\theta}_n = \hat{\sigma}_n^2$ for any n. The results for $n \in \{10, 25, 50, 100\}$ are summarized in Figure 2.1. Notice that as n increases, the sampling distributions become more concentrated around $\theta = 3$.

Unbiased estimators generally are not invariant under transformations [i.e., in general, if $\hat{\theta}$ is unbiased for θ, then $g(\hat{\theta})$ is not unbiased for $g(\theta)$], but consistent estimators do have such a property, a consequence of the so-called *Continuous Mapping Theorem* (basically Theorem 5.1.4 in HMC).

Theorem 2.2 (Continuous Mapping Theorem)**.** *Let g be a continuous function on Θ. If $\hat{\theta}_n$ is consistent for θ, then $g(\hat{\theta}_n)$ is consistent for $g(\theta)$.*

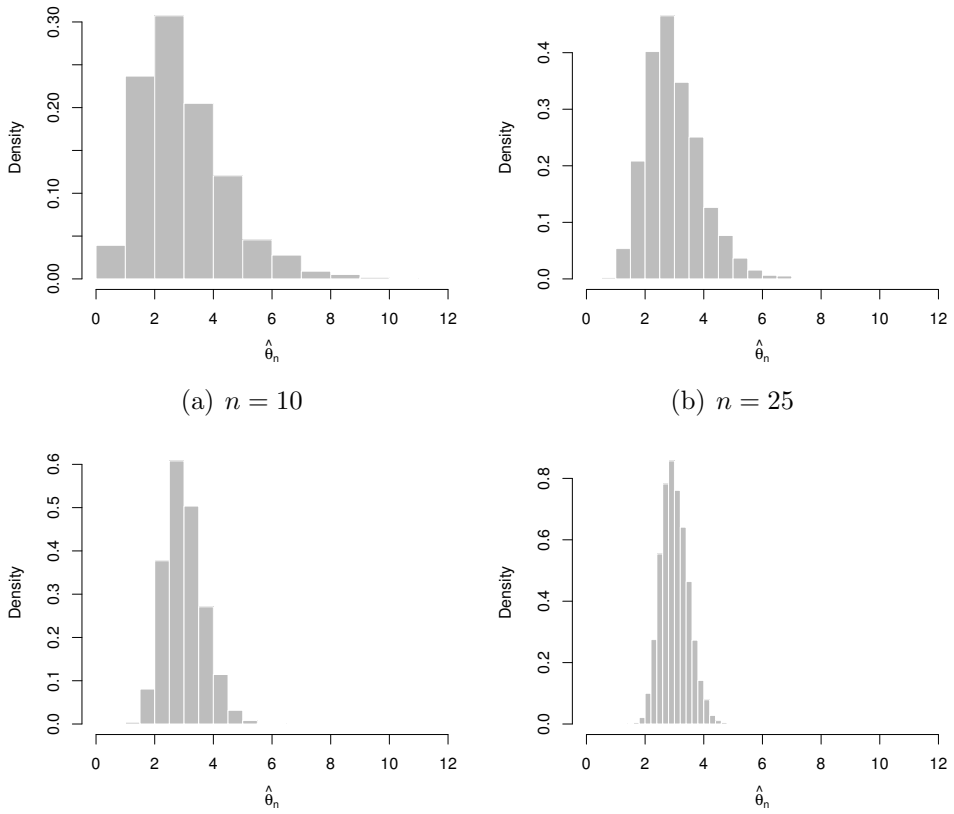
Proof. Fix a particular θ value. Since g is a continuous function on Θ, it's continuous at this particular θ. For any $\varepsilon > 0$, there exists a $\delta > 0$ (depending on ε and θ) such that

$$|g(\hat{\theta}_n) - g(\theta)| > \varepsilon \quad \text{implies} \quad |\hat{\theta}_n - \theta| > \delta.$$

Then the probability of the event on the left is no more than the probability of the event on the right, and this latter probability vanishes as $n \to \infty$ by assumption. Therefore

$$\lim_{n\to\infty} \mathsf{P}_\theta\{|g(\hat{\theta}_n) - g(\theta)| > \varepsilon\} = 0.$$

Since ε was arbitrary, the proof is complete. □



(a) $n = 10$ (b) $n = 25$

(c) $n = 50$ (d) $n = 100$

Figure 2.1: Plots of the sampling distribution of $\hat{\theta}_n$, the sample variance, for several values of n in the $\mathsf{Pois}(\theta)$ problem with $\theta = 3$.

Example 2.7. Let $X_1, \ldots, X_n$ be iid $\mathsf{Pois}(\theta)$. Since θ is both the mean and the variance for the Poisson distribution, it follows that both $\hat{\theta}_n = \bar{X}$ and $\tilde{\theta}_n = S^2$ are unbiased and consistent for θ by the results in Examples 2.3 and 2.6. Another comparison of these two estimators is given in Example 2.10. Here consider a new estimator $\dot{\theta}_n = (\bar{X}S^2)^{1/2}$. Define the function $g(x_1, x_2) = (x_1 x_2)^{1/2}$. Clearly g is continuous (why?). Since the pair $(\hat{\theta}_n, \tilde{\theta}_n)$ is a consistent estimator of (θ, θ), it follows from the continuous mapping theorem that $\dot{\theta}_n = g(\hat{\theta}_n, \tilde{\theta}_n)$ is a consistent estimator of $\theta = g(\theta, \theta)$.

Like with unbiasedness, consistency is a nice property for an estimator to have. But consistency alone is not enough to make an estimator a good one. Next is an exaggerated example that makes this point clear.

Example 2.8. Let $X_1, \ldots, X_n$ be iid $\mathsf{N}(\theta, 1)$. Consider the estimator

$$\hat{\theta}_n = \begin{cases} 10^7 & \text{if } n < 10^{750}, \\ \bar{X}_n & \text{otherwise.} \end{cases}$$

Let $N = 10^{750}$. Although N is very large, its ultimately finite and can have no effect on the limit. To see this, fix $\varepsilon > 0$ and define

$$a_n = \mathsf{P}_\theta\{|\hat{\theta}_n - \theta| > \varepsilon\} \quad \text{and} \quad b_n = \mathsf{P}_\theta\{|\bar{X}_n - \theta| > \varepsilon\}.$$

Since $b_n \to 0$ by the LLN, and $a_n = b_n$ for all $n \geq N$, it follows that $a_n \to 0$ and, hence, $\hat{\theta}_n$ is consistent. However, for any reasonable application, where the sample size is finite, estimating θ by a constant 10^7 is an absurd choice.

2.3.3 Mean-square error

Measuring closeness of an estimator $\hat{\theta}$ to its estimand θ via consistency assumes that the sample size n is very large, actually infinite. As a consequence, many estimators which are "bad" for any finite n (like that in Example 2.8) can be labelled as "good" according to the consistency criterion. An alternative measure of closeness is called the *mean-square error* (MSE), and is defined as

$$\mathsf{MSE}_\theta(\hat{\theta}) = \mathsf{E}_\theta\{(\hat{\theta} - \theta)^2\}. \tag{2.1}$$

This measures the average (squared) distance between $\hat{\theta}(X_1, \ldots, X_n)$ and θ as the data $X_1, \ldots, X_n$ varies according to F_θ. So if $\hat{\theta}$ and $\tilde{\theta}$ are two estimators of θ, we say that $\hat{\theta}$ is better than $\tilde{\theta}$ (in the mean-square error sense) if $\mathsf{MSE}_\theta(\hat{\theta}) < \mathsf{MSE}_\theta(\tilde{\theta})$.

Next are some properties of the MSE. The first relates MSE to the variance and bias of an estimator.

Proposition 2.1. $\mathsf{MSE}_\theta(\hat{\theta}) = \mathsf{V}_\theta(\hat{\theta}) + b_\theta(\hat{\theta})^2$. *Consequently, if $\hat{\theta}$ is an unbiased estimator of θ, then* $\mathsf{MSE}_\theta(\hat{\theta}) = \mathsf{V}_\theta(\hat{\theta})$.

Proof. Let $\bar{\theta} = \mathsf{E}_\theta(\hat{\theta})$. Then

$$\mathsf{MSE}_\theta(\hat{\theta}) = \mathsf{E}_\theta\{(\hat{\theta} - \theta)^2\} = \mathsf{E}_\theta\{[(\hat{\theta} - \bar{\theta}) + (\bar{\theta} - \theta)]^2\}.$$

Expanding the quadratic inside the expectation gives

$$\mathsf{MSE}_\theta(\hat{\theta}) = \mathsf{E}_\theta\{(\hat{\theta} - \bar{\theta})^2\} + 2(\bar{\theta} - \theta)\mathsf{E}_\theta\{(\hat{\theta} - \bar{\theta})\} + (\bar{\theta} - \theta)^2.$$

The first term is the variance of $\hat{\theta}$; the second term is zero by definition of $\bar{\theta}$; and the third terms is the squared bias. □

Often the goal is to find estimators with small MSEs. From Proposition 2.1, this can be achieved by picking $\hat{\theta}$ to have small variance and small squared bias. But it turns out that, in general, making bias small increases the variance, and vice versa. This it what is called the *bias–variance trade-off*. In some cases, if minimizing MSE is the goal, it can be better to allow a little bit of bias if it means a drastic decrease in the variance. In fact, many common estimators are biased, at least partly because of this trade-off.

Example 2.9. Let $X_1, \ldots, X_n$ be iid $\mathsf{N}(\mu, \sigma^2)$ and suppose the goal is to estimate σ^2. Define the statistic $T = \sum_{i=1}^n (X_i - \bar{X})^2$. Consider a class of estimators $\hat{\sigma}^2 = aT$ where a is a positive number. Reasonable choices of a include $a = (n-1)^{-1}$ and $a = n^{-1}$. Let's find the value of a that minimizes the MSE.

First observe that $(1/\sigma^2)T$ is a chi-square random variable with degrees of freedom $n-1$; see Theorem 3.6.1 in the text. It can then be shown, using Theorem 3.3.1 of the text, that $\mathsf{E}_{\sigma^2}(T) = (n-1)\sigma^2$ and $\mathsf{V}_{\sigma^2}(T) = 2(n-1)\sigma^4$. Write $R(a)$ for $\mathsf{MSE}_{\sigma^2}(aT)$. Using Proposition 2.1 we get

$$\begin{aligned} R(a) &= \mathsf{E}_{\sigma^2}(aT - \sigma^2)^2 = \mathsf{V}_{\sigma^2}(aT) + b_{\sigma^2}(aT)^2 \\ &= 2a^2(n-1)\sigma^4 + [a(n-1)\sigma^2 - \sigma^2]^2 \\ &= \sigma^4\big\{2a^2(n-1) + [a(n-1)-1]^2\big\}. \end{aligned}$$

To minimize $R(a)$, set the derivative equal to zero, and solve for a. That is,

$$0 \stackrel{\text{set}}{=} R'(a) = \sigma^4\big\{4(n-1)a + 2(n-1)^2 a - 2(n-1)\big\}.$$

From here it's easy to see that $a = (n+1)^{-1}$ is the only solution (and this must be a minimum since $R(a)$ is a quadratic). Therefore, among estimators of the form $\hat{\sigma}^2 = a\sum_{i=1}^n (X_i - \bar{X})^2$, the one with smallest MSE is $\hat{\sigma}^2 = (n+1)^{-1}\sum_{i=1}^n (X_i - \bar{X})^2$. Note that this estimator is not unbiased since $a \neq (n-1)^{-1}$. To put this another way, the classical estimator S^2 pays a price (larger MSE) for being unbiased.

Proposition 2.2 below helps to justify the approach of choosing $\hat{\theta}$ to make the MSE small. Indeed, if the choice is made so that the MSE vanishes as $n \to \infty$, then the estimator turns out to be consistent.

Proposition 2.2. *If $\mathsf{MSE}_\theta(\hat{\theta}_n) \to 0$ as $n \to \infty$, then $\hat{\theta}_n$ is a consistent estimator of θ.*

Proof. Fix $\varepsilon > 0$ and note that $\mathsf{P}_\theta\{|\hat{\theta}_n - \theta| > \varepsilon\} = \mathsf{P}_\theta\{(\hat{\theta}_n - \theta)^2 > \varepsilon^2\}$. Applying Markov's inequality to the latter term gives an upper bound of $\varepsilon^{-2}\mathsf{MSE}_\theta(\hat{\theta}_n)$. Since this goes to zero by assumption, $\hat{\theta}_n$ is consistent. □

The next example compares two unbiased and consistent estimators based on their respective MSEs. The conclusion actually gives a preview of some of the important results to be discussed later in Stat 411.

Example 2.10. Suppose $X_1, \ldots, X_n$ are iid $\mathsf{Pois}(\theta)$. We've looked at two estimators of θ in the context, namely, $\hat{\theta}_1 = \bar{X}$ and $\hat{\theta}_2 = S^2$. Both of these are unbiased and consistent. To decide which we like better, suppose we prefer the one with the smallest variance.[5] The variance of $\hat{\theta}_1$ is an easy calculation: $\mathsf{V}_\theta(\hat{\theta}) = \mathsf{V}_\theta(\bar{X}) = \mathsf{V}_\theta(X_1)/n = \theta/n$. But the variance of $\hat{\theta}_2$ is trickier, so we'll resort to an approximation, which relies on the following general fact.[6]

[5] In Chapter 1, we looked at this same problem in an example on the Monte Carlo method.

[6] A. DasGupta, *Asymptotic Theory of Statistics and Probability*, Springer, 2008, Theorem 3.8.

Let $X_1, \ldots, X_n$ be iid with mean μ. Define the sequence of population and sample central moments:

$$\mu_k = \mathsf{E}(X_1 - \mu)^k \quad \textit{and} \quad M_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^k, \quad k \geq 1.$$

Then, for large n, the following approximations hold:

$$\mathsf{E}(M_k) \approx \mu_k$$
$$\mathsf{V}(M_k) \approx n^{-1}\{\mu_{2k} - \mu_k^2 - 2k\mu_{k-1}\mu_{k+1} + k^2\mu_2\mu_{k-1}^2\}.$$

In the Poisson case, $\mu_1 = 0$, $\mu_2 = \theta$, $\mu_3 = \theta$, and $\mu_4 = \theta + 3\theta^2$; these can be verified directly by using the Poisson moment-generating function. Plugging these values into the above approximation ($k = 2$), gives $\mathsf{V}_\theta(\hat{\theta}_2) \approx (\theta + 2\theta^2)/n$. This is more than $\mathsf{V}_\theta(\hat{\theta}_1) = \theta/n$ so we conclude that $\hat{\theta}_1$ is better than $\hat{\theta}_2$ (in the mean-square error sense). In fact, it can be shown (via the Lehmann–Scheffe theorem in Chapter 7 in HMC) that, among all unbiased estimators, $\hat{\theta}_1$ is the best in the mean-square error sense.

Exercise 2.3. (a) Verify the expressions for μ_1, μ_2, μ_3, and μ_4 in Example 2.10. (b) Look back to Example 7 in Notes 01 and compare the Monte Carlo approximation of $\mathsf{V}(\hat{\theta}_2)$ to the large-n approximation $\mathsf{V}(\hat{\theta}_2) \approx (\theta + 2\theta^2)/n$ used above. Recall that, in the Monte Carlo study, $\theta = 3$ and $n = 10$. Do you think $n = 10$ is large enough to safely use a large-n approximation?

It turns out that, in a typical problem, there is no estimator which can minimize the MSE uniformly over all θ. If there was such a estimator, then this would clearly be the best. To see that such an ideal cannot be achieved, consider the silly estimator $\hat{\theta} \equiv 7$. Clearly $\mathsf{MSE}_7(\hat{\theta}) = 0$ and no other estimator can beat that; of course, there's nothing special about 7. However, if we restrict ourselves to the class of estimators which are unbiased, then there is a lower bound on the variance of such estimators and theory is available for finding estimators that achieve this lower bound.

2.4 Where do estimators come from?

In the previous sections we've simply discussed properties of estimators—nothing so far has been said about the origin of these estimators. In some cases, a reasonable choice is obvious, like estimating a population mean by a sample mean. But there are situations where this choice is not so obvious. There are some general methods for constructing estimators. Here I simply list the various methods with a few comments.

- Perhaps the simplest method of constructing estimators is the *method of moments*. This approach is driven by the unbiasedness property. The idea is to start with some statistic T and calculate its expectation $h(\theta) = \mathsf{E}_\theta(T)$; now set $T = h(\theta)$ and use the

solution $\hat{\theta}$ as an estimator. For example, if $X_1, \ldots, X_n$ are iid $\mathsf{N}(\theta, 1)$ and the goal is to estimate θ^2, a reasonable starting point is $T = \bar{X}^2$. Since $\mathsf{E}_\theta(T) = \theta^2 + 1/n$, an unbiased estimator of θ^2 is $\bar{X}^2 - 1/n$.

- Perhaps the most common way to construct estimator is via the *method of maximum likelihood*. We will spend a considerable amount of time discussing this approach. There are other related approaches, such as *M-estimation* and *least-squares estimation*,[7] which we will not discuss here.

- As alluded to above, one cannot, for example, find an estimator $\hat{\theta}$ that minimizes the MSE uniformly over all θ. But but restricting the class of estimators to those which are unbiased, a uniformly best estimator often exists. Such an estimator is called the *uniformly minimum variance unbiased estimates* (UMVUE) and we will spend a lot of time talking about this approach.

- *Minimax estimation* takes a measure of closeness of an estimator $\hat{\theta}$ to θ, such as $\mathsf{MSE}_\theta(\hat{\theta})$, but rather than trying to minimize the MSE pointwise over all θ, as in the previous point, one first maximizes over θ to give a pessimistic "worst case" measure of the performance of $\hat{\theta}$. Then one tries to find the $\hat{\theta}$ that MINImizes the MAXimum MSE. This approach is interesting, and relates to game theory and economics, but is somewhat out of style in the statistics community.

- *Bayes estimation* is an altogether different approach. We will discuss the basics of Bayesian inference, including estimation, in Chapter 6.

[7] Students may have heard of the least-square approach in other courses, such as applied statistics courses or linear algebra/numerical analysis.

Chapter 3

Likelihood and Maximum Likelihood Estimation

3.1 Introduction

Previously we have discussed various properties of estimator—unbiasedness, consistency, etc—but with very little mention of where such an estimator comes from. In this part, we shall investigate one particularly important process by which an estimator can be constructed, namely, *maximum likelihood*. This is a method which, by and large, can be applied in any problem, provided that one knows and can write down the joint PMF/PDF of the data. These ideas will surely appear in any upper-level statistics course.

Observable data $X_1, \ldots, X_n$ has a specified model, say, a collection of distribution functions $\{f_\theta : \theta \in \Theta\}$ indexed by the parameter space Θ. Data is observed, but we don't know which of the models F_θ it came from. We shall assume that the model is correct, i.e., that there is a θ value such that $X_1, \ldots, X_n$ are iid f_θ.[1] The goal, then, is to identify the "best" model—the one that explain the data the best. This amounts to identifying the true but unknown θ value. Hence, our goal is to estimate the unknown θ.

In the sections that follow, I shall describe this so-called likelihood function and how it is used to construct point estimators. The rest of the chapter will develop general properties of these estimators; these are important classical results in statistical theory. Focus is primarily on the single parameter case; Section 3.7 extends the ideas to the multi-parameter case.

3.2 Likelihood

Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f_\theta$, where θ is unknown. For the time being, we assume that θ resides in a subset Θ of $\mathbb{R}$. By the assumed independence, the joint distribution of $(X_1, \ldots, X_n)$ is

[1] This is a *huge* assumption. It can be relaxed, but then the details get much more complicated—there's some notion of geometry on the collection of probability distributions, and we can think about projections onto the model. We won't bother with this here.

characterized by

$$f_\theta(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_\theta(x_i),$$

i.e., "independence means multiply." From a probability point of view, we understand the above expression to be a function of $(x_1, \ldots, x_n)$ for fixed θ. In the statistics context, we flip this around. That is, we will fix $(x_1, \ldots, x_n)$ at the observed $(X_1, \ldots, X_n)$, and imagine the above expression as a function of θ only.

Definition 3.1. If $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f_\theta$, then the *likelihood function* is

$$L(\theta) = f_\theta(X_1, \ldots, X_n) = \prod_{i=1}^{n} f_\theta(X_i), \tag{3.1}$$

treated as a function of θ. In what follows, I may occasionally add subscripts, i.e., $L_X(\theta)$ or $L_n(\theta)$, to indicate the dependence of the likelihood on data $X = (X_1, \ldots, X_n)$ or on sample size n. Also write

$$\ell(\theta) = \log L(\theta), \tag{3.2}$$

for the log-likelihood; the same subscript rules apply to $\ell(\theta)$.

So clearly $L(\theta)$ and $\ell(\theta)$ depend on data $X = (X_1, \ldots, X_n)$, but they're treated as functions of θ only. How can we interpret this function? The first thing to mention is a warning—*the likelihood function is NOT a PMF/PDF for θ!* So it doesn't make sense to integrate over θ values like you would a PDF.[2] We're mostly interested in the shape of the likelihood curve or, equivalently, the relative comparisons of the $L(\theta)$ for different θ's. This is made more precise below:

> If $L(\theta_1) > L(\theta_2)$ (equivalently, if $\ell(\theta_1) > \ell(\theta_2)$), then θ_1 is more likely to have been responsible for producing the observed $X_1, \ldots, X_n$. In other words, f_{θ_1} is a better model than f_{θ_2} in terms of how well it fits the observed data.

So, we can understand likelihood (and log-likelihood) of providing a sort of *ranking* of the θ values in terms of how well they match with the observations.

Exercise 3.1. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Ber}(\theta)$, with $\theta \in (0, 1)$. Write down an expression for the likelihood $L(\theta)$ and log-likelihood $\ell(\theta)$. On what function of $(X_1, \ldots, X_n)$ does $\ell(\theta)$ depend. Suppose that $n = 7$ and T equals 3, where T is that function of $(X_1, \ldots, X_n)$ previously identified; sketch a graph of $\ell(\theta)$.

Exercise 3.2. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(\theta, 1)$. Find an expression for the log-likelihood $\ell(\theta)$.

[2]There are some exceptions to this point that we won't discuss here; but see Chapter 6.

3.3 Maximum likelihood estimators (MLEs)

In light of our interpretation of likelihood as providing a ranking of the possible θ values in terms of how well the corresponding models fit the data, it makes sense to estimate the unknown θ by the "highest ranked" value. Since larger likelihood means higher rank, the idea is to estimate θ by the maximizer of the likelihood function, if possible.

Definition 3.2. Given $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f_\theta$, let $L(\theta)$ and $\ell(\theta)$ be the likelihood and log-likelihood functions, respectively. Then the maximum likelihood estimator (MLE) of θ is defined as

$$\hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta) = \arg\max_{\theta \in \Theta} \ell(\theta), \tag{3.3}$$

where "arg" says to return the argument at which the maximum is attained. Note that $\hat{\theta}$ implicitly depends on $(X_1, \ldots, X_n)$ because the (log-)likelihood does.

Thus, we have defined a process by which an estimator of the unknown parameter can be constructed. I call this a "process" because it can be done in the same way for (essentially) any problem: write down the likelihood function and then maximize it. In addition to the simplicity of the process, the estimator also has the nice interpretation as being the "highest ranked" of all possible θ values, given the observed data, as well as nice properties.

I should mention that while I've called the construction of the MLE "simple," I mean that only at a fundamental level. Actually doing the maximization step can be tricky, and sometimes requires sophisticated numerical methods (see Section 3.8). In the nicest of cases, the estimation problem reduces to solving the *likelihood equation*,

$$(\partial/\partial\theta)\ell(\theta) = 0.$$

This, of course, only makes sense if $\ell(\theta)$ is differentiable, as in the next two examples.

Exercise 3.3. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Ber}(\theta)$, for $\theta \in (0, 1)$. Find the MLE of θ.

Exercise 3.4. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(\theta, 1)$, for $\theta \in (0, 1)$. Find the MLE of θ.

It can happen that extra considerations can make an ordinarily nice problem not so nice. These extra considerations are typically in the form of constraints on the parameter space Θ. The next example gives a couple illustrations.

Exercise 3.5. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Pois}(\theta)$, where $\theta > 0$.

(a) Find the MLE of θ.

(b) Suppose that we know $\theta \geq b$, where b is a known positive number. Using this additional information, find the MLE of θ.

(c) Suppose now that θ is known to be an integer. Find the MLE of θ.

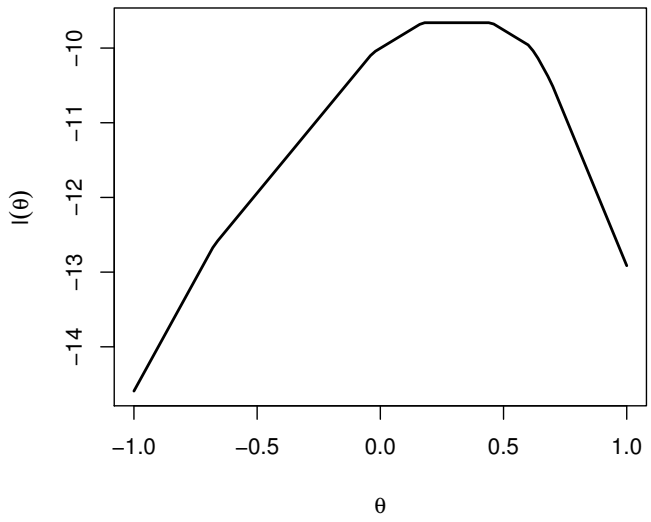


Figure 3.1: Graph of the Laplace log-likelihood function for a sample of size $n = 10$.

It may also happen the the (log-)likelihood is not differentiable at one or more points. In such cases, the likelihood equation itself doesn't make sense. This doesn't mean the problem can't be solved; it just means that we need to be careful. Here's an example.

Exercise 3.6. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Unif}(0, \theta)$ find the MLE of θ.

I should also mention that, even if the likelihood equation is valid, it may be that the necessary work to solve it cannot be done by hand. In such cases, numerical methods are needed. Some examples are given in the supplementary notes.

Finally, in some cases, the MLE is not unique (more than one solution to the likelihood equation) and in others no MLE exists (the likelihood function is unbounded). Example 3.1 demonstrates the former. The simplest example of the latter is in cases where the likelihood is continuous and there is an open set constraint on θ. An important practical example is in mixture models, which we won't discuss here.

Example 3.1. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f_\theta(x) = e^{-|x-\theta|}/2$; this distribution is often called the shifted Laplace or double-exponential distribution. For illustration, I consider a sample of size $n = 10$ from the Laplace distribution with $\theta = 0$. In Figure 3.1 we see that the log-likelihood flattens out, so there is an entire interval where the likelihood equation is satisfied; therefore, the MLE is not unique. (You should write R code to recreate this example.)

3.4 Basic properties

3.4.1 Invariance

In the context of unbiasedness, recall the claim that, if $\hat{\theta}$ is an unbiased estimator of θ, then $\hat{\eta} = g(\hat{\theta})$ is not necessarily and unbiased estimator of $\eta = g(\theta)$; in fact, unbiasedness holds if and only if g is a linear function. That is, unbiasedness is not invariant with respect to transformations. However, MLEs are invariant in this sense—if $\hat{\theta}$ is the MLE of θ, then $\hat{\eta} = g(\hat{\theta})$ is the MLE of $\eta = g(\theta)$.

Theorem 3.1 (HMC, Theorem 6.1.2)**.** *Suppose $\hat{\theta}$ is the MLE of θ. Then, for specified function g, $\hat{\eta} = g(\hat{\theta})$ is the MLE of $\eta = g(\theta)$.*

Proof. The result holds for any function g, but to see the main idea, suppose that g is one-to-one. Then our familiar likelihood, written as a function of η, is simply $L(g^{-1}(\eta))$. The largest this function can be is $L(\hat{\theta})$. Therefore, to maximize, choose $\hat{\eta}$ such that $g^{-1}(\hat{\eta}) = \hat{\theta}$, i.e., take $\hat{\eta} = g(\hat{\theta})$. □

This is a very useful result, for it allows us to estimate lots of different characteristics of a distribution. Think about it: since f_θ depends on θ, any interesting quantity (expected values, probabilities, etc) will be a function of θ. Therefore, if we can find the MLE of θ, then we can easily produce the MLE for any of these quantities.

Exercise 3.7. If $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Ber}(\theta)$, find the MLE of $\eta = \sqrt{\theta(1-\theta)}$. What quantity does η represent for the $\mathsf{Ber}(\theta)$ distribution?

Exercise 3.8. Let $X \sim \mathsf{Pois}(\theta)$. Find the MLE of $\eta = e^{-2\theta}$. How does the MLE of η here compare to the estimator given in Example 4 of Notes 02?

This invariance property is nice, but there is a somewhat undesirable consequence: *MLEs are generally NOT unbiased*. Both of the exercises above demonstrate this. For a simpler example, consider $X \sim \mathsf{N}(\theta, 1)$. The MLE of θ is $\hat{\theta} = X$ and, according to Theorem 3.1, the MLE of $\eta = \theta^2$ is $\hat{\eta} = \hat{\theta}^2 = X^2$. However, $\mathsf{E}_\theta(X^2) = \theta^2 + 1 \neq \theta^2$, so the MLE is biased.

Before you get too discouraged about this, recall the remarks made in Chapter 2 that unbiasedness is not such an important property. In fact, we will show below that MLEs are, at least for large n, the best one can do.

3.4.2 Consistency

In certain examples, it can be verified directly that the MLE is consistent, e.g., this follows from the law of large numbers if the distribution is $\mathsf{N}(\theta, 1)$, $\mathsf{Pois}(\theta)$, etc. It would be better, though, if we could say something about the behavior of MLEs in general. It turns out that this is, indeed, possible—it is a consequence of the process of maximizing the likelihood function, not of the particular distributional form.

We need a bit more notation. Throughout, θ denotes a generic parameter value, while $\theta^\star$ is the "true" but unknown value; HMC use the notation θ_0 instead of $\theta^\star$.[3] The goal is to demonstrate that the MLE, denoted now by $\hat{\theta}_n$ to indicate its dependence on n, will be close to $\theta^\star$ in the following sense:

For any $\theta^\star$, the MLE $\hat{\theta}_n$ converges to $\theta^\star$ in $\mathsf{P}_{\theta^\star}$-probability as $n \to \infty$, i.e.,

$$\lim_{n\to\infty} \mathsf{P}_{\theta^\star}\{|\hat{\theta}_n - \theta^\star| > \varepsilon\} = 0, \quad \forall\, \varepsilon > 0.$$

We shall also need to put forth some general assumptions about the model, etc. These are generally referred to as *regularity conditions*, and we will list this as R0, R1, etc. Several of these regularity conditions will appear in our development below, but we add new ones to the list only when they're needed. Here's the first three:

R0. If $\theta \neq \theta'$, then f_θ and $f_{\theta'}$ are different distributions.

R1. The support of f_θ, i.e., $\text{supp}(f_\theta) := \{x : f_\theta(x) > 0\}$, is the same for all θ.

R2. $\theta^\star$ is an interior point of Θ.

R0 is a condition called "identifiability," and it simply means that it is possible to estimate θ based on only a sample from f_θ. R1 ensures that ratios $f_\theta(X)/f_{\theta'}(X)$ cannot equal ∞ with positive probability. R2 ensures that there is an open subset of Θ that contains $\theta^\star$; R2 will also help later when we need a Taylor approximation of log-likelihood.

Exercise 3.9. Can you think of any familiar distributions that *do not* satisfy R1?

The first result provides a taste of why $\hat{\theta}$ should be close to $\theta^\star$ when n is large. It falls short of establishing the required consistency, but it does give some nice intuition.

Proposition 3.1 (HMC, Theorem 6.1.1)**.** *If R0 and R1 hold, then, for any $\theta \neq \theta^\star$,*

$$\lim_{n\to\infty} \mathsf{P}_{\theta^\star}\{L_X(\theta^\star) > L_X(\theta)\} = 1.$$

Sketch of the proof. Note the equivalence of the events:

$$\begin{aligned} L_X(\theta^\star) > L_X(\theta) &\iff L_X(\theta^\star)/L_X(\theta) > 1 \\ &\iff K_n(\theta^\star, \theta) := \frac{1}{n}\sum_{i=1}^n \log \frac{f_{\theta^\star}(X_i)}{f_\theta(X_i)} > 0. \end{aligned}$$

Define the quantity[4]

$$K(\theta^\star, \theta) = \mathsf{E}_{\theta^\star}\Big\{\log \frac{f_{\theta^\star}(X)}{f_\theta(X)}\Big\},$$

[3]Note that there is nothing special about any particular $\theta^\star$ value—the results to be presented hold for any such value. It's simply for convenience that we distinguish this value in the notation and keep it fixed throughout the discussion.

[4]This is known as the *Kullback–Leibler divergence*, a sort of measure of the distance between two distributions $f_{\theta^\star}$ and f_θ.

From Jensen's inequality (HMC, Theorem 1.10.5), it follows that $K(\theta^\star, \theta) \geq 0$ with equality iff $\theta = \theta^\star$; in our case, $K(\theta^\star, \theta)$ is strictly positive. From the LLN:

$$K_n(\theta^\star, \theta) \to K(\theta^\star, \theta) \quad \text{in } \mathsf{P}_{\theta^\star}\text{-probability.}$$

That is, $K_n(\theta^\star, \theta)$ is near $K(\theta^\star, \theta)$, a positive number, with probability approaching 1. The claim follows since the event of interest is equivalent to $K_n(\theta^\star, \theta) > 0$. □

The intuition is that the likelihood function at the "true" $\theta^\star$ tends to be larger than any other likelihood value. So, if we estimate θ by maximizing the likelihood, that maximizer ought to be close to $\theta^\star$. To get the desired consistency, there are some technical hurdles to overcome—the key issue is that we're maximizing a random function, so some kind of uniform convergence of likelihood is required.

If we add R2 and some smoothness, we can do a little better than Proposition 3.1.

Theorem 3.2 (HMC, Theorem 6.1.3)**.** *In addition to R0–R2, assume that $f_\theta(x)$ is differentiable in θ for each x. Then there exists a consistent sequence of solutions of the likelihood equation.*

The proof is a bit involved; see p. 325 in HMC. This is very interesting fact but, being an existence result alone, it's not immediately clear how useful it is. For example, as we know, the likelihood equation could have many solutions for a given n. For the question "which sequence of solutions is consistent?" the theorem provides no guidance. But it does suggest that the process of solving the likelihood equation is a reasonable approach. There is one special case in which Theorem 3.2 gives a fully satisfactory answer.

Corollary 3.1 (HMC, Corollary 6.1.1)**.** *In addition to the assumptions of Theorem 3.2, suppose the likelihood equation admits a unique solution $\hat{\theta}_n$ for each n. Then $\hat{\theta}_n$ is consistent.*

I shall end this section with a bit of history. Much of the ideas (though not the proofs) were developed by Sir Ronald A. Fisher, arguably the most influential statistician in history. At the time (1920s), the field of statistics was very new and without a formal mathematical framework. Fisher's ideas on likelihood and maximum likelihood estimation set the stage for all the theoretical work that has been done since then. He is also responsible for the ideas of information and efficiency in the coming sections, as well as the notion of sufficiency to be discussed later in the course. The p-value in hypothesis testing and randomization in designed experiments can be attributed to Fisher. Two of Fisher's other big ideas, which are less understood, are conditional inference (conditioning on ancillary statistics) and fiducial inference. Besides being one of the fathers of statistics, Fisher was also an extraordinary geneticist and mathematician. Personally, Fisher was a bit of a fiery character—there are well-documented heated arguments between Fisher, Neyman, and others about the philosophy of statistics. This "hot-headedness" was likely a result of Fisher's passion for the subject, as I have heard from people who knew him that he was actually very kind.

3.5 Fisher information and the Cramer–Rao bound

To further study properties of MLEs, we introduce a concept of *information*. Before we can do this, however, we need two more regularity conditions.

R3. $f_\theta(x)$ is twice differentiable in θ for each x;

R4. $\int f_\theta(x)\,dx$ in the continuous case, or $\sum_x f_\theta(x)$ in the discrete case, is twice differentiable in θ, and the derivative can be evaluated by interchanging the order of differentiation and integration/summation.

The first condition is to guarantee that the problem is sufficiently smooth. R4 is the first condition that's really technical. It holds for most problems we'll encounter herein, but it really has nothing to do with statistics or probability. For completeness, Section 3.10.9 gives some details about interchange of derivatives and integrals/sums.

In what follows I will work with the case of continuous distributions with PDF $f_\theta(x)$. The discrete case is exactly the same, but with summation over x where integration over x appears below. For moment, consider a single $X \sim f_\theta(x)$. Here's a simple calculus identity that will help simplify some notation, etc:

$$\frac{\partial}{\partial\theta} f_\theta(x) = \frac{\frac{\partial}{\partial\theta} f_\theta(x)}{f_\theta(x)} \cdot f_\theta(x) = \frac{\partial}{\partial\theta} \log f_\theta(x) \cdot f_\theta(x).$$

Using the fact that $1 = \int f_\theta(x)\,dx$ for all θ, if we differentiate both sides with respect to θ and apply R4 we get

$$0 = \int \frac{\partial}{\partial\theta} f_\theta(x)\,dx = \int \frac{\partial}{\partial\theta} \log f_\theta(x) \cdot f_\theta(x)\,dx = \mathsf{E}_\theta\Big\{ \frac{\partial}{\partial\theta} \log f_\theta(X) \Big\}.$$

The random variable $U_\theta(X) := \frac{\partial}{\partial\theta} \log f_\theta(X)$ is called the *score function*, and depends on both X and θ. We have shown that the score function has mean zero.

Differentiate the fundamental identity $1 = \int f_\theta(x)\,dx$ a second time and apply R4 once more we get

$$\begin{aligned} 0 &= \int \frac{\partial}{\partial\theta} \Big[\frac{\partial}{\partial\theta} \log f_\theta(x) \cdot f_\theta(x) \Big]\,dx \\ &= \cdots \\ &= \mathsf{E}_\theta\Big\{ \frac{\partial^2}{\partial\theta^2} \log f_\theta(X) \Big\} + \mathsf{E}_\theta\Big\{ \Big(\frac{\partial}{\partial\theta} \log f_\theta(X) \Big)^2 \Big\}. \end{aligned}$$

It follows that the latter two expectations are equal in magnitude—one negative, the other positive. This magnitude is called the *Fisher information*; that is,

$$I(\theta) = \mathsf{E}_\theta\Big\{ \Big(\frac{\partial}{\partial\theta} \log f_\theta(X) \Big)^2 \Big\} = -\mathsf{E}_\theta\Big\{ \frac{\partial^2}{\partial\theta^2} \log f_\theta(X) \Big\}. \tag{3.4}$$

This definition is understood that the Fisher information $I(\theta)$ can be evaluated with either of the two expressions on the right-hand side. You may use whichever is most convenient.

It is clear that the first expression for $I(\theta)$ in (3.4) is positive (why?) and, therefore, defines the magnitude mentioned above. So the second expectation is negative and multiplication by -1 makes it positive.

If you recall the score function $U_\theta(X)$ defined above, then you'll notice that $I(\theta) = \mathsf{E}_\theta\{U_\theta(X)^2\}$. If you also recall that $U_\theta(X)$ has mean zero, then you'll see that the Fisher information is simply the variance $\mathsf{V}_\theta\{U_\theta(X)\}$. But despite this simple expression for $I(\theta)$ in terms of a variance of the score, it turns out that it's usually easier to evaluate $I(\theta)$ using the version with second derivatives.

Exercise 3.10. Find $I(\theta)$ when X is $\mathsf{Ber}(\theta)$, $\mathsf{Pois}(\theta)$, and $\mathsf{Exp}(\theta)$.

Exercise 3.11. Let $X \sim \mathsf{N}(\theta, \sigma^2)$ where $\sigma > 0$ is a known number. Find $I(\theta)$.

Exercise 3.12. Let $X \sim f_\theta(x)$, where the PDF is of the form $f_\theta(x) = g(x - \theta)$, with g an arbitrary PDF. Show that $I(\theta)$ is a constant, independent of θ. (Hint: In the integration, make a change of variable $z = x - \theta$.)

Exercise 3.13. Let $I(\theta)$ be the Fisher information defined above. Let $\eta = g(\theta)$ be a reparametrization, where g is a one-to-one differentiable function. If $\tilde{I}(\eta)$ is the Fisher information for the new parameter η, show that $\tilde{I}(\eta) = I(g^{-1}(\eta)) \cdot [\{g^{-1}(\eta)\}']^2$.

So far we've considered only a single observation $X \sim f_\theta(x)$. What happens if we have a sample $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f_\theta(x)$? We simply replace $f_\theta(x)$ in the calculations above with the likelihood function $L(\theta) = L_X(\theta)$. Fortunately, since the model is iid, we don't have to redo all the calculations. For the score function, $U_\theta(X) = U_\theta(X_1, \ldots, X_n)$, we have

$$\begin{aligned} U_\theta(X_1, \ldots, X_n) &= \frac{\partial}{\partial\theta} \log L_X(\theta) \\ &= \frac{\partial}{\partial\theta} \sum_{i=1}^n \log f_\theta(X_i) \quad \text{(by independence)} \\ &= \sum_{i=1}^n \frac{\partial}{\partial\theta} \log f_\theta(X_i) \quad \text{(linearity of derivative)} \\ &= \sum_{i=1}^n U_\theta(X_i), \quad \text{(definition of } U_\theta(X_i)\text{)} \end{aligned}$$

the sum of the individual score functions. The Fisher information in the sample of size n is still defined as the variance of the score function. However, since we have a nice representation of the score as a sum of individual scores, we have

$$\begin{aligned} \mathsf{V}_\theta\{U_\theta(X_1, \ldots, X_n)\} &= \mathsf{V}_\theta\{U_\theta(X_1) + \cdots + U_\theta(X_n)\} \\ &= [\text{missing details}] \\ &= nI(\theta). \end{aligned}$$

Exercise 3.14. Fill in the missing details in the expression above.

We have, therefore, shown that the information in a sample of size n is simply n times the information in a single sample. This derivation depends critically on the iid assumption, but that's the only case we'll consider here; but know that in dependent or non-iid data problems the Fisher information would be different.

I have so far deferred the explanation of why $I(\theta)$ is called "information." A complete understanding cannot be given yet—wait until we discuss *sufficient statistics*—but the derivation above gives us some guidance. Intuitively, we expect that, as n increases (i.e., more data is collected), we should have more "information" about what distribution data was sample from and, therefore, we should be able to estimate θ better, in some sense. Our derivation shows that, since $I(\theta)$ is non-negative, as sample size n increases, the information about θ in that sample $nI(\theta)$ increases (linearly). So our intuition is satisfied in this case. For dependent-data problems, for example, information in the sample will still increase, but slower than linear. The Cramer–Rao lower bound result that follows should also help solidify this intuition.

In Chapter 2 we discussed point estimation and the idea of making mean-square error small. Of course, mean-square error is closely related to the variance of the estimator. The result that follows helps relate the variance of an estimator to the Fisher information. The message is that, if information is large, then better estimation should be possible.

Theorem 3.3 (Cramer–Rao; Theorem 6.2.1 in HMC)**.** *Let $X_1, \ldots, X_n \overset{iid}{\sim} f_\theta(x)$, and assume R0–R4 hold. Let $T_n = T_n(X_1, \ldots, X_n)$ be a statistic, with $\mathsf{E}_\theta(T_n) = \tau(\theta)$. Then*

$$\mathsf{V}_\theta(T_n) \geq \frac{[\tau'(\theta)]^2}{nI(\theta)}, \quad \forall\, \theta,$$

where $\tau'(\theta)$ denotes the derivative of $\tau(\theta)$.

Proof. Recall that for two random variables X and Y, the covariance (if it exists) is defined as $\mathsf{C}(X,Y) = \mathsf{E}(XY) - \mathsf{E}(X)\mathsf{E}(Y)$. The Cauchy–Schwartz inequality (you may have seen this in a linear algebra course) that says $|\mathsf{C}(X,Y)| \leq \sqrt{\mathsf{V}(X)\mathsf{V}(Y)}$.

Here I will work with the case $n = 1$; write $X = X_1$, $T = T(X)$ for the statistic in question, and $U = U_\theta(X)$ for the score function. The first goal is to evaluate the covariance $\mathsf{C}_\theta(T,U)$. For this, recall that U has zero mean, so $\mathsf{C}_\theta(T,U) = \mathsf{E}_\theta(TU)$. Recall that $\frac{\partial}{\partial\theta}\log f_\theta(x) \cdot f_\theta(x) = \frac{\partial}{\partial\theta} f_\theta(x)$; then the expectation of TU can be written as

$$\begin{aligned}
\mathsf{E}_\theta(TU) &= \int T(x) U_\theta(x) f_\theta(x)\, dx \\
&= \int T(x) \frac{\partial}{\partial\theta} \log f_\theta(x) f_\theta(x)\, dx \\
&= \int T(x) \frac{\partial}{\partial\theta} f_\theta(x)\, dx \\
&= \frac{\partial}{\partial\theta} \int T(x) f_\theta(x)\, dx \quad \text{(by R4)} \\
&= \tau'(\theta).
\end{aligned}$$

Now we know that $\mathsf{V}_\theta(U) = I(\theta)$, so the Cauchy–Schwartz inequality above gives

$$|\tau'(\theta)| \leq \sqrt{\mathsf{V}_\theta(T) I(\theta)}.$$

Squaring both sides and solving for $\mathsf{V}_\theta(T)$ gives the desired result. □

The following corollary helps us better understand the message of the Cramer–Rao inequality. Here we focus on the case where T_n is an unbiased estimator of θ.

Corollary 3.2 (HMC, Corollary 6.2.1)**.** *Let T_n be an unbiased estimator of θ. Then under the assumptions of Theorem 3.3, $\mathsf{V}_\theta(T_n) \geq [nI(\theta)]^{-1}$.*

Therefore, in this special case, the Cramer–Rao inequality can be understood as giving a lower bound on the variance of an unbiased estimator of θ. From a practical point of view, this provides us a gauge for measuring the quality of unbiased estimators. For example, if we find an unbiased estimator whose variance is exactly equal to the Cramer–Rao bound, then we know that no other unbiased estimator can do better than this one. We follow up on this idea in Section 3.6.

Exercise 3.15. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Pois}(\theta)$. Find the Cramer–Rao lower bound for unbiased estimators of θ. Find the variance of $\bar{X}$ and compare to this lower bound. We've seen before that S^2 is also an unbiased estimator of θ. What does your comparison of the Cramer–Rao lower bound and $\mathsf{V}_\theta(\bar{X})$ say about the relative performance of of $\bar{X}$ and S^2? You don't have to evaluate the variance of S^2, just explain how Corollary 3.2 helps with your comparison.

3.6 Efficiency and asymptotic normality

To follow up, more formally, on the notion of measuring performance of estimators by comparing their variance to the Cramer–Rao lower bound, we define a notion of efficiency. If $\hat{\theta}_n$ is an unbiased estimator of θ, then the *efficiency* of $\hat{\theta}_n$ is

$$\mathsf{eff}_\theta(\hat{\theta}_n) = \text{LB}/\mathsf{V}_\theta(\hat{\theta}_n), \quad \text{where} \quad \text{LB} = 1/nI(\theta).$$

An estimator is *efficient* if $\mathsf{eff}_\theta(\hat{\theta}_n) = 1$.

Exercise 3.16. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(0, \theta)$, where $\theta > 0$ denotes the variance.

(a) Let $\hat{\theta}_n^{(1)}$ be the sample variance. Find $\mathsf{eff}_\theta(\hat{\theta}_n^{(1)})$.

(b) Find the MLE of θ, and write this as $\hat{\theta}_n^{(2)}$. Find $\mathsf{eff}_\theta(\hat{\theta}_n^{(2)})$.

(c) Compare $\hat{\theta}_n^{(1)}$ and $\hat{\theta}_n^{(2)}$ based on their efficiencies.

We are particularly interested in the efficiency of MLEs, but there's not so many problems where the MLE has a nice expression, and even fewer of these cases can we write down a formula for its variance. So it would be nice to have some idea about the efficiency of MLEs without having to write down its variance. The next theorem, a fundamental result in statistics, gives us such a result. Indeed, a consequence of this theorem is that the MLE *asymptotically efficient* in the sense that, as $n \to \infty$, the efficiency of the MLE approaches 1. We need one more regularity condition:

R5. $f_\theta(x)$ is thrice differentiable in θ for each x, and there exists a constant $c > 0$ and a function $M(x) > 0$ such that $\mathsf{E}_\theta[M(X)] < \infty$ and, for "true value" $\theta^\star$, $|\frac{\partial^3}{\partial \theta^3} \log f_\theta(x)| \leq M(x)$ for all x and for all $\theta \in (\theta^\star - c, \theta^\star + c)$.

This assumption allows us to write a two-term Taylor approximation for $\ell'(\theta)$, which is the driving part of the proof, sketched below.

Theorem 3.4 (HMC, Theorem 6.2.2)**.** *Let $X_1, \ldots, X_n \overset{iid}{\sim} f_\theta(x)$. If R0–R5 hold, and $I(\theta) \in (0, \infty)$, then for any consistent sequence of solutions $\hat{\theta}_n$ of the likelihood equation $\sqrt{n}(\hat{\theta}_n - \theta) \to \mathsf{N}(0, I(\theta)^{-1})$ in distribution as $n \to \infty$.*

We can understand the result as saying that, when n is large, the MLE $\hat{\theta}_n$ is approximately normal with mean θ and variance $[nI(\theta)]^{-1}$. So the claim about asymptotic efficiency of the MLE is clear, since the variance is exactly the Cramer–Rao lower bound. Theorem 3.4 is fundamental for applied statistics. It says that no matter how the MLE is obtained—closed form expression, complicated numerical algorithms, etc—the sampling distribution is approximately normal when n is large. Many statistical computing packages report hypothesis tests and confidence intervals in relatively complex problems, such as logistic regression, and these are based on the sampling distribution result in Theorem 3.4.

Sketch of the proof of Theorem 3.4. The basic idea of the proof is fairly simple, although carrying out the precise details is a bit tedious. So, here I'll just give a sketch to communicate the ideas. First, do a Taylor approximation of $\ell_n'(\hat{\theta}_n)$ in a neighborhood of $\hat{\theta}_n = \theta$. Since $\hat{\theta}_n$ is a solution to the likelihood equation, we know that $\ell_n'(\hat{\theta}_n) = 0$. Therefore, this Taylor approximation looks like

$$0 = \ell_n'(\hat{\theta}_n) = \ell_n'(\theta) + \ell_n''(\theta_n)(\hat{\theta}_n - \theta) + \text{error},$$

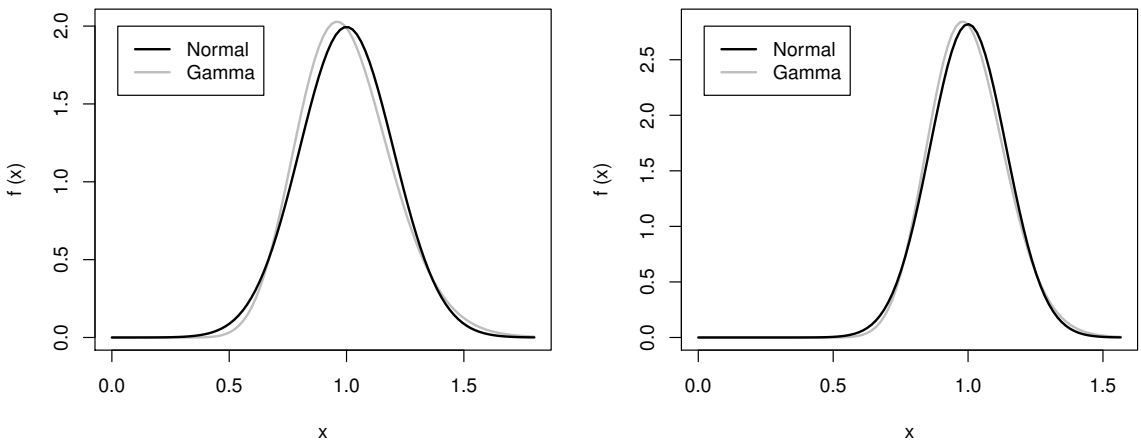
where θ_n is some value between $\hat{\theta}_n$ and θ. Since $\hat{\theta}_n$ is consistent, it follows that θ_n is too. Ignoring the error and rearranging the terms in the Taylor approximation gives

$$\sqrt{n}(\hat{\theta}_n - \theta) = -\frac{n^{1/2}\ell_n'(\theta)}{\ell_n''(\theta_n)} = -\frac{n^{-1/2}\ell_n'(\theta)}{n^{-1}\ell_n''(\theta_n)}.$$

Now we'll look at the numerator and denominator separately. We can apply the usual CLT to study the numerator. Indeed, note that

$$\bar{U}_n := \frac{1}{n}\ell_n(\theta) = \frac{1}{n}\sum_{i=1}^n U_\theta(X_i)$$

is an average of iid mean-zero, variance-$I(\theta)$ random variables. So the usual CLT says $n^{-1/2}\ell_n(\theta) = \sqrt{n}(\bar{U}_n - 0) \to \mathsf{N}(0, I(\theta))$ in distribution. For the denominator, we'll do a bit of fudging. Recall that θ_n is close to θ for large n. So we'll just replace $n^{-1}\ell_n''(\theta_n)$ in the denominator with $n^{-1}\ell_n''(\theta)$. A careful argument using the regularity conditions can make



(a) $n = 25$ (b) $n = 50$

Figure 3.2: Exact and approximate sampling distributions for the MLE in Example 3.2.

this step rigorous. Now $n^{-1}\ell_n''(\theta)$ is an average of iid mean-$I(\theta)$ random variables, so the usual LLN says $n^{-1}\ell_n''(\theta)$ converges in probability to $I(\theta)$. Slutsky's theorem gives

$$\sqrt{n}(\hat{\theta}_n - \theta) = -\frac{n^{1/2}\ell_n'(\theta)}{\ell_n''(\theta_n)} = -\frac{n^{-1/2}\ell_n'(\theta)}{n^{-1}\ell_n''(\theta_n)} \to I(\theta)^{-1} \cdot \mathsf{N}(0, I(\theta)), \quad \text{in distribution.}$$

But multiplying a normal random variable by a number changes the variance by the square of that number. That is,

$$I(\theta)^{-1} \cdot \mathsf{N}(0, I(\theta)) \equiv \mathsf{N}(0, I(\theta)^{-1}).$$

This completes the (sketch of the) proof. □

Example 3.2. An interesting question is: how accurate is the normal approximation for finite n? Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Exp}(\theta)$. If $\theta = 1$, then Theorem 3.4 says the MLE $\bar{X}_n$ is approximately normal with mean 1 and variance n^{-1}. However, it can be shown that $\bar{X}_n \sim \mathsf{Gamma}(n, n^{-1})$. Figure 3.2 shows the exact distribution of $\bar{X}_n$ and the normal approximation for two relatively small values of n. At $n = 25$ there's some noticeable differences between the two distributions, but for $n = 50$ there's hardly any difference.

Exercise 3.17. Show that if $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Exp}(\theta)$, then $\bar{X}_n \sim \mathsf{Gamma}(n, n^{-1}\theta)$.

Theorem 3.4 is much more broad than it looks initially. As it's stated, it applies only to the MLE of θ (specifically, consistent solutions of the likelihood equation). But in light of the invariance of MLE (Theorem 3.1) and the Delta Theorem (Theorem 1.2), we can develop a similar asymptotic normality result for any function of the MLE.

Exercise 3.18. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Exp}(\theta)$. The MLE is $\hat{\theta}_n = \bar{X}_n$. Use Theorem 3.4 and the Delta Theorem to find the limiting distribution of $\log \bar{X}_n$.

Asymptotic normality of MLEs, in combination with the Delta Theorem, is very useful in the construction of confidence intervals. Unfortunately, we don't have sufficient time to cover this important application in detail. But some supplementary material on maximum likelihood confidence intervals is provided in a separate document.

This consideration of the asymptotic efficiency of MLEs is effectively a comparison of the *asymptotic variance* of the MLE, which according to Theorem 3.4, is $I(\theta)^{-1}$. This is just like the "v_θ" in the Delta Theorem statement in Notes 01. So a way to compare two estimators is look at the ratio of their respective asymptotic variances. That is, the *asymptotic relative efficiency* of $\hat{\theta}_n^{(1)}$ and $\hat{\theta}_n^{(2)}$ is

$$\mathsf{are}_\theta(\hat{\theta}_n^{(1)}, \hat{\theta}_n^{(2)}) = \frac{\mathsf{aV}_\theta(\hat{\theta}_n^{(1)})}{\mathsf{aV}_\theta(\hat{\theta}_n^{(2)})},$$

where aV denotes the asymptotic variance. If this ratio is bigger (resp. smaller) than 1, then $\hat{\theta}_n^{(2)}$ is "better" (resp. "worse") than $\hat{\theta}_n^{(1)}$.

Example 3.3. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(\theta, \sigma^2)$, with σ known. The MLE is $\hat{\theta}_n^{(1)} = \bar{X}_n$, and it's easy to check that the MLE is efficient. An alternative estimator is $\hat{\theta}_n^{(2)} = M_n$, the sample median. The exact variance of M_n is difficult to get, so we shall compare these two estimators based on asymptotic relative efficiency. For this, we need a sort of CLT for M_n (the 50th percentile):

> (CLT for percentiles) *Let* $X_1, \ldots, X_n \overset{iid}{\sim} f(x) = F'(x)$. *For any* $p \in (0,1)$, *let* η_p *be the* $100p$*th percentile, i.e.,* $F(\eta_p) = p$. *Likewise, let* $\hat{\eta}_p$ *be the* $100p$*th sample percentile. If* $f(\eta_p) > 0$, *then*
>
> $$\sqrt{n}(\hat{\eta}_p - \eta_p) \to \mathsf{N}\big(0, p(1-p)/f(\eta_p)^2\big), \quad \textit{in distribution.}$$

In this case, the asymptotic variance of M_n is

$$\mathsf{aV}_\theta(\hat{\theta}_n^{(2)}) = \frac{0.5 \cdot 0.5}{\big(\sqrt{1/2\pi\sigma^2}\big)^2} = \frac{\pi\sigma^2}{2}.$$

Since $\mathsf{aV}_\theta(\hat{\theta}_n^{(1)}) = \sigma^2$, the asymptotic relative efficiency is

$$\mathsf{are}_\theta(\hat{\theta}_n^{(1)}, \hat{\theta}_n^{(2)}) = \frac{\sigma^2}{\pi\sigma^2/2} = \frac{2}{\pi} < 1.$$

This ratio is less than 1, so we conclude that $\hat{\theta}_n^{(1)}$ is "better" asymptotically.

3.7 Multi-parameter cases

Now suppose that $\theta \in \Theta \subseteq \mathbb{R}^d$, for integer $d \geq 1$. An important example is $\Theta = \{\theta = (\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+\}$ for the normal distribution where both mean μ and variance σ^2 are unknown. In general, let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f_\theta(x)$. Then we may define the likelihood, and log-likelihood functions just as before:

$$L(\theta) = \prod_{i=1}^{n} f_\theta(X_i) \quad \text{and} \quad \ell(\theta) = \log L(\theta).$$

Likelihood still can be understood as providing a ranking of the possible parameter values and, therefore, maximizing the likelihood function to estimate the unknown θ still makes sense. That is, the MLE $\hat{\theta}$ is still defined as

$$\hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta) = \arg\max_{\theta \in \Theta} \ell(\theta).$$

Conceptually, everything is the same as in the one-dimensional parameter case. Technically, however, things are messier, e.g., we need vectors, matrices, etc. We can immediately see how things get more technically involved, by considering the analogue of the likelihood equation: $\hat{\theta}$ is the solution to

$$\nabla \ell(\theta) = 0.$$

Here ∇ is the gradient operator, producing a vector of component wise partial derivatives,

$$\nabla \ell(\theta) = \Big(\frac{\partial \ell(\theta)}{\partial \theta_1}, \ldots, \frac{\partial \ell(\theta)}{\partial \theta_d} \Big)^\top,$$

and superscript $\top$ being the transpose operator.

Exercise 3.19. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(\mu, \sigma^2)$, with $\theta = (\mu, \sigma^2)$ unknown. Find the MLE.

For multiple parameters, it is less likely that a closed-form solution to the likelihood equation is available. Typically, some kind of numerical methods will be needed to find the MLE. Next is a simple example of this scenario.

Exercise 3.20. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Gamma}(\alpha, \beta)$, with $\theta = (\alpha, \beta)$ unknown. Write down the likelihood equation and confirm that no closed-form solution is available.

For a single observation $X \sim f_\theta(x)$, the score vector is

$$U_\theta(X) = \nabla \log f_\theta(X) = \Big(\frac{\partial}{\partial \theta_1} \log f_\theta(X), \ldots, \frac{\partial}{\partial \theta_d} \log f_\theta(X) \Big)^\top.$$

In this case, the score is a $d \times 1$ (column) random vector. Recall that, for random vectors, there are notions of a mean vector and a covariance matrix. In particular, if Z is a d-dimensional random vector, then

$$\mathsf{E}(Z) = \big(\mathsf{E}(Z_1), \ldots, \mathsf{E}(Z_d) \big)^\top \quad \text{and} \quad \mathsf{C}(Z) = \mathsf{E}(ZZ^\top) - \mathsf{E}(Z)\mathsf{E}(Z)^\top.$$

So the mean of a random vector is a $d \times 1$ vector and the covariance is a $d \times d$ matrix (provided these quantities exist). Under versions of the regularity conditions in the one-parameter case, it can be shown that

$$\mathsf{E}_\theta[U_\theta(X)] = 0 \quad \text{(a } d\text{-vector of zeros)}.$$

Just like in the one-parameter case, we define the Fisher information as the (co)variance of the score, i.e., $I(\theta) = \mathsf{C}_\theta[U_\theta(X)]$, which is a $d \times d$ matrix, rather than a number. Under regularity conditions, each component of this matrix looks like a one-dimensional information; in particular, its (j,k)th element satisfies

$$\begin{aligned} I(\theta)_{jk} &= \mathsf{E}_\theta\Big\{ \frac{\partial}{\partial \theta_j} \log f_\theta(X) \cdot \frac{\partial}{\partial \theta_k} \log f_\theta(X) \Big\} \\ &= -\mathsf{E}_\theta\Big\{ \frac{\partial^2}{\partial \theta_j \partial \theta_k} \log f_\theta(X) \Big\}. \end{aligned}$$

$I(\theta)$ is a symmetric matrix (i.e., $I(\theta) = I(\theta)^\top$). This means you only need to evaluate $d(d+1)/2$ of the d^2 total matrix entries.

Exercise 3.21. For $X \sim \mathsf{N}(\mu, \sigma^2)$, with $\theta = (\mu, \sigma^2)$, find $I(\theta)$.

What about if we have an iid sample $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f_\theta(x)$ of size n? Everything goes just as before, except we're working with vectors/matrices. In particular, we replace the density $f_\theta(X)$ in the definition of the score vector with the likelihood function $L_X(\theta)$, and just as before, the Fisher information matrix for a sample of size n is just n times the information matrix $I(\theta)$ for a single observation.

For brevity, I shall summarize the d-dimensional analogues of the large-sample results derived above with care for one-dimensional problems. Here I will not explicitly state the regularity conditions, but know that they are essentially just higher-dimensional versions of R0–R5 listed above.

- Under regularity conditions, there exists a consistent sequence $\hat{\theta}_n$ (a d-vector) of solutions of the likelihood equation.
- Under regularity conditions, for any consistent sequence of solutions $\hat{\theta}_n$,

  $$\sqrt{n}(\hat{\theta}_n - \theta) \to \mathsf{N}_d(0, I(\theta)^{-1}) \quad \text{in distribution (for all } \theta),$$

  where $\mathsf{N}_d(0, I(\theta)^{-1})$ denotes a d-dimensional normal distribution with mean vector 0 and covariance matrix $I(\theta)^{-1}$, the $d \times d$ inverse of the Fisher information matrix.
- (Delta Theorem) Let $g : \mathbb{R}^d \to \mathbb{R}^k$ have continuous partial derivatives, and define the $k \times d$ matrix

  $$D = \big(\partial g(\theta)_i / \partial \theta_j\big)_{i=1,\ldots,k; j=1,\ldots,d}.$$

  Then, under regularity conditions,

  $$\sqrt{n}[g(\hat{\theta}_n) - g(\theta)] \to \mathsf{N}_k(0, DI(\theta)^{-1}D^\top).$$

For example, take $g : \mathbb{R}^d \to \mathbb{R}$ so that $g(\theta) = \theta_j$. Then D is a $1 \times d$ matrix of all zeros except a 1 appearing in the $(1, j)$ position. With this choice,

$$\sqrt{n}(\hat{\theta}_{n,j} - \theta_j) \to \mathsf{N}(0, I(\theta)^{-1}_{jj}),$$

which is the one-dimensional counterpart, like Theorem 3.4.

Exercise 3.22. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Gamma}(\alpha, \beta)$, where $\theta = (\alpha, \beta)^\top$ is unknown. Denote the MLE by $\hat{\theta}_n = (\hat{\alpha}_n, \hat{\beta}_n)^\top$; there's no closed-form expression for the the MLE, but it can be readily evaluated numerically (Example 3.6). State the limiting distribution of $\hat{\theta}_n$.

3.8 MLE computation

Suppose that sample data $X_1, \ldots, X_n$ are iid with common distribution having PMF/PDF $f_\theta(x)$. The goal is to estimate the unknown parameter θ. In simple examples, there is a closed-form expression for the MLE $\hat{\theta}$. But, in many practical problems, there may be more than one solution to the likelihood equation, and there may be no nice formula for those solutions. In such cases, one will need to use some numerical methods to compute the MLE. Here I will briefly discuss computation. These are important throughout all of applied and theoretical statistics. Focus here is on just one optimization strategy, namely Newton's method. A course on computational statistics would discuss other methods, such as the EM algorithm, simulated annealing, and iteratively re-weighted least squares.

3.8.1 Newton's method

Suppose the goal is to solve $g(x) = 0$ for some function g. Newton's method is one useful technique, and the basics are presented in early calculus courses. The idea is based on the fact that, locally, any differentiable function can be suitably approximated by a linear function. This linear function is then used to define a recursive procedure that will, under suitable conditions, eventually find the desired solution.

Consider, first, the case where the unknown parameter θ is a scalar. An example is where the underlying distribution is $\mathsf{Pois}(\theta)$. The vector parameter case is dealt with below. Write $\ell(\theta)$ for the log-likelihood $\log L(\theta)$. Assume that $\ell(\theta)$ is twice differentiable with respect to θ; this is not really a practical restriction, in fact, our good theoretical properties of MLEs assumes this and more.

The goal is to solve the likelihood equation $\ell'(\theta) = 0$; here, and throughout, the "prime" will denote differentiation with respect to θ. So now we can identify ℓ' with the generic function g above, and apply Newton's method. The idea is as follows. Pick some guess $\theta^{(0)}$ of $\hat{\theta}$. Now approximate $\ell'(\theta)$ by a linear function at $\theta^{(0)}$. That is,

$$\ell'(\theta) = \ell'(\theta^{(0)}) + \ell''(\theta^{(0)})(\theta - \theta^{(0)}) + \text{error to be ignored.}$$

Now solve for θ, and call the solution $\theta^{(1)}$:

$$\theta^{(1)} = \theta^{(0)} - \ell'(\theta^{(0)})/\ell''(\theta^{(0)}).$$

If $\theta^{(0)}$ is close to the solution of the likelihood equation, then so will $\theta^{(1)}$ (draw a picture!). The idea is to iterate this process until the solutions converge. So the method is to pick a "reasonable" starting value $\theta^{(0)}$ and, at iteration $t \geq 0$ set

$$\theta^{(t+1)} = \theta^{(t)} - \ell'(\theta^{(t)})/\ell''(\theta^{(t)}).$$

Then stop the algorithm when t is large and/or $|\theta^{(t+1)} - \theta^{(t)}|$ is small. General R code to implement Newton's method (for one or many parameters) is presented in Section 3.10.1. Then two examples are considered.

Example 3.4. Suppose $X_1, \ldots, X_n$ be an iid sample from a mean-θ exponential distribution with PDF $f_\theta(x) = \theta^{-1}e^{-x/\theta}$. We've seen already that the MLE for θ in this problem is the sample mean: $\hat{\theta} = \bar{X}$. For comparison, we can run Newton's algorithm to see what answer it produces. In this case, the log-likelihood is

$$\ell(\theta) = -n \log \theta - n\bar{X}/\theta.$$

The first and second derivatives of $\ell(\theta)$ are:

$$\ell'(\theta) = -\frac{n}{\theta} + \frac{n\bar{X}}{\theta^2} \quad \text{and} \quad \ell''(\theta) = \frac{n}{\theta^2} - \frac{2n\bar{X}}{\theta^3}.$$

We now have what we need to run Newton's algorithm. Suppose that the following data is a sample of size $n = 20$ from the mean-θ exponential distribution:

```
0.27  0.83  0.27  1.52  0.04  0.43  0.92  0.58  0.20  0.32
0.82  0.91  0.66  0.01  0.56  1.21  1.44  0.64  0.53  0.30
```

From the data we have $\hat{\theta} = \bar{X} = 0.623$. If we take $\theta^{(0)} \in (0, 1)$, then the iterates $\theta^{(t)}$ converge quickly to $\hat{\theta} = 0.623$, the same as we get from direct calculations; see Section 3.10.2. If $\theta^{(0)} > 1$, then Newton's method apparently does not converge to the correct solution.

Example 3.5. Refer to Example 6.1.4 in HMC. That is, let $X_1, \ldots, X_n$ be iid observations from a logistic distribution with PDF

$$f_\theta(x) = \frac{e^{-(x-\theta)}}{(1 + e^{-(x-\theta)})^2}, \quad x \in \mathbb{R}, \quad \theta \in \mathbb{R}.$$

The log-likelihood can be written as

$$\ell(\theta) = \sum_{i=1}^{n} \log f_\theta(X_i) = n\theta - n\bar{X} - 2\sum_{i=1}^{n} \log\{1 + e^{-(X_i-\theta)}\}.$$

Differentiating with respect to θ gives the likelihood equation:

$$0 = \ell'(\theta) = n - 2\sum_{i=1}^{n} \frac{e^{-(X_i-\theta)}}{1 + e^{-(X_i-\theta)}}.$$

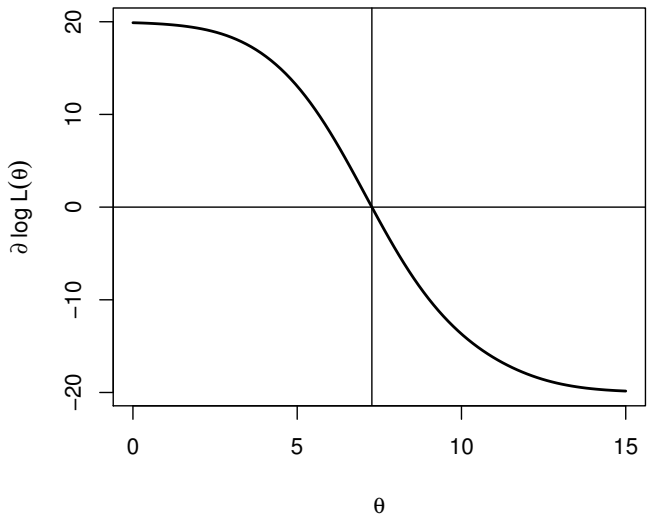


Figure 3.3: Graph of $\ell'(\theta)$ in Example 3.5 and the MLE $\hat{\theta} = 7.273$.

Although there is no formula for the solution, HMC shows that there is a unique solution. To find this solution, we shall employ Newton's method. We already have a formula for $\ell'(\theta)$, and we need one for $\ell''(\theta)$ as well:

$$\ell''(\theta) = -2\sum_{i=1}^{n} \frac{e^{-(X_i-\theta)}}{(1+e^{-(X_i-\theta)})^2}.$$

Now we have what we need to use Newton's estimate to find the MLE. Suppose that the data below are obtained by sampling from the logistic distribution:

```
6.37  12.01  7.34  6.28  7.09  7.51  8.24  7.35  6.70  4.95
5.14  10.72  3.67  6.35  9.71  7.20  9.21  7.88  5.80  8.27
```

From Newton's method, we find that $\hat{\theta} = 7.273$; see Section 3.10.3. Figure 3.3 shows a graph of $\ell'(\theta)$ and the solution to the likelihood equation $\ell'(\theta) = 0$. Incidentally, the data was simulated from a logistic distribution with $\theta = 7$, so the MLE is close to the truth.

Moving from a scalar to vector parameter, the only thing that changes is $\ell(\theta)$ is a function of $p > 1$ variables instead of just one, i.e., $\theta = (\theta_1, \ldots, \theta_p)^\top \in \mathbb{R}^p$; here I'm using the notation $x^\top$ instead of x' for vector/matrix transpose, so there's no confusion with differentiation. So, in this context, $\ell'(\theta)$ corresponds to a vector of partial derivatives and $\ell''(\theta)$ a matrix of mixed second-order partial derivatives. Therefore, the Newton's iterations are

$$\theta^{(t+1)} = \theta^{(t)} - [\ell''(\theta^{(t)})]^{-1}\ell'(\theta^{(t)}), \quad t \geq 0,$$

where $[\cdot]^{-1}$ denotes matrix inverse. This matches up conceptually with what was done in the single parameter problem.

Example 3.6. Suppose that $X_1, \ldots, X_n$ is an iid sample from a $\mathsf{Gamma}(\alpha, \beta)$ distribution where $\theta = (\alpha, \beta)^\top$ is unknown. The likelihood and log-likelihood functions are given by

$$L(\alpha, \beta) = \prod_{i=1}^n \frac{1}{\beta^\alpha \Gamma(\alpha)} X_i^{\alpha-1} e^{-X_i/\beta} = \Big(\frac{1}{\beta^\alpha \Gamma(\alpha)}\Big)^n e^{(\alpha-1)\sum_{i=1}^n \log X_i} e^{-(1/\beta)\sum_{i=1}^n X_i},$$

$$\ell(\alpha, \beta) = -n\alpha \log \beta - n \log \Gamma(\alpha) + (\alpha - 1) \sum_{i=1}^n \log X_i - \frac{1}{\beta} \sum_{i=1}^n X_i,$$

where $\Gamma(\alpha) = \int_0^\infty z^{\alpha-1} e^{-z}\, dz$ denotes the gamma function. Differentiating $\ell(\alpha, \beta)$ with respect to α and β, we get

$$\frac{\partial}{\partial \alpha} \ell(\alpha, \beta) = -n \log \beta - n\psi(\alpha) + \sum_{i=1}^n \log X_i$$

$$\frac{\partial}{\partial \beta} \ell(\alpha, \beta) = -\frac{n\alpha}{\beta} + \frac{n\bar{X}}{\beta^2},$$

where $\psi(\alpha) = \Gamma'(\alpha)/\Gamma(\alpha)$ is called the digamma function. There's no closed-form solution to the likelihood equation $\nabla \ell(\alpha, \beta) = 0$, so we must resort to a numerical method, like Newton's. For this we need the matrix of second derivatives:

$$\frac{\partial^2}{\partial \alpha^2} \ell(\alpha, \beta) = -n\psi'(\alpha),$$

$$\frac{\partial^2}{\partial \beta^2} \ell(\alpha, \beta) = \frac{n\alpha}{\beta^2} - \frac{2n\bar{X}}{\beta^3},$$

$$\frac{\partial^2}{\partial \alpha \partial \beta} \ell(\alpha, \beta) = -\frac{n}{\beta}.$$

The derivative $\psi'(\alpha)$ of the digamma function is called the trigamma function; both of these are special functions in R. Finally, the matrix of second partial derivatives is

$$\ell''(\alpha, \beta) = -n \begin{pmatrix} \psi'(\alpha) & 1/\beta \\ 1/\beta & 2\bar{X}/\beta^3 - \alpha/\beta^2 \end{pmatrix}.$$

To implement Newton's method, we'd need to calculate the inverse of this matrix. For 2×2 matrices, there's a relatively simple formula. But here we'll leave matrix inversion for R to do, via the function `ginv`, for (generalized) inverse.

How might we initialize Newton's method, i.e., how to choose $\theta^{(0)} = (\alpha^{(0)}, \beta^{(0)})$? A useful technique is the *method of moments*. In this case, since $\mathsf{E}_\theta(X_1) = \alpha\beta$ and $\mathsf{V}_\theta(X_1) = \alpha\beta^2$, the method of moments sets $\bar{X} = \alpha\beta$ and $S^2 = \alpha\beta^2$ and solves for α and β. In this case, $\beta^{(0)} = S^2/\bar{X}$ and $\alpha^{(0)} = \bar{X}^2/S^2$. These are the starting values we'll use the following calculations.

Suppose that the data below are obtained by sampling from the gamma distribution:

```
6.00  5.98  7.81  6.77  10.64  13.63  10.53  8.92  3.77  5.78
6.32  4.44  5.34  1.54   4.42   4.71   6.12  2.57  9.47  9.03
```

The starting values are $(\alpha^{(0)}, \beta^{(0)}) = (S^2/\bar{X}, \bar{X}^2/S^2) = (5.05, 1.32)$, based on method of moments. After running Newton's method (see Section 3.10.4), we find the MLE $(\hat{\alpha}, \hat{\beta}) = (4.75, 1.41)$. Incidentally, the true (α, β) used to simulated the data above was $(5, 1.4)$, so both the method of moments and MLE are close to the truth in this case.

3.8.2 Estimation of the Fisher information

For defining confidence intervals for θ, one typically needs and estimate of the Fisher information $I(\theta)$. One possible estimator is $I(\hat{\theta})$, but this is not always the best choice. It can be argued that $n^{-1}\ell''(\hat{\theta})$ is a better estimator. This is typically called the *observed information* while $I(\hat{\theta})$ is called the *expected information*. In some cases, these two are the same but not always. The advantage of the observed information $n^{-1}\ell''(\hat{\theta})$ is that it comes as a natural by-product of Newton's algorithm, whereas $I(\hat{\theta})$ requires some extra calculations. There are other, more fundamental, reasons for preferring the observed over expected information, but these are a bit beyond the scope of Stat 411.

3.8.3 An aside: one-step estimators

The presentation in this note has been mostly practical/computational. But there is an interesting theoretical concept related to Newton's method, that's worth mentioning; the book makes a remark along these lines at the bottom of page 329.

Suppose $\hat{\theta}_n$ is a consistent sequence of estimators of θ. This sequence may or may not be asymptotically efficient, however. The so-called *one-step estimator* is a trick by which a consistent sequence of estimators can be transformed into an asymptotically efficient one. Specifically, the sequence $\tilde{\theta}_n$ defined as

$$\tilde{\theta}_n = \hat{\theta}_n - [\ell''(\hat{\theta}_n)]^{-1}\ell'(\hat{\theta}_n)$$

is asymptotically efficient. That is, by applying one step of Newton's method, any consistent sequence of estimators can be turned into an asymptotically efficient sequence.

An interesting (albeit extreme) example is the normal mean problem. Let $\hat{\theta}_n$ be the sample median based on $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(\theta, 1)$. We know that this sequence is consistent. In this case, $\ell'(\theta) = n(\bar{X} - \theta)$ and $\ell''(\theta) = -n$, so the one-step estimator is

$$\tilde{\theta}_n = \hat{\theta}_n - \frac{\ell'(\hat{\theta}_n)}{\ell''(\hat{\theta}_n)} = \hat{\theta}_n + \frac{n(\bar{X} - \hat{\theta}_n)}{n} = \bar{X}.$$

We already know that the MLE $\bar{X}$ is (asymptotically) efficient.

3.8.4 Remarks

There are lots of tools available for doing optimization, the Newton method described above is just one simple approach. Fortunately, there are good implementations of these methods

already available in the standard software. For example, the routine `optim` in R is a very powerful and simple-to-use tool for generic optimization. For problems that have a certain form, specifically, problems that can be written in a "latent variable" form, there is a very clever tool called the EM algorithm for maximizing the likelihood. Section 6.6 in HMC gives a good but brief description of this very important method.

An interesting and unexpected result is that sometimes optimization can be used to do integration. The technical result I'm referring to is the *Laplace Approximation*. Some further comments on this will be made in Chapter 6.

3.9 Confidence intervals

Point estimation is an important problem in statistics. However, a point estimator alone is not really a good summary of the information in data. In particular, we know that, with large probability, $\hat{\theta} \neq \theta$, so the chance that our point estimator is "correct" is very small. So, in addition to a point estimator, it is helpful to provide some convenient summary of the variability of $\hat{\theta}$. This information is contained in the sampling distribution of $\hat{\theta}$, but the full sampling distribution is *not* a convenient summary; also the full sampling distribution may be unknown. One could report, for example, the point estimator and its variance. Alternatively, it is common to report what is called a *confidence interval* for θ. The basic idea is to report an interval which we believe to have large probability under the sampling distribution of $\hat{\theta}$. In these notes I will explain what confidence intervals are and how maximum likelihood estimators (MLEs) can be used to construct them.

Let $X = (X_1, \ldots, X_n)$ be an iid sample from a distribution with PDF/PMF $f_\theta(x)$, where $\theta \in \Theta$ is unknown. Let $\alpha \in (0, 1)$ be a specified (small) number, like $\alpha = 0.05$; in an applied statistics course, this number is the significance level. A $100(1-\alpha)\%$ confidence interval for θ, denoted by $\mathcal{C}_{n,\alpha} = \mathcal{C}_{n,\alpha}(X)$, is an interval such that

$$\mathsf{P}_\theta\{\mathcal{C}_{n,\alpha} \ni \theta\} = 1 - \alpha, \quad \forall\, \theta \in \Theta. \tag{3.5}$$

The event in the probability statement is "the interval $\mathcal{C}_{n,\alpha}$ contains θ." We want this probability to be large, as it gives us some *confidence* that the interval $\mathcal{C}_{n,\alpha}$ (which we can calculate with given data X) contains the parameter we're trying to estimate.

The most common example is that of a confidence interval for the mean of a normal distribution. Let $X = (X_1, \ldots, X_n)$ be an iid $\mathsf{N}(\theta, 1)$ sample, and consider the interval

$$\mathcal{C}_{n,\alpha}(X) = \big[\bar{X} - z^\star_\alpha n^{-1/2},\, \bar{X} + z^\star_\alpha n^{-1/2}\big], \tag{3.6}$$

where $z^\star_\alpha$ satisfies $\Phi(z^\star_\alpha) = 1 - \alpha/2$. In Homework 01 you showed that this interval satisfies (3.5) and, hence, is a $100(1-\alpha)\%$ confidence interval for θ. This is sometimes called a z-confidence interval. There are other popular confidence intervals, e.g., the t-interval, that are introduced in applied statistics courses. Here the goal is not to enumerate examples; instead, I want to discuss how MLEs, and their good properties, can be used to construct (approximate) confidence intervals.

Before I go on to these details, I should make some comments on the interpretation of confidence intervals.

- We *do not* interpret (3.5) as "there is $1-\alpha$ probability that θ is inside the confidence interval." The point here is that θ is not a random variable, so there are no meaningful probability statements for events concerning the location of θ. This is why I write the event as "$\mathcal{C}_{n,\alpha} \ni \theta$" as opposed to "$\theta \in \mathcal{C}_{n,\alpha}$"—the latter seems to indicate that θ is random and $\mathcal{C}_{n,\alpha}$ is fixed, while the former (correctly) indicates that $\mathcal{C}_{n,\alpha}$ is random and θ is fixed.

- Once we observe $X = x$, the interval $\mathcal{C}_{n,\alpha}(x)$ is *fixed* and can be calculated. Since θ is unknown, I cannot say for sure if $\mathcal{C}_{n,\alpha}(x)$ contains θ or not—property (3.5) is of no help. All that (3.5) tells us is that the *procedure* is good, i.e., the "sampling distribution" of $\mathcal{C}_{n,\alpha}(X)$, as a function of random vector X, has a desirable property. So one must be careful when interpreting (3.5) in a real data analysis problem.

Where does $\mathcal{C}_{n,\alpha}$ come from? Often, confidence intervals take the form of $\mathcal{C}_{n,\alpha} = \hat{\theta} \pm \text{SD}$, where $\hat{\theta}$ is some estimator of θ, and SD is something like a standard deviation of the sampling distribution of $\hat{\theta}$. If $\hat{\theta}$ is the MLE, then we may not know the sampling distribution exactly, but for large n we have a general result. Theorem 3.4 says that, if n is large, then $\hat{\theta}_n$ is approximately normal with mean θ and variance $[nI(\theta)]^{-1}$. So, if we assume n is large enough for this approximation to hold, then we are back to that normal distribution problem from before, and the z-interval (3.6) looks like

$$\mathcal{C}_{n,\alpha} = \hat{\theta}_n \pm z_\alpha^\star [nI(\theta)]^{-1/2}. \tag{3.7}$$

Asymptotic normality of the MLE (Theorem 3.4) implies that

$$\lim_{n\to\infty} \mathsf{P}_\theta\{\mathcal{C}_{n,\alpha} \ni \theta\} = 1 - \alpha, \quad \forall\, \theta, \tag{3.8}$$

so, if n is large, we have coverage probability $\mathsf{P}_\theta\{\mathcal{C}_{n,\alpha} \ni \theta\} \approx 1 - \alpha$. Therefore, $\mathcal{C}_{n,\alpha}$ is what is called an *asymptotically correct* confidence interval for θ.

The calculations above are fine; however, observed that the right-hand side of (3.7) depends, in general, on the unknown θ. That is, we cannot evaluate the right-hand side in practice! To deal with this problem, there are two common fixes. The first and simplest is to plug in $\hat{\theta}_n$ for θ on the right-hand side of (3.7). The second is based on a suitably chosen transformation and an application of the delta theorem.

Wald-style plug-in

Clearly, a simple fix to the problem of dependence on θ is to plug-in an estimator for θ. That is, the plug-in approach, named after Abraham Wald, uses the interval

$$\mathcal{C}_{n,\alpha} = \hat{\theta}_n \pm z_\alpha^2 [nI(\hat{\theta}_n)]^{-1/2}.$$

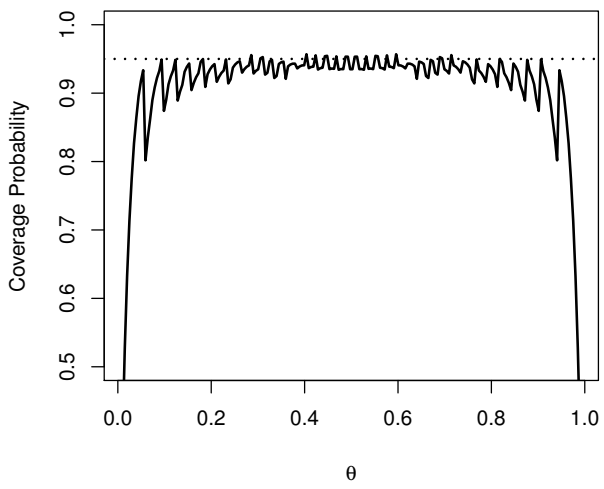


Figure 3.4: Coverage probability $\mathsf{P}_\theta\{\mathcal{C}_{n,\alpha} \ni \theta\}$ in Example 3.7 as a function of θ for $n = 50$ and $\alpha = 0.05$.

While this interval usually works fairly well, a result like (3.8) is not an immediate consequence of asymptotic normality. In fact, there are known cases where this procedure works rather poorly.[5] However, the method's simplicity has kept it in common use, despite its shortcomings.

Example 3.7. Let $X_1, \ldots, X_n$ be iid $\mathsf{Ber}(\theta)$. We know that $\hat{\theta}_n = \bar{X}$ and $I(\theta) = 1/\theta(1-\theta)$. Therefore, the plug-in confidence interval is

$$\mathcal{C}_{n,\alpha} = \bar{X} \pm z_\alpha^\star \sqrt{\frac{\bar{X}(1-\bar{X})}{n}}.$$

A plot of the coverage probability $\mathsf{P}_\theta\{\mathcal{C}_{n,\alpha} \ni \theta\}$, as a function of θ, for $n = 50$ and $\alpha = 0.05$, is shown in Figure 3.4; R code is given in Section 3.10.5. Here we see that the coverage probability tends to be too low (less than 0.95) for most θ values.

Example 3.8. Let $X_1, \ldots, X_n$ be iid $\mathsf{Exp}(\theta)$. In this case, we know that $\hat{\theta}_n = \bar{X}$ and $I(\theta) = 1/\theta^2$. Therefore, the plug-in style confidence interval is

$$\mathcal{C}_{n,\alpha} = \bar{X} \pm z_\alpha^\star \bar{X} n^{-1/2}.$$

The coverage probability $\mathsf{P}_\theta\{\mathcal{C}_{n,\alpha} \ni \theta\}$ can be calculated by using the CDF of the $\mathsf{Gamma}(n, \theta)$ distribution. Moreover, one can show that the coverage probability does not depend on θ. So, in this case, Figure 3.5 shows the coverage probability as a function of n for $\theta = 1$ and $\alpha = 0.05$; see Section 3.10.6. Here we see that, even for $n \approx 1500$, the coverage probability is still below the target 0.95. However, it is close and continues to get closer as n increases.

[5] Brown, Cai, and DasGupta, "Interval estimation for a Binomial proportion" *Statistical Science*, 2001

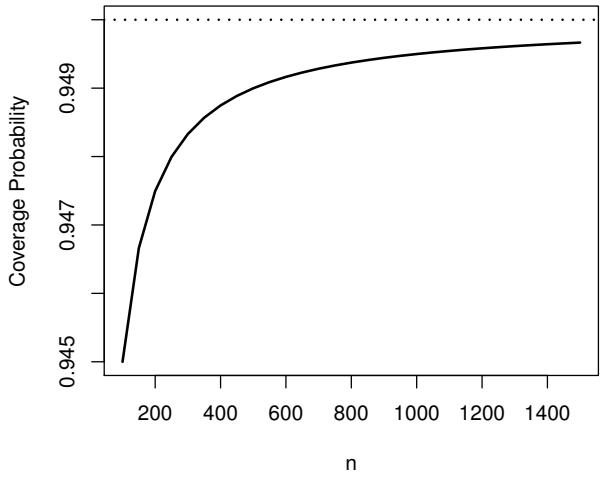


Figure 3.5: Coverage probability $\mathsf{P}_\theta\{\mathcal{C}_{n,\alpha} \ni \theta\}$ in Example 3.8 as a function of n for $\theta = 1$ and $\alpha = 0.05$.

Variance-stabilizing transformations

The second approach to correct for the dependence on θ in (3.7) is based on the idea of transformations. The Delta Theorem provides a way to get the asymptotic distribution of $g(\hat{\theta}_n)$ from the asymptotic distribution of $\hat{\theta}_n$ when g is a nice enough function. In particular, if we can choose a function g such that the asymptotic variance of $g(\hat{\theta}_n)$ is free of θ, then we can construct an asymptotically correct confidence interval for $g(\theta)$. Then by undoing the transformation g, we can get an asymptotically correct confidence interval for θ. The catch, here, is that the interval will no longer be of the nice form $\hat{\theta}_n \pm \text{SD}$ like before.

A function g is said to be *variance-stabilizing* if it satisfies the differential equation $[g'(\theta)]^2 = c^2 I(\theta)$, where $c > 0$ is some constant. If we apply the Delta Theorem now, for variance-stabilizing g, then we get

$$n^{1/2}[g(\hat{\theta}_n) - g(\theta)] \to \mathsf{N}(0, c^2), \quad \text{in distribution.}$$

Notice that the asymptotic variance of $g(\hat{\theta}_n)$ is a constant free of θ; that is, the variance has been "stabilized."

Now an asymptotically correct confidence interval for $g(\theta)$ is $g(\hat{\theta}_n) \pm z_\alpha^\star c n^{-1/2}$. From here, we get a confidence interval for θ:

$$\mathcal{C}_{n,\alpha} = \{\theta : g(\theta) \in g(\hat{\theta}_n) \pm z_\alpha^\star c n^{-1/2}\}.$$

One should choose g such that this interval is, indeed, an interval.

Example 3.9. Let $g(\theta) = \arcsin\sqrt{\theta}$. From calculus you might recall that

$$g'(\theta) = \frac{1}{2}\frac{1}{\sqrt{\theta(1-\theta)}} \quad \Longrightarrow \quad [g'(\theta)]^2 = \frac{1}{4} I(\theta).$$

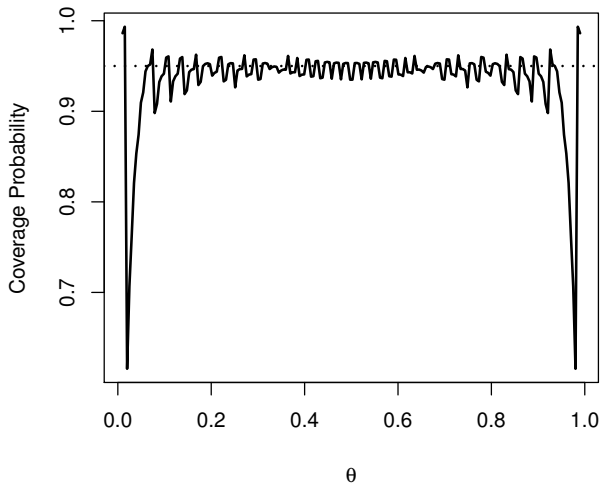


Figure 3.6: Coverage probability $\mathsf{P}_\theta\{\mathcal{C}_{n,\alpha} \ni \theta\}$ in Example 3.9 as a function of θ for $n = 50$ and $\alpha = 0.05$.

Therefore, g is variance-stabilizing with $c = 1/2$, and an asymptotically correct confidence interval for $g(\theta)$ is $\arcsin\sqrt{\bar{X}} \pm z^\star_\alpha(4n)^{-1/2}$. A little work reveals a confidence interval $\mathcal{C}_{n,\alpha}$ for θ:

$$\Big[\sin^2\{\arcsin\sqrt{\bar{X}} + z^\star_\alpha(4n)^{-1/2}\},\ \sin^2\{\arcsin\sqrt{\bar{X}} - z^\star_\alpha(4n)^{-1/2}\}\Big].$$

A plot of the coverage probability, as a function of θ, for $n = 50$ and $\alpha = 0.5$, is shown in Figure 3.6; see Section 3.10.7. There we see the coverage probability is, overall, a bit closer to the target 0.95 compared to that of the plug-in interval above.

Example 3.10. Let $g(\theta) = \log\theta$. Then $g'(\theta) = 1/\theta$, so $[g'(\theta)]^2 = I(\theta)$. Hence g is variance-stabilizing with $c = 1$. So an asymptotically correct confidence interval for $\log\theta$ is $\log\bar{X} \pm z^\star_\alpha n^{-1/2}$. Changing back to the θ scale gives a confidence interval for θ:

$$\mathcal{C}_{n,\alpha} = \Big[\bar{X}e^{-z^\star_\alpha n^{-1/2}}, \bar{X}e^{z^\star_\alpha n^{-1/2}}\Big].$$

A plot of coverage probability as a function of n, for $\theta = 1$ and $\alpha = 0.05$, is shown in Figure 3.7; see Section 3.10.8. We see that it's closer to the target than was the plug-in interval coverage probability, especially for relatively small n.

3.10 Appendix

3.10.1 R code implementation Newton's method

```
newton <- function(f, df, x0, eps=1e-08, maxiter=1000, ...) {
```

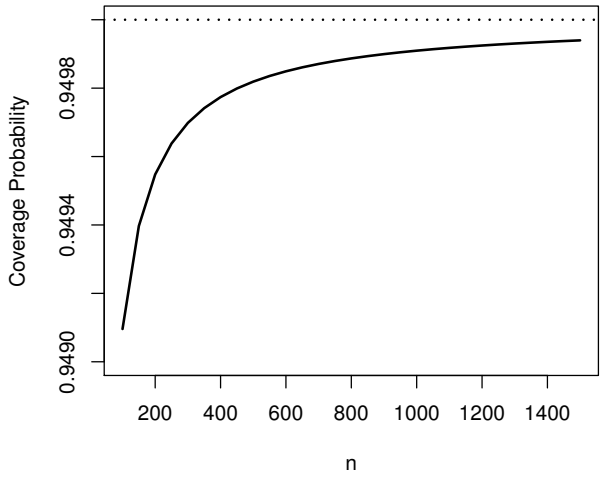


Figure 3.7: Coverage probability $\mathsf{P}_\theta\{\mathcal{C}_{n,\alpha} \ni \theta\}$ in Example 3.10 as a function of n for $\theta = 1$ and $\alpha = 0.05$.

```
  if(!exists("ginv")) library(MASS)
  x <- x0
  t <- 0
  repeat {

    t <- t + 1
    x.new <- x - as.numeric(ginv(df(x, ...)) %*% f(x, ...))
    if(mean(abs(x.new - x)) < eps | t >= maxiter) {

      if(t >= maxiter) warning("Maximum number of iterations reached!")
      break

    }
    x <- x.new

  }
  out <- list(solution=x.new, value=f(x.new, ...), iter=t)
  return(out)

}
```

3.10.2 R code for Example 3.4

```
X <- c(0.27, 0.83, 0.27, 1.52, 0.04, 0.43, 0.92, 0.58, 0.20, 0.32,
       0.82, 0.91, 0.66, 0.01, 0.56, 1.21, 1.44, 0.64, 0.53, 0.30)
dl <- function(theta) length(X) * (mean(X) / theta - 1) / theta
ddl <- function(theta) length(X) * (1 - 2 * mean(X) / theta) / theta**2
exp.newton <- newton(dl, ddl, 0.5)
```

```
print(exp.newton$value)
print(cbind(xbar=mean(X), newton.soln=exp.newton$solution))
```

3.10.3 R code for Example 3.5

```
X <- c(6.37, 12.01, 7.34, 6.28, 7.09, 7.51, 8.24, 7.35, 10.72, 4.95,
       5.14, 6.70, 3.67, 6.35, 9.71, 7.20, 9.21, 7.88, 5.80, 8.27)
n <- length(X)
dl <- function(theta) n - 2 * sum(exp(theta - X) / (1 + exp(theta - X)))
ddl <- function(theta) -2 * sum(exp(theta - X) / (1 + exp(theta - X))**2
logis.newton <- newton(dl, ddl, median(X))
print(logis.newton$solution)
```

3.10.4 R code for Example 3.6

```
X <- c(6.00, 5.98, 7.81, 6.77, 10.64, 13.63, 10.53, 8.92, 3.77, 5.78,
       6.32, 4.44, 5.34, 1.54, 4.42, 4.71, 6.12, 2.57, 9.47, 9.03)
dl <- function(theta) {

  alpha <- theta[1]
  beta <- theta[2]
  n <- length(X)
  o1 <- -n * log(beta) - n * digamma(alpha) + sum(log(X))
  o2 <- -n * alpha / beta + n * mean(X) / beta**2
  return(c(o1, o2))

}
ddl <- function(theta) {

  alpha <- theta[1]
  beta <- theta[2]
  n <- length(X)
  o11 <- -n * trigamma(alpha)
  o12 <- -n / beta
  o22 <- -n * (2 * mean(X) / beta**3 - alpha / beta**2)
  return(matrix(c(o11, o12, o12, o22), 2, 2, byrow=TRUE))

}
theta0 <- c(mean(X)**2 / var(X), var(X) / mean(X))
gamma.newton <- newton(dl, ddl, theta0)
print(cbind(start=theta0, finish=gamma.newton$solution))
```

3.10.5 R code for Example 3.7

```
binom.wald.cvg <- function(theta, n, alpha) {

  z <- qnorm(1 - alpha / 2)
  f <- function(p) {

    t <- 0:n
```

```
    s <- sqrt(t * (n - t) / n)
    o <- (t - z * s <= n * p & t + z * s >= n * p)
    return(sum(o * dbinom(t, size=n, prob=p)))

  }
  out <- sapply(theta, f)
  return(out)

}
n <- 50
alpha <- 0.05
theta <- seq(0.01, 0.99, len=200)
plot(theta, binom.wald.cvg(theta, n, alpha), ylim=c(0.5, 1), type="l", lwd=2,
     xlab=expression(theta), ylab="Coverage Probability")
abline(h=1-alpha, lty=3, lwd=2)
```

3.10.6 R code for Example 3.8

```
expo.wald.cvg <- function(N, alpha) {

  z <- qnorm(1 - alpha / 2)
  theta <- 1
  f <- function(n) {

    f1 <- 1 - pgamma(n * theta / (1 - z / sqrt(n)), shape=n, rate=1/theta)
    f2 <- pgamma(n * theta / (1 + z / sqrt(n)), shape=n, rate=1/theta)
    return(1 - f1 - f2)

  }
  out <- sapply(N, f)
  return(out)

}
alpha <- 0.05
n <- seq(100, 1500, by=50)
plot(n, expo.wald.cvg(n, alpha), ylim=c(0.945, 0.95), type="l", lwd=2,
     xlab="n", ylab="Coverage Probability")
abline(h=1-alpha, lty=3, lwd=2)
```

3.10.7 R code for Example 3.9

```
binom.vst.cvg <- function(theta, n, alpha) {

  z <- qnorm(1 - alpha / 2)
  f <- function(p) {

    t <- 0:n
    a <- asin(sqrt(t / n))
    s <- z / 2 / sqrt(n)
    o <- (a - s <= asin(sqrt(p)) & a + s >= asin(sqrt(p)))
```

```
    return(sum(o * dbinom(t, size=n, prob=p)))

  }
  out <- sapply(theta, f)
  return(out)

}
n <- 50
alpha <- 0.05
theta <- seq(0.01, 0.99, len=200)
plot(theta, binom.vst.cvg(theta, n, alpha), type="l", lwd=2,
     xlab=expression(theta), ylab="Coverage Probability")
abline(h=1-alpha, lty=3, lwd=2)
```

3.10.8 R code for Example 3.10

```
expo.vst.cvg <- function(N, alpha) {

  z <- qnorm(1 - alpha / 2)
  theta <- 1
  f <- function(n) {

    f1 <- 1 - pgamma(n * theta * exp(z / sqrt(n)), shape=n, rate=1/theta)
    f2 <- pgamma(n * theta * exp(-z / sqrt(n)), shape=n, rate=1/theta)
    return(1 - f1 - f2)

  }
  out <- sapply(N, f)
  return(out)

}

alpha <- 0.05
n <- seq(100, 1500, by=50)
plot(n, expo.vst.cvg(n, alpha), ylim=c(0.949, 0.95), type="l", lwd=2,
     xlab="n", ylab="Coverage Probability")
abline(h=1-alpha, lty=3, lwd=2)
```

3.10.9 Interchanging derivatives and sums/integrals

Condition R4 requires that derivatives of integrals/sums can be evaluated by differentiating the integrand/summand. The question of whether these operations can be interchanged has nothing to do with statistics, these are calculus/analysis issues. But, for completeness, I wanted to give a brief explanation of what's going on.

Let $f(x, \theta)$ be a function of two variables, assumed to be differentiable with respect to θ for each x. Here f need not be a PDF/PMF, just a function like in calculus. Let's consider the simplest case: suppose x ranges over a finite set, say, $\{1, 2, \ldots, r\}$. Then it's a trivial

result from calculus that

$$\frac{d}{d\theta}\sum_{x=1}^{r} f(x,\theta) = \sum_{x=1}^{r}\frac{\partial}{\partial\theta}f(x,\theta).$$

This is referred to as the linearity property of differentiation. Similarly, suppose x ranges over a bounded interval $[a,b]$, where neither a nor b depends on θ (this last assumption can easily be relaxed). Then the famous Leibnitz formula gives

$$\frac{d}{d\theta}\int_a^b f(x,\theta)\,dx = \int_a^b \frac{\partial}{\partial\theta}f(x,\theta)\,dx.$$

In these two cases, derivative and sum/integral can be interchanged with essentially no conditions. The common feature of these two situations is that summation/integration is over a bounded range. Things are messier when "infinities" are involved.

Both the summation and integration problems over bounded and unbounded ranges can be lumped together under one umbrella in a measure-theoretic context, and the question of interchange with differentiation can be answered with the Lebesgue Dominated Convergence Theorem. The general details are too technical, so I'll work on the two cases separately.

Start with the summation problem. That is, we want to know when

$$\frac{d}{d\theta}\sum_{x=1}^{\infty} f(x,\theta) = \sum_{x=1}^{\infty}\frac{\partial}{\partial\theta}f(x,\theta). \tag{3.9}$$

The three sufficient conditions are

S1. $\sum_{x=1}^{\infty} f(x,\theta)$ converges for all θ in an interval (a,b);

S2. $\frac{\partial}{\partial\theta}f(x,\theta)$ is continuous in θ for all x;

S3. $\sum_{x=1}^{\infty}\frac{\partial}{\partial\theta}f(x,\theta)$ converges uniformly on every compact subset of (a,b).

That is, if S1–S3 hold, then (3.9) is valid.

In the integration problem, we want to know when

$$\frac{d}{d\theta}\int_{-\infty}^{\infty} f(x,\theta)\,dx = \int_{-\infty}^{\infty}\frac{\partial}{\partial\theta}f(x,\theta)\,dx. \tag{3.10}$$

In this case, there is just one sufficient condition, with two parts. Suppose that there exists a function $g(x,\theta)$ and a number $\delta > 0$ such that

$$\left|\frac{f(x,\theta+\delta') - f(x,\theta)}{\delta'}\right| \le g(x,\theta) \quad \text{for all } x \text{ and all } |\delta'| \le \delta,$$

and

$$\int_{-\infty}^{\infty} g(x,\theta)\,dx < \infty.$$

Then statement (3.10) is valid.

Chapter 4

Sufficiency and Minimum Variance Estimation

4.1 Introduction

We have discussed various properties of point estimators. In particular, we have discussed the idea of mean-square error (MSE) and that it is desirable for an estimator to have small MSE. Unfortunately, it is generally not possible to construct "optimal" estimators that have smallest MSE over all parameter values—typically, given two estimators, the MSE for one estimator will be smaller than that of the other estimator at some parameter values, and larger for other parameter values. Maximum likelihood estimators (MLEs) have a variety of desirable properties, the most striking being that they are *asymptotically efficient*, i.e., MLEs are asymptotically unbiased and their asymptotic variance is the smallest possible among unbiased estimators. This is a really nice result, but one could still ask for some kind of optimality for finite n. No such results are available for MLEs.

In many cases, if one focuses attention strictly to unbiased estimators, there is such an estimator with smallest MSE—smallest variance, in this case—uniformly over all parameter values. Typically, this variance-minimizing estimator will be unique, and we shall call it the *minimum variance unbiased estimator*, or MVUE for short.[1] These MVUEs will be our version of "optimal" estimators. Unbiased estimation is a fundamental development in the theory of statistical inference. Nowadays there is considerably less emphasis on unbiasedness in statistical theory and practice, particularly because there are other more pressing concerns in modern high-dimensional problems (e.g., regularization). Nevertheless, it is important for students of statistics to learn and appreciate these classical developments.

Before we get to this optimality stuff, we must first discuss the fundamentally important notion of *sufficiency*. Besides as a tool for constructing optimal estimators, the notion of sufficiency helps to solidify our understanding of Fisher information.

Along the way, we shall also encounter an important class of distributions, known as the

[1] In some texts, the adjective "uniformly" is appended to the front, making the acronym UMVUE, to emphasize that the minimum variance is *uniform* over the parameter values; see Definition 4.2.

exponential family. Distributions in this class are "nice" for a variety of reasons. First, all those annoying regularity conditions (R0–R5) in Chapter 3 hold for distributions belonging to the exponential family. Second, for each distribution in this family, construction of optimal estimators is relatively straightforward. Finally, the exponential family is quite broad—essentially all the distributions we encounter here are members—so we automatically get nice properties of estimators, etc in a wide range of examples from one general result.

4.2 Sufficiency

4.2.1 Intuition

Suppose we have collected a sample $X_1, \ldots, X_n$ and stored the observed values in the computer. Unfortunately, your laptop crashed and you lost the file. Is there some way to produce an "equivalent" set of data, say, $\tilde{X}_1, \ldots, \tilde{X}_n$, without redoing the experiment? In particular, is there some characteristic of the X-sample such that knowledge of this characteristic would allow us to simulate a $\tilde{X}$-sample with the same basic properties? Of course, the values in the X- and $\tilde{X}$-samples would be different, but we want both samples to contain the same amount of *information* about the underlying parameter of interest.

The basic idea is that, if we remember the observed value t of a "sufficient" statistic T, calculated from the X-sample, then, at least in principle, it is possible to generate a $\tilde{X}$-sample from the conditional distribution of $(X_1, \ldots, X_n)$ given $T = t$, and this $\tilde{X}$-sample will contain the same information about the parameter θ as did the original X-sample.

The idea is not to prepare ourselves for crashing computers, though it was useful in pre-computer times. Instead, the importance is that we should always base our estimators on the "sufficient" statistic T, since, otherwise, we're wasting some of the information in the sample. This fundamental idea is due to Fisher, the same guy the Fisher information is named after. The technical point being that the Fisher information based on the full sample $X_1, \ldots, X_n$ is exactly the same as that based on the reduced sample T *if and only if* T is a "sufficient" statistic.

4.2.2 Definition

Recall the usual setup: $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f_\theta(x)$, where θ is the unknown parameter of interest, and $f_\theta(x)$ is the PDF/PMF that describes the distribution. The formal definition of a sufficient statistic is as follows.

Definition 4.1. A statistic $T = T(X_1, \ldots, X_n)$ is sufficient for θ (or for f_θ) if the joint conditional distribution of $(X_1, \ldots, X_n)$, given $T = t$, does not depend on θ.

To relate this to the intuition from before, note that if T is sufficient for θ, then the $\tilde{X}$-sample can be simulated from the conditional distribution of $(X_1, \ldots, X_n)$, given $T = t$. If T were not sufficient, then one could not simulate from this distribution because it would depend on the unknown parameter θ.

Exercise 4.1. Are sufficient statistics unique?

A technical point. We will not be careful about the definition of conditional distributions in Stat 411. Complete rigor requires concepts of measure-theoretic probability which is beyond the prerequisites for the course. So, in continuous problems, our calculations will be somewhat sloppy; when the problem is discrete, there's no measure-theoretic challenges to overcome, so our solutions are perfectly fine.

Exercise 4.2. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Ber}(\theta)$. Show that $T = \sum_{i=1}^n X_i$ is a sufficient statistic.

Exercise 4.3. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Pois}(\theta)$. Show that $T = \sum_{i=1}^n X_i$ is a sufficient statistic.

Example 4.1. Let $X_1, X_2 \overset{\text{iid}}{\sim} \mathsf{N}(\theta, 1)$, and let $T = X_1 + X_2$. We know know that $T \sim \mathsf{N}(2\theta, 2)$. Let's figure out the conditional distribution of (X_1, X_2) given $T = t$. Here, for this continuous problem, we do a bit of fudging with the conditional distribution.

$$\begin{aligned}
\text{“}\mathsf{P}_\theta(X_1 = x_1, X_2 = x_2 \mid T = t)\text{”} &= \frac{\mathsf{P}_\theta(X_1 = x_1, X_2 = x_2, T = t)}{\mathsf{P}_\theta(T = t)} \\
&= \frac{\mathsf{P}_\theta(X_1 = x_1, X_2 = t - x_1)}{\mathsf{P}_\theta(T = t)} \\
&= \frac{\mathsf{P}_\theta(X_1 = x_1)\mathsf{P}_\theta(X_2 = t - x_1)}{\mathsf{P}_\theta(T = t)} \\
&= \frac{\frac{1}{2\pi}\exp\left\{-\frac{1}{2}[(x_1 - \theta)^2 + (t - x_1 - \theta)^2]\right\}}{\frac{1}{2\sqrt{\pi}}\exp\left\{-\frac{1}{4}(t - 2\theta)^2\right\}}.
\end{aligned}$$

This last expression is messy but, fortunately, we don't need to evaluate it. A little extra work shows that this expression is free of θ. Therefore, the conditional distribution of (X_1, X_2), given $T = t$, does not depend on θ, so T is a sufficient statistic for θ. Note, also, that there is nothing special about $n = 2$; the claim for general n, i.e., $T = \sum_{i=1}^n X_i$ is a sufficient statistic for θ based on a sample of size n.

An important remark. We use the terminology "T is a sufficient statistic for θ" even though this is a bit misleading. For example, suppose we are interested not in θ but, say, e^θ—do we need to consider finding a sufficient statistic for e^θ? The answer is NO. When we say that a statistic T is sufficient, we mean that T contains all the information in data relevant to distinguishing between f_θ and $f_{\theta'}$. So, if T is sufficient for θ, then it's also sufficient for e^θ or any other function of θ. The same remark goes for minimal sufficiency and completeness, discussed below. You will want to remember this point when we get to finding MVUEs for functions of θ in Sections 4.4–4.5. In summary, we should actually say "T is sufficient for the family $\{f_\theta : \theta \in \Theta\}$," to make clear that it's not specific to a particular parametrization, but it is not customary to do so.

It is apparent from the exercises, example, and discussion above that the definition is not so convenient for finding a sufficient statistic. It would be nice if there was a trick that allowed us to not only quickly check if something is sufficient but also help to find a candidate statistic in the first place. Fortunately, there is such a result...

4.2.3 Neyman–Fisher factorization theorem

It turns out that sufficient statistics can be found from an almost immediate inspection of the likelihood function. Recall that, in our iid setting, the likelihood function is

$$L(\theta) = \prod_{i=1}^{n} f_\theta(X_i),$$

with notations $L_X(\theta)$ or $L_n(\theta)$ to indicate, where needed, that the likelihood depends also on $X = (X_1, \ldots, X_n)$ and n.

Theorem 4.1 (Neyman–Fisher)**.** *Let $X_1, \ldots, X_n \overset{iid}{\sim} f_\theta(x)$. A statistic $T = T(X_1, \ldots, X_n)$ is sufficient for θ if and only if there are functions K_1 and K_2 such that*

$$L(\theta) = K_1(T(X_1, \ldots, X_n); \theta) K_2(X_1, \ldots, X_n). \tag{4.1}$$

Here K_1 depends on θ but K_2 does not.

Proof. Idea is simple, but, to be rigorous, one must be careful in dealing with conditional distributions; we won't be careful here. Note that $K_1(T; \theta)$ is (essentially) just the PDF/PMF of T, so, in the conditional distribution ratio, the parts with θ cancel out. □

Alternatively: T is a sufficient statistic if and only if the *shape* of the likelihood function depends on $(X_1, \ldots, X_n)$ only through the value of $T = T(X_1, \ldots, X_n)$.

Example 4.2. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(\theta, 1)$. As we have seen previously, the likelihood function can be written as

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-(X_i-\theta)^2/2} = \Big(\frac{1}{\sqrt{2\pi}}\Big)^n e^{-\sum_{i=1}^n (X_i-\theta)^2/2} \\ &= \underbrace{(2\pi)^{-n/2} e^{-\sum_{i=1}^n X_i^2/2}}_{K_2(X_1,\ldots,X_n} \cdot \underbrace{e^{\theta \sum_{i=1}^n X_i - \theta^2/2}}_{K_1(\sum_{i=1}^n X_i; \theta)}. \end{aligned}$$

Therefore, by the factorization theorem, $T = \sum_{i=1}^n X_i$ is sufficient.

Note that a sufficient statistic in this problem is $T = \sum_{i=1}^n X_i$, and we know that the MLE is $\bar{X} = T/n$. It is not a coincidence that the MLE is a function of the sufficient statistic. In fact, in many cases, we can take the MLE as the sufficient statistic.

Exercise 4.4. Redo the sufficiency calculations in the $\mathsf{Ber}(\theta)$ and $\mathsf{Pois}(\theta)$ examples above using the Neyman–Fisher factorization theorem.

Exercise 4.5. Use the Neyman–Fisher factorization theorem to find a sufficient statistic T if (a) $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Unif}(0, \theta)$, and (b) if $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(0, \theta)$, with $\theta > 0$ the variance.

4.3 Minimum variance unbiased estimators

As we discussed earlier in the course, unbiasedness is a good property for an estimator to have; however, I gave several examples which showed that unbiasedness does not necessarily make an estimator good. But if one insists on an unbiased estimator of θ, then it is natural to seek out the best among them. Here "best" means smallest variance. That is, the goal is to find the *minimum variance unbiased estimator* (MVUE) of θ.

Definition 4.2. An estimator $\hat{\theta}$ is the minimum variance unbiased estimator (MVUE) if, for all θ, $\mathsf{E}_\theta(\hat{\theta}) = \theta$ and $\mathsf{V}_\theta(\hat{\theta}) \leq \mathsf{V}_\theta(\tilde{\theta})$ for any other unbiased estimator $\tilde{\theta}$.

First, note that $\hat{\theta}$ in the definition is called *the* MVUE. This choice of word suggests that MVUEs are unique. Let's check this now. Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be two MVUEs; that is, both are unbiased and both have smallest possible variance, say v. Define a new estimator $\hat{\theta}_3 = (\hat{\theta}_1 + \hat{\theta}_2)/2$. Clear $\hat{\theta}_3$ is also unbiased. Its variance is

$$\mathsf{V}_\theta(\hat{\theta}_3) = \mathsf{V}_\theta(\hat{\theta}_1)/4 + \mathsf{V}_\theta(\hat{\theta}_2)/4 + 2\mathsf{C}_\theta(\hat{\theta}_1, \hat{\theta}_2)/4 = v/2 + \mathsf{C}_\theta(\hat{\theta}_1, \hat{\theta}_2)/2.$$

Since $\mathsf{V}_\theta(\hat{\theta}_3) \geq v$, it follows that $\mathsf{C}_\theta(\hat{\theta}_1, \hat{\theta}_2) \geq v$. Now consider the difference $\hat{\theta}_1 - \hat{\theta}_2$:

$$\mathsf{V}_\theta(\hat{\theta}_1 - \hat{\theta}_2) = \mathsf{V}_\theta(\hat{\theta}_1) + \mathsf{V}_\theta(\hat{\theta}_2) - 2\mathsf{C}_\theta(\hat{\theta}_1, \hat{\theta}_2) \leq 2v - 2v = 0.$$

The only random variable with zero variance is a constant; therefore, $\hat{\theta}_1 - \hat{\theta}_2$ must be a constant, and the constant must be zero, so $\hat{\theta}_1 = \hat{\theta}_2$. Indeed, there is only one MVUE.

Second, let's recall what we know about the variance of unbiased estimators. The Cramer–Rao lower bound tells us that, if $\hat{\theta}$ is an unbiased estimator of θ, then its variance cannot be less than $[nI(\theta)]^{-1}$, where $I(\theta)$ is the Fisher information. So, if we can find an estimator with variance equal to the Cramer–Rao lower bound, then we know that no other unbiased estimator can beat it. However, there are two things to consider here.

- In some problems, it may be too difficult to find an exact formula for the variance of a candidate unbiased estimator. In that case, it is not possible to check if the estimator is efficient.
- It can happen that the smallest possible variance of an unbiased estimator is strictly bigger than the Cramer–Rao lower bound, i.e., the lower bound is not always attainable. In such cases, there can be an MVUE but its efficiency is not equal to 1, so our old method for ranking estimators won't work.

In light of these challenges, it would be desirable to find techniques or special results that will allow us to identify MVUEs without explicitly calculating variances, efficiencies, etc. The next two sections describe two powerful techniques.

4.4 Rao–Blackwell theorem

Based on the calculations above, we see some connection between sufficient statistics and the "good" estimators (e.g., MLEs) we are familiar with from the previous notes. Is this a coincidence? The answer is NO—it turns out that, in some sense, only functions of sufficient statistics are good estimators.

The first result, the Rao–Blackwell theorem, makes this claim precise. Specifically, if you give me an unbiased estimator, then the Rao–Blackwell theorem tells me how I can improve upon the one you gave me, i.e., maintain the unbiasedness property, but potentially reduce the variance.

Theorem 4.2 (Rao–Blackwell)**.** *Let T be a sufficient statistic and $\tilde{\theta}$ an unbiased estimator of θ. Define the function $g(t) = \mathsf{E}(\tilde{\theta} \mid T = t)$, the conditional expected value of $\tilde{\theta}$ given the value t of the sufficient statistic T. Then $\hat{\theta} = g(T)$ is an unbiased estimator and $\mathsf{V}_\theta(\hat{\theta}) \leq \mathsf{V}_\theta(\tilde{\theta})$ for all θ, with equality if and only if $\tilde{\theta}$ and $\hat{\theta}$ are the same.*

Proof. Write $\varphi(a) = (a - \theta)^2$. This function is a parabola opening up and, therefore, is clearly convex (concave up). It follows from Jensen's inequality that

$$\varphi(\mathsf{E}_\theta(\tilde{\theta} \mid T = t)) \leq \mathsf{E}_\theta[\varphi(\tilde{\theta}) \mid T = t],$$

or, equivalently, $[\mathsf{E}_\theta(\tilde{\theta} \mid T = t) - \theta]^2 \leq \mathsf{E}_\theta[(\tilde{\theta} - \theta)^2 \mid T = t]$. Replacing t with it's random variable version T gives

$$[\mathsf{E}_\theta(\tilde{\theta} \mid T) - \theta]^2 \leq \mathsf{E}_\theta[(\tilde{\theta} - \theta)^2 \mid T],$$

with equality if and only if $\tilde{\theta} = \mathsf{E}_\theta(\tilde{\theta} \mid T)$, i.e., if $\tilde{\theta}$ is itself a function of T. Taking expected value on both sides gives $\mathsf{V}_\theta(\hat{\theta}) \leq \mathsf{V}_\theta(\tilde{\theta})$, completing the proof. □

The proof given here is different from the one using iterated expectation given in Theorem 2.3.2 in HMC; see, also, the exercise below. The two proofs are equivalent in the case considered here. However, if one wants to change from comparing estimators based on variance to comparing based on $\mathsf{E}_\theta\{d(\hat{\theta} - \theta)\}$ for a convex function d, then the proof based on Jensen's inequality is easier to modify.

Exercise 4.6. Let X, Y be random variables with finite variances.

(a) Show that $\mathsf{E}(Y) = \mathsf{E}\{\mathsf{E}(Y \mid X)\}$.

(b) Show that $\mathsf{V}(Y) = \mathsf{E}\{\mathsf{V}(Y|X)\} + \mathsf{V}\{\mathsf{E}(Y|X)\}$.

(c) Let $X = T$ and $Y = \tilde{\theta}$. Use the two results above to prove Theorem 4.2.

The utility of the Rao–Blackwell theorem is that one can take a "naive" unbiased estimator and improve upon it by taking conditional expectation given a sufficient statistic. This has two important consequences. First, it says that if one cares about unbiasedness, then there is no reason to consider an estimator that is not a function of the sufficient statistic. Second, it gives an explicit recipe for improving upon a given unbiased estimator; this is

different/better than typical existence theorems that say "there exists a better estimator..." but do not explain how to find it.

To apply the Rao–Blackwell theorem, one should try to find the simplest possible unbiased estimator to start with. That way, the conditional expectation will be relatively easy to calculate. For example, it can help to use only a subset of the original sample. The examples that follow should help illustrate the technique.

Example 4.3. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(\theta, 1)$. The goal is to estimate θ. We know that $\hat{\theta} = \bar{X}$ is a good estimator. Let's try to apply the Rao–Blackwell theorem. Start with a naive unbiased estimator: $\tilde{\theta} = X_1$ is a reasonable choice. We know, by the factorization theorem, that $T = \sum_{i=1}^n X_i$ is a sufficient statistic. Therefore, we need to calculate $g(t) = \mathsf{E}\{X_1 \mid T = t\}$. But if I tell you that the sum of iid random variables is t, then, at least on average, each X_1 contributes equally to the total. Therefore, the expected value of X_1 given $T = t$ must be $g(t) = t/n$. It follows from the Rao–Blackwell theorem that $g(T) = \bar{X}$ is a better estimator than X_1 in terms of variance. We already knew this, but note that here we didn't need to do any variance calculations to confirm it.

Exercise 4.7. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(\theta, 1)$ and $T = \sum_{i=1}^n X_i$. Show that the conditional distribution of X_1, given $T = t$, is $\mathsf{N}(\frac{t}{n}, \frac{n-1}{n})$.

Example 4.4. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Pois}(\theta)$. The goal is to estimate $\eta = e^{-\theta}$. Here's an example we've encountered before. We know that the MLE $e^{-\bar{X}}$ is not unbiased, but we don't know how to construct a good unbiased estimator. The Rao–Blackwell theorem tells us how to do it. We need first a naive unbiased estimator of η. Note that $\eta = \mathsf{P}_\theta(X_1 = 0)$, so a good choice is $\tilde{\eta} = I_0(X_1)$, i.e., $\tilde{\eta} = 1$ if $X_1 = 0$, and $\tilde{\eta} = 0$ otherwise. Clearly $\mathsf{E}(\tilde{\eta}) = \eta$. We also know from the factorization theorem that $T = \sum_{i=1}^n X_i$ is a sufficient statistic. Let's calculate the conditional expectation $g(t)$ directly:

$$\begin{aligned}
\mathsf{E}_\theta(I_0(X_1) \mid T = t) &= \mathsf{P}_\theta(X_1 = 0 \mid T = t) \\
&= \frac{\mathsf{P}_\theta(X_1 = 0, T = t)}{\mathsf{P}_\theta(T = t)} \\
&= \frac{\mathsf{P}_\theta(X_1 = 0, X_2 + \cdots + X_n = t)}{\mathsf{P}_\theta(T = t)} \\
&= \frac{\mathsf{P}_\theta(X_1 = 0)\mathsf{P}_\theta(X_2 + \cdots + X_n = t)}{\mathsf{P}_\theta(T = t)}.
\end{aligned}$$

If we remember that $T \sim \mathsf{Pois}(n\theta)$ and $X_2 + \cdots + X_n \sim \mathsf{Pois}((n-1)\theta)$, both derived using moment-generating functions, then we can plug in Poisson PMF formulas in the numerator and denominator. This gives

$$\mathsf{E}_\theta(I_0(X_1) \mid T = t) = \frac{e^{-\theta} e^{-(n-1)\theta}[(n-1)\theta]^t}{e^{-n\theta}[n\theta]^t} = \Big(1 - \frac{1}{n}\Big)^t.$$

Therefore, $\hat{\eta} = (1 - \frac{1}{n})^T$ is, by the Rao–Blackwell theorem, unbiased and has smaller variance than $\tilde{\eta} = I_0(X_1)$. It will be seen later that $\hat{\eta}$ is actually the MVUE for $\eta = e^{-\theta}$.

To re-iterate, the important aspect of the Rao–Blackwell theorem is that it gives a particular recipe for improving a given, or "naive," unbiased estimator by making it a function of a sufficient statistic. As a result, we find that there is no reason to consider unbiased estimators that are not functions of a sufficient statistic since conditioning on a sufficient statistic leads to an estimator that is no worse in terms of variance.

The down-side, however, to the Rao–Blackwell theorem is that the conditional expectation can be difficult to calculate, even in relatively simple problems. So it would be nice if there was a special feature of the problem which, if we can see it, will tell us if an estimator is the MVUE. There is such a special feature but this too can be tricky.

4.5 Completeness and Lehmann–Scheffe theorem

To be able to jump directly to MVUE conclusions without appealing to the Rao–Blackwell theorem and conditional expectation calculations, we need something more than just sufficiency of T. The new additional property is called completeness.

Definition 4.3. A statistic T is complete if $\mathsf{E}_\theta\{h(T)\} = 0$ for all θ implies the function $h(t)$ is (almost everywhere) the zero-function.

The definition of completeness in HMC is a bit different, but equivalent for us. They present completeness as a property of a family of distributions. Here, we are just taking that family to be the sampling distribution of T as θ varies over Θ.

Completeness is a tricky property, and checking it often requires some very special techniques. Before we look at examples, let's see why we should care about completeness.

Theorem 4.3 (Lehmann–Scheffe)**.** *Let T be a complete sufficient statistic for θ. If there is a function g such that $\hat{\theta} = g(T)$ is an unbiased estimator of θ, then $\hat{\theta}$ is the MVUE.*

Proof. See Section 4.9.2. □

The Lehmann–Scheffe theorem goes a bit further than Rao–Blackwell in the sense that it gives a sufficient condition for finding the unique unbiased estimator with smallest variance. The key ingredient is that the sufficient statistic must also be complete. In other words, if we can find an unbiased estimator that is a function of a complete sufficient statistic, then we know we have found the MVUE—no conditional expectation calculations are necessary.

The definition of completeness is "mathy," what does it mean intuitively? If a statistic T is sufficient, then we understand that it contains all the information about θ in the original sample. However, T may not summarize the information efficiently, e.g., if $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(\theta, 1)$, then $T = (\bar{X}, X_1)$ is also sufficient. Somehow, the statistic $T = (\bar{X}, X_1)$ contains *redundant* information. The intuitive meaning of completeness, together with sufficiency, is that the statistic contains *exactly* all the information in the original data concerning θ—no more and no less.

Exercise 4.8. For each distribution, show that T is complete and find the MVUE.

1. $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Pois}(\theta)$; $T = \sum_{i=1}^n X_i$.
2. $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Unif}(0, \theta)$; $T = X_{(n)}$.
3. $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Ber}(\theta)$; $T = \sum_{i=1}^n X_i$.

Exercise 4.9. $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Ber}(\theta)$. The statistic $T = \sum_{i=1}^n X_i$ is complete and sufficient. Use the Rao–Blackwell theorem to find the MVUE of $\mathsf{V}_\theta(X_1) = \theta(1-\theta)$.

Exercise 4.10. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(\theta, \theta^2)$, $\theta \in \mathbb{R}$.

1. Use the factorization theorem to show that $T = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ is a sufficient statistic for θ.
2. Show that T is *not* complete.

The down-side to the Lehmann–Scheffe theorem is that it is not too easy to show that a statistic is complete. We saw that special techniques and results (e.g., Laplace transforms, power series, properties of solutions to polynomials) are needed, and there is no general way to know ahead of time what techniques are needed. It would be great if there was a broad class of distributions for which there is a simple way to identify a complete sufficient statistic. Fortunately, such a class exists, and it is broad enough to contain almost all the distributions we consider here.

4.6 Exponential families

The exponential family is a broad class of distributions—both discrete and continuous—that has many nice properties. Most of the distributions we encounter here belong to this class.

Definition 4.4. Let $f_\theta(x)$ be a PDF/PMF with support $\mathcal{S} = \{x : f_\theta(x) > 0\}$ and parameter space Θ an open interval in $\mathbb{R}$. Then $f_\theta(x)$ is a (regular) exponential family if

$$f_\theta(x) = \exp\{p(\theta)K(x) + S(x) + q(\theta\} \cdot I_{\mathcal{S}}(x), \tag{4.2}$$

with the following properties:

(i) $\mathcal{S}$ does not depend on θ,

(ii) $p(\theta)$ is non-constant and continuous, and

(iii) if X is continuous, then K' and S are continuous functions; if X is discrete, then K is non-constant.

Exercise 4.11. Show that $\mathsf{Ber}(\theta)$, $\mathsf{Pois}(\theta)$, $\mathsf{Exp}(\theta)$, and $\mathsf{N}(\theta, 1)$ are exponential families.

It can be shown that the regularity conditions R0–R5 from our study of sampling distributions of MLEs hold for (regular) exponential families. Therefore, we know, for example, in exponential families, MLEs are asymptotically efficient. Here are two more important properties of exponential families.

- If $X \sim f_\theta(x)$ and $f_\theta(x)$ is an exponential family, then

$$\mathsf{E}_\theta[K(X)] = -\frac{q'(\theta)}{p'(\theta)} \quad \text{and} \quad \mathsf{V}_\theta[K(X)] = \frac{p''(\theta)q'(\theta) - p'(\theta)q''(\theta)}{p'(\theta)^3}.$$

  To prove this, differentiate under the integral sign, like how we used R4 before.

- If $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f_\theta(x)$ and $f_\theta(x)$ is an exponential family, then $T = \sum_{i=1}^n K(X_i)$ has distribution $g_\theta(t)$ an exponential family, with

$$g_\theta(t) = \exp\{p(\theta)t + R(t) + nq(\theta)\}$$

  and p, q are the same as for $f_\theta(x)$. The easiest way to prove this is via moment-generating functions; see Exercise 4.13.

Exercise 4.12. Verify the general formula for $\mathsf{E}_\theta[K(X)]$ for each of the four exponential family distributions in Exercise 4.11.

Exercise 4.13. For an exponential family distribution $f_\theta(x)$ as in (4.2), find the moment-generating function $M_\theta(s) = \mathsf{E}_\theta(e^{sX})$.

Besides that exponential families are so common and have a lot of nice properties, they are particularly convenient in the context of sufficient statistics, etc. In particular, (regular) exponential families admit an easy-to-find complete sufficient statistic.

Theorem 4.4. *Let $X_1, \ldots, X_n \overset{iid}{\sim} f_\theta(x)$ where $f_\theta(x)$ is a (regular) exponential family distribution* (4.2)*. Then $T = \sum_{i=1}^n K(X_i)$ is a complete sufficient statistic for θ.*

Theorem 4.4 shows how one can find a complete sufficient statistic in (regular) exponential families. Therefore, in a particular example, if we recognize that the distribution in question is an exponential family, then we can bypass the difficult step of verifying that the sufficient statistic is complete, saving ourselves time and energy. Moreover, once we find our complete sufficient statistic T, the Lehmann–Scheffe theorem says as soon as we find an unbiased estimator that's a function of T, then we've found the MVUE.

Exercise 4.14. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(0, \theta)$, where $\theta > 0$ is the variance. Use the results of this section to find the MVUE of θ.

4.7 Multi-parameter cases

Here we shall discuss extension of the main ideas in the previous sections to the case where $\theta = (\theta_1, \ldots, \theta_d)^\top$ is a d-dimensional (column) vector, for $d \geq 1$. That is, we now assume Θ is a subset of $\mathbb{R}^d$, the d-dimensional Euclidean space. Conceptually everything stays the same here as before. In this more general case, the notation and technical details are more tedious.

It helps to start out by recalling some important multi-parameter problems. The most important is probably $\mathsf{N}(\theta_1, \theta_2)$ where θ_1 is the mean and θ_2 is the variance. Other problems include $\mathsf{Unif}(\theta_1, \theta_2)$, $\mathsf{Beta}(\theta_1, \theta_2)$, $\mathsf{Gamma}(\theta_1, \theta_2)$, etc.

The first important result presented above for the one-parameter case was sufficiency. One can define sufficiency very much like in Definition 4.1. The one key difference is the addition of an adjective "jointly." That is, we say a statistic $T = (T_1, \ldots, T_m)^\top$ is *jointly sufficient* for $\theta = (\theta_1, \ldots, \theta_d)^\top$ if the conditional distribution of $(X_1, \ldots, X_n)$, given the observed vector $T = t$, is free of θ. Note that the number of coordinates in T may be different from d, the number of parameters. Moreover, even if $m = d$, it may not be true that T_1 is sufficient for θ_1, T_2 sufficient for θ_2, etc. In general, one needs the whole of T to summarize all the information in $(X_1, \ldots, X_n)$ about θ.

Like before, the definition of sufficient statistics is not particularly convenient for finding them. We have a natural extension of the Neyman–Fisher factorization theorem, which goes as follows. A statistic $T = (T_1, \ldots, T_m)^\top$ is (jointly) sufficient for $\theta = (\theta_1, \ldots, \theta_d)^\top$ if and only if there exists functions K_1 and K_2 such that the likelihood factors as

$$L(\theta) = K_1(T(X_1, \ldots, X_n); \theta) K_2(X_1, \ldots, X_n).$$

Exercise 4.15. Find a (jointly) sufficient statistic T for the two problems below:

1. $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(\theta_1, \theta_2)$;
2. $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Beta}(\theta_1, \theta_2)$, where $f_\theta(x) \propto x^{\theta_1 - 1}(1 - x)^{\theta_2 - 1}$.

The two key theorems on MVUEs, namely, the Rao–Blackwell and Lehmann–Scheffe theorems, to not depend at all on the dimension of the parameter or statistic (though the proofs given here focus on the one-dimensional cases). Anyway, there are versions of such theorems for multi-parameter cases. For example, we can say that if T is a (jointly) complete and sufficient statistic for θ and $g(T)$ is an unbiased estimator of $\eta = h(\theta)$, then $\hat{\eta} = g(T)$ is the MVUE of η.

Example 4.5. Here is an example of a sort of nonparametric estimation problem. Let $X_1, \ldots, X_n$ be iid with common CDF $F(x)$, and corresponding PDF $f(x) = F'(x)$. Here the CDF $F(x)$ is completely unknown. The likelihood function is

$$L(F) = \prod_{i=1}^{n} f(X_i) = \prod_{i=1}^{n} f(X_{(i)}).$$

Therefore, $T = (X_{(1)}, \ldots, X_{(n)})$ is a (jointly) sufficient statistic for F; note that no dimension reduction is possible in a nonparametric problem. It turns out that T is also complete, but I will not show this calculation. Now consider estimating $\theta = F(c)$ for a fixed constant c. Towards finding an MVUE, consider the "naive" unbiased estimator $\tilde{\theta} = I_{(-\infty, c]}(X_1)$. The conditional distribution of X_1, given $T = t$ is

$$X_1 \mid T = t \sim \mathsf{Unif}\{t_1, \ldots, t_n\}, \quad t_i = \text{observed } X_{(i)}.$$

Now apply the Rao–Blackwell theorem to get an improved estimator:

$$\hat{\theta} = \mathsf{E}\{I_{(-\infty,c]}(X_1) \mid T\} = \mathsf{P}\{X_1 \leq c \mid T\} = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,c]}(X_{(i)}).$$

This is the sample proportion of observations less than or equal to c, an intuitively natural choice. Because $\hat{\theta}$ is unbiased and a function of the complete sufficient statistic T, it follows from the Lehmann–Scheffe theorem that $\hat{\theta}$ is the MVUE of $\theta = F(c)$.

Just like in the one-parameter case, verifying that a (jointly) sufficient statistic is also complete is a difficult task. Fortunately, there is a large and convenient class of distributions for which we can immediately find a (jointly) complete sufficient statistic, directly from the likelihood function. This is the multi-parameter exponential family of distributions.

To be consistent with HMC, I will now use m for the dimension of θ. The distribution $f_\theta(x)$, $\theta \in \Theta \subseteq \mathbb{R}^m$, is an exponential family if

$$f_\theta(x) = \exp\{\textstyle\sum_{j=1}^{m} p_j(\theta)K_j(x) + S(x) + q(\theta)\}, \quad x \in \mathcal{S}.$$

We say that $f_\theta(x)$ is a *regular* exponential family if $\mathcal{S}$ is free of θ, Θ contains an open rectangle, and the $p_j(\theta)$'s are continuous and linearly independent.[2]

We were not careful about the distinction between exponential families and regular exponential families in the one-parameter case. Here, however, we need to be more careful. There are examples of multi-parameter exponential families which are not regular. One such example is $\mathsf{N}(\theta, \theta^2)$. This is essentially a two-parameter problem, but there is a link between the two parameters. This fails to be regular because the effective parameter space is $\Theta = \{(\theta, \theta^2) : \theta \in \mathbb{R}\}$ which is a curve in $\mathbb{R}^2$ and, therefore, does not contain an open rectangle. Hence $\mathsf{N}(\theta, \theta^2)$ is an exponential family, but it's not regular. This is one example of a general class called *curved exponential families*.

Exercise 4.16. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(\theta, \theta^2)$. Verify that this is a two-parameter exponential family and find a (jointly) sufficient statistic.

Many distributions we are familiar with belong to the (regular) exponential family. An important example is $\mathsf{N}(\theta_1, \theta_2)$. A key result is an extension of Theorem 4.4 to multi-parameter problems. In particular, in regular exponential families, the statistic $T = (T_1, \ldots, T_m)^\top$, with $T_j = \sum_{i=1}^{n} K_j(X_i)$, $j = 1, \ldots, m$, is a (jointly) complete sufficient statistic for θ. Here it is important to note that T and θ have the same dimension, namely, m. A somewhat surprising fact is that this efficient reduction of data—a m-dimensional complete sufficient statistic for a m-dimensional parameter—is available only for regular exponential families. Compare this to the calculation in the curved exponential family example mentioned above.

Exercise 4.17. For $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Beta}(\theta_1, \theta_2)$, find a (jointly) complete sufficient statistic for (θ_1, θ_2).

Exercise 4.18. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(\theta_1, \theta_2)$. Find the MVUE of (θ_1, θ_2).

[2]Linear independence of functions means that there is no pair (j, k) such that $c_j p_j(\theta) + c_k p_k(\theta) \equiv 0$ for constants (c_j, c_k); this is similar to the notion of linearly independent vectors in linear algebra.

4.8 Minimal sufficiency and ancillarity

We have seen that sufficient statistics provide a summary of the full data, but there are lots of such summaries; for example, $(X_1, \ldots, X_n)$ is always a sufficient statistic. When a complete sufficient statistic is available, then we know we have the most efficient summary of data—a complete sufficient statistic contains exactly the information in data about parameter, no more and no less. But there are cases where a perfectly efficient summary is not available. One example is the curved exponential family $\mathsf{N}(\theta, \theta^2)$ described above. Other examples include a Student-t model with known degrees of freedom but unknown mean. In such cases, there is no perfectly efficient summary, but perhaps we can describe what is the most efficient summary. This is the concept of *minimal sufficient statistics*.

Definition 4.5. A sufficient statistic T is minimal if, for any other sufficient statistic T', there is a function h such that $T = h(T')$.

In less mathematical terms, a statistic T is minimal sufficient if knowing the value of any other sufficient statistic T' is enough to know the value of T. It is not immediately clear how to find a minimal sufficient statistic, in general. For us in Stat 411, here are some convenient rules of thumb:

- If T is complete and sufficient, then it's minimal (converse is *false*);
- If the MLE $\hat{\theta}$ exists and is both unique and sufficient, then it's minimal sufficient.

Exercise 4.19. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Unif}(-\theta, \theta)$, $\theta > 0$. Show that $\hat{\theta} = \max\{-X_{(1)}, X_{(n)}\}$, the MLE, is a minimal sufficient statistic.

Example 4.6. Consider the model $X_i = \theta + Z_i$ where $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} f(z)$ and $f(z)$ is a known PDF. If $f(z)$ is a standard normal PDF, then the MLE $\hat{\theta} = \bar{X}$ is both complete and sufficient; therefore, it's also minimal sufficient. If $f(z) = e^{-z}/(1 + e^{-z})^2$, a logistic distribution, then the MLE exists (and can be computed numerically) but it is not sufficient; therefore, it's not minimal sufficient.

On one end of the spectrum are sufficient statistics, functions of $X_1, \ldots, X_n$ that retain all the available information about θ. On the other end are functions that retain none of the information about θ. At first sight, these functions would seem to be useless for statistics; however, that is not the case. First a definition.

Definition 4.6. A statistic $U = U(X_1, \ldots, X_n)$ is called *ancillary* if its sampling distribution does not depend on the parameter θ.

In words, we can understand ancillary statistics as containing no information about θ. This is because the sampling distribution is where the information about θ is stored, so a sampling distribution free of θ carries no such information.

It turns out that ancillary statistics are actually not useless. In fact, there's two primary applications of ancillary statistics in statistical theory. One application is via Basu's theorem

(below), and the other is what's called *conditional inference*. Some remarks will be given below about the latter, but this concept is a bit too advanced for Stat 411.

Before we get to the applications, let's first consider some special examples where ancillary statistics arise quite naturally.

Location parameter problems.

Suppose the model can be expressed as $X_i = \theta + Z_i$, i.e., the parameter θ controls the "location" of the distribution. The $\mathsf{N}(\theta, 1)$ problem is one example. Then a statistic $U = U(X_1, \ldots, X_n)$ is ancillary if

$$U(x_1 + a, \ldots, x_n + a) = U(x_1, \ldots, x_n) \quad \forall\, a \in \mathbb{R}, \quad \forall\, (x_1, \ldots, x_n).$$

Examples of such statistics are ones that involve differences, e.g., $\sum_{i=1}^n (X_i - \bar{X})^2$.

Scale parameter problems.

Suppose the model can be expressed as $X_i = \theta Z_i$, i.e., the parameter θ controls the "scale" of the distribution. The $\mathsf{Exp}(\theta)$, where θ is the mean, is one example. In this case, a statistic $U = U(X_1, \ldots, X_n)$ is ancillary if

$$U(ax_1, \ldots, ax_n) = U(x_1, \ldots, x_n), \quad \forall\, a > 0, \quad \forall (x_1, \ldots, x_n).$$

Examples of such statistics are ones that involve ratios, e.g., $\bar{X}/(\sum_{i=1}^n X_i^2)^{1/2}$.

Location–scale parameter problems.

Suppose the model can be expressed by $X_i = \theta_1 + \theta_2 Z_i$, i.e., θ_1 controls the "location" of the and θ_2 controls the "scale." A statistic $U = U(X_1, \ldots, X_n)$ is ancillary if

$$U(a + bx_1, \ldots, a + bx_n) = U(x_1, \ldots, x_n), \quad \forall\, a, \forall\, b, \forall\, (x_1, \ldots, x_n).$$

Any function of $\{\frac{X_1 - \bar{X}}{S}, \ldots, \frac{X_n - \bar{X}}{S}\}$ is an ancillary statistic.

Those of you familiar with algebra, in particular, *group theory*, might find it interesting that these three classes of problems all fall under the umbrella of what are called *group transformation problems* where the model exhibits a certain kind of invariance under a group of transformations on the sample space. We won't discuss this in Stat 411.

Our first application of ancillary statistics is in their relationship with complete sufficient statistics. This is summarized in the famous result of D. Basu.

Theorem 4.5 (Basu)**.** *Let T be a complete sufficient statistic, and U an ancillary statistic. Then T and U are statistically independent.*

Proof. Here I will follow our usual convention and not be precise about conditional probabilities, etc. Pick some arbitrary event A for U. Since U is an ancillary statistic, the probability $p_A = \mathsf{P}_\theta(U \in A)$ does not depend on θ. Define $\pi_A(t) = \mathsf{P}_\theta(U \in A \mid T = t)$. By the iterated expectation property, we have $\mathsf{E}_\theta\{\pi_A(T)\} = p_A$ for all θ, i.e.,

$$\mathsf{E}_\theta\{\pi_A(T) - p_A\} = 0 \quad \forall\, \theta.$$

Since T is complete, it must be that $\pi_A(t) = p_A$ for all t, i.e., $\mathsf{P}_\theta(U \in A \mid T = t) = \mathsf{P}_\theta(U \in A)$ for all t. Since the event A was arbitrary, the independence claim follows. □

Basu's theorem can be used to simplify a number of calculations. Below are some particular applications of this phenomenon.

Exercise 4.20. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(\theta_1, \theta_2^2)$, where θ_2 is the standard deviation. Use Basu's theorem to show that $\bar{X}$ and S^2 are independent.[3]

Exercise 4.21. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(\theta, 1)$. Use Basu's theorem to evaluate the covariance between $\bar{X}$ and M_n, the sample median.

Exercise 4.22. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(0, \theta^2)$, where $\theta > 0$ is the standard deviation. Use Basu's theorem to evaluate $\mathsf{E}_\theta\{X_1^2/(X_1^2 + \cdots + X_n^2)\}$.

I will end this section with some remarks about the use of ancillary statistics for what's called "conditional inference." The basic idea is this: since an ancillary statistic does not depend on θ, we can sometime develop better statistical methods by looking at the conditional sampling distribution of estimates, given the observed value of the ancillary statistics. Example 7.9.5 in HMC gives some description of the basic idea. Let me give a much simpler and exaggerated example for illustration. Suppose X_1, X_2 are independent observations from a discrete uniform distribution, $\mathsf{Unif}\{\theta - 1, \theta + 1\}$. That is, each X_i equals $\theta - 1$ or $\theta + 1$, both with probability 0.5. Since θ is the mean of this distribution, $\bar{X}$ is a reasonable choice of estimator. However, we can write

$$\bar{X} = \begin{cases} \theta & \text{if } |X_1 - X_2| = 1, \\ \text{either } \theta - 1 \text{ or } \theta + 1 & \text{if } |X_1 - X_2| = 0. \end{cases}$$

Note that $U = |X_1 - X_2|$ is ancillary, in fact, $U \sim \mathsf{Ber}(0.5)$, free of θ. If we ignore U, then $\bar{X}$ is still a good estimate, i.e., consistent, etc. But if look at the observed value of U, then we know that $\bar{X}$ is exactly θ when $U = 1$; when $U = 0$, we don't really get any help. So, in some sense, the observed value of U can help us sharpen our claims about the sampling distribution of estimators.

Conditional inference was another big idea of Fisher's, but this has been much slower to develop than his ideas on maximum likelihood, efficiency, and randomization in experimental design. Perhaps the reason is that ancillary statistics are not well understood, e.g., it is not always clear how to find a good one, and the calculations with conditional distributions are generally much more difficult. One particular instance where this can be useful is if, say, the MLE $\hat{\theta}$ is not minimal sufficient, but $(\hat{\theta}, U)$ is minimal sufficient, for an ancillary statistic U. In this case, the conditional sampling distribution of $\hat{\theta}$, given $U = u$, is a better choice for inference (constructing confidence intervals, etc), than the marginal distribution of $\hat{\theta}$. The details behind these arguments are complicated, but quite nice. A more advanced course in statistical theory could present some of these ideas.

[3]Other standard proofs of this fact make extensive use of linear algebra and properties of the multivariate normal distribution.

4.9 Appendix

4.9.1 Rao–Blackwell as a complete-class theorem

In this chapter, the Rao–Blackwell theorem was used basically as a tool that can be used to create a good unbiased estimator from any not-so-good unbiased estimator. However, the result is deeper than this, which I will explain briefly here. There is a notion of a *complete class theorem*, which we will touch on briefly in Chapter 6. The point of a complete class theorem is to identify all those estimators (or tests, or whatever) which are "equivalent" in the sense that no other estimator is better for all possible θ values. The discussion in Chapter 2 about how, say, the mean square error of an estimator is a function of θ and, for any two estimators, usually the mean square error curves would intersect, meaning that one is better than the other for some θ but not all θ. These estimators could be considered "equivalent" in the sense here.

What the Rao–Blackwell theorem says is that, any estimator which is not a function of a sufficient statistic, can be beaten by an estimator that is a function of a sufficient statistic; moreover, the theorem describes how to construct that "better" estimator. So, the conclusion is that the collection of all estimators that are functions of a sufficient statistic only forms a complete class. A subtle point is that, even among estimators that depend on a sufficient statistic, there may be some which are everywhere worse than another estimator depending on a sufficient statistic. Technically, there may be some estimators which are functions of a sufficient statistic which are *inadmissible*. Anyway, the deeper implication of Rao–Blackwell about sufficient statistics should be clear now.

4.9.2 Proof of Lehmann–Scheffe Theorem

I will start by proving a more general result from which the main theorem will follow. Perhaps the reason why this theorem is not the "main" theorem is because the conditions required are too difficult to check. Surprisingly though, the proof is rather straightforward. For this I will need a bit of new notation. Let $\mathbb{Z}$ denote the set of all "unbiased estimators of 0," i.e.,

$$\mathbb{Z} = \{Z : \mathsf{E}_\theta(Z) = 0 \text{ for all } \theta\}.$$

Then a characterization of the MVUE $\hat{\theta}$ can be given based on the correlation, or covariance $\mathsf{C}_\theta(\hat{\theta}, Z)$, between $\hat{\theta}$ and unbiased estimators Z of zero.

Lemma 4.1. *$\hat{\theta}$ is a MVUE if and only if $\mathsf{C}_\theta(\hat{\theta}, Z) = 0$ for all θ and for all $Z \in \mathbb{Z}$.*

Proof. I'll prove the "if" and "only if" parts separately.

"If" part. Suppose that $\mathsf{C}_\theta(\hat{\theta}, Z) = 0$ for all θ and for all $Z \in \mathbb{Z}$. Set $\tilde{\theta} = \hat{\theta} + aZ$ for any

number a and any $Z \in \mathbb{Z}$. Clearly $\tilde{\theta}$ is an unbiased estimator of θ. Moreover,

$$\begin{aligned}\mathsf{V}_\theta(\tilde{\theta}) &= \mathsf{V}_\theta(\hat{\theta} + aZ) \\ &= \mathsf{V}_\theta(\hat{\theta}) + a^2\mathsf{V}_\theta(Z) + 2a\mathsf{C}_\theta(\hat{\theta}, Z) \\ &= \mathsf{V}_\theta(\hat{\theta}) + a^2\mathsf{V}_\theta(Z) \\ &\geq \mathsf{V}_\theta(\hat{\theta}).\end{aligned}$$

The last "$\geq$" is "$=$" if and only if $a = 0$ or $Z \equiv 0$. Since all unbiased estimators $\tilde{\theta}$ of θ can be written in the form $\tilde{\theta} = \hat{\theta} + aZ$ for some a and some $Z \in \mathbb{Z}$, it follows that $\mathsf{V}_\theta(\tilde{\theta}) \geq \mathsf{V}_\theta(\hat{\theta})$ for all θ; therefore, $\hat{\theta}$ is a MVUE.

"Only if" part. Suppose $\hat{\theta}$ is a MVUE. Then for any number a and any $Z \in \mathbb{Z}$, $\tilde{\theta} = \hat{\theta} + aZ$ is an unbiased estimator and, by assumption, $\mathsf{V}_\theta(\tilde{\theta}) \geq \mathsf{V}_\theta(\hat{\theta})$. We can express $\mathsf{V}_\theta(\hat{\theta} + aZ)$ in terms of variances and covariances; that is,

$$\mathsf{V}_\theta(\hat{\theta} + aZ) \geq \mathsf{V}_\theta(\hat{\theta}) \quad \Longrightarrow \quad \mathsf{V}_\theta(\hat{\theta}) + a^2\mathsf{V}_\theta(Z) + 2a\mathsf{C}_\theta(\hat{\theta}, Z) \geq \mathsf{V}_\theta(\hat{\theta}).$$

Canceling out $\mathsf{V}_\theta(\hat{\theta})$ on both sides above gives $a^2\mathsf{V}_\theta(Z) + 2a\mathsf{C}_\theta(\hat{\theta}, Z) \geq 0$ for all θ. This is a quadratic in a so there exists some value of a for which the left-hand side is negative *unless* $\mathsf{C}_\theta(\hat{\theta}, Z) = 0$ for all θ. So if $\hat{\theta}$ is a MVUE, then $\mathsf{C}_\theta(\hat{\theta}, Z) = 0$ for all θ and all $Z \in \mathbb{Z}$, proving the "only if" part. □

The above lemma gives a necessary and sufficient condition for $\hat{\theta}$ to be the MVUE. However, the condition is not so easy to check—how could we proceed to verify that $\mathsf{C}_\theta(\hat{\theta}, Z) = 0$ *for all* $Z \in \mathbb{Z}$? The Lehmann–Scheffe theorem gives a more convenient condition to check, one that depends on the original model and the choice of sufficient statistic. Essentially, the condition reduces the set $\mathbb{Z}$ to a single element for which the covariance condition is obvious.

Proof of Theorem 4.3. The MVUE must be a function of T; otherwise, it could be improved via conditioning by the Rao–Blackwell theorem. By restricting to functions of T, the set $\mathbb{Z}$ shrinks down to

$$\mathbb{Z}_T = \{Z : Z \text{ is a function of } T \text{ and } \mathsf{E}_\theta(Z) = 0 \text{ for all } \theta\} \subset \mathbb{Z}.$$

Since T is complete, the only $Z \in \mathbb{Z}_T$ is the trivial statistic $Z \equiv 0$ (constant equal to zero). It is clear that $\mathsf{C}_\theta(\hat{\theta}, 0) = 0$ and so, by Lemma 4.1, $\hat{\theta}$ is the unique MVUE. □

4.9.3 Connection between sufficiency and conditioning

No doubt that sufficiency has been fundamental to our understanding of statistics and what are the relevant portions of the data. Here let's consider the dimension-reduction aspect of sufficiency. That is, when we have n iid samples and a single scalar parameter, we hope that we can reduce our data down to a single (minimal) sufficient statistic. That the data dimension can be reduced to that of the parameter is a desirable feature.

What I want to demonstrate here is that this same dimension-reduction can be achieved by using *conditioning*, which we discussed briefly at the end of Section 4.8. To keep things clear and simple, let's consider an example. Let $X_1, X_2 \overset{\text{iid}}{\sim} \mathsf{N}(\theta, 1)$. Of course, we can keep $T = \bar{X}$ as the (minimal) sufficient statistic and, for example, a 95% confidence interval for θ based on this is $T \pm 1.96/\sqrt{2}$. As an alternative, take $T = X_1$ as our statistic, also one-dimensional. Now consider the distribution of $T = X_1$ given $X_2 - X_1$, the latter being observable and also ancillary. Then

$$T \mid (X_2 - X_1) \sim \mathsf{N}(\theta + \tfrac{X_2 - X_1}{2}, \tfrac{1}{2}),$$

which can be easily found by doing the conditional distribution calculations directly, or by making use of special properties of the multivariate normal distribution. Anyway, if we ask for a 95% confidence interval for θ based on this conditional distribution we get

$$X_1 - \frac{X_2 - X_1}{2} \pm \frac{1.96}{\sqrt{2}} = \bar{X} \pm \frac{1.96}{\sqrt{2}},$$

which is exactly the same as what we got from sufficiency.

The point here is that sufficiency may not be all that important, since the same dimension-reduction can be achieved by other means, namely, by conditioning. The example used here is too simple to be particularly convincing, but these same ideas can be applied in many other (but probably not all) problems. I'm not sure exactly how well this point is understood by statisticians, and could be worth exploring.

Chapter 5

Hypothesis Testing

5.1 Introduction

Thus far in the course, we have focused on the problem where the parameter of interest θ is unknown, and the goal is to estimate this unknown quantity based on observable data. However, one could consider a somewhat simpler goal, namely, to determine if the unknown θ belongs to one subset Θ_0 or another subset Θ_1. In other words, we don't really care about the actual value of θ, we only want to know where it lives in the parameter space Θ. In this sense, the hypothesis testing problem is simpler than the estimation problem. Probably the reason why we present the simpler hypothesis testing problem *after* the more difficult estimation problem is that the important concepts of likelihood, sufficiency, etc are easier to understand in the estimation problem.

In this chapter, we will consider first the motivation and terminology in the hypothesis testing problem. Terminology can be a difficult hurdle to overcome in these problems, in part because there's some inconsistencies across books, papers, etc. We then proceed to describe the kinds of properties we like a hypothesis test to satisfy. This parallels our earlier discussion on the basic properties of estimators, in that these are all characteristics of the sampling distribution of some statistic. The main characteristics of tests we are interested in are the size and power. We will also discuss the somewhat elusive p-value.

We move on to take a more formal approach to hypothesis testing. Section 5.4 describes the ideal situation where a "best possible" test is available. There is a very general result (Neyman–Pearson lemma[1]) that allows us to find a test of given size with the highest power. Unfortunately, such an optimal test is only available in relatively simple problems. A more flexible approach to construct tests is via likelihood ratios. This theory overlaps with what we discussed regarding maximum likelihood estimators (MLEs). In fact, the important theorem of Wilks says that likelihood ratios are asymptotically chi-square; this is similar to the asymptotic normality theorem for MLEs, and it allows us to construct tests with approximately the correct size.

[1]Same Neyman as in the factorization theorem.

5.2 Motivation and setup

To understand the testing problem setup, consider the following example. I offer you the opportunity to play a game in which you win \$1 if a coin lands on Heads and you lose \$1 if the coin lands on Tails. If the coin is fair, then you might be willing to play; otherwise, probably not. So, the goal is to determine if the coin is fair or, more precisely, to determine if the probability θ that the coin lands on Heads is at least 0.5 or not. This can be done by watching a sequence of plays of the game: conclude that the coin is unfair ($\theta < 0.5$) if and only if you observe too many tails in the sequence of plays. This is a very natural strategy; things can get technical is in setting the "too many" cutoff.

To be specific, let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f_\theta(x)$, where θ is the unknown parameter, assumed to lie in a parameter space Θ. For us, Θ will usually be a subset of $\mathbb{R}$ and occasionally a subset of $\mathbb{R}^d$ for some $d > 1$; in general, Θ can be basically anything. The hypothesis testing problem is specified by splitting Θ, i.e., $\Theta = \Theta_0 \cup \Theta_1$, with $\Theta_0 \cap \Theta_1 = \varnothing$. This decomposition is then written in the following form:

$$H_0 : \theta \in \Theta_0 \quad \text{vs.} \quad H_1 : \theta \in \Theta_1.$$

The statement H_0 is called the *null hypothesis* and H_1 is called the *alternative hypothesis*. In the motivating example above, the null hypothesis is $H_0 : \theta \geq 0.5$ and the alternative hypothesis is $H_1 : \theta < 0.5$, corresponding to $\Theta_0 = [0.5, 1]$ and $\Theta_1 = [0, 0.5)$.

The goal is to use observable data $X_1, \ldots, X_n$ to decide between H_0 and H_1. Formally, a test is a function $\varphi(\cdot)$ that maps $(x_1, \ldots, x_n)$ to the interval $[0, 1]$; in almost all cases, the test can be chosen to take only values 0 and 1. Suppose, for the moment, that $\varphi(\cdot)$ takes only values 0 and 1. Then the possible conclusions of the hypothesis testing problem are

$$\begin{aligned} \varphi(X_1, \ldots, X_n) = 1 &\iff \text{Reject } H_0 \\ \varphi(X_1, \ldots, X_n) = 0 &\iff \text{Do not reject } H_0. \end{aligned}$$

Note that the conclusions (i) never say anything about directly H_1 and (ii) never say that one of the two hypothesis is true. In particular, we cannot conclude from sample data that either H_0 or H_1 is true. In what follows (and in class), for simplicity of presentation, I may write (or say), e.g., "accept H_0" but please understand this to mean "Do not reject H_0." Practically there is no difference, but logically the difference is drastic: it is impossible for an experiment to *confirm* a theory, but when enough evidence exists to suggests that a theory is false, it is standard to *reject* that theory and develop a new one.[2][3] Tests $\varphi(\cdot)$ are like estimators in the estimation problem, and our primary focus is how to choose a "good" test.

[2]Such questions fall under the umbrella of "Philosophy of Science." In particular, Cournot's principle helps to connect probability to how decisions can be made, which is the backbone of statistical inference.

[3]This is exactly the process by which scientists have recently reached the conclusion that the Higgs' Boson particle exists.

	H_0 is true	H_0 is false
accept H_0	correct	Type II error
reject H_0	Type I error	correct

Table 5.1: Four possible outcomes in a hypothesis testing problem. Remember that "accept H_0" actually means "Do not reject H_0."

5.3 Basics

5.3.1 Definitions

In a hypothesis testing problem, there are four possible outcomes; these are described in Table 5.1. Two of the four possible outcomes result in a correct decision, while the other two result in an error. The two errors have names:

$$\text{Type I error} = \text{Rejecting } H_0 \text{ when it's true}$$
$$\text{Type II error} = \text{Accepting } H_0 \text{ when it's false}$$

We want to choose the test φ such that the probability of making is small in a certain sense to be described below.

Before we discuss the choice of test, I want to present an alternative, but equivalent, way to formulate the test. Before we defined a function φ defined on the set of all $(x_1, \ldots, x_n)$ that takes values 0 or 1. The alternative strategy is to specify direct the set of all $(x_1, \ldots, x_n)$ such that the function φ takes value 1. This set, which I will denote by C, is called the *critical region*. To see that the two are equivalent, note that

$$C = \{(x_1, \ldots, x_n) : \varphi(x_1, \ldots, x_n) = 1\} \iff \varphi(x_1, \ldots, x_n) = I_C(x_1, \ldots, x_n),$$

where $I_C(\cdot)$ is the indicator function of C. Whether one works with a test function φ or a critical region C is really just a matter of taste. I personally like to use the test function because the definitions below are easier to state in terms of φ.

The two primary characteristics of a test φ (or C) are *size* and *power*.

Definition 5.1. Given Θ_0, the size of the test φ (or C), denote by $\mathsf{size} = \mathsf{size}_\varphi$ is

$$\mathsf{size} = \max_{\theta \in \Theta_0} \mathsf{E}_\theta\{\varphi(X_1, \ldots, X_n)\} = \max_{\theta \in \Theta_0} \mathsf{P}_\theta\{(X_1, \ldots, X_n) \in C\}.$$

In words, the size of the test is the probability of a Type I error.

Definition 5.2. The power of the test φ (or C), denoted by $\mathsf{pow}(\cdot) = \mathsf{pow}_\varphi(\cdot)$, is a function

$$\mathsf{pow}(\theta) = \mathsf{E}_\theta\{\varphi(X_1, \ldots, X_n)\} = \mathsf{P}_\theta\{(X_1, \ldots, X_n) \in C\}.$$

In words, $\mathsf{pow}(\theta')$ is the probability of rejecting H_0 when $\theta = \theta'$.

Since size is an error probability, it is clear that we want it to be small. For power, we're primarily interested in how it looks for θ outside Θ_0. In that case, $\mathsf{pow}(\theta)$ denotes the probability of correctly rejecting H_0 when it's false. Alternatively, for $\theta \notin \Theta_0$, $1 - \mathsf{pow}(\theta)$ denotes the probability of making a Type II error. In this light, we see that we want the power to be large. Things are often difficult because power is function, not a number.

With this terminology, we are now in a position to formally state our goal. In particular, we want to fix the size at some small value, say $\alpha = 0.05$, and then seek a test φ (or C) that maximizes the power $\mathsf{pow}(\cdot)$ in some sense. The reason we fix size and try to maximize pow is that the two quantities are competing in the following sense: if we force the test to have small size, then we indirectly reduce the power, and vice versa. That is, improving in one area hurts in the other. So the accepted strategy is to allow a little chance of a Type I error with the hopes of making the power large.

5.3.2 Examples

Before we get into formalities about "best" tests, etc, we should do some simple intuitive examples. At this point, it may not be clear why the various steps are taken.

Exercise 5.1. (*z-test*) Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(\theta, 1)$, and consider $H_0 : \theta = 0$ vs. $H_1 : \theta > 0$.

(a) Define a test with critical region $C = \{(x_1, \ldots, x_n) : \bar{x} > k\}$. Find k such that the size of the test is $\alpha = 0.05$.

(b) For the test you derived in (a), find the power function $\mathsf{pow}(\theta)$.

Exercise 5.2. (*t-test*) A particular car is advertised to have gas mileage 30mpg. Suppose that the population of gas mileages for all cars of this make and model looks like a normal distribution $\mathsf{N}(\mu, \sigma^2)$, but with both μ and σ unknown. To test the advertised claim, sample $n = 10$ cars at random and measure their gas mileages $X_1, \ldots, X_{10}$. Use these data to test the claim, i.e., test $H_0 : \mu = 30$ vs. $H_1 : \mu < 30$. Assume $\bar{x} = 26.4$ and $s = 3.5$ are the observed sample mean and standard deviation, respectively. Use $\alpha = 0.05$. (Hint: Use a Student-t distribution.)

Example 5.1. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Ber}(\theta)$, and consider $H_0 : \theta = \theta_0$ vs. $H_1 : \theta < \theta_0$, for fixed θ_0. Define a critical region $C = \{(x_1, \ldots, x_n) : \sum_{i=1}^n x_i \leq k\}$. This is like the coin-tossing example from before—we should reject the claim that the coin is fair ($\theta_0 = 0.5$) if too few heads are observed in the sample, i.e., that the total number of heads $\sum_{i=1}^n X_i$ is below a threshold k. Given k, the size of such a test is

$$\mathsf{P}_{\theta_0}\Big\{\sum_{i=1}^n X_i \leq k\Big\} = \mathsf{P}\{\mathsf{Bin}(n, \theta_0) \leq k\} = \sum_{t=0}^k \binom{n}{t} \theta_0^t (1-\theta_0)^{n-t}.$$

There is no nice expression for this quantity on the right-hand side, call it $F_{n,\theta_0}(k)$, but it can be easily evaluated with software (e.g., R). However, recall that $\mathsf{Bin}(n, \theta_0)$ is a discrete distribution, so $F_{n,\theta_0}(k)$ has only finitely many values possible values. So, it is quite possible

that there is no k that solves the equation, $F_{n,\theta_0}(k) = \alpha$, for a particular α, such as $\alpha = 0.05$. Therefore, for given α, there may not exist[4] a test of exact size α.

To avoid such a difficulty, one might consider an approximation, in particular, an approximation of the discrete $\mathsf{Bin}(n, \theta)$ distribution of $Y = \sum_{i=1}^{n} X_i$ with a continuous one for which an exact size-α test is available. In particular, recall the CLT-based approximation: $Y \sim \mathsf{N}(n\theta, n\theta(1-\theta))$. In this case, the size can be approximated by

$$\mathsf{P}_{\theta_0}(Y \le k) = \mathsf{P}_{\theta_0}\Big(\frac{Y - n\theta_0}{\sqrt{n\theta_0(1-\theta_0)}} \le \frac{k - n\theta_0}{\sqrt{n\theta_0(1-\theta_0)}}\Big) \approx \Phi\Big(\frac{k - n\theta_0}{\sqrt{n\theta_0(1-\theta_0)}}\Big).$$

If we set the right-hand side equal to α we can solve for k, that is,

$$k = n\theta_0 + z_\alpha^\star \sqrt{n\theta_0(1-\theta_0)}, \quad \text{where} \quad \Phi(z_\alpha^\star) = \alpha.$$

Thus, the CLT-based approximate size-α test rejects H_0 iff $Y \le n\theta_0 + z_\alpha^\star \sqrt{n\theta_0(1-\theta_0)}$. Power calculations for this test can be carried out similar to that in Exercise 5.1(b).

5.3.3 Remarks

First I give some "philosophical" remarks. Recall our discussion of properties of estimators earlier in the course, e.g., unbiasedness, consistency, etc. These are properties of the sampling distribution of the estimator and, in particular, they do not guarantee that, for a particular data set, the estimator obtained is "good." The same is true in the hypothesis testing problem. That is, for a given data set, even the best test can lead to a wrong decision (Type I or Type II error). Having exact size α and a large power function are properties to help one select a test to use—once the data set is fixed, these properties are meaningless in terms of explaining the uncertainty that a correct decision was made. So one should be careful in interpreting size and $1 - \mathsf{pow}(\theta)$. These are *not* probabilities that a Type I and Type II error is committed for the given data.

In practice, we usually focus on null hypotheses that consist of a single point, i.e., $H_0 : \theta = \theta_0$. This is essentially without loss of generality. For example, suppose we want to test $H_0 : \theta \le \theta_0$ versus $H_1 : \theta > \theta_0$. In most cases, the power function $\mathsf{pow}(\theta)$ is continuous and monotone in θ, in this case, monotone increasing. Since the size of the test is $\max_{\theta \le \theta_0} \mathsf{pow}(\theta)$, monotonicity of pow implies that the size is simply $\mathsf{pow}(\theta_0)$, the power at the upper end point. This is actually the same as the size of the test of $H_0 : \theta = \theta_0$. Therefore, we usually will formulate the test in terms of just the point θ_0 in Θ_0 that is closest to Θ_1. These "point null hypotheses" are also intuitively easier.

5.3.4 P-values

There are a couple of issues that arise in the testing problem that are somewhat undesirable. First, as we saw above, not all problems admit an exact (non-randomized) size-α test.

[4] An exact size-α test always exists if one considers *randomized* tests; see Appendix 5.7.3. These randomized tests, however, are rather strange, so people rarely use them in practice.

Second, the neither the size nor the power are directly related to the data that is actually observed. One tool that can help to rectify this is the *p-value*.

Definition 5.3. Let $T = T(X_1, \ldots, X_n)$ be a statistic, and consider $H_0 : \theta = \theta_0$ vs. $H_1 : \theta < \theta_0$ (or $H_1 : \theta > \theta_0$). Suppose the test rejects H_0 if $T \leq k$ (or $T > k$). Let $t = T(x_1, \ldots, x_n)$ for the given sample $x_1, \ldots, x_n$. Then the p-value is

$$\mathsf{pval}(t) = \mathsf{P}_{\theta_0}(T \leq t) \quad \text{or} \quad \mathsf{pval}(t) = \mathsf{P}_{\theta_0}(T > t).$$

The p-value is sometimes called the "observed size" because the calculation looks similar; in my opinion, this is a potentially misleading name. Note also that there's no reason to worry about T having discrete instead of continuous sampling distribution as there was in the previous binomial example.

How do we interpret the p-value? Formally, p-value represents the probability, assuming H_0 is true, of a "more extreme" observation (with respect to H_1) compared to what was actually observed. To understand how such a quantity relates to H_0 and H_1, consider two possible scenarios:

- *Large p-value* means the observed t sits near the middle of the sampling distribution of T under f_{θ_0}, i.e., the observed t is consistent with H_0. This is evidence consistent with H_0 being true.
- *Small p-value* means the observed t sits in one of the tails of the sampling distribution of T under f_{θ_0}, i.e., the observed t is inconsistent with H_0. This suggests one of two things: either H_0 is false or t is an outlier/extreme observation. Since the latter is deemed unlikely, one tends to believe the former—that is, small p-value is evidence inconsistent with H_0 being true.[5]

At a high level, p-values are "measures of evidence" in the truthfulness of H_0. That is, small p-value means little evidence in favor of H_0 and large p-value means substantial evidence in favor of H_0. I view $\mathsf{pval}(t)$ as representing the *plausibility* of H_0 given observation t.

From an operational point of view, one might consider defining a test based on the p-value. In particular, one can test $H_0 : \theta = \theta_0$ with the p-value:

$$\text{reject } H_0 \text{ iff } \mathsf{pval}(t) \leq \alpha. \tag{5.1}$$

When T has a continuous sampling distribution, the test above has exact size α; in general, its size is bounded above by α. This is the standard use of p-values, but I personally feel (as did Fisher, the inventor of p-value) that p-values are more valuable than just as a tool to define a testing rule like in (5.1).

Example 5.2. Recall the setup in Exercise 5.2. That is, $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(\mu, \sigma^2)$ and the goal is to test $H_0 : \mu = 30$ versus $H_1 : \mu < 30$. Let $T = n^{1/2}(\bar{X} - 30)/S$ be the t-statistic; its observed value, based on $\bar{x} = 26.4$ and $s = 3.5$ is

$$t = 10^{1/2}(26.4 - 30)/3.5 = -3.25.$$

[5]This reasoning is consistent with Cournot's principle for the use of probability to test theories.

Because the alternative has "<" the p-value is defined as

$$\mathsf{pval}(-3.25) = \mathsf{P}_{30}(T < -3.25) = G_{10-1}(-3.25) \approx 0.005.$$

This calculation is based on the fact that, when $\mu = 30$, T has a Student-t distribution with $10 - 1 = 9$ degrees of freedom; G_9 is this distribution's CDF. The value 0.005 can be obtained from software (in R, use `pt(-3.25, df=9)`) or from Table IV in HMC, p. 660. Since the p-value is small, data in this case provides evidence *against* H_0. Formally, if we take $\alpha = 0.05$, then an exact size-0.05 test would, in this case, reject H_0 because the p-value is less than 0.05.

Exercise 5.3. Find an expression for the p-value, as a function of the observed $Y = y$, for the CLT-based test in Example 5.1. State your conclusions if $n = 100$, $\theta_0 = 0.3$, and $y = 33$, i.e., which hypothesis is supported by data?

To conclude the discussion, let me list two common *misinterpretations*. If you remember anything about p-values from Stat 411, you should remember these warnings.

- *The p-value is not the probability that H_0 is true.* The null is fixed and either true or not true, so there are no non-trivial probabilities here. However, one can introduce non-trivial probabilities by taking a Bayesian point of view.
- *The p-value is not the probability a Type I error is made, given data.* The p-value calculation has nothing to do with accept/reject decisions. It's simply a tail probability for the sampling distribution of the test statistic T, which can be used to measure the amount of support in data for H_0. One can, of course, use p-values to make decisions but then the p-value obviously cannot be the probability of making a Type I error.

5.4 Most powerful tests

5.4.1 Setup

The material in this section parallels that in the section on minimum variance unbiased estimation. Indeed, we will focus on tests which have a given size α, and seek the one with the largest power. Fixing the size to be α is like requiring the estimators to be unbiased, and looking for the size-α test with largest power is like looking for the unbiased estimator with the smallest variance.

A general theory of most powerful tests exists only for the relatively simple problem of testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$, where $\theta_1 \neq \theta_0$; this is called a *simple-versus-simple* hypothesis testing problem. We can now define what is meant by most powerful test.

Definition 5.4. Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$ ($\theta_1 \neq \theta_0$) based on $X = (X_1, \ldots, X_n)$ iid samples from $f_\theta(x)$. Given $\alpha \in (0, 1)$, let

$$\mathbb{T}_\alpha = \{\varphi : \mathsf{E}_{\theta_0}[\varphi(X)] = \alpha\}$$

be the collection of all tests φ with size α. Then the *most powerful* size-α test $\varphi^\star$ satisfies (i) $\varphi^\star \in \mathbb{T}_\alpha$, and (ii) for any $\varphi \in \mathbb{T}_\alpha$, $\mathsf{E}_{\theta_1}[\varphi(X)] \leq \mathsf{E}_{\theta_1}[\varphi^\star(X)]$.

The point here is that by focusing just on two points, θ_0 and θ_1, and insisting on size-α, we only need to worry about the power function at $\theta = \theta_1$, i.e., just the number $\mathsf{E}_{\theta_1}[\varphi(X)]$. In general, power is a function not just a number, and we know that comparing functions is much harder than comparing numbers. We will see below that, in some cases, it is possible to extend the most powerful claim to more general alternatives.

5.4.2 Neyman–Pearson lemma

The idea of most powerful tests is simple, but it is not at all clear how to find $\varphi^\star$ for a given problem. Fortunately, the Neyman–Pearson lemma, given next, is at our disposal.

Theorem 5.1 (Neyman–Pearson)**.** *Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$ based on continuous data X with likelihood function $L_X(\theta)$. Then the most powerful size-α test is given by*

$$\varphi^\star(X) = \begin{cases} 1 & \text{if } L_X(\theta_0) \,/\, L_X(\theta_1) \leq k \\ 0 & \text{if } L_X(\theta_0) \,/\, L_X(\theta_1) > k, \end{cases}$$

where k is chosen such that $\mathsf{E}_{\theta_0}[\varphi^\star(X)] = \alpha$.

Here's a few remarks about this important result:

- An equivalent formulation of the most powerful test, in terms of a critical region $C_\star$, writes $\varphi^\star(X) = I_{C_\star}(X)$, where

  $$C_\star = \{x : L_x(\theta_0) \,/\, L_x(\theta_1) \leq k\},$$

  and k is just as in the statement of the theorem. It turns out that this latter formulation is a bit more convenient for proving the result; see Appendix 5.7.2.
- Also note the version of the Neyman–Pearson lemma stated above only handles continuous distribution data. This is not true. It applies for discrete problems as well, but with a slight modification. Recall that, in discrete problems, it may not be possible to define a test with exact size α. The remedy is to allow what's called *randomized* tests. Appendix 5.7.3 gives the extended version of the Neyman–Pearson lemma. This is a bit trickier and, frankly, quite impractical, so we won't focus on this here. Here, we agree that, for discrete problems, randomization can be ignored *unless you are specifically asked to consider it.*
- The Neyman–Pearson lemma makes no claims on the dimension of data or parameter θ. So, the result holds for vector or even function parameters without changing a word. The only limitation of the result is that it holds only for simple-vs-simple hypotheses; but see the next remark.
- In general, the most powerful test will depend on θ_1. However, in many cases, it can be shown that the test actually does not depend on θ_1. In those cases, the optimality of the test can be extended to a sort of *uniform* optimality for a more general kind of alternative hypothesis; see Section 5.4.3.

Exercise 5.4. For each case below, write down the form of the most-powerful size-α test of $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$, with $\theta_1 > \theta_0$. Please simplify the test as much as possible. For discrete problems, you may ignore randomization.

(a) $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Ber}(\theta)$;

(b) $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Pois}(\theta)$;

(c) $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(\theta, 1)$;

(d) $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(0, \theta)$.

Exercise 5.5. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Unif}(0, \theta)$. Use the Neyman–Pearson lemma to find the most powerful size-α test of $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$, with $\theta_1 < \theta_0$.

5.4.3 Uniformly most powerful tests

Observe that, in Exercise 5.4, none of the tests derived based on the Neyman–Pearson lemma depend on the particular θ_1 value; all that mattered was that θ_1 was greater than θ_0, although the direction of the inequality isn't really important either. In such cases, it is possible to extend the claimed optimality to more general kinds of alternatives hypotheses.

Definition 5.5. A size-α test is uniformly most powerful (UMP) for $H_0 : \theta = \theta_0$ versus $H_1 : \theta \in \Theta_1$ if it's most powerful for $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$ for each $\theta_1 \in \Theta_1$.

For the examples in Exercise 5.4, the set Θ_1 is $\{\theta_1 : \theta_1 > \theta_0\}$. So, there, we can make a stronger conclusion and say that those tests derived from the Neyman–Pearson lemma are *uniformly* most powerful size-α tests for $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$. One should keep in mind that these are special examples (exponential families) and that, in general, there may be no UMP test. When a UMP test exists, the following definition and result gives a convenient tool for find it.

Definition 5.6. A distribution with likelihood function $L(\theta)$ has the monotone likelihood ratio (MLR) property in a statistic $T = T(X_1, \ldots, X_n)$ if, for $\theta_0 < \theta_1$, $L(\theta_0)/L(\theta_1)$ is a monotone function of T.

Proposition 5.1. *If the distribution has the MLR property in T, then the UMP test of $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$ or $H_1 : \theta < \theta_0$ exists and can be expressed in terms of T instead of the full likelihood ratio.*

Proof. Use the Neyman–Pearson lemma and use MLR property to simplify. □

So, the MLR property provides a convenient way to identify when a UMP test exists and, when it does, how to actually write it. The nice thing is, we are already very familiar with likelihood ratios from our investigation of MLEs, sufficient statistics, etc.

Exercise 5.6. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f_\theta(x)$, where $f_\theta(x)$ is a regular one-parameter exponential family with PDF/PMF $f_\theta(x) = \exp\{p(\theta)K(x) + S(x) + q(\theta)\}$. Show that the likelihood has the MLR property in $T = \sum_{i=1}^n K(X_i)$. Apply this result to identify UMP tests for the examples in Exercise 5.4.

Exercise 5.7. For the $\mathsf{Unif}(0, \theta)$ problem in Exercise 5.5, is the test derived a UMP test?

To conclude this section, I will give some remarks on non-existence of UMP tests. We have focused on tests of simple nulls versus simple or "one-sided" alternatives, e.g., $H_1 : \theta = \theta_1$, $H_1 : \theta > \theta_0$, or $H_1 : \theta < \theta_0$. We have left out the "two-sided" alternative $H_1 : \theta \neq \theta_0$. In general, there are no UMP tests for two-sided alternatives. The argument goes like this, but you should draw a picture to make this clear. Consider a $\mathsf{N}(\theta, 1)$ example. We have already found UMP tests for $H_1 : \theta < \theta_0$ and $H_1 : \theta > \theta_0$, and these two are clearly different from one another. If we assume that there exists a UMP test for $H_1 : \theta \neq \theta_0$, then its power function is everywhere $\geq$ than the two UMP test power functions, with $>$ for some θ values. But this contradicts the claim that the two one-sided tests are UMP; therefore, there cannot be a UMP test for the two-sided alternative.

5.5 Likelihood ratio tests

5.5.1 Motivation and setup

As we saw in the previous section, optimality results can be applied only to some relatively simple problems, such as simple-vs-simple hypotheses. Therefore, there is some reason to consider a more flexible approach. This parallels the material presented in the section on MLEs; that is, we consider a general approach for constructing a test (based on likelihood), and show that the resulting tests have good properties. The key to the broad applicability of estimation via MLE is that there is a general asymptotic normality theory that unifies all problems. There is something similar here in the context of testing.

Let $X_1, \ldots, X_n$ be iid with some some PDF/PMF $f_\theta(x)$. The likelihood function is $L(\theta) = \prod_{i=1}^n f_\theta(X_i)$; the MLE $\hat{\theta}$ is the value that maximizes the likelihood function. In what follows, we will consider a general two-sided hypothesis testing problem, $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. But, if θ is a vector, then "$\theta \neq \theta_0$" is not really "two-sided."

5.5.2 One-parameter problems

Define the likelihood ratio statistic

$$\Lambda = \Lambda(X_1, \ldots, X_n; \theta_0) = L(\theta_0)/L(\hat{\theta}).$$

By definition of $\hat{\theta}$, note that $\Lambda \leq 1$. Recall our interpretation of likelihood: it provides a ranking of parameter values in terms of how well the postulated model fits the observed data, i.e., $L(\theta') > L(\theta'')$ means the model $f_{\theta'}$ gives a better fit to the observed data compared to

$f_{\theta''}$. So the intuition is that, if H_0 is true, then we expect Λ to be close to 1. With this intuition, we define a critical region for the test as

$$C = \{(x_1, \ldots, x_n) : \Lambda(x_1, \ldots, x_n; \theta_0) \leq k\},$$

and the test function $\varphi = I_C$ with k chosen such that the size of the test is a specified $\alpha \in (0, 1)$. In words, we reject H_0 iff the observed likelihood ratio statistic Λ is too small.

Exercise 5.8. For each case below, write down the likelihood ratio statistic Λ.

(a) $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Ber}(\theta)$;
(b) $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Pois}(\theta)$;
(c) $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(\theta, 1)$;
(d) $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(0, \theta)$;
(e) $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Unif}(0, \theta)$;
(f) $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f_\theta(x) = e^{-(x-\theta)} I_{[\theta,\infty)}(x)$.

Take the $\mathsf{N}(\theta, 1)$ problem above as an example. It was shown that $\Lambda = e^{-\frac{n}{2}(\bar{X}-\theta_0^2)}$, where $\bar{X}$ is the sample mean and θ_0 is the hypothesized mean. Then the likelihood ratio test is determined by $\Lambda \leq k$ for a suitably chosen k. Let's work on this:

$$\Lambda := e^{-n(\bar{X}-\theta_0)^2/2} \leq k \iff -2\log\Lambda := n(\bar{X} - \theta_0)^2 \geq -2\log k.$$

Set $k' = -2\log k$. Then we want to choose k' such that

$$\mathsf{P}_{\theta_0}\{n(\bar{X} - \theta_0)^2 \geq k'\} = \alpha.$$

Recall that, if $Z \sim \mathsf{N}(0, 1)$, then $Z^2 \sim \mathsf{ChiSq}(1)$. In our case, $n^{1/2}(\bar{X} - \theta_0) \sim \mathsf{N}(0, 1)$, so the probability in question is a right-tail probability for a $\mathsf{ChiSq}(1)$ random variable. That is, if G_1 is the CDF of $\mathsf{ChiSq}(1)$, then we want to find k' such that $1 - G_1(k') = \alpha$ or, equivalently, $G_1(k') = 1 - \alpha$. Therefore, k' is the $(1 - \alpha)$-quantile of $\mathsf{ChiSq}(1)$, denoted by $\chi^2_{1,1-\alpha}$, which can be obtained from the chi-square table (HMC, Table II, p. 658). In particular, if $\alpha = 0.05$, then $k' = \chi^2_{1,0.95} = 3.841$. If we wanted to, we could convert k' back to k, but this is generally not necessary.

Exercise 5.9. Find the likelihood ratio test for the $\mathsf{N}(0, \theta)$ problem above. In this case, an exact test can be derived, also using chi-square distribution. (Hint: If $X_i \sim \mathsf{N}(0, \theta)$, then $X_i^2/\theta \sim \mathsf{ChiSq}(1)$; also, $\frac{1}{\theta}\sum_{i=1}^n X_i^2 \sim \mathsf{ChiSq}(n)$.)

The cutoff k (or k' or whatever) can, in principle, be found exactly like in the normal example above. In the normal example, $-2\log\Lambda$ had a nice distributional form, namely, $-2\log\Lambda \sim \mathsf{ChiSq}(1)$. In general, however, Λ or $-2\log\Lambda$ may not have a convenient sampling distribution, so solving for k or k' can be difficult.[6] Therefore, it may be of interest to consider a more convenient approximation. The following theorem, due to Wilks, shows that the chi-square distribution of $-2\log\Lambda$ in the normal example above serves as a good approximation in general, provided that n is sufficiently large.

[6] You can use Monte Carlo to get it in general.

Theorem 5.2 (Wilks, one parameter)**.** *Assume those regularity conditions* R0–R5 *from the material on MLEs. If $H_0 : \theta = \theta_0$ is true, then $-2 \log \Lambda \to \mathsf{ChiSq}(1)$ in distribution.*

Proof. This will just be a quick sketch of the proof; details in HMC, p. 344. Let $\ell(\theta) = \log L(\theta)$, and take a Taylor approximation of $\ell(\theta)$ around $\theta = \theta_0$, evaluated at $\theta = \hat{\theta}$:

$$\ell(\hat{\theta}) = \ell(\theta_0) + \ell'(\theta_0)(\hat{\theta} - \theta_0) + \ell''(\theta_0)(\hat{\theta} - \theta_0)^2/2 + \text{error},$$

where the derivatives are given by

$$\ell'(\theta_0) = \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f_\theta(X_i)\Big|_{\theta=\theta_0} \quad \text{and} \quad \ell''(\theta_0) = \sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \log f_\theta(X_i)\Big|_{\theta=\theta_0}.$$

We will ignore the error term in what follows. Then $-2 \log \Lambda$ is approximately

$$\begin{aligned} -2 \log \Lambda &= 2\{\ell(\hat{\theta}) - \ell(\theta_0)\} \\ &= 2\ell'(\theta_0)(\hat{\theta} - \theta_0) + \ell''(\theta_0)(\hat{\theta} - \theta_0)^2 \\ &= 2n^{-1/2}\ell'(\theta_0) \cdot n^{1/2}(\hat{\theta} - \theta_0) + n^{-1}\ell''(\theta_0) \cdot \left[n^{1/2}(\hat{\theta} - \theta_0)\right]^2. \end{aligned}$$

By the LLN, $n^{-1}\ell''(\theta_0) \to -I(\theta_0)$ in probability. Also, from the proof of asymptotic normality of MLEs, we have that

$$n^{-1/2}\ell'(\theta_0) = n^{1/2} I(\theta_0)(\hat{\theta} - \theta_0) + Z_n,$$

where $Z_n \to 0$ in probability. Plugging these things in above we get

$$-2 \log \Lambda \approx 2nI(\theta_0)(\hat{\theta} - \theta_0) - I(\theta_0)[n^{1/2}(\hat{\theta} - \theta_0)]^2 = \{[nI(\theta_0)]^{1/2}(\hat{\theta} - \theta_0)\}^2.$$

Since $[nI(\theta_0)]^{1/2}(\hat{\theta} - \theta_0) \to \mathsf{N}(0, 1)$ in distribution, the square converges in distribution to $\mathsf{ChiSq}(1)$, completing the proof. □

Again, the point of Wilks theorem is that we can construct an approximate size-α test based on likelihood ratio Λ, without considering its exact sampling distribution. The theorem says, if n is large, then $-2 \log \Lambda$ is approximately $\mathsf{ChiSq}(1)$ when H_0 is true. So, all we need is to calculate Λ and find the chi-square quantile like in the normal example.

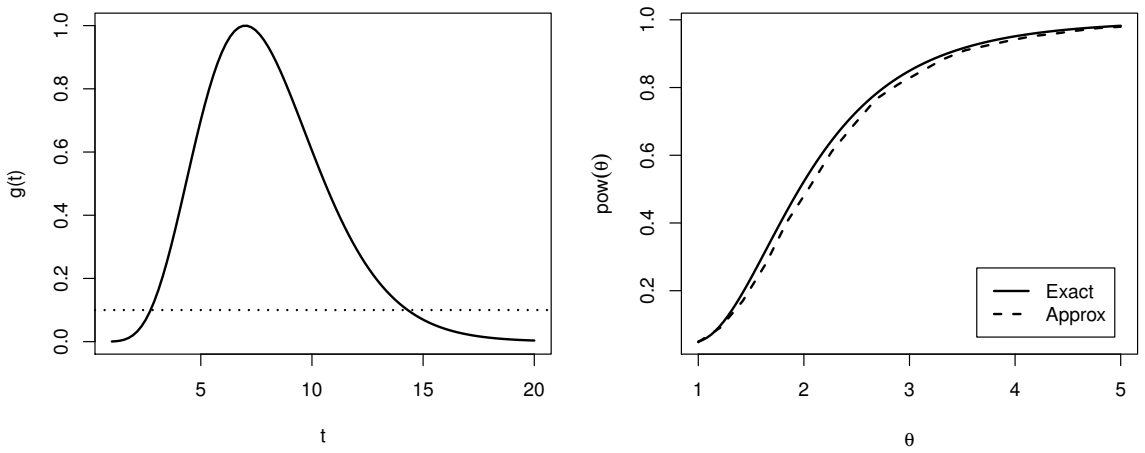
Exercise 5.10. For the Bernoulli and Poisson problems above, find an approximate size-α likelihood ratio test based on Wilks theorem.

Example 5.3. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Exp}(\theta)$, where θ is the mean. The likelihood function is $L(\theta) = \theta^{-n} e^{-n\bar{X}/\theta}$, so that the MLE is $\hat{\theta} = \bar{X}$. The corresponding likelihood ratio statistic is

$$\Lambda = L(\theta_0)/L(\hat{\theta}) = (\bar{X}/\theta_0)^n e^{-n(\bar{X}/\theta_0 - 1)}.$$

Write $T = n\bar{X}/\theta_0$; we know that, under the stated model, $T \sim \mathsf{Gamma}(n, 1)$. Then the likelihood ratio, as a function of T, is

$$\Lambda = g(T) = n^{-n} T^n e^{-(T-n)}.$$



(a) Plot of $g(t)$ versus t, with $n = 7$

(b) Comparison of power functions

Figure 5.1: Figures for the exponential mean likelihood ratio tests in Example 5.3.

It is easy to check that $\Lambda \leq k$ iff $T \leq k_1$ or $T \geq k_2$ for suitable constants k_1, k_2; see Figure 5.1(a). Therefore, the exact size-α likelihood ratio test is

$$\text{Reject } H_0 \text{ iff } T \leq k_1 \text{ or } T \geq k_2,$$

where k_1, k_2 are chosen to make the size of this test equal to α. We can take $k_1 = \gamma_{n,\alpha/2}$ and $k_2 = \gamma_{n,1-\alpha/2}$, where $\gamma_{n,p}$ is the pth quantile of the $\mathsf{Gamma}(n, 1)$ distribution. (This is the same test as in Example 6.3.1 in HMC.) As an alternative, we can apply Wilks theorem to get an approximate size-α test. Simply, the test is

$$\text{Reject } H_0 \text{ iff } -2n(\log \bar{X} - \log \theta_0) + 2n(\bar{X}/\theta_0 - 1) \geq \chi^2_{1,\alpha}.$$

This is much easier to write down compared to the exact size-α test; however, we are sacrificing a bit on size and also on power; see Figure 5.1(b). R codes are given in Section 5.7.1.

5.5.3 Multi-parameter problems

When θ is a vector, i.e., $\Theta \subseteq \mathbb{R}^d$ for $d > 1$, the problem is the same, at least in principle, though the details become more cumbersome. The problem is formulated as follows:

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \notin \Theta_0.$$

This looks the same as before, though, there is a special feature hidden in the notation. The set Θ_0 will have effective dimension $d_0 < d$. That is, the null hypothesis specifies a lower-dimensional subset of the full parameter space. For example, in the $\mathsf{N}(\theta_1, \theta_2)$ example, we might want to test if θ_1, the mean, is equal to zero. In this case, Θ_0 is a one-dimensional

subset $\{(\theta_1, \theta_2) : \theta_1 = 0, \theta_2 > 0\}$ of the two-dimensional plane. As we will see below, the dimension of the null hypothesis parameter space is important.

As before, we can write the likelihood function $L(\theta)$ for θ based on the given data $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f_\theta(x)$. For the testing problem above, the likelihood ratio statistic is

$$\Lambda = \frac{\max_{\theta \in \Theta_0} L(\theta_0)}{\max_{\theta \in \Theta} L(\hat{\theta})};$$

the denominator is just $L(\hat{\theta})$ where $\hat{\theta}$ is the global MLE, while the numerator is the *constrained* maximum of $L(\theta)$. In most cases, it is best to think carefully about the structure of the problem when H_0 is true—in those cases, it may be possible to borrow our previous simpler calculations of MLEs.

Exercise 5.11. Let $X_{j1}, \ldots, X_{jn} \overset{\text{iid}}{\sim} \mathsf{Ber}(\theta_j)$, for $j = 1, 2$; that is, we have two Bernoulli samples of the same size n but with potentially different success probabilities. The goal is to test $H_0 : \theta_1 = \theta_2$ versus $H_1 : \theta_1 \neq \theta_2$. Find the likelihood ratio statistic Λ. (Hint: if $\theta_1 = \theta_2$, then we have a sample of size $2n$ from a Bernoulli distribution with parameter θ being the common value of θ_1, θ_2.)

Exercise 5.12. Let $X_{j1}, \ldots, X_{jn} \overset{\text{iid}}{\sim} \mathsf{N}(\theta_j, \sigma^2)$, for $j = 1, 2$, with $\sigma > 0$ known. The goal is to test $H_0 : \theta_1 = \theta_2$ versus $H_1 : \theta_1 \neq \theta_2$. Find the likelihood ratio statistic Λ.

The interpretation of Λ here is the same as in the one-parameter case, i.e., if H_0 is true then we expect Λ to be close to 1. Therefore, it is reasonable to define the size-α likelihood ratio test as follows:

$$\text{Reject } H_0 \text{ iff } \Lambda \leq k$$

where k is chosen to make the size of the test equal to α. It is often the case that the exact distribution of Λ is not available, so finding k is not possible. Then it helps to have a result by which we can derive an approximate test. In the one-parameter case we had Wilks theorem and, as it turns out, we have a version of Wilks theorem in this more general case too.

Theorem 5.3 (Wilks, multi-parameter)**.** *Under certain regularity conditions, if $H_0 : \theta = \theta_0$ is true, then $-2 \log \Lambda \to \mathsf{ChiSq}(d - d_0)$ in distribution.*

The (assumptions and) conclusions of this theorem are essentially the same as those in the previous Wilks theorem. The main difference is in the degrees of freedom in the chi-square approximation. To see that this is a generalization of the one-parameter case, note that, in the latter case, H_0 is equivalent to $\Theta_0 = \{\theta_0\}$, which is a $d_0 = 0$-dimensional subset of $\mathbb{R}$; then $d - d_0 = 1 - 0 = 1$, just like in the first Wilks theorem.

Exercise 5.13. Use Wilks theorem to find an approximation size-α test in Exercise 5.12.

5.6 Likelihood ratio confidence intervals

A useful tool for constructing confidence intervals/regions for unknown parameters is to "invert" a hypothesis test. That is, for a given test of size α, define a subset of the parameters (depending on data X) as follows:

$$\mathcal{C}_\alpha(X) = \{\theta_0 : \text{the test does not reject } H_0 : \theta = \theta_0\}.$$

Then it is easy to see that, for any θ_0,

$$\mathsf{P}_{\theta_0}\{\mathcal{C}_\alpha(X) \ni \theta\} = \mathsf{P}_{\theta_0}(\text{test does not reject } H_0 : \theta = \theta_0) = 1 - \mathsf{size} = 1 - \alpha.$$

Therefore, $\mathcal{C}_\alpha(X)$ is a $100(1-\alpha)\%$ confidence interval/region for θ.

Since the likelihood ratio test is a general tool for constructing tests, it makes sense that it would also be used for constructing (approximate) confidence intervals. Define

$$\mathcal{C}_\alpha^{\text{LR}}(X) = \{\theta_0 : -2\log\Lambda(\theta_0) \le k_\alpha\},$$

where $\Lambda(\theta_0) = L(\theta_0)/L(\hat\theta)$ and k_α is either the cutoff required for the likelihood ratio test to have exactly size α, or the chi-square percentile as in the case of Wilks theorem. In the latter case, $\mathcal{C}_\alpha^{\text{LR}}(X)$ is only an asymptotically approximate confidence interval, i.e.,

$$\lim_{n\to\infty} \mathsf{P}_{\theta_0}\{\mathcal{C}_\alpha^{\text{LR}}(X) \ni \theta_0\} = 1 - \alpha.$$

That is, when n is large, the coverage probability for $\mathcal{C}_\alpha^{\text{LR}}(X)$ is approximately $1-\alpha$.

Example 5.4. Let $X_1, \ldots, X_n$ be iid $\mathsf{Beta}(\theta, 1)$, with density $f_\theta(x) = \theta x^{\theta-1}$, $x \in (0,1)$. Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \ne \theta_0$ for some fixed θ_0. We have

$$L(\theta) = \theta^n \Big(\prod_{i=1}^n X_i\Big)^{\theta-1},$$

and, by direct calculation, the MLE is $\hat\theta = -\{n^{-1}\sum_{i=1}^n \log X_i\}^{-1}$. Then

$$-2\log\Lambda(\theta_0) = 2n\log(\hat\theta/\theta_0) - 2(\theta_0 - \hat\theta)\sum_{i=1}^n \log X_i.$$

I simulated $n = 25$ observations from a $\mathsf{Beta}(5,1)$ distribution, and a plot of $-2\log\Lambda(\theta_0)$ as a function of θ_0 is shown in Figure 5.2. The horizontal line at $\chi^2_{1,0.95} = 3.84$, the Wilks test cutoff, determines the (asymptotically approximate) 95% confidence interval on the θ_0-axis. The particular interval in this case is $(3.53, 7.76)$, which contains the true $\theta = 5$.

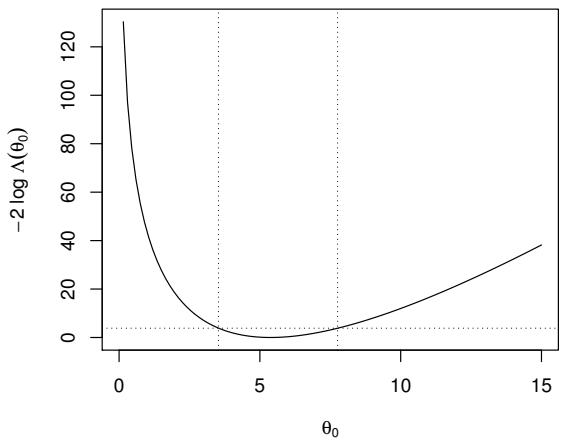


Figure 5.2: Plot of $-2\log\Lambda(\theta_0)$ based on simulated data for the beta distribution problem in Example 5.4.

5.7 Appendix

5.7.1 R code for Example 5.3

```
power.exact <- function(theta) {

  p1 <- pchisq(theta0 * qchisq(alpha / 2, 2 * n) / theta, 2 * n)
  p2 <- 1 - pchisq(theta0 * qchisq(1 - alpha / 2, 2 * n) / theta, 2 * n)
  return(p1 + p2)

}
power.approx <- function(theta) {

  M <- 1e4
  pow <- 0 * theta
  for(i in 1:length(theta)) {

    Xbar <- rgamma(M, shape=n, rate=1 / theta[i]) / n
    lrt <- 2 * n * ((Xbar / theta0 - 1) - log(Xbar) + log(theta0))
    pow[i] <- mean(lrt >= qchisq(1-alpha, 1))

  }
  return(pow)

}
n <- 7
theta0 <- 1
alpha <- 0.05
g <- function(t) n**(-n) * t**n * exp(-(t - n))
curve(g, xlim=c(1, 20), lwd=2, xlab="t", ylab="g(t)")
```

```
abline(h=0.1, lty=3, lwd=2)
Theta <- seq(1, 5, len=20)
curve(power.exact, xlim=range(Theta), lwd=2, xlab=expression(theta),
      ylab=expression(pow(theta)))
lines(Theta, power.approx(Theta), lwd=2, lty=2)
legend(x="bottomright", inset=0.05, lwd=2, lty=1:2, c("Exact", "Approx"))
```

5.7.2 Proof of Neyman–Pearson lemma

Take any test $\varphi = I_C$ in $\mathbb{T}_\alpha$, and let $\varphi^\star$ be the test described in the Neyman–Pearson lemma. Then the power functions, at generic θ, for φ and $\varphi^\star$ are given by

$$\begin{aligned}
\mathsf{pow}(\theta) &= \int_C L_x(\theta)\,dx = \int_{C\cap C_\star} L_x(\theta)\,dx + \int_{C\cap C_\star^c} L_x(\theta)\,dx,\\
\mathsf{pow}^\star(\theta) &= \int_{C_\star} L_x(\theta)\,dx = \int_{C_\star\cap C} L_x(\theta)\,dx + \int_{C_\star\cap C^c} L_x(\theta)\,dx,
\end{aligned}$$

which is a straightforward property of the integral and the decomposition, e.g., $C = (C \cap C_\star) \cup (C \cap C_\star^c)$. Note that the common term—with integration over $C \cap C_\star$—will cancel out if we take a difference between these two. Next note that if we take $\theta = \theta_0$, then both expressions above must equal α by the size assumption, so taking a difference of the two expressions reveals that

$$\int_{C^c\cap C_\star} L_x(\theta_0)\,dx = \int_{C\cap C_\star^c} L_x(\theta_0)\,dx. \tag{5.2}$$

Now we take a difference $\mathsf{pow}^\star(\theta_1) - \mathsf{pow}(\theta_1)$, with the goal of showing that this difference is non-negative. The key is the observation that, if $x \in C_\star$, then $L_x(\theta_1) \geq L_x(\theta_0)/k$ and, likewise, if $x \in C_\star^c$, then $L_x(\theta_1) < L_x(\theta_0)/k$. Therefore,

$$\begin{aligned}
\mathsf{pow}^\star(\theta_1) - \mathsf{pow}(\theta_1) &= \int_{C^c\cap C_\star} L_x(\theta_1)\,dx - \int_{C\cap C_\star^c} L_x(\theta_1)\,dx\\
&\geq \frac{1}{k}\int_{C^c\cap C_\star} L_x(\theta_0)\,dx - \frac{1}{k}\int_{C\cap C_\star^c} L_x(\theta_0)\,dx\\
&= \frac{1}{k}\Big[\int_{C^c\cap C_\star} L_x(\theta_0)\,dx - \int_{C\cap C_\star^c} L_x(\theta_0)\,dx\Big].
\end{aligned}$$

By (5.2), the term inside the brackets is zero and, hence, the difference in powers is non-negative. This completes the proof.

5.7.3 Randomized tests

In some cases, the likelihood ratio has a discrete distribution, e.g., when the X_i's are discrete. Then it may not be possible to achieve exact size α with the test in Theorem 5.1. In practice,

people generally don't concern themselves with this, mostly because the remedy for this size restriction is silly in real problems. The particular remedy is to consider a test which sometimes relies on a flip of a coin to choose between H_0 and H_1. Obviously, basing decisions on the result of a coin flip, even after resources have been spent to collect data $X_1, \ldots, X_n$, is generally unsatisfactory in applications. Nevertheless, the notion of randomized tests is interesting, if only for historical purposes.

The generalization of the Neyman–Pearson most powerful test is to add a constant $p \in (0, 1)$ to the mix, which can be determines based on the size condition. That is, the most powerful size-α randomized test is given by

$$\varphi^\star(X) = \begin{cases} 1 & \text{if } L_X(\theta_0) \,/\, L_X(\theta_1) < k, \\ 0 & \text{if } L_X(\theta_0) \,/\, L_X(\theta_1) > k, \\ p & \text{if } L_X(\theta_0)/L_X(\theta_1) = k, \end{cases}$$

where (k, p) satisfy

$$\alpha = \mathsf{P}_{\theta_0}\Big\{\frac{L_X(\theta_0)}{L_X(\theta_1)} < k\Big\} + p\mathsf{P}_{\theta_0}\Big\{\frac{L_X(\theta_0)}{L_X(\theta_1)} = k\Big\}.$$

The calculation of (k, p) for a given problem starts by finding the largest k such that the first term is $\leq \alpha$; then p can be chosen after k.

The way one interprets a randomized test is as follows. After X is observed, one flips a count with probability $\varphi^\star(X)$ probability of landing on heads; then reject H_0 iff this coin lands on heads. To my knowledge, randomized tests are never used in practice. A perhaps more reasonable strategy is to simply adjust the desired size α to some value that can be achieved in the given problem.

Chapter 6

Bayesian Statistics

6.1 Introduction

Up to know, our focus in Stat 411 has been on what's called *frequentist* statistics. That is, our main objective was sampling distribution properties of our estimators, tests, etc. Why the name "frequentist?" Recall that the sampling distribution of, say, an estimator $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ describes the distribution of $\hat{\theta}$ as the sample $X_1, \ldots, X_n$ varies according to its law. In particular, the sampling distribution describes the *frequency* of certain events concerning $\hat{\theta}$ in repeated sampling.

Here we want to shift gears and briefly discuss the other dominant school of thought in statistics, namely, the *Bayesian* school. The name comes from Reverend Thomas Bayes, who developed *Bayes theorem* you learn in a basic probability course. It is maybe not clear why such a simple result could have paved the way for the substantial amount of work on Bayesian analysis that's been done to date. Presentations of Bayes theorem in basic probability courses usually side-step the philosophical importance of a result like this—it's a fundamental tool that describes how uncertainties can be updated when new information becomes available. So one can think of Bayes' theorem as a statement about how information is processed and beliefs are updated. In the statistics context, Bayes theorem is used to take prior or initial beliefs about the parameter of interest and, after data is observed, those beliefs are updated to reflect what has been learned. Expressed in this way, one should see that Bayes theorem is a bit more than just a simple manipulation of the symmetry in $\mathsf{P}(A \cap B)$.

What sets the Bayesian framework apart from what we've previously seen is the way that uncertainty is defined and represented. The usual setup we've encountered is that observable data $X_1, \ldots, X_n$ is available from a distribution $f_\theta(x)$. The starting point is that θ is unknown and to be estimated/tested from the data. But we've not really said what it means that θ is "unknown." Do we really know nothing about it, or do we not know how to summarize what knowledge we have, or are we uneasy using this knowledge? It seems unrealistic that we actually know *nothing* about θ, e.g., when θ is the mean income in Cook county to be estimated, we know that θ is positive and less than \$1 billion; we'd also likely believe that $\theta \in (\$40K, \$60K)$ is more likely that $\theta \in (\$200K, \$220K)$. In what

we've discussed so far in Stat 411, such information is not used. The jumping off point for a Bayesian analysis is the following belief:

The only way to describe uncertainties is with probability.

Such a concept permeates even our everyday lives. For example, suppose you and several friends have been invited to a party next saturday. When your friend Kevin asks if you will attend, you might respond with something like "there's a 90% chance that I'll go." Although not on the scale of probabilities, this has such an interpretation. The same thing goes for weather reports, e.g., "there's a 30% chance of rain tomorrow." What's particularly interesting is the nature of these probabilities. We're used to using probabilities to describe the uncertain results of a random experiment (e.g., rolling dice). The two scenarios above (party and weather) are not really random; moreover, these are singular events and not something that can repeated over and over. And yet probabilities—in particular, *subjective probabilities*—are introduced and can be used. In the statistics problem, to that thing θ about which we are uncertain, we must assign probabilities to certain events, such as $\{\theta > 7\}$, $\{-0.27 \leq \theta < 0.98\}$, etc. Once this probability assignment has been made to all such events, what we have is a probability distribution, what's called the *prior distribution* for θ. This is effectively the same as assuming that the unknown parameter itself is a random variable with a specified distribution. It is a common misconception to say that Bayesian analysis assumes the parameter is a random variable. On the contrary, a Bayesian starts by assigning probabilities to all such things which uncertain; that this happens to be equivalent to taking θ to be a random variable is just a (convenient or unfortunate?) consequence.

There are reasons to take a Bayesian approach, other than these rather philosophical reasons mentioned above. In fact, there is a remarkable theorem of deFinetti which says that, if data are exchangeable, i.e., if permuting the data does not change its joint distribution, then there exists a likelihood and prior like assumed in the Bayesian setting. Also, there is a very surprising result that says, roughly, given any estimator, there exists a Bayes (or approximate Bayes) estimator that's as good or better in terms of mean-square error. In other words, one cannot do too bad by using Bayes estimators. Finally, there's some advantage to the Bayesian approach in the high-dimensional problems because a suitable Bayes model will result in some automatic penalties on models with higher dimensionality. Outside the Bayesian context, one must actively introduce such penalties. Some additional details about these and other aspects of Bayesian analysis can be found in Section 6.4.

As you can probably tell already, Bayesian analysis is not so easy to come to grips with, at least at first. Here, we will not dwell on these philosophical matters. My focus in these notes is to give you a basic introduction to the ideas and terminology in the Bayesian setting. In particular, I hope that students will understand the basic steps in a Bayesian analysis—choosing a prior distribution, updating the prior to a posterior distribution using data and Bayes theorem, and summarizing this posterior. Some additional important points are mentioned in the last section.

6.2 Mechanics of Bayesian analysis

6.2.1 Ingredients

The Bayesian problem starts with an important additional ingredient compared to the frequentist problem we've considered so far. This additional ingredient is the *prior distribution* for the parameter θ. Here I will modify our familiar notation a bit. Now Θ will denote a random variable version of the parameter θ—we'll let the parameter space (our usual use of the notation Θ) be implicit and determined by context.

The prior distribution for the parameter is a statement that $\Theta \sim \pi(\theta)$, where $\pi(\theta)$ is a distribution (i.e., a PDF/PMF) defined on the parameter space. The idea is that the distribution $\pi(\theta)$ encodes our uncertainty about the parameter. For example, if θ is the mean income in Cook County, my belief that this value is between \$25K and \$35K is given by the probability calculation $\int_{25K}^{35K} \pi(\theta)\, d\theta$. Where this prior distribution comes from is an important question—see Section 6.3—but for now just take π as given.

In addition to our prior beliefs about the parameter, we get to observe data just like in our previous settings. That is, given $\Theta = \theta$, we get $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f_\theta(x)$. Note the emphasis here that $f_\theta(x)$ is the conditional distribution of X_1 given the random parameter Θ happens to equal the particular value θ. From this sample, we can again define a likelihood function $L(\theta) = \prod_{i=1}^n f_\theta(X_i)$. I may add a subscript $L_X(\theta)$ to remind us that the likelihood is a function of θ but it depends on data $X = (X_1, \ldots, X_n)$.

To summarize, the Bayesian model specifies a joint distribution for $(\Theta, X_1, \ldots, X_n)$, with a "density"[1] $\pi(\theta)L(\theta)$. But this specification is done in two stages: first a marginal distribution for Θ and then a conditional distribution for $(X_1, \ldots, X_n)$ given $\Theta = \theta$. Such a model could be called a hierarchical model, because it's done in stages; I'll call it a *Bayes model*.

6.2.2 Bayes theorem and the posterior distribution

The key feature of a Bayesian analysis is that the prior distribution for Θ is updated after seeing data to what is called the *posterior distribution*. Bayesian inference is based entirely on this posterior distribution—see Section 6.2.3. The key to this transformation is the simple Bayes' formula: for given a probability space, consisting of a collection of events and a probability P, if A and B are events with positive P-probability, the conditional probability $\mathsf{P}(A \mid B)$, defined as $\mathsf{P}(A \cap B)/\mathsf{P}(B)$ satisfies

$$\mathsf{P}(A \mid B) = \mathsf{P}(B \mid A)\mathsf{P}(A)/\mathsf{P}(B).$$

In our case, A is some event concerning Θ and B corresponds to our observable data. To make the connection more precise, consider a simple discrete problem involving only PMFs. Let Θ take two values—0.25 and 0.75—with equal probability and, given $\Theta = \theta$, let $X \sim \mathsf{Ber}(\theta)$.

[1] Quotes here indicate that $\pi(\theta)L(\theta)$ may not be a genuine PDF for $(\Theta, X_1, \ldots, X_n)$ because the data may be discrete and parameter may be continuous or vice-versa. But this is not really a problem.

The posterior probability that $\Theta = 0.25$, given $X = x$, can be obtained via Bayes formula:

$$\mathsf{P}(\Theta = 0.25 \mid X = x) = \frac{f_{0.25}(x) \cdot 0.5}{f_{0.25}(x) \cdot 0.5 + f_{0.75}(x) \cdot 0.5}.$$

Depending on which x is used, this can be easily evaluated. In general, $\mathsf{P}(\Theta = 0.25 \mid X = x)$ will be different from $\mathsf{P}(\Theta = 0.25)$, so Bayes formula is, indeed, updating our prior beliefs about the possible values of Θ.

This is a very simple version of the problem. Fortunately, there is a version of Bayes formula for the case where both data and parameter are continuous. The proof is more difficult than the simple Bayes formula, but the formula itself is very similar. Essentially, one just pretends that $\pi(\theta)$ and $L(\theta)$ are PMFs instead of PDFs, and apply the familiar Bayes' formula. In particular, the posterior distribution of Θ, given $X = x$, has a PDF/PMF $\pi(\theta \mid x)$ given by

$$\pi(\theta \mid x) = \frac{L_x(\theta)\pi(\theta)}{\int L_x(\theta)\pi(\theta)\, d\theta},$$

where, in the denominator, integration is over the entire parameter space. If Θ is a discrete random variable, i.e., it's prior is a PMF instead of a PDF, the formula looks the same but the denominator has a sum over all possible θ values instead of an integral.

Exercise 6.1. Suppose that, given $\Theta = \theta$, $X \sim \mathsf{Bin}(n, \theta)$. If $\Theta \sim \mathsf{Unif}(0, 1)$, find the posterior distribution $\pi(\theta \mid x)$.

6.2.3 Bayesian inference

Bayesian analysis is really focused on the posterior distribution, which is assumed to describe all uncertainty about the parameter after data $X = x$ is observed. But it is often of interest to answer questions like those we've encountered previously in the course. For example, if we want to produce an estimator or do a hypothesis test, we can do such things from a Bayesian perspective as well.

Point estimation

The most basic problem in statistics is one of producing an estimator of unknown parameter θ. Of course, in the Bayes context, the parameter is random, not fixed, so it's not immediately clear what we're trying to estimate. The way to understand this from a Bayesian point of view is that the goal is to report the "center" of the posterior distribution $\pi(\theta \mid x)$, which is some function of the observed $X = x$, as statistic. This "center" is often the mean of the posterior distribution, but it doesn't have to be.

To properly describe how Bayes methods are derived, one should first introduce what is called a loss function $\ell(\delta(x), \theta)$, which measures the penalty one incurs by using, say, an procedure generically written $\delta(x)$ when the true parameter value is θ. Once this loss function is specified, the Bayes method is derived by choosing $\delta(x)$ to minimize the expectation $\int \ell(\delta(x), \theta)\pi(\theta \mid x)\, d\theta$, called the posterior Bayes risk. In the estimation problem, the loss

function is usually $\ell(\hat{\theta}(x), \theta) = (\theta - \hat{\theta}(x))^2$, called squared-error loss, so the goal is to choose $\hat{\theta}$ such that

$$\int (\theta - \hat{\theta}(x))^2 \pi(\theta \mid x)\, dx \quad \text{is minimized.}$$

It is a relatively simple calculus exercise to show that the minimizer is the mean of the posterior distribution, i.e., the Bayes estimator of θ is

$$\hat{\theta}(x) = \mathsf{E}(\Theta \mid x) = \int \theta \pi(\theta \mid x)\, d\theta.$$

If the loss function is different from squared-error, then the Bayes estimator would be something different. For Stat 411, we will only consider squared-error loss.

Exercise 6.2. Reconsider the binomial problem in Exercise 6.1. Find the posterior mean $\mathsf{E}(\Theta \mid x)$ and calculate its mean-square error. Compare it to that of the MLE $\hat{\theta} = \bar{X}$.

Hypothesis testing

The hypothesis testing problem also has a loss function type of description, but I'll not get into that here because it's a bit more difficult to explain in this context. I'll also use some notation that's a bit different from what we've used before, though it should be clear. Start with the hypotheses $H_0 : \theta \in U_0$ versus $H_1 : \theta \in U_1$, where $U_1 = U_0^c$. With a prior distribution π for Θ, we can calculate prior probabilities for H_0 and H_1, which are $\pi(U_0) = \int_{U_0} \pi(\theta)\, d\theta$ and $\pi(U_1) = 1 - \pi(U_0)$, respectively. Now we just want to update these probabilities in light of the observed data. In other words, we calculate the posterior probabilities of H_0 and H_1, which are readily available once we have the posterior distribution $\pi(\theta \mid x)$. These probabilities are

$$\pi(U_0 \mid x) = \int_{U_0} \pi(\theta \mid x)\, d\theta \quad \text{and} \quad \pi(U_1 \mid x) = 1 - \pi(U_0 \mid x)$$

Finally, to decide whether to reject H_0 or not, we just compare the relative magnitudes of $\pi(U_0 \mid x)$ and $\pi(U_1 \mid x)$ and choose the larger of the two. Because $U_1 = U_0^c$, it is easy to see that we reject H_0 if and only if $\pi(U_0 \mid x) < 1/2$. This is the Bayes test.

One remark to make here is that, in a certain sense, we can accomplish here, in the Bayesian setting, what we could not in the frequentist setting. That is, we know have measures of certainty that H_0 is true/false given data. The frequentist approach has nothing like this—size and power are only sampling distribution properties and have nothing to do with observed data. The trade-off is that, in the Bayesian setting, one requires a prior distribution, and if the prior distribution is "wrong," then the Bayes test may give unsatisfactory answers. Also, note that the Bayes test has no "α" involved, so it makes no attempt to control the tests size at a particular level. Of course, the test has a size and power associated with it, but one must calculate these separately. For example, the size of the Bayes test would look like

$$\max_{\theta \in U_0} \mathsf{P}_\theta\{\pi(U_0 \mid X) < 1/2\}.$$

If the posterior probability $\pi(U_0 \mid x)$ has a nice looking form, then perhaps this calculation would not be too difficult. Alternatively, one could use Monte Carlo to evaluate this. But keep in mind that size and power and not really meaningful to a Bayesian, so there's no need to do these things unless it's of interest to compare the performance of a Bayes test with that of a frequentist test.

Notice that the explanation above implicitly requires that $\pi(U_0) > 0$. If $H_0 : \theta = \theta_0$ and Θ is continuous, then this will automatically be zero, so we need to adjust the methodology. The standard approach here is to use Bayes factors to perform the test. For the nice problems we consider in Stat 411, the necessary Bayes factor calculations are relatively simple, and the result resembles a likelihood ratio test. But the interpretation of Bayes factors is less straightforward than that of the posterior probabilities $\pi(U_0 \mid x)$ from above, so I'll not discuss this here. A formal course on Bayesian analysis would not ignore this issue.

Credible intervals

Bayesians have an analog of the frequentist's confidence intervals, which are called *credible intervals*. There are a couple of ways to do this, which I'll explain only briefly. Throughout, take $\alpha \in (0, 1)$ fixed.

- *Equal-tailed credible interval.* Let $\Pi(\theta \mid x)$ denote the posterior CDF. An equal-tailed $100(1-\alpha)\%$ credible interval looks like

$$\{\theta : \alpha/2 \leq \Pi(\theta \mid x) \leq 1 - \alpha/2\}.$$

  This is just the central $1 - \alpha$ region for the posterior distribution.

- *Highest posterior density credible interval.* For a cutoff $c > 0$, define an interval[2]

$$H(c) = \{\theta : \pi(\theta \mid x) \geq c\}.$$

  Note that, as c changes, the posterior probability of $H(c)$ changes, in particular, this posterior probability increases as c decreases, and vice versa. If it varies continuously, then by the intermediate value theorem, there is a point $c = c_\alpha$ such that the posterior probability of $H(c_\alpha)$ equals $1 - \alpha$. The set $H(c_\alpha)$ is the $100(1-\alpha)\%$ highest posterior density credible interval.

Note that these credible intervals are not understood the same way as confidence intervals. In particular, they may not have coverage probability equal to $1 - \alpha$. In this case, $1 - \alpha$ represents the amount of probability the posterior assigns to the interval. But in many cases, Bayesian credible intervals do have reasonable frequentist coverage probabilities, though this is not guaranteed by the construction.

[2]This set may not be an interval if the posterior PDF $\pi(\theta \mid x)$ has multiple bumps.

6.2.4 Marginalization

An important point that was glossed over in the earlier chapters concerns marginal inference. That is, suppose $\theta = (\theta_1, \theta_2)$ is a pair of unknowns, but interest is only in θ_1. In that case, we consider θ_2 a *nuisance parameter*, and inference about θ_1 in the presence of a nuisance parameter θ_2 is called *marginal inference*. Consider, for example, a $\mathsf{Gam}(\theta_1, \theta_2)$ problem. We can produce the MLE for the pair (θ_1, θ_2) numerically (we did this in Chapter 3) and we know that the limiting distribution is normal with covariance matrix characterized by the inverse of the Fisher information matrix. However, how might we construct a confidence interval for θ_1? A natural choice would be "$\hat{\theta}_1 \pm$ something" but it's not exactly clear how to pick the "something." The main challenge is that the θ_1-component of the asymptotic covariance matrix depends on both parameters (θ_1, θ_2). So, a natural choice is to plug in $(\hat{\theta}_1, \hat{\theta}_2)$ into that covariance and do the confidence interval thing as usual. (Another idea is to use some variance stabilizing transformation, but this is tricky and may not always work.) The problem is that there is no way to be sure that plug-in estimators of the confidence interval margin of error will be reliable in terms of coverage probabilities, etc. This question of how to do marginal inference turns out to be rather difficult in classical statistics. However, it is straightforward in the Bayesian setting.

If I have a prior density $\pi(\theta_1, \theta_2)$, then I can easily get a posterior density $\pi(\theta_1, \theta_2 \mid x)$ using Bayes formula. Now, if we want to do marginal inference on θ_1 as a Bayesian, all we need is the *marginal posterior* density $\pi(\theta_1 \mid x)$, which is given by

$$\pi(\theta_1 \mid x) = \int \pi(\theta_1, \theta_2 \mid x)\, d\theta_2,$$

which is exactly the usual formula for the marginal density given in basic probability courses. So, the basic rules of probability tell the Bayesian exactly how to do marginal inference. Once the marginal posterior density is available, carry out the summaries just as before, e.g., get a marginal credible interval for θ_1.

The point to be made here is that the Bayesian marginalization operations deals with the uncertainty about θ_2 in a natural way. In the classical setting, the naive strategy is to plug in an estimator of θ_2, which can perhaps over- or under-estimate the variability, giving unreliable marginal confidence intervals. This is not to say that the Bayesian version is without shortcomings, but at least it is simple and not obviously flawed.

6.3 Choice of prior

In the previous discussion, the prior distribution seem to appear from thin air. Indeed, if a prior is provided, then the statistics problem really reduces to some simple algebra/calculus. In real life, however, one must select the prior. Here I discuss the three ways in which one can come up with the prior.

6.3.1 Elicitation from experts

The most natural way to get a prior distribution is to go and ask experts about the problem at hand what they expect. For example, if $\theta \in (0,1)$ represents the probability that some widget manufacturing machine will produce a defective widget, then the statistician might go to speak to the experts, the people who designed and built the machine, to get their opinions about reasonable values of θ. These experts might be able to tell the statistician some helpful information about the shape or percentiles of the prior distribution, which can be used to make a specific choice of prior. However, this can be expensive and time-consuming, and may not result in reasonable prior information, so this is rarely done in practice.

6.3.2 Convenient priors

Before we had high-powered computers readily available, Bayesian analysis was mostly restricted to the use of "convenient" priors. The priors which are convenient for a particular problem, may not be so realistic. This is perhaps the biggest reason it took so long for Bayesian analysis to catch on in the statistics community. Although computation with realistic priors is possible these days, these convenient are still of interest and can be used as prior distributions for high-level hyperparameters in hierarchical models.

The most popular class of "convenient priors" are called *conjugate priors*. A class of priors is said to be conjugate for a given problem if, when combined with the likelihood in Bayes' formula, the resulting posterior distribution is also a member of this class. This is convenient because calculations for the posterior distribution can often be done in closed-form or with very easy numerical methods. Here are a few examples.

Example 6.1. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Ber}(\theta)$ given $\Theta = \theta$. We shall consider a prior distribution $\Theta \sim \mathsf{Beta}(a, b)$, where a and b are some fixed positive numbers. Here the likelihood is

$$L(\theta) = c\theta^{\sum_{i=1}^n X_i}(1-\theta)^{n-\sum_{i=1}^n X_i},$$

where c is a constant depending on n and X_i's only. Also, the prior density for Θ is

$$\pi(\theta) = c\theta^{a-1}(1-\theta)^{b-1},$$

where c is a (different) constant depending on a and b only. We can see some similarities in these two formulas. Indeed, if we multiply them together, we see that the the posterior distribution $\pi(\theta \mid x)$ satisfies

$$\pi(\theta \mid x) = C\theta^{a+\sum_{i=1}^n x_i - 1}(1-\theta)^{n+b-\sum_{i=1}^n x_i - 1},$$

where C is some constant depending on n, x_i's, a and b. This is clearly of the same form as as the prior density $\pi(\theta)$, but with different parameters. That is, the posterior distribution for Θ, given $X = x$, is also a beta distribution but with parameters $a' = a + \sum_{i=1}^n x_i$ and $b' = b + n - \sum_{i=1}^n x_i$. So, the Bayes estimate of θ can be found easily from the standard formula for the mean of a beta distribution, i.e.,

$$\hat{\theta}(x) = \mathsf{E}(\Theta \mid x) = \frac{a'(x)}{a'(x) + b'(x)} = \frac{a + n\bar{x}}{a + b + n}.$$

This estimator depends on both the prior parameters (a, b) and the observable data. In fact, the posterior mean is just a weighted average of the prior mean $a/(a+b)$ and the data mean $\bar{x}$. This is a fairly general phenomenon.

Exercise 6.3. Suppose the model is $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{N}(\theta, 1)$ given $\Theta = \theta$. Show that the class of normal distributions for Θ is conjugate.

Exercise 6.4. Suppose the model is $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Pois}(\theta)$ given $\Theta = \theta$. Show that the class of gamma distributions for Θ is conjugate.

Conjugate priors are nice to work with but, as mentioned above, may not be realistic representations of prior uncertainty. Some additional flexibility can be achieved, while keeping the convenience, by working with priors that are "mixtures" of conjugate priors.

6.3.3 Non-informative priors

Since a "real" prior is rarely available, there can be a number of reasonable priors to choose from and, generally, the posterior distributions for these different priors will have some differences. How can one justify a particular choice of prior? One popular strategy is to choose the prior that, in some sense, influences the posterior as little as possible, so that the data drives the Bayesian analysis more than the choice of prior. Such a prior is called *non-informative*[3]—it gives data the freedom to find the "best" parameters.

It is not even clear how to define non-informative priors. It is easy, however, to give an example of an informative prior: take $\Theta = \theta_0$ with prior probability 1. In this case, the posterior also assigns probability 1 to the value θ_0, so the data can do nothing to identify a different parameter value. This is an extreme example, but the general idea is similar—a prior is informative if it imposes too much restriction on the posterior. Formal definitions of non-informative priors are complicated and somewhat technical,[4] so I won't go into these. I will try to give some intuition, and a good technique for choosing a non-informative prior in standard priors.

One might think that some kind of uniform prior is, in some sense, non-informative because all values are equally likely. For along time, this was the case—both Bayes and Laplace, the first "Bayesians," used such a prior in their examples. But Fisher astutely criticized this approach. Here is my summary of Fisher's argument, say, for Θ the success probability for a binomial experiment:

> Suppose we don't know where Θ is likely to be, and for this reason we choose to take $\Theta \sim \mathsf{Unif}(0, 1)$. Then we also don't know anything about $\Lambda = -\log \Theta$ either, so, if we express the problem in terms of Λ, then we should also take a uniform prior for Λ. But there's an inconsistency in the logic, because we should also be

[3] Such priors are also sometimes called *objective*—I think this is a potentially misleading name.

[4] The formal definitions require some notion of an infinite sequence of experiments and prior is chosen to maximize the limiting distance between prior and posterior or, alternatively, by some sort of "probability matching."

able to apply standard results from probability theory, in particular, transformation formulas, to get the prior for Λ from the prior for Θ. But it's easy to see that a uniform prior for Θ does not correspond to a uniform prior for Λ obtained from the transformation theorem of PDFs. Hence, a logical inconsistency.

So, this suggests uniform is not a good definition of non-informative. However, with an adjustment to how we understand "uniformity," we can make such a definition. In fact, the priors we will discuss next are uniform in this different sense.[5]

A standard form of non-informative prior is *Jeffreys' prior*. This has a close connection to some of the calculations we did when studying properties of maximum likelihood estimators. In particular, these priors depend critically on the Fisher information. Suppose that, given $\Theta = \theta$, data are modeled as $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f_\theta(x)$. From the distribution $f_\theta(x)$, we can calculate the Fisher information:

$$I(\theta) = -\mathsf{E}_\theta\Big\{\frac{\partial^2}{\partial\theta^2}\log f_\theta(X_1)\Big\}.$$

Then Jeffreys' prior is defined as

$$\pi_J(\theta) = cI(\theta)^{1/2},$$

where $c > 0$ is a constant to make π_J integrate to 1; sometimes there is no such constant c, so it can be chosen arbitrarily. I cannot explain the properties π_J has here, but a very compelling case for Jeffreys' prior in low-dimensional problems is given in Chapter 5 of Ghosh, Delampady, and Samanta, *Introduction to Bayesian Analysis*, Springer 2006.

Example 6.2. Suppose, given $\Theta = \theta$, data $X_1, \ldots, X_n$ are iid $\mathsf{Ber}(\theta)$. The Fisher information for this problem is $I(\theta) = \{\theta(1-\theta)\}^{-1}$, so Jeffreys' prior is

$$\pi_J(\theta) = \frac{c}{\theta^{1/2}(1-\theta)^{1/2}}.$$

It is easy to check that this is a special case of the $\mathsf{Beta}(a, b)$ priors used in Example 6.1 above, with $a = b = 1/2$. Therefore, the posterior distribution of Θ, given $X = x$, is simply $\Theta \mid x \sim \mathsf{Beta}(n\bar{x} + 1/2, n - n\bar{x} + 1/2)$, and, for example, the posterior mean is $\mathsf{E}(\Theta \mid x) = (n\bar{x} + 1/2)/(n+1)$, which is very close to the MLE $\bar{x}$. Notice, as we could anticipate from the previous discussion, the non-informative prior is *not* uniform!

Exercise 6.5. Find the Jeffreys' prior π_J for Θ when, given $\Theta = \theta$, data are iid $\mathsf{N}(\theta, 1)$. Calculate the corresponding posterior distribution.

Exercise 6.6. Find the Jeffreys' prior π_J for Θ when, given $\Theta = \theta$, data are iid $\mathsf{Pois}(\theta)$. Calculate the corresponding posterior distribution.

[5] One should use a different sort of geometry defined on the parameter space, not necessarily the usual Euclidean geometry. With this adjusted geometric, basically a Riemannian metric induced by the Fisher information, one can recover the Jeffreys' prior as a sort of uniform distribution in this geometry.

There are some potential difficulties with Jeffreys' prior that should be mentioned. First, it can happen that $I(\theta)^{1/2}$ does not integrate to a finite number, which means that Jeffreys' prior may not be a *proper prior*. People generally do not concern themselves with this issue, provided that the posterior is proper. So, when using non-informative improper priors, one should always check that the corresponding posterior is proper before using it for inference. If it's not proper, then posterior means, for example, are meaningless. The second point is a more philosophical one: by using Jeffreys' prior, one assumes that the prior for Θ depends on the model for data. Ideally, the prior is based on our beliefs about Θ, which has no direct connection with the particular model we use for observable data.

6.4 Other important points

In this last section, I shall give some brief comments on some important aspects of Bayesian analysis that we don't have time to discuss in detail in Stat 411. In a formal Bayesian analysis course, one would surely spend a bit of time on each of these items.

6.4.1 Hierarchical models

The particular models considered here are so simple that it really doesn't matter what kind of analysis one uses, Bayesian or non-Bayesian, the result will generally be the same. However, in more complicated problems, the advantages of a Bayesian analysis become more pronounced. A now classical problem is the so-called "many-normal-means" problem. Suppose $X_1, \ldots, X_n$ are independent (not iid) with $X_i \sim \mathsf{N}(\theta_i, 1)$, $i = 1, \ldots, n$. The maximum likelihood (or least-squares) estimator of the mean vector $\theta = (\theta_1, \ldots, \theta_n)^\top$ is the observed data vector $X = (X_1, \ldots, X_n)^\top$. However, there is a famous result of C. Stein which says that this estimator is bad (for a certain kind of loss function) whenever $n \geq 3$, in particular, there is another estimator that is at least as good for all possible θ values. For this reason, one must make some kind of adjustments to the MLE, and the particular adjustments are often of the form of "shrinking" X towards some fixed point in $\mathbb{R}^n$. This shrinkage is somewhat ad hoc, so it might be nice to have a more formal way to accomplish this.

In a Bayesian setting, this can be accomplished quite easily with a hierarchical model. Consider the following model:

$$\begin{aligned} X_i \mid (\theta_i, v) &\sim \mathsf{N}(\theta_i, 1), \quad i = 1, \ldots, n \quad \text{(independent)} \\ \Theta_1, \ldots, \Theta_n \mid v &\overset{\text{iid}}{\sim} \mathsf{N}(0, v) \\ V &\sim \pi(v). \end{aligned}$$

The idea is that there are several layers of priors—one for the main parameters $\theta = (\theta_1, \ldots, \theta_n)^\top$ and one for the "hyperparameter" v. The key to the success of such a model is that one can initially marginalize away the high-dimensional parameter θ. Then there is lots of data which contains considerable information for inference on V. That is, the posterior distribution $\pi(v \mid x)$ is highly informative. The goal then is to calculate the posterior mean of Θ_i as

an estimate:

$$\mathsf{E}(\Theta_i \mid x_i) = \int \mathsf{E}(\Theta_i \mid x_i, v)\pi(v \mid x)\, dv.$$

The inner expectation has a closed-form expression, $vx_i/(1+v)$, so it would be easy to approximate the full expectation by Monte Carlo. The point, however, is that the resulting estimator with have certain "shrinkage" features, since $v/(1+v) \in (0,1)$, which occur automatically in the Bayesian setup—no ad hoc shrinkage need be imposed.

As an alternative to the "full Bayes" analysis described above, one could perform an *empirical Bayes* analysis and estimate v from the data, say, by $\hat{v}$; typically, $\hat{v}$ is found via marginal maximum likelihood. Then one would estimate θ_i via

$$\hat{\theta}_i = \mathsf{E}(\Theta_i \mid x_i, \hat{v}) = \frac{\hat{v}x_i}{1+\hat{v}}, \quad i = 1, \ldots, n.$$

Such empirical Bayes strategies are popular nowadays.

6.4.2 Complete-class theorems

There are a collection of results, which fall under the general umbrella of "complete-class theorems," which gives some frequentist justification for Bayesian methods. Recall that in a frequentist estimation problem, one is looking for estimators which have small mean-square error. Complete-class theorems identify the collection of estimators which are *admissible*—not uniformly worse than another estimator. In other words, admissible estimators are the only estimators one should consider, if one cares about mean-square error. There is one such theorem which says that, roughly, Bayes estimators (and limits thereof) form a complete class. That is, one can essentially restrict their frequentist attention to estimators obtained by taking a prior distribution for Θ and produce the posterior mean as an estimator. So, you see, Bayesian methods are not only of interest to Bayesians.

6.4.3 Computation

We did not discuss this matter in these notes, but computation is an important part of Bayesian analysis, perhaps more important than for frequentist analysis. Typically, unless one uses a convenient conjugate prior, some kind of numerical simulations are needed to carry out a Bayesian analysis. There are so many such techniques available now, but they all fall under the general class of a Monte Carlo method, the most popular are Gibbs samplers and Markov chain Monte Carlo (MCMC). One could devote an entire course to MCMC, so I'll not get into any details here. Instead, I'll just show why such methods are needed.

Suppose I want to estimate θ. I have a Bayes model with a prior π and, at least formally, I can write down what the posterior distribution $\pi(\theta \mid x)$ looks like. But to estimate θ, I need to calculate the posterior mean $\mathsf{E}(\Theta \mid x)$, which is just the integral $\int \theta\pi(\theta \mid x)\, d\theta$. One way to approximate this is by Monte Carlo. That is,

$$\mathsf{E}(\Theta \mid x) \approx \frac{1}{T}\sum_{t=1}^{T} \theta^{(t)},$$

where $\theta^{(1)}, \ldots, \theta^{(T)}$ is an (approximately) independent sample from the posterior distribution $\pi(\theta \mid x)$. If such a sample is can be obtained, and if T is sufficiently large, then we know by the law of large numbers that this sample average will be a good approximation to the expectation. These MCMC approaches strive to obtain this sample of $\theta^{(t)}$'s efficiently and with as little input as possible about the (possibly complex) posterior distribution $\pi(\theta \mid x)$.

6.4.4 Asymptotic theory

We discussed briefly the question about how to choose the prior. But one can ask if the choice of prior really matters that much. The answer is "no, not really" in low-dimensional problems because there is an asymptotic theory in Bayesian analysis which says that, under suitable conditions, the data wash away the effect of the prior when n is large. These results can be rather technical, so here I'll only briefly mention the main result. It parallels the asymptotic normality of MLEs discussed earlier in the course.

Let $I(\theta)$ be the Fisher information and $\hat{\theta}_n$ the MLE. Then the Bayesian asymptotic normality, or the Bernstein–von Mises theorem, says that the posterior distribution of Θ, given $X = x$, is approximately normal with mean $\hat{\theta}_n$ and variance $[nI(\hat{\theta}_n)]^{-1}$. More formally, under suitable conditions on likelihood and prior,

$$[nI(\hat{\theta}_n)]^{1/2}(\Theta - \hat{\theta}_n) \to \mathsf{N}(0,1), \quad \text{in distribution.} \tag{6.1}$$

(This is not the only version, but I think it's the simplest.) Notice here that the statement looks similar to that of the asymptotic normality of the MLE, except that the order of the terms is rearranged. Here the random variable is Θ and the distribution we're talking about is it's posterior distribution. Stripping away all the details, what the Bernstein–von Mises theorem says is that, no matter the prior, the posterior distribution will look like a normal distribution when n is large. From this, one can then reach the conclusion that the choice of prior does not really matter, provided n is sufficiently large.

An interesting tool that can be used to prove this Bernstein–von Mises theorem is what is called the *Laplace Approximation*. This is a general method for analytic approximation of integrals. The result holds for integrals of functions over $\mathbb{R}^p$, though here we will focus on the case $p = 1$. Consider an integral of the form

$$\text{integral} = \int q(\theta) e^{nh(\theta)}\, d\theta,$$

where both q and h are smooth functions of a scalar θ. Here it is assumed that n is large or increasing to ∞. Let $\hat{\theta}$ be the unique maximum of h. Then the Laplace approximation provides a way to calculate the integral without integration—only optimization!

Theorem 6.1 (Laplace approximation)**.** *Let h' and h'' denote derivatives of h. Then, as $n \to \infty$,*

$$\textit{integral} = q(\hat{\theta}) e^{nh(\hat{\theta})} (2\pi)^{1/2} n^{-1/2} \{-h''(\hat{\theta})\}^{-1/2} \{1 + O(n^{-1})\}.$$

(Note that $h''(\hat{\theta})$ is negative, by assumption, so $-h''(\hat{\theta})$ is positive.)

Sketch of the proof. The first observation is that, if h has a unique maximum at $\hat{\theta}$ and n is very large, then the primary contribution to the integral is in a small interval around $\hat{\theta}$, say $\hat{\theta} \pm a$. Second, since this interval is small, and $q(\theta)$ is smooth, it is reasonable to approximate $q(\theta)$ by the constant function $q(\hat{\theta})$ for $\theta \in (\hat{\theta} - a, \hat{\theta} + a)$. Now the idea is to use a Taylor approximation of $h(\theta)$ up to order two around $\theta = \hat{\theta}$:

$$h(\theta) = h(\hat{\theta}) + h'(\hat{\theta})(\theta - \hat{\theta}) + (1/2)h''(\hat{\theta})(\theta - \hat{\theta})^2 + \text{error}.$$

Since $h'(\hat{\theta}) = 0$ by definition of $\hat{\theta}$, plugging this into the exponential term in the integral (and ignoring the error terms) gives

$$\begin{aligned}
\text{integral} &\approx \int_{\hat{\theta}-a}^{\hat{\theta}+a} q(\theta) \exp\{n[h(\hat{\theta}) + (1/2)h''(\hat{\theta})(\theta - \hat{\theta})^2]\}\, d\theta \\
&\approx \int_{\hat{\theta}-a}^{\hat{\theta}+a} q(\hat{\theta}) \exp\{n\hat{\theta} - (\theta - \hat{\theta})^2/2\sigma^2\}\, d\theta \\
&= q(\hat{\theta}) e^{nh(\hat{\theta})} \int_{\hat{\theta}-a}^{\hat{\theta}+a} e^{-(\theta-\hat{\theta})^2/2\sigma^2}\, d\theta,
\end{aligned}$$

where $\sigma^2 = [-nh''(\hat{\theta})]^{-1}$, small. The last integrand looks almost like a normal PDF, except that it's missing $(2\pi\sigma^2)^{-1/2}$. Multiply and divide by this quantity to get

$$\text{integral} \approx q(\hat{\theta}) e^{nh(\hat{\theta})} (2\pi)^{1/2} n^{-1/2} [-h''(\hat{\theta})]^{-1/2}. \qquad \square$$

One simple yet interesting application of the Laplace approximation is the famous *Stirling's approximation* of $n!$. By writing $n! = \Gamma(n+1)$ in integral form, we get

$$n! = \Gamma(n+1) = \int_0^\infty \theta^n e^{-\theta}\, d\theta = \int_0^\infty e^{n(\log\theta - \theta/n)}\, d\theta.$$

In this case, $h(\theta) = \log\theta - \theta/n$, $q(\theta) = 1$, $\ddot{h}(\theta) = -1/\theta^2$, and, obviously, $\hat{\theta} = n$. Therefore, Laplace approximation with $p = 1$ gives

$$n! \approx e^{n(\log n - n/n)} (2\pi)^{1/2} n^{-1/2} n = (2\pi)^{1/2} n^{n+1/2} e^{-n},$$

which is Stirling's approximation.

The connection to the normal distribution we found in the above proof sketch is the key to posterior normality. The idea is to approximate the log-likelihood by a quadratic function via Taylor approximation. Let $U = n^{1/2}(\theta - \hat{\theta})$ be the rescaled parameter value. Then

$$\Pi_n(-a < U < a) = \Pi_n(\hat{\theta} - an^{-1/2} < \theta < \hat{\theta} + an^{-1/2}) = \frac{\text{num}}{\text{den}}.$$

Write $L_n(\theta)$ for likelihood. Letting

$$q(\theta) = \pi(\theta) \quad \text{and} \quad h(\theta) = \frac{1}{n} \log L_n(\theta) = \frac{1}{n} \sum_{i=1}^n \log f_\theta(X_i),$$

then the denominator above can be approximated, via Theorem 6.1, by

$$\text{den} = \int L_n(\theta)\pi(\theta)\,d\theta = \int \pi(\theta)e^{nh(\theta)}\,d\theta \approx L_n(\hat\theta)\pi(\hat\theta)(2\pi/nv)^{1/2},$$

where $v = -h''(\hat\theta) = -n^{-1}\sum_{i=1}^n (\partial^2/\partial\theta^2)\log f_\theta(X_i)\big|_{\theta=\hat\theta}$; this is called the "observed" Fisher information.[6] The numerator can be similarly approximated:

$$\text{num} = \int_{\hat\theta - an^{-1/2}}^{\hat\theta + an^{-1/2}} L_n(\theta)\pi(\theta)\,d\theta \approx L_n(\hat\theta)\pi(\hat\theta)n^{-1/2}\int_{-a}^{a} e^{-vu^2/2}\,du.$$

Taking the ratio of num to den gives

$$\Pi_n(-a < U < a) = \frac{\text{num}}{\text{den}} \approx \frac{L_n(\hat\theta)\pi(\hat\theta)n^{-1/2}\int_{-a}^{a} e^{-vu^2/2}\,du}{L_n(\hat\theta)\pi(\hat\theta)(2\pi/nv)^{1/2}} = \int_{-a}^{a}\frac{\sqrt{v}}{\sqrt{2\pi}}e^{-vu^2/2}\,du,$$

and this latter expression is the probability that a normal random variable with mean zero and variance v^{-1} is between $-a$ and a, which was what we set out to show.

[6]Here we are replacing $I(\hat\theta_n)$ in (6.1) with this observed Fisher information; the two are asymptotically equivalent, so there is no contradiction here.

Chapter 7

What Else is There to Learn?

7.1 Introduction

The first six chapters of these notes covers at least a large chunk of what would be considered classical statistical theory. A natural question for students who have completed this or a similar course is: *what else is there to learn?* I think it is important for students to get an answer to this question, even if the answer is not a precise one. In this chapter I give just a taste of some things not covered in these notes, with the hope that seeing a bit about some of these more advanced topics might inspire some students to take more advanced courses on statistics to further their knowledge. Of course, there are probably other important topics that are not included here—this just a (biased) sample of some topics that students would see in later courses in statistics.

7.2 Sampling and experimental design

It was briefly mentioned at the beginning of these notes that we will take the availability of an iid sample from the reference population as given. It turns out, however, that this is a very strong assumption, i.e., it is not particularly easy to get a good sample. Issues one would encounter in trying to get a "representative sample"[1] include

- Not knowing the *sampling frame*, i.e., if we don't know who all is a member of the population of interest, then how can we be sure to give them and appropriate chance of being included in the sample?
- Even if the sampling frame is known, how to be sure that different subgroups of the population are represented in the sample? Such an issue would be important because, arguably, a sample of individuals that contains only males, while perhaps being a

[1] An interesting point is that, to my knowledge, there is no real definition for a representative sample. For sure, no matter what the definition would be, one could not tell whether a given sample satisfies that definition without knowing the population in question; and if the population is known, then there is no point of doing sampling...

random sample, should not be considered "representative" in any sense. *Stratified sampling* is one way to avoid this.

- What to do if the units sampled do not respond to the survey or whatever? This is an issue of *non-response*.
- etc, etc, etc.

These are just a few of the issues that one could encounter in trying to construct a good representative sample from a population. A very nice book devoted to sampling is Hedayat and Sinha, *Design and Inference in Finite Population Sampling*, Wiley, 1991.

A related issue is, in certain cases, our data is to be collected based off a suitable experiment. For such cases, we would like to select the settings of this experiment so that the data collected will be as informative as possible. I expect that you're wondering how we can design an experiment to get good results if you don't already know what we're looking for. This is a good question, and for "linear models" (see below) there is a lot already known about how to design good experiments. Outside of the linear model case, in general, arguably not so much is known. There are very deep results about optimal designs for "polynomial models" and for a few other models. A book that was recommended to me is Silvey, *Optimal Designs*, Chapman & Hall, 1980; surely there are some more modern textbook references, but I don't know this area well enough to make any recommendations.

7.3 Non-iid models

In Stat 411 we focused exclusively on models where the data points are iid. However, there are lots of more general kinds of structures that one could encounter in practice. Perhaps the most common is independent but not iid, that is, the data points are independent, but each has its own distribution, which might be different from that of another data point. A standard example of this situation is in the *linear regression model*. That is

$$Y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

where x_i is a fixed number, $\varepsilon_1, \ldots, \varepsilon_n$ are iid $\mathsf{N}(0, \sigma^2)$, and $(\alpha, \beta, \sigma^2)$ are parameters. In that case, we could write

$$Y_i \sim \mathsf{N}(\alpha + \beta x_i, \sigma^2), \quad i = 1, \ldots, n,$$

and we see that, while Y_i and Y_j are independent, in general they have different distributions. This linear model, along with various extensions, is arguably the most common statistical model used in applications. At UIC, both Stat 481 and Stat 521 would cover the linear model in depth. A comprehensive book on applied linear models is Kutner, Nachtsheim, Neter, and Li's *Applied Linear Statistical Models*, Wiley 2004, and a good book I know on the theory of linear models is Seber and Lee's *Linear Regression Analysis*, Wiley 2003. So-called *generalized linear models* are hugely important in applications of statistics, and for this material I would recommend Agresti's *Categorical Data Analysis*, Wiley 2013, and McCullagh and Nelder's *Generalized Linear Models*, Chapman–Hall 1989.

There are other models where the data points are actually dependent. Such data can occur if measurements on the same subject are taken at multiple time points. For example, if you have a set of patients and you measure the blood pressure $Y_{i,t}$ of patient i at time t for a range of t values, then $Y_{i,t}$ and $Y_{i,t'}$ are generally dependent. Examples like these fall under the umbrella of *time series* models, and are obviously of great importance. Another example that one is familiar with is the pricing of stocks, etc, in finance. Obviously, the price X_{t+1} of a given stock at time $t+1$ will depend on the price X_t at time t, so again we see some dependence. Famous models used in such applications, such as the Black–Scholes model, are basically just some normal distribution model that allows for some kind of dependence. A nice book on dependent-data models, with a slant towards financial applications, is Ruppert's *Statistics and Data Analysis for Financial Engineering*, Springer 2011.

7.4 High-dimensional models

In Stat 411, we have focused on models with only a few parameters, so that, typically, we have much more samples than parameters to estimate. However, there are real applications where this is not the case. A canonical example is the following independent but not iid model:

$$X_i \sim \mathsf{N}(\theta_i, 1), \quad i = 1, \ldots, n.$$

In this case, every sample has its own parameter so, even if n is large, we cannot expect to be able to do a good job at estimating all the parameters, because ultimately we only have one sample for each parameter. This may seem like an unrealistic example and while it is oversimplified, it does match up with the challenges faced in many examples. One is in genetics where the goal is to identify which genes, among thousands of possible genes, are relevant in distinguishing between individuals with and without a particular disease, like cancer. The basic idea is to construct two sets of patients, one with the disease and one without. Then each individuals genetic structure is measured, and the two groups are tested for significant differences on each particular gene. So, if a particular gene has a significant difference across the two groups, then the conclusion is that that gene can help to distinguish between people who will and will not get cancer. With this information, doctors can identify at-risk patients early in life and hopefully provide some preventative treatments. For some details on this, I recommend the books: Dudoit and van der Laan, *Multiple Testing Procedures with Applications to Genomics*, Springer, 2008, and Efron, *Large-Scale Inference*, Cambridge, 2010.

There are other versions of the high-dimensional problem, some of them are related to the linear model described above. I will not go into any details here, but I'll recommend the book: Bühlmann and van de Geer, *Statistics for High-Dimensional Data*, Springer, 2011.

7.5 Nonparametric models

Nonparametrics means two different things. Classically, the name refers to methods that make minimal assumptions about the underlying model. For example, suppose that f_θ is a density that is symmetric about the point θ. This could be a $\mathsf{N}(\theta, 1)$ density or something else. The goal then is to estimate the point of symmetry θ. The challenge is that the point of symmetry can have different interpretations depending on the underlying distribution, which is not being specified. In the normal case, θ is the mean, so estimating it with the sample mean is reasonable (and optimal). However, in the Cauchy case, there is no mean, so opting to estimate θ by $\bar{X}$ will be terrible. So, the point is to find an estimator which is not sensitive—or *robust*—to the shape of the distribution. Nonparametric methods, in this classical sense, are those which can give reasonable results with only minimal assumptions (e.g., symmetry was the only assumption in the illustration above) on the underlying model. Of course, these methods are not as good as the model-specific methods when the model is correctly specified. So, there is a trade-off between efficiency and robustness that one must consider in deciding which methods to apply in a given problem.

The second, more modern interpretation of "nonparametrics" is for those problems where there are infinitely many parameters. For a simple example, suppose $X_1, \ldots, X_n$ are iid from a density f, but f itself is unknown and to be estimated. This is a nonparametric problem because the unknown parameter, f in this case, is infinite-dimensional. To see this, note that in order to fully characterize f, in the absence of any structural assumptions, one must know $f(x)$ for each $x \in \mathbb{R}$. This is (uncountably) infinitely many quantities, so we say f is infinite-dimensional and the problem is a nonparametric one. The histograms we drew occasionally in this course are examples of estimators of a density function f; we normally think of these just as pictures, but there is a formula that describes that picture, and this defines an estimator of f. Other density estimators include kernel estimators, and mixture models; these have both Bayesian and non-Bayesian versions too. What is essential in these problems are notions of smoothness of the function being estimated. In a certain sense, knowing that f is smooth means that there are "fewer" parameters to be estimated. For example, if you know that f is differentiable, then knowing f at a value x means that we also effectively know f at all x' close to x, which makes the problem simpler. For more on nonparametrics, I'd recommend the book: Wasserman, *All of Nonparametric Statistics*, Springer, 2010.

7.6 Advanced asymptotic theory

We spent a considerable amount of time in Stat 411 discussing the asymptotic theory for maximum likelihood estimators and likelihood ratio tests. However, these results are just "baby" versions of what is known. Here is a short list of things that can be considered beyond the results presented in these notes.

- Recall the long list of regularity conditions needed for the MLE asymptotic normality result. It turns out that most of those conditions can be removed, being replace

by just one condition, namely that f_θ is *differentiable in quadratic mean* (DQM). Obviously, all regular exponential families would satisfy this condition, but there are other distributions which are DQM but do not satisfy the regularity conditions discuss here. One example is the Laplace location model, with density $f_\theta(x) \propto e^{-|x-\theta|}$. The usual regularity conditions fail because $f_\theta(x)$ is not everywhere differentiable in θ, but this model is DQM and the MLE is asymptotically normal.

- One can even develop theory for models in which asymptotic normality cannot hold. Recall the uniform and shifted exponential models (which don't satisfy the regularity conditions because their support depends on the parameter) have asymptotic distribution results, but that the limiting distribution is not normal. This can be formalized to some general results about asymptotics for non-regular models. One can also give versions of the Cramer–Rao inequality for cases where the Fisher information is not well-defined due to irregularity of the model.
- Finally, all this talk about asymptotic normality for the MLE actually has nothing to do with the MLE at all! These are properties of the underlying model, not of the choice of estimator. Certain models, such as those satisfying the DQM property, can be lumped under the category of *local asymptotically normal* (LAN) which means, roughly, that any choice of estimator (MLE, Bayes, or something else), if it has a limiting distribution, then it must inherit the limiting normal from the underlying structure. Such results are of considerable help in unifying the various kinds of asymptotic results, and for identifying asymptotically optimal estimators, tests, etc.

All of these things are discussed in considerable detail in the excellent (albeit technical) book: van der Vaart, *Asymptotic Statistics*, Cambridge, 1998.

There are other other extensions of the theory presented in Stat 411, one is to give more refined approximations. In the proof of the asymptotic normality result, we used a second-order Taylor approximation. However, it is clear that some more precise approximations are available if we keep some higher-order terms, but this comes at the expense of additional technicalities. Such extensions are called *higher-order asymptotics*. Actually, just keeping higher-order terms in the Taylor approximation is not exactly how this is done. The main tools used are saddle-point and Edgeworth approximations. A nice introduction to these ideas is in the book: Young and Smith, *Essentials of Statistical Inference*, Cambridge, 2004.

Finally, I should mention that *empirical process theory* is now a powerful and widely-used tool in statistics. To illustrate these ideas, I will use a simple but fundamentally important example. Let $X_1, \ldots, X_n$ be iid with distribution function F. If we pick some fixed value, say, x_0, then we can easily get a good estimator of $F(x_0)$ by thinking of the problem as a binomial, i.e., write $Y_i = I_{X_i \leq x_0}$, so that $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \mathsf{Bin}(n, F(x_0))$. But this trick works only for a fixed x_0, what if we want to estimate the entire distribution function F? An natural estimator (same as what one gets from the binomial trick) is the *empirical distribution function*:

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{X_i \leq x},$$

which is just the proportion of the sample less than or equal to x. The first result from empirical process theory is the following:

$$\sup_x |\hat{F}_n(x) - F(x)| \to 0, \quad \text{with probability 1 as } n \to \infty.$$

This result is a special case of the *Glivenko–Cantelli* theorem, often called the "fundamental theorem of statistics." The reason why this result might have such a strong name is that it shows that, with enough data, it is possible to learn *everything* about a distribution F. Empirical process theory concerns various generalizations and extensions of the above result. van der Vaart's book is probably the best introduction to these ideas.

7.7 Computational methods

In Stat 411, the focus was primarily on estimators, tests, etc, that we can write down a simple formula for, such as $\bar{X}$. However, there are actually many cases where there are no simple formulae available; we saw a few of these examples in Chapters 3 and 6. It is important for students to be aware of the fact that, those problems with simple formulae are the exception, rather than the rule. That is, essentially all real problems will require some non-trivial computation, which means that students must have some exposure to statistical software and the various algorithms that are employed.

We talked briefly about Monte Carlo methods and bootstrap in Chapter 1, about optimization methods in Chapter 3, and about some Monte Carlo methods in Chapter 6. I should mention that there are folks whose research focuses on developing efficient algorithms and software for carrying out these computations in these and other more sophisticated problems. Stat 451 at UIC would cover some of the basics of these computational methods. I would recommend the following books: Givens and Hoeting, *Computational Statistics*, Wiley, 2013 and Lange, *Numerical Analysis for Statisticians*, Springer, 2010. There are some books devoted (almost) entirely to Monte Carlo methods: Robert and Casella, *Monte Carlo Statistical Methods*, Springer, 2004 and Liang, Liu, and Carroll, *Advanced Markov Chain Monte Carlo Methods*, Wiley, 2010.

7.8 Foundations of statistics

Finally, in Stat 411, the focus is primarily on answering the questions: what methods are available and what are their properties? We did not address the question of *why* are these the approaches taken. The two mainstream approaches to statistics are *frequentist* and *Bayesian*, and we have discussed both in these notes. The frequentist approach, often attributed to Neyman, is based on the idea that one should choose estimators, tests, etc with good repeated-sample properties, such as unbiasedness, consistency, etc. This is what motivated our study of sampling distribution properties of statistics. The Bayesian approach, on the other hand, is based on the idea that probability is the correct tool for summarizing uncertainty, and, in order to have a posterior distribution for inference, one must have a prior

distribution to start with. Later is was learned that Bayes methods often have desirable frequentist properties as well, e.g., the complete-class theorems from decision theory.

Although these two are the mainstream approaches, it is important for students to know that there is no solid justification for these being the "correct" way to do statistical inference. That is, there is really no solid foundations of statistics. This doesn't mean that statistics is not a scientifically sound subject, just that it is still not yet fully developed. As an analogy, in mathematics, calculus is several hundred years old, while statistics as a subject is only about 100 years old. Anyway, the point is that we should not just accept what is now mainstream statistics as "correct" and not look to other ideas that might be better.

You might ask what else is there besides Bayesian and frequentist statistics. Although Fisher's name has come up a lot in these notes, he was part of neither of these two camps. Fisher was adamantly non-Bayesian because he didn't like the influence the prior distribution had on the conclusions; he also was non-frequentist because he thought the focus on repeated-sample properties was only for mathematical convenience and not for any scientific relevancy. Fisher was a scientist, not a mathematician or a statistician, so he was interested in solving problems, not in putting down and following some rigid set of rules of how a solution should be obtained. For that reason, much of Fisher's ideas about the foundations of statistical inference are not completely clear. Some might say that Fisher had his own school of thought on statistics—the *Fisherian* school—but I think it's hard to say exactly what that is.

Fisher had one idea that he was very serious about throughout his career, what he called the *fiducial argument*. The basic goal is to use the sampling model and data alone, no prior, to produce a distribution on the parameter space to be used for inference.[2] He called this the *fiducial distribution*. This is a very appealing idea but, unfortunately, Fisher did not lay out a clear explanation of how the framework should go, just basically a list of examples. Because Fisher's ideas on fiducial inference were not clear, it never really gained any traction as an approach for statistics, and eventually almost disappeared. But Neyman's idea for confidence intervals, which is a cornerstone of modern statistics, is basically his interpretation of Fisher's fiducial idea. So, even though fiducial itself has not yet had a major impact, it has had an indirect impact through confidence intervals.

Anyway, Fisher's thoughts on fiducial inference have inspired a number of other not-yet-mainstream ideas. *Dempster–Shafer theory* (due to Art Dempster and Glenn Shafer), developed in the 1960s and has a considerable following in computer science, can be considered as an extension of Fisher's fiducial, and introduced some interesting concepts about random sets and belief functions. More recently, there has been a surge of interest in statistics on prior-free distributional inference. One example of this is "objective" Bayes, where priors are chosen not based on prior beliefs but based on a goal to get good sampling distribution properties; this is now probably the norm in application of Bayesian methods. There is also a *generalized fiducial inference*, championed by Jan Hannig, which has successfully been applied in a number of interesting and important problems. *Confidence distributions*

[2]The word "fiducial" means *based on trust or faith*. To see the connection with Fisher idea, consider the following simple example. For an observation $X \sim \mathsf{N}(\theta, 1)$, write $X = \theta + U$, for $U \sim \mathsf{N}(0, 1)$. If $X = x$ is observed, and we *trust* that observing $X = x$ does not change our belief that $U \sim \mathsf{N}(0, 1)$, then we can write $\theta = x - U$ and take a distribution $\mathsf{N}(x, 1)$ as a summary of our uncertainty about θ after observing $X = x$.

have recently been proposed by Kesar Singh, Min-ge Xie, and others. Most recently, my collaborator, Chuanhai Liu, and I have developed a new framework, which we call *inferential models* (IMs). This framework provides meaningful prior-free probabilistic inference, and is based off of the use of predictive random sets and belief/plausibility functions. A book (Liu and Martin, *Inferential Models: Reasoning with Uncertainty*, Chapman & Hall) on IMs is scheduled to be published in 2015.