Properties of Estimators

BS2 Statistical Inference, Lecture 2
Michaelmas Term 2004

Steffen Lauritzen, University of Oxford; October 15, 2004

Notation and setup

$\mathcal{X}$ denotes *sample space*, typically either finite or countable, or an open subset of $\mathcal{R}^k$.

We have observed *data* $x \in \mathcal{X}$ which are assumed to be a realisation $X = x$ of a random variable X.

The probability mass function (or density) of X is partially unknown, i.e. of the form $f(x;\theta)$ where θ is a *parameter*, varying in the *parameter space* Θ.

This lecture is concerned with principles and methods for estimating (guessing) θ on the basis of having observed $X = x$.

Unbiased estimators

An estimator $\hat{\theta} = t(x)$ is said to be *unbiased* for a function θ if it equals θ in expectation:

$$\mathbf{E}_\theta\{t(X)\} = \mathbf{E}\{\hat{\theta}\} = \theta.$$

Intuitively, an unbiased estimator is 'right on target'.

The *bias* of an estimator $\hat{\theta} = t(X)$ of θ is

$$\mathrm{bias}(\hat{\theta}) = \mathbf{E}\{t(X) - \theta\}.$$

If $\mathrm{bias}(\hat{\theta})$ is of the form $c\theta$, $\tilde{\theta} = \hat{\theta}/(1+c)$ is unbiased for θ. We then say that $\tilde{\theta}$ is a *bias-corrected* version of $\hat{\theta}$.

Unbiased functions

More generally $t(X)$ is *unbiased for a function* $g(\theta)$ if

$$\mathbf{E}_\theta\{t(X)\} = g(\theta).$$

Note that even if $\hat{\theta}$ is an unbiased estimator of θ, $g(\hat{\theta})$ will generally *not* be an unbiased estimator of $g(\theta)$ unless g is linear or affine.

This limits the importance of the notion of unbiasedness. It might be at least as important that an estimator is *accurate* so its distribution is highly concentrated around θ.

If an unbiased estimator of $g(\theta)$ has mimimum variance among all unbiased estimators of $g(\theta)$ it is called a *minimum variance unbiased estimator* (MVUE).

Is unbiasedness a good thing?

Unbiasedness is important when combining estimates, as *averages of unbiased estimators are unbiased* (sheet 1).

When combining standard deviations $s_1, \ldots, s_k$ with d..o.f. $d_1, \ldots, d_k$ we *always average their squares*

$$\bar{s} = \sqrt{\frac{d_1 s_1^2 + \cdots + d_k s_k^2}{d_1 + \cdots + d_k}}$$

as each of these are unbiased estimators of the variance σ^2, whereas s_i are *not* unbiased estimates of σ.

Be careful when averaging biased estimators! It may well be appropriate to make a bias-correction before averaging.

Mean Square Error

One way of measuring the accuracy of an estimator is via its *mean square error* (MSE):

$$\mathrm{mse}(\hat{\theta}) = \mathbf{E}(\hat{\theta} - \theta)^2.$$

Since it holds for any Y that $\mathbf{E}(Y^2) = \mathbf{V}(Y) + \{\mathbf{E}(Y)\}^2$, the MSE can be decomposed as

$$\mathrm{mse}(\hat{\theta}) = \mathbf{V}(\hat{\theta} - \theta) + \{\mathbf{E}(\hat{\theta} - \theta)\}^2 = \mathbf{V}(\hat{\theta}) + \{\mathrm{bias}(\theta)\}^2,$$

so getting a small MSE often involves a *trade-off* between variance and bias. By not insisting on $\hat{\theta}$ being unbiased, the variance can sometimes be drastically reduced.

For *unbiased* estimators, the MSE is obviously equal to the variance, $\mathrm{mse}(\hat{\theta}) = \mathbf{V}(\hat{\theta})$, so no trade-off can be made.

Asymptotic consistency

An estimator $\hat{\theta}$ (more precisely a sequence of estimators $\hat{\theta}_n$) is said to be (weakly) *consistent* if it converges to θ in probability, i.e. if for all $\epsilon > 0$,

$$\lim_{n \to \infty} P\{|\hat{\theta} - \theta| > \epsilon\} = 0.$$

It is *consistent in mean square error* if $\lim_{n \to \infty} \mathrm{mse}(\hat{\theta}) = 0$.

Both of these notions refer to the *asymptotic* behaviour of $\hat{\theta}$ and expresses that, as data accumulates, $\hat{\theta}$ gets closer and closer to the true value of θ.

Asymptotic consistency is a good thing. However, in a given case, for fixed n it may only be modestly relevant. Asymptotic *inconsistency* is generally worrying.

Fisher consistency

An estimator is *Fisher consistent* if the estimator is the same functional of the empirical distribution function as the parameter of the true distribution function:

$$\hat{\theta} = h(F_n), \quad \theta = h(F_\theta)$$

where F_n and F_θ are the empirical and theoretical distribution functions:

$$F_n(t) = \frac{1}{n}\sum_{1}^{n} 1\{X_i \leq t\}, \quad F_\theta(t) = P_\theta\{X \leq t\}.$$

Examples are $\hat{\mu} = \bar{X}$ which is Fisher consistent for the mean μ and $\hat{\sigma}^2 = SSD/n$ which is Fisher consistent for σ^2. Note $s^2 = SSD/(n-1)$ is *not* Fisher consistent.

Consistency relations

If an estimator is mean square consistent, it is weakly consistent.

This follows from Chebyshov's inequality:

$$P\{|\hat{\theta} - \theta| > \epsilon\} \leq \frac{\mathbf{E}(\hat{\theta} - \theta)^2}{\epsilon^2} = \frac{\mathrm{mse}(\hat{\theta})}{\epsilon^2},$$

so if $\mathrm{mse}(\hat{\theta}) \to 0$ for $n \to \infty$, so does $P\{|\hat{\theta} - \theta| > \epsilon\}$.

The relationship between Fisher consistency and asymptotic consistency is less clear. It is generally true that

$$\lim_{n \to \infty} F_n(t) = F_\theta(t) \text{ for continuity points } t \text{ of } F_\theta,$$

so $\hat{\theta} = h(F_n) \to F_\theta$ *if h is a suitably continuous functional*.

Score statistic

For $X = x$ to be informative about θ, the density (and therefore the likelihood function) must vary with θ.

If $f(x;\theta)$ is smooth and differentiable, this change is quantified to first order by the *score function*:

$$s(x;\theta) = \frac{\partial}{\partial\theta}\log f(x;\theta) = \frac{f'(x;\theta)}{f(x;\theta)}.$$

If differentiation w.r.t. θ and integration w.r.t. x can be interchanged, the score statistic has expectation zero

$$\begin{aligned}\mathbf{E}\{S(\theta)\} &= \int \frac{f'(x;\theta)}{f(x;\theta)} f(x;\theta)\,dx = \int f'(x;\theta)\,dx \\ &= \frac{\partial}{\partial\theta}\left\{\int f(x;\theta)\,dx\right\} = \frac{\partial}{\partial\theta}1 = 0.\end{aligned}$$

Fisher information

The variance of $S(\theta)$ is the *Fisher information* about θ:

$$i(\theta) = \mathbf{E}\{S(\theta)^2\}.$$

If integration and differentiation can be interchanged

$$i(\theta) = \mathbf{V}\{S(\theta)\} = -\mathbf{E}\left\{\frac{\partial}{\partial\theta}S(\theta)\right\} = -\mathbf{E}\left\{\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\right\},$$

since then

$$\begin{aligned}
&\mathbf{E}\frac{\partial^2}{\partial\theta^2}\log f(X;\theta) = \\
&\qquad \int \frac{f''(x;\theta)}{f(x;\theta)} f(x;\theta)\,dx - \int \left\{\frac{f'(x;\theta)}{f(x;\theta)}\right\}^2 f(x;\theta)\,dx \\
&= \quad 0 - \mathbf{E}\{S(\theta)\}^2 = -i(\theta).
\end{aligned}$$

The normal case

It may be illuminating to consider the special case when $X \sim \mathcal{N}(\theta, \sigma^2)$ with σ^2 known and θ unknown. Then

$$\log f(x;\theta) = -\frac{1}{2}\log(2\pi\sigma^2) - (x-\theta)^2/(2\sigma^2)$$

so the score statistic and information are

$$s(x;\theta) = (x-\theta)/\sigma^2, \quad i(\theta) = \mathbf{E}(1/\sigma^2) = 1/\sigma^2.$$

So the score statistic can be seen as a linear approximation to the normal case, with the information determining the scale, here equal to the inverse of the variance.

Cramér–Rao's inequality

The Fisher information yields a lower bound on the variance of an unbiased estimator:

We assume suitable smoothness conditions, including that

- *The region of positivity of $f(x;\theta)$ is constant in θ*;
- *Integration and differentiation can be interchanged.*

Then for any unbiased estimator $T = t(X)$ of $g(\theta)$ it holds

$$\mathbf{V}(T) = \mathbf{V}(\hat{g}(\theta)) \geq \{g'(\theta)\}^2 / i(\theta).$$

Note that for $g(\theta) = \theta$ the lower bound is simply the *inverse Fisher information* $i^{-1}(\theta)$.

Proof of Cramér–Rao's inequality

Since $\mathbf{E}\{S(\theta)\} = 0$, the Cauchy–Schwarz inequality yields

$$|\mathrm{Cov}\{T, S(\theta)\}|^2 \leq \mathbf{V}(T)\mathbf{V}\{S(\theta)\} = \mathbf{V}(T)i(\theta). \qquad (1)$$

Now, since $\mathbf{E}\{S(\theta)\} = 0$,

$$\begin{aligned} \mathrm{Cov}\{T, S(\theta)\} &= \mathbf{E}\{T\, S(\theta)\} = \int t(x)\frac{f'(x;\theta)}{f(x;\theta)} f(x;\theta)\, dx \\ &= \int t(x) f'(x;\theta)\, dx = \frac{\partial}{\partial\theta}\mathbf{E}\{T\} = g'(\theta), \end{aligned}$$

inserting this into the inequality (1) and dividing both sides with $i(\theta)$ yields the result.

Attaining the lower bound

It is rarely possible to find an estimator which attains the bound. In fact (under the usual conditions)

An unbiased estimator of $g(\theta)$ *with variance* $\{g'(\theta)\}^2/i(\theta)$ *exists if and only if the score statistic has the form*

$$s(x;\theta) = \frac{i(\theta)\{t(x) - g(\theta)\}}{g'(\theta)}.$$

In the special case where $g(\theta) = \theta$ we have

$$s(x;\theta) = i(\theta)\{t(x) - g(\theta)\}.$$

Proof of the expression for the score statistic

Cauchy–Schwarz inequality is sharp unless T is an affine function of $S(\theta)$ so

$$t(x) = \hat{g}(\theta) = a(\theta)s(x;\theta) + b(\theta) \tag{2}$$

for some $a(\theta), b(\theta)$.

Since $t(X)$ is unbiased for θ and $\mathbf{E}\{S(\theta)\} = 0$, we have $b(\theta) = g(\theta)$. From the proof of the inequality we have

$$\mathrm{Cov}\{T, S(\theta)\} = g'(\theta).$$

Combining with the linear expression in (2) gives

$$g'(\theta) = \mathrm{Cov}\{T, S(\theta)\} = a(\theta)\mathbf{V}\{S(\theta)\} = a(\theta)i(\theta)$$

and the result follows.

Efficiency

If an unbiased estimator attains the Cramér–Rao bound, it it said to be *efficient*.

An efficient unbiased estimator is clearly also MVUE.

The *Bahadur efficiency* of an unbiased estimator is the inverse of the ratio between its variance and the bound:

$$0 \leq \operatorname{beff} \hat{g}(\theta) = \frac{\{g'(\theta)\}^2}{i(\theta)\mathbf{V}\{\hat{g}(\theta)\}} \leq 1.$$

Since the bound is rarely attained, it is sometimes more reasonable to compare with the smallest obtainable

$$0 \leq \operatorname{eff} \hat{g}(\theta) = \frac{\inf_{\{T:\mathbf{E}(T)=g(\theta)\}} \mathbf{V}(T)}{\mathbf{V}\{\hat{g}(\theta)\}} \leq 1.$$