Topic 13: Unbiased Estimation

November 3, 2009

When we look to estimate the distribution mean μ, we use the sample mean $\bar{x}$. For the variance σ^2, we have seen two choices:

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2.$$

One criterion for choosing is **statistical bias**.

Definition 1. *A statistic d is called an* **unbiased estimator** *for a function of the parameter $g(\theta)$ provided that for every choice of θ,*

$$E_\theta d(X) = g(\theta).$$

Any estimator that not unbiased is called **biased**. *The* **bias** *is the difference*

$$b_d(\theta) = E_\theta d(X) - g(\theta).$$

We can assess the quality of an estimator by computing its **mean square error**.

$$\begin{aligned} E_\theta[(d(X)-g(\theta))^2] &= E_\theta[(d(X)-E_\theta d(X)+b_d(\theta))^2] \\ &= E_\theta[(d(X)-E_\theta d(X))^2] + 2b_d(\theta)(E_\theta[(d(X)-E_\theta d(X)] + b_d(\theta)^2 \\ &= \text{Var}_\theta(d(X)) + b_d(\theta)^2 \end{aligned}$$

Note that the mean square error for an unbiased estimator is its variance. Bias increases the mean square error.

Example 2. *Let $X_1, X_2, \ldots,$ be Bernoulli trials with success parameter p and set $d(X) = \bar{X}$,*

$$E_\theta \bar{X} = \frac{1}{n}(p+\cdots+p) = p$$

Thus, $\bar{X}$ is an unbiased estimator for p. In this circumstance, we generally write $\hat{p}$ instead of $\bar{X}$. In addition,

$$\text{Var}(\hat{p}) = \frac{1}{n^2}(p(1-p)+\cdots+p(1-p)) = \frac{1}{n}p(1-p).$$

1 Computing Bias

Example 3. *If a simple random sample $X_1, X_2, \ldots,$ has unknown finite variance σ^2, then, we can consider the sample variance*

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2.$$

To find the mean of S^2, we begin with the identity

$$\begin{aligned}\sum_{i=1}^{n}(X_i-\mu)^2 &= \sum_{i=1}^{n}((X_i-\bar{X})+(\bar{X}-\mu))^2\\ &= \sum_{i=1}^{n}(X_i-\bar{X})^2+\sum_{i=1}^{n}(X_i-\bar{X})(\bar{X}-\mu)+n(\bar{X}-\mu)^2\\ &= \sum_{i=1}^{n}(X_i-\bar{X})^2+n(\bar{X}-\mu)^2\end{aligned}$$

Then,

$$\begin{aligned}ES^2 &= E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2-(\bar{X}-\mu)^2\right]\\ &= \frac{1}{n}n\sigma^2-\frac{1}{n}\sigma^2=\frac{n-1}{n}\sigma^2.\end{aligned}$$

This shows that S^2 is a biased estimator for σ^2. We can see that it is biased downwards.

$$b(\sigma^2)=\frac{n-1}{n}\sigma^2-\sigma^2=-\frac{1}{n}\sigma^2.$$

In addition,

$$E\left[\frac{n}{n-1}S^2\right]=\sigma^2$$

and

$$S_u^2=\frac{n}{n-1}S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$$

is an unbiased estimator for σ^2. As we shall learn in the next example, because the square root is concave downward, S_u as an estimator for σ is **downwardly biased**.

Example 4. *If $X_1,\ldots,X_n$ form a simple random sample with unknown finite mean μ, then $\bar{X}$ is an unbiased estimator of μ. If the X_i have variance σ^2, then*

$$Var(\bar{X})=\frac{\sigma^2}{n}.$$

In the methods of moments estimation, we have used $g(\bar{X})$ as an estimator for $g(\mu)$. If g is a **convex function**, *we can say something about the bias of this estimator. In Figure 1, we see the method of moments estimator for the estimator g for a parameter in the Pareto distribution. The choice of $\beta=3$ corresponds to a mean of $\mu=3/2$ for the Pareto random variables. The central limit theorem states that the sample mean $\bar{X}$ is nearly normally distributed with mean 3/2. Thus, the distribution is nearly symmetric around 3/2. From the figure, we can see that the interval from 1.4 to 1.5 under the function g maps into a shorter interval than the interval from 1.5 to 1.6. Thus, a sample mean above 1.5 is stretched away from $\beta=3$ under the mapping g more than a sample mean below 1.5. Consequently, we anticipate that the estimator $\hat{\beta}$ will be* **upwardly biased**.

One way to characterize a convex function is that its graph lies above any tangent line. If we look at the point $x=\mu$, then this statement becomes

$$g(x)-g(\mu)\geq g'(\mu)(x-\mu).$$

Now replace x with the random variable $\bar{X}$ and take expectations.

$$E_\mu[g(\bar{X})-g(\mu)]\geq E_\mu[g'(\mu)(\bar{X}-\mu)]=g'(\mu)E_\mu[\bar{X}-\mu]=0.$$

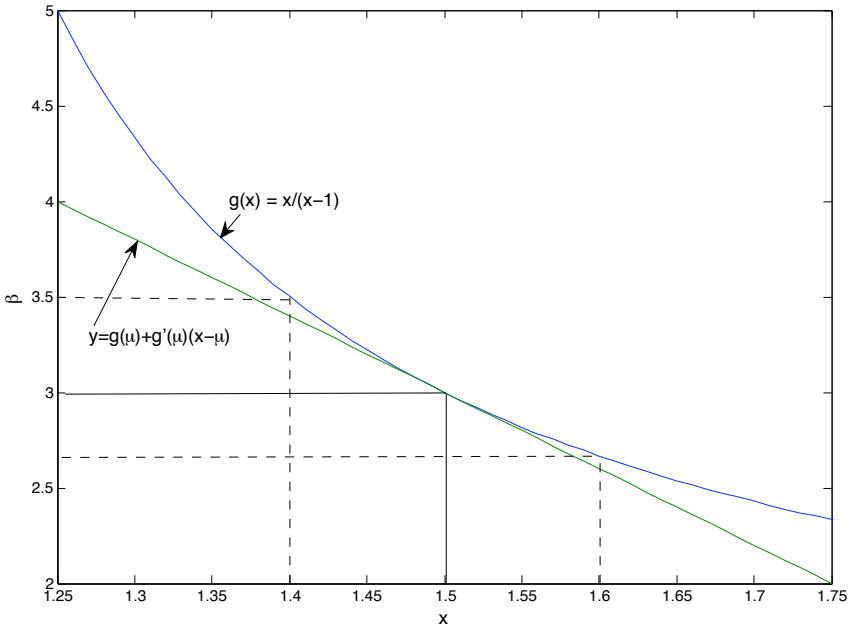

Figure 1: Graph of a convex function. Note that the tangent line is below the graph of g.

Consequently,

$$E_\mu g(X) \geq g(\mu) \tag{1}$$

and $g(\bar{X})$ is **biased upwards**. *The expression in (1) is known as* **Jensen's inequality**.

For the method of moments estimator for the Pareto random variable, we determined that

$$g(\mu) = \frac{\mu}{\mu - 1}.$$

By taking the second derivative, we see that $g''(\mu) = 2(\mu-1)^{-3} > 0$ and, because $\mu > 1$, g is a convex function . To estimate the size of the bias, we look at a quadratic approximation for g

$$g(x) - g(\mu) \approx g'(\mu)(x-\mu) + \frac{1}{2}g''(\mu)(x-\mu)^2.$$

Again, replace x with the random variable $\bar{X}$ and take expectations.

$$b_g(\mu) = E_\mu[g(\bar{X})] - g(\mu) \approx E_\mu[g'(\mu)(\bar{X}-\mu)] + \frac{1}{2}E[g''(\mu)(\bar{X}-\mu)^2] = \frac{1}{2}g''(\mu)\text{Var}(\bar{X}) = .\frac{1}{2n}g''(\mu)\sigma^2$$

Returning to the Pareto example, we have that

$$g''\left(\frac{\beta}{\beta-1}\right) = 2(\beta-1)^3.$$

Thus, the bias

$$b_g(\beta) \approx \frac{\beta(\beta-1)}{n(\beta-2)}$$

So, for $\beta = 3$ and $n = 100$, the bias is approximately 0.06. Compare this to the estimated value of 0.053 from simulation.

2 Cramér-Rao Bound

So, among unbiased estimators, one important goal is to find an estimator that has as small a variance as possible, A more precise goal would be to find an unbiased estimator d that has **uniform minimum variance**. In other words, $d(X)$ has finite variance for every value θ of the parameter and for any other unbiased estimator $\tilde{d}$,

$$\text{Var}_\theta d(X) \le \text{Var}_\theta \tilde{d}(X).$$

The **efficiency** of unbiased estimator $\tilde{d}$,

$$e(\tilde{d}) = \frac{\text{Var}_\theta d(X)}{\text{Var}_\theta \tilde{d}(X)}.$$

Thus, the efficiency is between 0 and 1.

The Cramér-Rao bound tells us how small a variance is ever possible. The formula is a bit mysterious at first, but can be seen after we review a bit on correlation. Recall that for two random variables Y and Z, the correlation

$$\rho(Y,Z) = \frac{\text{Cov}(Y,Z)}{\sqrt{\text{Var}(Y)\text{Var}(Z)}}. \tag{1}$$

The correlation takes values $-1 \le \rho(Y,Z) \le 1$ and takes the extreme values ± 1 if and only if Y and Z are linearly related, i.e., $Z = aY + b$ for some constants a and b. Consequently,

$$\text{Cov}(Y,Z)^2 \le \text{Var}(Y)\text{Var}(Z).$$

One of the random variables we will encounter on the way to finding the Cramér-Rao bound will have mean zero. In this case, we can make a small simplification.

If the random variable Z has mean zero, then $\text{Cov}(Y,Z) = E[YZ]$ and

$$E[YZ]^2 \le \text{Var}(Y)\text{Var}(Z) = \text{Var}(Y)EZ^2. \tag{2}$$

We begin with data $X = (X_1, \dots, X_n)$ drawn from an unknown probability P_θ. The paramater space $\Theta \subset \mathbb{R}$. Denote the joint density of these random variables

$$\mathbf{f}(\mathbf{x}|\theta), \quad \text{where } \mathbf{x} = (x_1 \dots, x_n).$$

In the case that the date comes from a simple random sample then the joint density is the product of the marginal densities.

$$\mathbf{f}(\mathbf{x}|\theta) = f(x_1|\theta)\cdots f(x_n|\theta). \tag{3}$$

For continuous random variables, we have

$$1 = \int_{\mathbb{R}^n} \mathbf{f}(\mathbf{x}|\theta)\, d\mathbf{x} \tag{4}$$

Now, let d be the unbiased estimator of $g(\theta)$, then

$$g(\theta) = E_\theta d(X) = \int_{\mathbb{R}^n} d(\mathbf{x})\mathbf{f}(\mathbf{x}|\theta)\, d\mathbf{x} \tag{5}.$$

If the functions in (4) and (5) are differentiable with respect to the parameter θ and we can pass the derivative through the integral, then

$$0 = \int_{\mathbb{R}^n} \frac{\partial \mathbf{f}(\mathbf{x}|\theta)}{\partial\theta}\, d\mathbf{x} = \int_{\mathbb{R}^n} \frac{\partial \ln \mathbf{f}(\mathbf{x}|\theta)}{\partial\theta}\mathbf{f}(\mathbf{x}|\theta)\, d\mathbf{x} = E_\theta\left[\frac{\partial \ln \mathbf{f}(X|\theta)}{\partial\theta}\right]. \tag{6}$$

From a similar calculation,

$$g'(\theta) = E_\theta \left[d(X) \frac{\partial \ln \mathbf{f}(X|\theta)}{\partial \theta} \right]. \tag{7}$$

Now, return to the review on correlation with $Y = d(X)$ and the **score function** $Z = \partial \ln \mathbf{f}(X|\theta)/\partial\theta$. Then, by equation (6), $EZ = 0$, and from equations (7) and (2), we find that

$$g'(\theta)^2 = E_\theta \left[d(X) \frac{\partial \ln \mathbf{f}(X|\theta)}{\partial \theta} \right]^2 \le \text{Var}_\theta(d(X)) E_\theta \left[\left(\frac{\partial \ln \mathbf{f}(X|\theta)}{\partial \theta} \right)^2 \right],$$

or,

$$\text{Var}_\theta(d(X)) \ge \frac{g'(\theta)^2}{I(\theta)}. \tag{8}$$

where

$$I(\theta) = E_\theta \left[\left(\frac{\partial \ln \mathbf{f}(X|\theta)}{\partial \theta} \right)^2 \right]$$

is called the **Fisher information**.

Equation (8), called the **Cramér-Rao lower bound** or the **information inequality**, states that the lower bound for the variance of an unbiased estimator is the reciprocal of the Fisher information. In other words, the higher the information, the lower is the possible value of the variance of an unbiased estimator.

If we return to the case of a simple random sample then

$$\ln \mathbf{f}(\mathbf{x}|\theta) = \ln f(x_1|\theta) + \cdots + \ln f(x_n|\theta).$$

$$\frac{\partial \ln \mathbf{f}(\mathbf{x}|\theta)}{\partial \theta} = \frac{\partial \ln f(x_1|\theta)}{\partial \theta} + \cdots + \frac{\partial \ln f(x_n|\theta)}{\partial \theta}.$$

Also, the random variables $\{\partial \ln f(x_k|\theta)/\partial\theta; 1 \le k \le n\}$ are independent and have the same distribution. Using the fact that the variance of th sum is the sum of the variances for independent random variables, we see that the Fisher information.

$$I(\theta) = nE[(\frac{\partial \ln f(X_1|\theta)}{\partial \theta})^2].$$

Example 5. *For independent Bernoulli random variable with unknown success probability θ,*

$$\ln f(x|\theta) = x \ln \theta + (1-x) \ln(1-\theta),$$

$$\frac{\partial}{\partial \theta} \ln f(x|\theta) = \frac{x}{\theta} - \frac{1-x}{1-\theta} = \frac{x-\theta}{\theta(1-\theta)},$$

$$E\left[\left(\frac{\partial}{\partial \theta} \ln f(X|\theta) \right)^2 \right] = \frac{1}{\theta^2(1-\theta)^2} E[(X-\theta)^2] = \frac{1}{\theta(1-\theta)}$$

and the information is the reciprocal of the variance. Thus, by the Cramér-Rao lower bound, any unbiased estimator based on n observations must have variance al least $\theta(1-\theta)/n$. However, if we take $d(\mathbf{x}) = \bar{x}$, then

$$Var_\theta d(X) = \frac{\theta(1-\theta)}{n}$$

and $\bar{x}$ is a uniformly minimum variance unbiased estimator.

Example 6. *For independent normal random variables with known variance σ_0^2 and unknown mean μ,*

$$\ln f(x|\mu) = -\ln(\sigma_0\sqrt{2\pi}) - \frac{(x-\mu)^2}{2\sigma_0^2}.$$

and

$$\frac{\partial}{\partial\mu}\ln f(x|\mu) = \frac{1}{\sigma_0^2}(x-\mu).$$

$$E\left[\left(\frac{\partial}{\partial\mu}\ln f(X|\mu)\right)^2\right] = \frac{1}{\sigma_0^4}E[(X-\mu)^2] = \frac{1}{\sigma_0^2}.$$

Again, the information is the reciprocal of the variance. Thus, by the Cramér-Rao lower bound, any unbiased estimator based on n observations must have variance al least σ_0^2/n. However, if we take $d(\mathbf{x}) = \bar{x}$, then

$$Var_\mu d(X) = \frac{\sigma_0^2}{n}.$$

and $\bar{x}$ is a uniformly minimum variance unbiased estimator.

Using integration by parts, we have an identity that is often an useful alternative to the Fisher Information.

$$I(\theta) = E_\theta\left[\left(\frac{\partial \ln \mathbf{f}(X|\theta)}{\partial\theta}\right)^2\right] = -E_\theta\left[\frac{\partial^2 \ln \mathbf{f}(X|\theta)}{\partial\theta^2}\right]$$

For an exponential random variable,

$$\ln f(x|\lambda) = \ln\lambda - \lambda x, \quad \frac{\partial^2 f(x|\lambda)}{\partial\lambda^2} = -\frac{1}{\lambda^2}.$$

This,

$$I(\lambda) = \frac{1}{\lambda^2}.$$

Now, $\bar{X}$ is an unbiased estimator for $g(\lambda) = 1/\lambda$ with variance

$$\frac{1}{n\lambda^2}.$$

Cramér-Rao lower bound, we have that

$$\frac{g'(\lambda)^2}{nI(\lambda)} = \frac{1/\lambda^4}{n\lambda^2} = \frac{1}{n\lambda^2}.$$

Because $\bar{X}$ has this variance, it is a uniformly minimum variance unbiased estimator.