Unbiased Estimation

February 7-12, 2008

We begin with a sample $X = (X_1, \ldots, X_n)$ of random variables chosen according to one of a family of probabilities P_θ where θ is element from the parameter space Θ.

For random variables, we shall use the term **density function** to refer to both continuous and discrete random variables. Thus, to each $\theta \in \Theta$, there exists and density function which we denote

$$\mathbf{f}(\mathbf{x}|\theta).$$

Example 1 (Parametric families of densities)**.**

1. *Binomial random variables with known number of trials n but unknown success probability parameter θ has density*
$$f(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}.$$
2. *Normal random variables with known variance σ_0 but unknown mean μ has density*
$$f(x|\mu) = .\frac{1}{\sigma_0\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma_0^2}\right).$$
3. *Normal random variables with unknown mean μ and variance σ has density*
$$f(x|\mu,\theta) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

Definition 2. *A* **statistic** *is a function of the random variable that does not depend on any unknown parameter.*

The goal of estimation is to determine which of the P_θ is the source of the data X. In this case the action space A is the same as the parameter space and the **estimator** is the decision function

$$d : \text{data} \to \Theta.$$

Example 3. *If $X = (X_1, \ldots, X_n)$ are independent $Ber(\theta)$ random variables, then the simple choice for estimating θ is*

$$d(x_1, \ldots, x_n) = \frac{1}{n}(x_1 + \cdots + x_n) = \bar{x}.$$

1 Unbiased Estimators

Definition 4. *A statistic d is called an* **unbiased estimator** *for a function of the parameter $g(\theta)$ provided that for every $\theta \in \Theta$*

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

Any estimator that not unbiased is called **biased**.
If the image of $g(\theta)$ is a vector space, then the **bias**

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

Exercise 5. *If $X_1, \dots, X_n$ 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*

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

If we choose the quadratic loss function $\mathcal{L}(\theta, a) = (a-\theta)^2$, then corresponding risk function

$$\begin{aligned} \mathcal{R}_2(g(\theta), d) &= E_\theta[(d(X) - g(\theta))^2] = E_\theta[(d(X) - E_\theta d(X) + b_d(\theta))^2] \\ &= E_\theta[(d(X) - E_\theta d(X))^2] + 2b_d(\theta)(E_\theta[(d(X) - E_\theta d(X)] + b_d(\theta)^2 \\ &= \mathrm{Var}_\theta(d(X)) + b_d(\theta)^2 \end{aligned}$$

Note that the risk is the variance of an unbiased estimator and the bias adds to the risk.

In the example above, with $d(X) = \bar{X}$,

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

Thus, $\bar{x}$ is an unbiased estimator for θ. In addition,

$$\mathrm{Var}(\bar{X}) = \frac{1}{n^2}(\theta(1-\theta) + \cdots + \theta(1-\theta)) = \frac{1}{n}\theta(1-\theta).$$

Example 6. *If, in addition, the simple random sample 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,

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

Thus,

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

and

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

is an unbiased estimator for σ^2.

Definition 7. *An unbiased estimator d is a* **uniformly minimum variance unbiased estimator (UMVUE)** *if $d(X)$ has finite variance for every value θ of the parameter and for every unbiased estimator $\tilde{d}$,*

$$Var_\theta d(X) \leq Var_\theta \tilde{d}(X).$$

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

$$e(\tilde{d})=\frac{Var_\theta d(X)}{Var_\theta \tilde{d}(X)}.$$

Thus, the efficiency is between 0 and 1.

2 Cramér-Rao Bound

First, we will review a bit on correlation. For two random variables Y and Z, the correlation

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

The correlation takes values $-1 \leq \rho(Y,Z) \leq 1$ 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,

$$\mathrm{Cov}(Y,Z)^2 \leq \mathrm{Var}(Y)\mathrm{Var}(Z).$$

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

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

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

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

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

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

or,

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

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

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

Also, the random variables $\{\ln f(x_k|\theta); 1 \le k \le n\}$ are independent and have the same distribution. Thus, the Fisher information.

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

Example 8. *For independent Bernoulli random variable with unknown success probability θ,*

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

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

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

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

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

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

Example 9. *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

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

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

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.

Recall that for the correlation to be ± 1, the estimator $d(X)$ and the score function $\partial \ln \mathbf{f}(X|\theta)/\partial\theta$. must be linearly related with probability 1.

$$\frac{\partial}{\partial\theta} \ln f(X|\theta) = a(\theta)d(X) + b(\theta)$$

After integrating, we obtain,

$$f(X|\theta) = c(\theta)h(x)\exp(\pi(\theta)d(X)). \tag{9}$$

We shall call density functions satisfying equation (9) an **exponential family** with **natural parameter** $\pi(\theta)$.

Example 10 (Poisson random variables)**.**

$$f(x|\lambda) = \frac{\lambda^x}{x!}e^{-\lambda} = e^{-\lambda}\frac{1}{x!}\exp(x\ln\lambda).$$

Thus, Poisson random variables are an exponential family. The score function

$$\frac{\partial}{\partial\lambda}f(x|\lambda) = \frac{\partial}{\partial\lambda}(x\ln\lambda - \ln x! - \lambda) = \frac{x}{\lambda} - 1.$$

The Fisher information

$$I(\lambda) = E_\lambda\left[\left(\frac{X}{\lambda} - 1\right)^2\right] = \frac{1}{\lambda^2}E_\lambda[(X-\lambda)^2] = \frac{1}{\lambda}.$$

If X is $Pois(\lambda)$, then $E_\lambda X = Var_\lambda(X) = \lambda$. For a simple random sample having n observations both $\bar{X}$ and $\sum_{k=1}^n (X_k\bar{X})^2/(n-1)$ are unbiased estimators. However,

$$Var_\lambda(\bar{X}) = \frac{\lambda}{n}$$

and $d(x) = \bar{x}$ has efficiency 1.

This could have been predicted. The density of n independent observations is

$$\mathbf{f}(\mathbf{x}|\lambda) = \frac{e^{-n\lambda}\lambda^{x_1\cdots+x_n}}{x_1!\cdots x_n!} = \frac{e^{-n\lambda}\lambda^{n\bar{x}}}{x_1!\cdots x_n!}$$

and so the score function

$$\frac{\partial}{\partial\lambda}\ln\mathbf{f}(\mathbf{x}|\lambda) = \frac{\partial}{\partial\lambda}(-n\lambda + n\bar{x}\ln\lambda) = -n + \frac{n\bar{x}}{\lambda}$$

showing that the estimator and the score function are linearly related.