8.6 Efficiency and the Cramer-Rao Lower Bound

Let X be a random variable with pdf $f(x;\theta)$, $\theta \in \Omega$ where the parameter space Ω is an interval. Note that,

$$\int_{-\infty}^{\infty} f(x;\theta)dx = 1$$

and, if we can differentiate under the integral sign,

$$\frac{\partial \int_{-\infty}^{\infty} f(x;\theta)dx}{\partial\theta} = \int_{-\infty}^{\infty} \frac{\partial f(x;\theta)}{\partial\theta}dx = 0.$$

Notice that,

$$\int_{-\infty}^{\infty} \frac{\partial f(x;\theta)}{\partial\theta}dx = \int_{-\infty}^{\infty} \frac{\partial ln[f(x;\theta)]}{\partial\theta} f(x;\theta)dx = 0.$$

Differentiating again, we have

$$\int_{-\infty}^{\infty} \left[\frac{\partial^2 ln[f(x;\theta)]}{\partial \theta^2} f(x;\theta) + \frac{\partial ln[f(x;\theta)]}{\partial \theta} \frac{\partial f(x;\theta)}{\partial \theta} \right] dx = 0.$$

Notice that for the second term on the left side of the equation above we can write

$$\int_{-\infty}^{\infty} \frac{\partial ln[f(x;\theta)]}{\partial \theta} \frac{\partial f(x;\theta)}{\partial \theta} dx$$

$$= \int_{-\infty}^{\infty} \left[\frac{\partial ln[f(x;\theta)]}{\partial \theta} \right]^2 f(x;\theta) dx.$$

This integral is called **Fisher information** and is denoted by $I(\theta)$. It is an expectation!

We can see from the work above that

$$I(\theta) = \int_{-\infty}^{\infty} \left[\frac{\partial ln[f(x;\theta)]}{\partial \theta} \right]^2 f(x;\theta) dx$$

or equivalently,

$$I(\theta) = -\int_{-\infty}^{\infty} \frac{\partial^2 ln[f(x;\theta)]}{\partial \theta^2} f(x;\theta) dx.$$

Example 1: Let X be $N(\theta, \sigma^2)$, where $-\infty < \theta < \infty$. and σ^2 is known. Then

$$f(x;\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left[-\frac{(x-\theta)^2}{2\sigma^2}\right]$$

and

$$ln[f(x;\theta)] = -\frac{1}{2} ln(2\pi\sigma^2) - \frac{(x-\theta)^2}{2\sigma^2}.$$

Differentiating with respect to θ we have

$$\frac{\partial ln[f(x;\theta)]}{\partial\theta} = \frac{x-\theta}{\sigma^2}$$

and

$$\frac{\partial^2 ln[f(x;\theta)]}{\partial\theta^2} = \frac{-1}{\sigma^2}.$$

No matter which version of $I(\theta)$ we use, we see that

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

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

Example 2: Let X be binomial $b(1, \theta)$. Then

$$f(x;\theta) = \theta^x (1-\theta)^{1-x}$$

and

$$ln[f(x;\theta)] = x ln(\theta) + (1-x) ln(1-\theta).$$

$$\frac{\partial ln[f(x;\theta)]}{\partial \theta} = \frac{x}{\theta} - \frac{1-x}{1-\theta}$$

$$\frac{\partial^2 ln[f(x;\theta)]}{\partial \theta^2} = \frac{-x}{\theta^2} - \frac{1-x}{(1-\theta)^2}.$$

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

$$= \frac{\theta}{\theta^2} + \frac{1-\theta}{(1-\theta)^2} = \frac{1}{\theta(1-\theta)}$$

Now suppose that we have a random sample $X_1, X_2, ..., X_n$ from a distribution with pdf $f(x;\theta)$. The likelihood function is given by

$$L(\theta) = f(x_1;\theta)f(x_2;\theta)\cdots f(x_n;\theta)$$

and

$$ln[L(\theta)] = \sum_{i=1}^{n} ln[f(x_i;\theta)]$$

which implies that

$$\frac{\partial ln[L(\theta)]}{\partial\theta} = \sum_{i=1}^{n} \frac{\partial ln[f(x_i;\theta)]}{\partial\theta}.$$

Thus, the natural definition of Fisher information in a sample of size n is

$$I_n(\theta) = E\left(\left[\frac{\partial ln[L(\theta)]}{\partial\theta}\right]^2\right).$$

Notice that for $i \neq j$, cross-product terms in this expectation are 0. By independence,

$$E\left[\frac{\partial ln[f(X_i;\theta)]}{\partial\theta}\frac{\partial ln[f(X_j;\theta)]}{\partial\theta}\right]$$

$$= E\left[\frac{\partial ln[f(X_i;\theta)]}{\partial\theta}\right] E\left[\frac{\partial ln[f(X_j;\theta)]}{\partial\theta}\right] = 0$$

It follows that

$$I_n(\theta) = \sum_{i=1}^{n} E\left(\left[\frac{\partial ln[f(X_i;\theta)]}{\partial\theta}\right]^2\right) = nI(\theta)$$

Theorem A Cramer-Rao Inequality
Let $X_1, \ldots, X_n$ be i.i.d with density function $f(x;\theta)$.

Let $T = u(X_1, X_2, ..., X_n)$ be an estimator of θ. We allow that T might be biased, and denote its expectation by

$$E[T] = E[u(X_1, ..., X_n)] = k(\theta).$$

It turns out that we can bound $Var(T)$ from below using the **Cramer-Rao inequality**,

$$Var(T) \geq \frac{[k'(\theta)]^2}{nI(\theta)}.$$

If $T = u(X_1, X_2, ..., X_n)$ is an unbiased estimator of θ, then $k(\theta) = \theta$ and $k'(\theta) = 1$. In this case, the Cramer-Rao inequality becomes

$$Var(T) \geq \frac{1}{nI(\theta)}.$$

Recall from Examples 1 and 2 that $\frac{1}{nI(\theta)}$ equals σ^2/n and $\theta(1-\theta)/n$, respectively. Thus, we see that in both cases the sample mean $\bar{X}$ achieves the Rao-Cramer lower bound.

Definition Let T be an unbiased estimator of θ. The statistic T is called an **efficient estimator** of θ if and only if the variance of T attains the Cramer-Rao lower bound.

Definition The ratio of the Rao-Cramer lower bound to the actual variance of an unbiased estimator of θ is called the **efficiency** of that estimator.

Example 3 Let $X_1, X_2, ..., X_n$ be a random sample from a Poisson distribution with mean $\theta > 0$. We have seen that $\bar{X}$ is the maximum likelihood estimator of θ.

$$f(x;\theta) = \frac{\theta^x e^{-\theta}}{x!}$$

$$ln[f(x;\theta)] = xln(\theta) - \theta - ln(x!)$$

$$\frac{\partial ln[f(x;\theta)]}{\partial\theta} = \frac{(x-\theta)}{\theta}$$

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

We see that the Rao-Cramer lower bound is θ/n, which is the variance of $\bar{X}$. Hence $\bar{X}$ is an efficient estimator of θ.

Consider a family of distributions $\{f(x;\theta) : \theta \in \Omega\}$ where Ω is an interval. Assuming that we can interchange differentiation with integration in the manner described above, we can determine the limiting distribution of the maximum likelihood estimator $\hat{\theta}$.

In particular, if $\hat{\theta}$ denotes the maximum likelihood estimator and

$$Z_n = \frac{\hat{\theta} - \theta}{\sqrt{\frac{1}{nI(\theta)}}}$$

then Z_n has a $N(0,1)$ limiting distribution.

This implies that $\hat{\theta}$ is **asymptotically unbiased** and the **asymptotic variance** of $\hat{\theta}$ is $1/[nI(\theta)]$, which implies that $\hat{\theta}$ is **asymptotically efficient**.

By using a Theorem, we can see that the statistic

$$\frac{\hat{\theta} - \theta}{\sqrt{\frac{1}{nI(\hat{\theta})}}}$$

also has limiting distribution $N(0, 1)$. This implies that a confidence interval for θ with confidence level of approximately $(1 - \alpha)100\%$ is given by

$$\hat{\theta} \pm \frac{z_{\alpha/2}}{\sqrt{nI(\hat{\theta})}}$$

8.7 Sufficiency

Suppose $X_1, X_2, ..., X_n$ is a random sample from a distribution that has pdf $f(x;\theta)$, $\theta \in \Omega$. A statistic $T = u(X_1, X_2, ..., X_n)$ can be viewed as a reduction of the data.
We will be concerned with when it is possible to reduce the data into a statistic that suffices for retaining all of the information in the sample about the parameter θ.

Suppose we have a statistic $T = u(X_1, X_2, ..., X_n)$ that partitions the sample space into

$$(X_1, X_2, ..., X_n) \in \{(x_1, x_2, ..., x_n) : u(x_1, x_2, .., x_n) = t\}$$

in such a way that the conditional probability distribution of $X_1, X_2, ..., X_n$ given $T = t$ no longer depends on θ.

In this respect, T contains all of the information in the sample about θ, and we call T a **sufficient statistic**.

Example 4: Let $X_1, X_2, ..., X_n$ be a random sample from the pdf

$$f(x; \theta) = \theta^x (1 - \theta)^{1-x}$$

for $x = 0, 1$; $0 < \theta < 1$.

The statistic $T = X_1 + X_2 + \cdots X_n$ has the pdf

$$f_T(t) = \frac{n!}{t!(n-t)!} \theta^t (n - \theta)^{1-t}$$

for $t = 0, 1, 2, ..., n$.

Consider the conditional probability

$$P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n | T = t)$$

This conditional probability obviously equals 0 when $t \neq \Sigma\, x_i$.
When $t = \Sigma\, x_i$,

$$P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n | T = t)$$

$$\frac{\Pi_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}}{\frac{n!}{t!(n-t)!}\theta^t(1-\theta)^{n-t}}$$

$$= \frac{\theta^{\sum x_i}(1-\theta)^{n-\sum x_i}}{\frac{n!}{(\sum x_i)!(n-(\sum x_i))!}\theta^{\sum x_i}(1-\theta)^{n-\sum x_i}}$$

$$= \frac{(\sum x_i)!(n-(\sum x_i))!}{n!}$$

Notice that this does not involve θ so $T = \sum X_i$ is a sufficient statistic for θ.

Theorem A (factorization theorem): Let $X_1, X_2, ..., X_n$ denote a random sample that has pdf $f(x;\theta)$, $\theta \in \Omega$. The statistic $T = u(X_1, X_2, ..., X_n)$ is a sufficient statistic for θ if and only if we can find two nonnegative functions, g and h such that

$$\prod_{i=1}^{n} f(x_i;\theta) = g[u(x_1, x_2, ..., x_n);\theta]h(x_1, x_2, ..., x_n)$$

where $h(x_1, x_2, ..., x_n)$ does not depend upon θ.

Example 5: Let $X_1, X_2, ..., X_n$ denote a random sample from a distribution with pdf

$$f(x;\theta) = \theta x^{\theta-1}$$

for $0 < x < 1$ and $\theta > 0$.

Use the factorization theorem to prove that

$$T = u(X_1, X_2, ..., X_n) = X_1 X_2 \cdots X_n$$

is a sufficient statistic for θ.

The joint pdf of $X_1, X_2, ..., X_n$ is

$$\theta^n (x_1 x_2 \cdots x_n)^{\theta-1} = [\theta^n (x_1 x_2 \cdots x_n)^{\theta}] \left(\frac{1}{x_1 x_2 \cdots x_n}\right)$$

for $0 < x_i < 1$. In the factorization theorem we let

$$g[u(x_1, x_2, ..., x_n);\theta] = \theta^n (x_1 x_2 \cdots x_n)^{\theta}$$

and let

$$h(x_1, x_2, ..., x_n) = \frac{1}{x_1 x_2 \cdots x_n}$$

Since h does not depend on θ, the product $X_1 X_2 \cdots X_n$ is a sufficient statistic for θ.

Example 6: Let $X_1, X_2, ..., X_n$ be a sample of size n from a Poisson distribution with mean θ, $0 < \theta < \infty$. Show that

$$T = \sum_{i=1}^{n} X_i$$

is sufficient for θ. The joint pdf of $X_1, X_2, ..., X_n$ is

$$\prod_{i=1}^{n} \frac{\theta^{x_i} e^{-\theta}}{x_i!}$$

$$= [\theta^{\sum x_i} e^{-n\theta}] \left(\frac{1}{x_1! x_2! \cdots x_n!} \right)$$

In the factorization theorem we let

$$g[u(x_1, x_2, ..., x_n); \theta] = [\theta^{T} e^{-n\theta}]$$

and let

$$h(x_1, x_2, ..., x_n) = \left(\frac{1}{x_1!x_2!\cdots x_n!}\right)$$

Since h does not depend on θ, T is a sufficient statistic for θ.

Exponential Families

One-parameter members of an exponential family have a density or frequency function of the form

$$f(x;\theta) = \exp[c(\theta)T(x) + d(\theta) + S(x)], x \in A$$

and equal to 0 for x not in A, where the set A does not depend on θ.

Let $X_1, X_2, ..., X_n$ be a random sample of size n and has a joint pdf

$$\prod_{i=1}^{n} f(x_i;\theta) = exp[c(\theta)\sum_{i=1}^{n} T(x_i) + nd(\theta)]exp[\sum_{i=1}^{n} S(x_i)]$$

We see from this result that $\Sigma_{i=1}^{n} T(X_i)$ is a sufficent statistics.

Note: normal, binomial, Poisson, Gamma are members of this family!

Corollary A: If T is sufficient for θ and if a maximum likelihood estimator $\hat{\theta}$ exists uniquely, then $\hat{\theta}$ is a function of T.

The following theorem of Rao and Blackwell implies that in searching for a best unbiased estimator, we may restrict our attention to functions of a sufficient statistic, if a sufficient statistic exists.

This is helpful because there is usually only one unbiased estimator of θ based on a sufficient statistic.

Theorem A Rao-Blackwell Theorem: Let $\hat{\theta}$ be an estimator of θ with $E(\hat{\theta}^2) < \infty$ for all θ. Suppose that T is sufficient for θ and let $\tilde{\theta} = E(\hat{\theta}|T)$. Then, for all θ,

$$E(\tilde{\theta} - \theta)^2 \leq E(\hat{\theta} - \theta)^2$$

If an estimator is not a function of the sufficient statistic, it can be improved!