Lecture 6

Efficient estimators. Rao-Cramer bound.

1 MSE and Sufficiency

Let $X = (X_1, ..., X_n)$ be a random sample from distribution f_θ. Let $\hat{\theta} = \delta(X)$ be an estimator of θ. Let $T(X)$ be a sufficient statistic for θ. As we have seen already, MSE provides one way to compare the quality of different estimators. In particular, estimators with smaller MSE are said to be more efficient. On the other hand, once we know $T(X)$, we can discard X. How do these concepts relate to each other? The theorem below shows that for any estimator $\hat{\theta} = \delta(X)$, there is another estimator which depends on data X only through $T(X)$ and is at least as efficient as $\hat{\theta}$:

Theorem 1 (Rao-Blackwell). *In the setting above, define $\phi(T) = E[\delta(X)|T]$. Then $\hat{\theta}_2 = \phi(T(X))$ is an estimator for θ and $MSE(\hat{\theta}_2) \leq MSE(\hat{\theta})$. In addition, if $\hat{\theta}$ is unbiased, then $\hat{\theta}_2$ is unbiased as well.*

Proof. To show that $\hat{\theta}_2$ is an estimator, we have to check that it does not depend on θ. Indeed, since T is sufficient for θ, the conditional distribution of X given T is independent of θ. So the conditional distribution of $\delta(X)$ given T is independent of θ as well. In particular, the conditional expectation $E[\delta(X)|T]$ does not depend on θ. Thus, $\phi(T(X))$ depends only on the data X and $\hat{\theta}_2$ is an estimator.

$$\begin{aligned}
\text{MSE}(\hat{\theta}) &= E[(\hat{\theta} - \hat{\theta}_2 + \hat{\theta}_2 - \theta)^2] \\
&= E[(\hat{\theta} - \hat{\theta}_2)^2] + 2E[(\hat{\theta} - \hat{\theta}_2)(\hat{\theta}_2 - \theta)] + E[(\hat{\theta}_2 - \theta)^2] \\
&= E[(\hat{\theta} - \hat{\theta}_2)^2] + 2E[(\hat{\theta} - \hat{\theta}_2)(\hat{\theta}_2 - \theta)] + \text{MSE}(\hat{\theta}_2) \\
&= E[(\hat{\theta} - \hat{\theta}_2)^2] + \text{MSE}(\hat{\theta}_2),
\end{aligned}$$

where in the last line we used

$$\begin{aligned}
E[(\hat{\theta} - \hat{\theta}_2)(\hat{\theta}_2 - \theta)] &= E[(\delta(X) - \phi(T(X)))(\phi(T(X)) - \theta)] \\
&= E[E[(\delta(X) - \phi(T(X)))(\phi(T(X)) - \theta)|T]] \\
&= E[(\phi(T(X)) - \theta)E[(\delta(X) - \phi(T(X)))|T]] \\
&= E[(\phi(T(X)) - \theta) \cdot (E[\delta(X)|T] - \phi(T(X)))] \\
&= 0,
\end{aligned}$$

since $E[\delta(X)|T] = \phi(T(X))$.

Cite as: Anna Mikusheva, course materials for 14.381 Statistical Methods in Economics, Fall 2011. MIT OpenCourseWare (http://ocw.mit.edu), Massachusetts Institute of Technology. Downloaded on [DD Month YYYY].

To show the last result, we have

$$E[\phi(T(X))] = E[E[\delta(X)|T]] = E[\delta(X)] = \theta$$

by the law of iterated expectation. □

Example Let $X_1, ..., X_n$ be a random sample from Binomial(p, k), i.e. $P\{X_j = m\} = (k!/(m!(k - m)!))p^m(1-p)^{k-m}$ for any integer $m \geq 0$. Suppose our parameter of interest is the probability of one success, i.e. $\theta = P\{X_j = 1\} = kp(1-p)^{k-1}$. One possible estimator is $\hat{\theta} = \sum_{i=1}^n I(X_i = 1)/n$. This estimator is unbiased, i.e. $E[\hat{\theta}] = \theta$. Let us find a sufficient statistic. The joint density of the data is

$$\begin{aligned} f(x_1, ..., x_n) &= \prod_{i=1}^n (k!/(x_i!(k-x_i)!))p^{x_i}(1-p)^{k-x_i} \\ &= \text{function}(x_1, ..., x_n)p^{\sum x_i}(1-p)^{nk-\sum x_i} \end{aligned}$$

Thus, $T = \sum_{i=1}^n X_i$ is sufficient. In fact, it is minimal sufficient.

Using the Rao-Blackwell theorem, we can improve $\hat{\theta}$ by considering its conditional expectation given T. Let $\phi = E[\hat{\theta}|T]$ denote this estimator. Then, for any nonnegative integer t,

$$\begin{aligned} \phi(t) &= E[\sum_{i=1}^n I(X_i = 1)/n | \sum_{i=1}^n X_i = t] \\ &= \sum_{i=1}^n P\{X_i = 1 | \sum_{j=1}^n X_j = t\}/n \\ &= P\{X_1 = 1 | \sum_{j=1}^n X_j = t\} \\ &= \frac{P\{X_1 = 1, \sum_{j=1}^n X_j = t\}}{P\{\sum_{j=1}^n X_j = t\}} \\ &= \frac{P\{X_1 = 1, \sum_{j=2}^n X_j = t-1\}}{P\{\sum_{j=1}^n X_j = t\}} \\ &= \frac{P\{X_1 = 1\}P\{\sum_{j=2}^n X_j = t-1\}}{P\{\sum_{j=1}^n X_j = t\}} \\ &= \frac{kp(1-p)^{k-1} \cdot (k(n-1))!/((t-1)!(k(n-1)-(t-1))!)p^{t-1}(1-p)^{k(n-1)-(t-1)}}{(kn)!/(t!(kn-t)!)p^t(1-p)^{kn-t}} \\ &= \frac{k(k(n-1))!/((t-1)!(k(n-1)-(t-1))!)}{(kn)!/(t!(kn-t)!)} \\ &= \frac{k(k(n-1))!(kn-t)!t}{(kn)!(kn-k+1-t)!} \end{aligned}$$

where we used the fact that X_1 is independent of $(X_2, ..., X_n)$, $\sum_{i=1}^n X_i \sim$ Binomial(kn, p), and $\sum_{i=2}^n X_i \sim$

Binomial$(k(n-1),p)$. So our new estimator is

$$\hat{\theta}_2 = \phi(X_1, ..., X_n) = \frac{k(k(n-1))!(kn - \sum_{i=1}^n X_i)!(\sum_{i=1}^n X_i)}{(kn)!(kn-k+1-\sum_{i=1}^n X_i)!}$$

By the theorem above, it is unbiased and at least as efficient as $\hat{\theta}$. The procedure we just applied is sometimes informally referred to as Rao-Blackwellization.

2 Fisher information

Let $f(x|\theta)$ with $\theta \in \Theta$ be some parametric family. For given $\theta \in \Theta$, let $Supp_\theta = \{x : f(x|\theta) > 0\}$. $Supp_\theta$ is usually called the support of distribution $f(x|\theta)$. Assume that $Supp_\theta$ does not depend on θ. As before, $l(x,\theta) = \log f(x|\theta)$ is called loglikelihood function. Assume that $l(x,\theta)$ is twice continuously differentiable in θ for all $x \in S$. Let X be some random variable with distribution $f(x|\theta)$. Then

Definition 2. $I(\theta) = E_\theta[(\partial l(X,\theta)/\partial\theta)^2]$ is called Fisher information.

Fisher information plays an important role in maximum likelihood estimation. The theorem below gives two information equalities:

Theorem 3. *In the setting above, (1)* $E_\theta[\partial l(X,\theta)/\partial\theta] = 0$ *and (2)* $I(\theta) = -E_\theta[\partial^2 l(X,\theta)/\partial\theta^2]$.

Proof. Since $l(x,\theta)$ is twice differentiable in θ, $f(x|\theta)$ is twice differentiable in θ as well. Differentiating identity $\int_{-\infty}^{+\infty} f(x|\theta)dx = 1$ with respect to θ yields

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

for all $\theta \in \Theta$. The second differentiation yields

$$\int_{-\infty}^{+\infty} \frac{\partial^2 f(x|\theta)}{\partial\theta^2} dx = 0 \tag{1}$$

for all $\theta \in \Theta$. In addition,

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

and

$$\frac{\partial^2 l(x,\theta)}{\partial\theta^2} = -\frac{1}{f^2(x|\theta)}\left(\frac{\partial f(x|\theta)}{\partial\theta}\right)^2 + \frac{1}{f(x|\theta)}\frac{\partial^2 f(x|\theta)}{\partial\theta^2}.$$

The former equality yields

$$E_\theta\left[\frac{\partial l(X,\theta)}{\partial\theta}\right] = E_\theta\left[\frac{1}{f(X|\theta)}\frac{\partial f(X|\theta)}{\partial\theta}\right] = \int_{-\infty}^{+\infty}\frac{1}{f(x|\theta)}\frac{\partial f(x|\theta)}{\partial\theta}f(x|\theta)dx = \int_{-\infty}^{+\infty}\frac{\partial f(x|\theta)}{\partial\theta}dx = 0,$$

which is our first result. The latter equality yields

$$E_\theta\left[\frac{\partial^2 l(X,\theta)}{\partial\theta^2}\right] = -\int_{-\infty}^{+\infty}\frac{1}{f(x|\theta)}\left(\frac{\partial f(x|\theta)}{\partial\theta}\right)^2 dx$$

in view of equation (1). So,

$$\begin{aligned} I(\theta) &= E_\theta\left[\left(\frac{\partial l(X,\theta)}{\partial\theta}\right)^2\right] \\ &= \int_{-\infty}^{+\infty}\left(\frac{1}{f(x|\theta)}\frac{\partial f(x|\theta)}{\partial\theta}\right)^2 f(x|\theta)dx \\ &= \int_{-\infty}^{+\infty}\frac{1}{f(x|\theta)}\left(\frac{\partial f(x|\theta)}{\partial\theta}\right)^2 dx \\ &= -E_\theta\left[\frac{\partial^2 l(X,\theta)}{\partial\theta^2}\right], \end{aligned}$$

which is our second result. □

Example Let us calculate Fisher information for an $N(\mu,\sigma^2)$ distribution where σ^2 is known. Thus, our parameter $\theta=\mu$. The density of a normal distribution is $f(x|\mu)=\exp(-(x-\mu)^2/(2\sigma^2))/\sqrt{2\pi}$. The log-likelihood is $l(x,\mu)=-\log(2\pi)/2-(x-\mu)^2/(2\sigma^2)$. So $\partial l(x,\mu)/\partial\mu=(x-\mu)/\sigma^2$ and $\partial^2 l(x,\mu)/\partial\mu^2=-1/\sigma^2$. So $-E_\theta[\partial^2 l(X,\mu)/\partial\mu^2]=1/\sigma^2$. At the same time,

$$I(\theta)=E_\mu[(\partial l(X,\mu)/\partial\mu)^2]=E_\mu[(X-\mu)^2/\sigma^4]=1/\sigma^2$$

So, as was expected in view of the theorem above, $I(\theta)=-E_\mu[\partial^2 l(X,\mu)/\partial\mu^2]$ in this example.

Example Let us calculate Fisher information for a Bernoulli(θ) distribution. Note that a Bernoulli distribution is discrete. So we use probability mass function (pms) instead of pdf. The pms of Bernoulli(θ) is $f(x|\theta)=\theta^x(1-\theta)^{1-x}$ for $x\in\{0,1\}$. The log-likelihood is $l(x,\theta)=x\log\theta+(1-x)\log(1-\theta)$. So $\partial l(x,\theta)/\partial\theta=x/\theta-(1-x)/(1-\theta)$ and $\partial^2 l(x,\theta)/\partial\theta^2=-x/\theta^2-(1-x)/(1-\theta)^2$. So

$$\begin{aligned} E_\theta[(\partial l(X,\theta)/\partial\theta)^2] &= E_\theta[(X/\theta-(1-X)/(1-\theta))^2] \\ &= E_\theta[X^2/\theta^2]-2E_\theta[X(1-X)/(\theta(1-\theta))]+E_\theta[(1-X)^2/(1-\theta)^2] \\ &= E_\theta[X/\theta^2]+E_\theta[(1-X)/(1-\theta)^2] \\ &= 1/(\theta(1-\theta)) \end{aligned}$$

since $x = x^2$, $x(1-x) = 0$, and $(1-x) = (1-x)^2$ if $x \in \{0,1\}$. At the same time,

$$\begin{aligned} -E_\theta[\partial^2 l(X,\theta)/\partial\theta^2] &= E_\theta[X/\theta^2 + (1-X)/(1-\theta)^2] \\ &= \theta/\theta^2 + (1-\theta)/(1-\theta)^2 \\ &= 1/\theta + 1/(1-\theta) \\ &= 1/(\theta(1-\theta)) \end{aligned}$$

So $I(\theta) = -E_\theta[\partial^2 l(X,\theta)/\partial\theta^2]$ as it should be.

2.1 Information for a random sample

Let us now consider Fisher information for a random sample. Let $X = (X_1, ..., X_n)$ be a random sample from distribution $f(x|\theta)$. Then the joint pdf is $f_n(x) = \prod_{i=1}^n f(x_i|\theta)$ where $x = (x_1, ..., x_n)$. The joint log-likelihood is $l_n(x,\theta) = \sum_{i=1}^n l(x_i,\theta)$. So Fisher information for the sample X is

$$\begin{aligned} I(\theta) &= E_\theta[(\partial l_n(X,\theta)/\partial\theta)^2] \\ &= E_\theta[\sum_{1\le i,j\le n}(\partial l(X_i,\theta)/\partial\theta)(\partial l(X_j,\theta)/\partial\theta)] \\ &= E_\theta[\sum_{i=1}^n(\partial l(X_i,\theta)/\partial\theta)^2] + 2E_\theta[\sum_{1\le i<j\le n}(\partial l(X_i,\theta)/\partial\theta)(\partial l(X_j,\theta)/\partial\theta)] \\ &= E_\theta[\sum_{i=1}^n(\partial l(X_i,\theta)/\partial\theta)^2] \\ &= nI(\theta) \end{aligned}$$

where we used the fact that for any $i < j$, $E_\theta[(\partial l(X_i,\theta)/\partial\theta)(\partial l(X_j,\theta)/\partial\theta)] = 0$ by independence and first information equality. Here $I(\theta)$ denotes Fisher information for distribution $f(x|\theta)$.

3 Rao-Cramer bound

An important question in the theory of statistical estimation is whether there is a nontrivial bound such that no estimator can be more efficient than this bound. The theorem below is a result of this sort:

Theorem 4 (Rao-Cramer bound)**.** *Let $X = (X_1, ..., X_n)$ be a random sample from distribution $f(x|\theta)$ with information $I_n(\theta)$. Let $W(X)$ be an estimator of θ such that (1) $dE_\theta[W(X)]/d\theta = \int W(x) df(x,\theta)/d\theta dx$ where $x = (x_1, ...x_n)$ and (2) $V(W) < \infty$. Then $V(W) \ge (dE_\theta[W(X)]/d\theta)^2/I_n(\theta)$. In particular, if W is unbiased for θ, then $V(W) \ge 1/I_\theta(\theta)$.*

Proof. The first information equality gives $E_\theta[\partial l(X,\theta)/\partial\theta]=0$. So,

$$\begin{aligned}
\operatorname{cov}(W(X),\partial l(X,\theta)/\partial\theta) &= E[W(X)\partial l(X,\theta)/\partial\theta] \\
&= \int W(x)\partial l(x,\theta)/\partial\theta f(x|\theta)dx \\
&= \int W(x)\partial f(x|\theta)/\partial\theta\cdot(1/f(x|\theta))f(x|\theta)dx \\
&= \int W(x)\partial f(x|\theta)/\partial\theta dx \\
&= dE_\theta[W(X)]/d\theta.
\end{aligned}$$

By the Cauchy-Schwarz inequality,

$$(\operatorname{cov}(W(X),\partial l(X,\theta)/\partial\theta)^2\leq V(W(X))V(\partial l(X,\theta)/\partial\theta)=V(W(X))I_n(\theta).$$

Thus,

$$V(W(X))\geq(dE_\theta[W(X)]/d\theta)^2/I_n(\theta).$$

If W is unbiased for θ, then $E_\theta[W(X)]=\theta$, $dE_\theta[W(X)]/d\theta=1$, and $V(W(X))\geq 1/I_n(\theta)$. □

Example Let us calculate the Rao-Cramer bound for random sample $X_1,...,X_n$ from Bernoulli(θ) distribution. We have already seen that $I(\theta)=1/(\theta(1-\theta))$ in this case. So Fisher information for the sample is $I_n(\theta)=n/(\theta(1-\theta))$. Thus, any unbiased estimator of θ, under some regularity conditions, has variance no smaller than $\theta(1-\theta)/n$. On the other hand, let $\hat\theta=\overline{X}_n=\sum_{i=1}^n X_i/n$ be an estimator of θ. Then $E_\theta[\hat\theta]=\theta$, i.e. $\hat\theta$ is unbiased, and $V(\hat\theta)=\theta(1-\theta)/n$ which coincides with the Rao-Cramer bound. Thus, $\overline{X}_n$ is the uniformly minimum variance unbiased (UMVU) estimator of θ. Word "uniformly" in this situation means that $\overline{X}_n$ has the smallest variance among unbiased estimators for *all* $\theta\in\Theta$.

Example Let us now consider a counterexample to the Rao-Cramer theorem. Let $X_1,...,X_n$ be a random sample from $U[0,\theta]$. Then $f(x_i|\theta)=1/\theta$ if $x_i\in[0,\theta]$ and 0 otherwise. So $l(x_i,\theta)=-\log\theta$ if $x_i\in[0,\theta]$. Then $\partial l/\partial\theta=-1/\theta$ and $\partial^2 l/\partial\theta^2=1/\theta^2$. So $I(\theta)=1/\theta^2$ while $-E_\theta[\partial^2 l(X_i,\theta)/\partial\theta^2]=-1/\theta^2\neq I(\theta)$. Thus, the second information equality does not hold in this example. The reason is that support of the distribution depends on θ in this example. Moreover, consider an estimator $\hat\theta=((n+1)/n)X_{(n)}$ of θ. Then $E_\theta[X_{(n)}]=\theta$ and

$$V(\hat\theta)=((n+1)^2/n^2)V(X_{(n)})=\theta^2/(n(n+2))$$

as we saw when we considered order statistics. So $\hat\theta$ is unbiased, but its variance is smaller than $1/I_n(\theta)=\theta^2/n^2$. Thus, the Rao-Cramer theorem does not work in this example as well. Again, the reason is that Rao-Cramer theorem assumes that support is independent of parameter.

14.381 Statistical Method in Economics
Fall 2013