Lecture 4

Xiaoping Shi
Department of Statistics, University of Toronto
xpshi@utstat.toronto.edu

July 21, 2013

- Efficiency and the Cramer-Rao Lower Bound
- Sufficiency, the Factorization Theorem and Exponential family
- The Rao-Blackwell Theorem

In most statistical estimation problems, there are a variety of possible parameter estimates.

Given a variety of possible estimates, how would we choose which to use? Two quantitative measures are specified: Mean squared error (MSE) and efficiency.

The mean squared error of $\hat{\theta}$ as an estimate of θ_0 is

$$MSE(\hat{\theta}) = E(\hat{\theta} - \theta_0)^2 = Var(\hat{\theta}) + (E(\hat{\theta}) - \theta_0)^2$$

Given two estimates, $\hat{\theta}$ and $\tilde{\theta}$, of a parameter θ_0, the efficiency of $\hat{\theta}$ relative to $\tilde{\theta}$ is defined to be

$$\text{eff}(\hat{\theta}, \tilde{\theta}) = \frac{Var(\tilde{\theta})}{Var(\hat{\theta})}$$

Recall: $\hat{\theta}$ is a consistent estimate of θ_0 in probability, that is, for any $\varepsilon > 0$,

$$P(|\hat{\theta} - \theta_0| > \varepsilon) \to 0, \quad \text{as} \quad n \to \infty$$

by **Chebyschev's Lemma**. Link *MSE* and Consistence by Chebyschev's Lemma:

$$P(|\hat{\theta} - \theta_0| > \varepsilon) \leq \frac{MSE(\hat{\theta})}{\varepsilon^2}$$

Question: $X_1, \cdots, X_n$ are iid $N(\mu, \sigma^2)$, is $\hat{\sigma}^2$ (mle of σ^2) a consistent estimate of σ^2?

Compare $MSE(\hat{\sigma}^2_{\text{mle}})$ and $MSE(\hat{\sigma}^2_{\text{S}})$, where $\hat{\sigma}^2_{\text{S}}$ is the sample variance.

To find optimal estimate with smallest *MSE* may be difficult. We could find it with smallest variance among unbiased estimates.

Definition An unbiased estimate whose variance achieves this lower bound (Cramér-Rao bound) is said to be **efficient**.

Theorem Cramér-Rao Inequality
Let $X_1, \cdots, X_n$ be iid with density function $f(x|\theta)$. Let $T = t(X_1, \cdots, X_n)$ be an unbiased estimate of θ. Then, under smoothness assumptions of $f(x|\theta)$,

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

- $\frac{1}{nI(\theta)}$ is called Cramér-Rao bound.
- The asymptotic variance of mle is equal to the lower bound, mle is said to be asymptotically efficient.

Proof of C-R inequality:

- i.e., prove that $Var(T)[nI(\theta)] \geq 1$.
- Let $Z = \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f(X_i|\theta) = \sum_{i=1}^{n} \frac{\frac{\partial}{\partial\theta} f(X_i|\theta)}{f(X_i|\theta)}$ with $Var(Z) = nI(\theta)$
- Cauchy-Schwartz inequality:

$$\begin{aligned} Cov^2(Z,T) &= \{E[(Z-E(Z))(T-E(T))]\}^2 \\ &\leq E[(Z-E(Z)]^2 E[T-E(T)]^2 = Var(Z)Var(T) \end{aligned}$$

- Show that $Cov(Z,T) = 1$ by $E(T) = \theta$.

Among the models encountered in practice, efficient estimators exist for: Poisson distribution, Bernoulli distribution and Normal distribution.

Example 7 continued: Poisson distribution $Pois(\lambda)$ and $I(\lambda) = 1/\lambda$.

Therefore, by C-R inequality, for any unbiased estimate T of λ, based on a sample of iid Poisson r.v.s, $X_1, \cdots, X_n$,

$$Var(T) \geq \frac{\lambda}{n}$$

The mle of λ was found to be $\bar{X}$ with $E(\bar{X}) = \lambda$ and $Var(\bar{X}) = \lambda/n$. In this sense, $\bar{X}$ is efficient.

Example 13 continued: Bernoulli distribution $B(1,\theta)$ and $I(\theta) = \frac{1}{\theta(1-\theta)}$.

Therefore, by C-R inequality, for any unbiased estimate T of θ, based on a sample of iid Bernoulli r.v.s, $X_1, \cdots, X_n$,

$$Var(T) \geq \frac{\theta(1-\theta)}{n}$$

The mle of θ was found to be $\bar{X}$ with $E(\bar{X}) = \theta$ and $Var(\bar{X}) = \theta(1-\theta)/n$. In this sense, $\bar{X}$ is efficient.

Example 1. Suppose X is a normally distributed $N(\mu, \sigma^2)$ with known μ and unknown variance σ^2. Consider the following two statistics:

$$T_1 = \frac{\sum_{i=1}^n (X_i - \mu)^2}{n}, \quad T_2 = \frac{\sum_{i=1}^n (X_i - \mu)^2}{n+2}$$

The Fish information is

$$I(\sigma^2) = -E\left(-\frac{(X-\mu)^2}{\sigma^6} + \frac{1}{2\sigma^4}\right) = \frac{1}{2\sigma^4}$$

$E(T_1) = \sigma^2$ and $Var(T_1) = 2\sigma^4/n$ which reaches the C-R lower bound, hence T_1 is efficient.

$E(T_2) = n\sigma^2/(n+2)$, $Var(T_2) = 2n\sigma^4/(n+2)^2$ and $MSE(T_2) = Var(T_2) + (E(T_2) - \sigma^2)^2 = \frac{2\sigma^4}{n+2}$, which is clearly less than $MSE(T_1) = Var(T_1) = 2\sigma^4/n$.
This shows that the biased estimator T_2 of σ^2 has a smaller mean squared error than T_1.

Definition A statistic $T(X_1, \cdots, X_n)$ is said to be **sufficient** for θ if the conditional distribution of $X_1, \cdots, X_n$, given $T = t$, does not depend on θ for any value of t.

In other words, the **sufficient statistic** T gives all knowledge about θ and we can gain no more knowledge about θ.

The preceding definition of sufficiency is hard to work with, because it does not indicate how to go about finding a sufficient statistic because of the difficulty in evaluating the conditional distribution. The following factorization theorem provides a convenient means of identifying sufficient statistics.

A Factorization Theorem A necessary and sufficient condition for $T(X_1, \cdots, X_n)$ to be sufficient for a parameter θ is that the joint probability function (density function or frequency function) factors in the form

$$f(x_1, \cdots, x_n|\theta) = g[T(x_1, \cdots, x_n), \theta]h(x_1, \cdots, x_n)$$

Proof:

$$\begin{aligned} P(\mathbf{X} = \mathbf{x}|\theta) &= P(T = t|\theta)P(\mathbf{X} = \mathbf{x}|T = t) \\ &= g(t, \theta)h(\mathbf{x}) \end{aligned}$$

where the conditional distribution of $\mathbf{X}$ given T is independent of θ due to the definition of sufficient statistic.

Example 2. Suppose the sample data $X_1, \cdots, X_n$ from the distribution with pdf

$$f(x|\theta = e^{-(x-\theta)}\mathbf{1}(\theta, x)$$

where $\mathbf{1}(a, b)$ is 1 or 0 if $a \le b$ or $a > b$, respectively. What's the sufficient statistic for θ?
The joint pdf of $X_1, \cdots, X_n$ is

$$\begin{aligned}\prod_{i=1}^{n}[e^{-(X_i-\theta)}]\mathbf{1}(\theta, X_i) &= [e^{n\theta}\mathbf{1}(\theta, \min\{X_1, \cdots, X_n\})][e^{-n\bar{X}}] \\ &= g(t, \theta)h(\mathbf{x})\end{aligned}$$

Thus $\min\{X_1, \cdots, X_n\}$ is the sufficient statistic for θ.

Question: Suppose pdf is Uniform distribution $U(0, \theta)$. What is the sufficient statistic for θ?

We can demonstrate the utility of the Factorization Theorem by introducing the **exponential family** of probability distributions.

Many common distribution, including the normal, the binomial, the Poisson, and the gamma, are members of this family.

One-parameter members of the exponential family have density or frequency functions of the form

$$f(x|\theta) = \exp[c(\theta)T(x) + d(\theta) + S(x)]$$

A k-parameter member of the exponential family has density or frequency functions of the form

$$f(x|\theta) = \exp[\sum_{i=1}^{k} c_i(\theta)T_i(x) + d(\theta) + S(x)]$$

Suppose that $X_1, \cdots, X_n$ is a sample from a member of the exponential family; the joint probability function is

$$\begin{aligned} f(\mathbf{X}|\theta) &= \prod_{i=1}^{n} \exp[c(\theta)T(X_i) + d(\theta) + S(x)] \\ &= \exp\left[c(\theta)\sum_{i=1}^{n} T(X_i) + nd(\theta)\right] + \exp\left[\sum_{i=1}^{n} S(X_i)\right] \end{aligned}$$

From this result, it is apparent by the factorization theorem that $\sum_{i=1}^{n} T(X_i)$ is a sufficient statistic.

Example 3. Consider a sequence of independent Bernoulli random variables $B(1, \theta)$, $X_1, \cdots, X_n$,

$$\begin{aligned} f(\mathbf{X}|\theta) &= \prod_{i=1}^{n} \theta^{X_i}(1-\theta)^{1-X_i} \\ &= \theta^{\sum_{i=1}^{n} X_i}(1-\theta)^{n-\sum_{i=1}^{n} X_i} \\ &= \left(\frac{\theta}{1-\theta}\right)^{\sum_{i=1}^{n} X_i}(1-\theta)^{n} \end{aligned}$$

This is a member of the exponential family with $c(\theta) = \frac{\theta}{1-\theta}$ and $T(x) = x$, and then $\sum_{i=1}^{n} T(X_i)$ is a sufficient statistic.

Example 4. Consider a sequence of independent Normal random variables $N(\mu, \sigma^2)$

$$
\begin{aligned}
f(\mathbf{X}|\mu, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[\frac{-1}{2\sigma^2}(X_i - \mu)^2\right] \\
&= \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left[\frac{-1}{2\sigma^2}\left(\sum_{i=1}^{n} X_i^2 - 2\mu\sum_{i=1}^{n} X_i + n\mu^2\right)\right]
\end{aligned}
$$

This expression is just a function of $\sum_{i=1}^{n} X_i$ and $\sum_{i=1}^{n} X_i^2$, which are therefor sufficient statistics.

In this example we have a two-dimensional sufficient statistic.

Theorem Rao-Blackwell Theorem
Let $\hat{\theta}$ be an unbiased estimator of θ. Suppose that T is sufficient for θ, and let $\tilde{\theta} = E(\hat{\theta}|T)$. Then this statistic $\tilde{\theta}$ is unbiased and

$$Var(\tilde{\theta}) \leq Var(\hat{\theta})$$

This theorem tells us that if we begin with an unbiased estimator $\hat{\theta}$ alone, then we can always improve on this by computing $\tilde{\theta}$ so that $\tilde{\theta}$ is an unbiased estimator with smaller variance that that of $\hat{\theta}$.
Proof:

$$E(\tilde{\theta}) = E[E(\hat{\theta}|T)] = E(\hat{\theta}) = \theta$$

by the property of iterated conditional expectation (Theorem A of Section 4.4.1), and

$$Var(\hat{\theta}) = Var(\tilde{\theta}) + E[Var(\hat{\theta}|T)]$$

by Theorem B of Section 4.4.1.

Exercises:

1. $X_n \geq 0$, $\mu \geq 0$, $X_n \to \mu$ in prob. Then $\sqrt{X_n} \to \sqrt{\mu}$ in prob.
$X_n \to \mu_1$ in prob. and $Y_n \to \mu_2$ in prob. Then $X_n + Y_n \to \mu_1 + \mu_2$.
2. Method of moment and MLE. For example $N(\mu, \sigma^2)$.
3. Bayesian estimator: Posterior mean.
4. Fish information and large sample theory.
5. C-R lower bound and efficient.
6. Sufficient. For example $U(\theta)$.