Probability and Statistics

Fall 2000

Lect. 16: Large Sample Properties of Maximum Likelihood Estimators, CB 7.3.4-7.4.2

If there is an Minimum Variance Unbiased Estimator with variance equal to the Cramer–Rao bound, then the MVUE is equal to the Maximum Likelihood Estimator (MLE). In this case everything works out fine. The question arises what to do when this is not true. We can still calculate the maximum likelihood estimator. It is no longer unbiased, so it cannot be the MVUE. Nevertheless, it is going to have very similar properties approximately, in a large sample sense.

Example

Let $X_1, \ldots, X_n$ be a random sample from an exponential distribution with arrival rate λ^*: $f_X(x; \lambda^*) = \lambda^* \exp(-x\lambda^*)$. The Cramér-Rao bound for the variance is λ^{*2}/N. The log likelihood function is

$$L(\lambda) = \sum_{i=1}^{N} \ln \lambda - x_i \lambda,$$

and the maximimum likelihood estimator is $\hat{\lambda} = 1/\bar{x}$. What can we say about the large sample properties of this estimator. Using a law of large numbers we have

$$\bar{x} \xrightarrow{p} E[X] = 1/\lambda^*,$$

so

$$\hat{\lambda} = 1/\bar{x} \xrightarrow{p} 1/E[X] = \lambda^*.$$

Using a central limit theorem we have

$$\sqrt{N} \cdot (\bar{x} - 1/\lambda^*) \xrightarrow{d} \mathcal{N}(0, 1/\lambda^{*2}).$$

Then we can use the delta method to establish that

$$\sqrt{N} \cdot (g(\bar{x}) - g(1/\lambda^*)) \xrightarrow{d} \mathcal{N}(0, g'(1/\lambda^*)^2/\lambda^{*2}).$$

Applying this with $g(a) = 1/a$, and thus $g'(a) = -1/a^2$, we get

$$\sqrt{N} \cdot (1/\bar{x}) - \lambda^*) \xrightarrow{d} \mathcal{N}(0, \lambda^{*4}/\lambda^{*2}) = \mathcal{N}(0, \lambda^{*2}).$$

Hence, approximately,

$$\hat{\lambda} \sim \mathcal{N}(\lambda^*, \lambda^{*2}/N).$$

So, approximately, in large samples, this maximum likelihood estimator is unbiased, and has variance approximately equal to the Cramér-Rao bound. This is true in general for maximum likelihood estimators. □

Result 1

Let $X_1, \ldots, X_n$ be a random sample from $f_X(x; \theta^*)$. Let $\hat{\theta}$ be the maximum likelihood estimator:

$$\hat{\theta} = \text{argmax}_\theta \sum_{i=1}^{N} \ln f_X(x_i; \theta).$$

Then $\hat{\theta}$ is <u>consistent</u> for θ^*:

$$\hat{\theta} \xrightarrow{p} \theta^*,$$

and $\hat{\theta}$ has asymptotically a normal distribution:

$$\sqrt{N}(\hat{\theta} - \theta^*) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}(\theta^*)^{-1}),$$

where $\mathcal{I}(\theta^*)$ is the single observation information matrix:

$$\mathcal{I}(\theta^*) = E\Big[\frac{\partial \ln f_X}{\partial \theta}(X; \theta^*) \cdot \frac{\partial \ln f_X}{\partial \theta}(X; \theta^*)\Big] = -E\Big[\frac{\partial^2 \ln f_X}{\partial \theta^2}(X; \theta^*)\Big].$$

□

First let us interpret this result relative to the Cramer–Rao bound. The CR bound implies that no unbiased estimator has a variance smaller than

$$\mathcal{I}(\theta^*)^{-1}/N.$$

The maximum likelihood estimator has an limiting normal distribution

$$\sqrt{N}(\hat{\theta} - \theta^*) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}(\theta^*)^{-1}),$$

implying that for fixed, large N,

$$\sqrt{N}(\hat{\theta} - \theta^*) \approx \mathcal{N}(0, \mathcal{I}(\theta^*)^{-1}).$$

This in turn implies that

$$\hat{\theta}_{mle} \approx \mathcal{N}(\theta^*, \mathcal{I}(\theta^*)^{-1}/N).$$

Now if this distribution was exact, the mle would be the minimum variance unbiased estimator. These distributions are however not exact, but only hold in large sample sense. In large samples, however, the maximum likelihood estimator cannot systematically been beaten: either the estimator must have a bias or a larger variance.

Example

To illustrate what this means consider an example we have looked at before, where the maximum likelihood estimator differs from the minimum variance unbiased estimator. Suppose $X_1, \ldots, X_N$ are a random sample from a normal distribution with unknown mean μ and unknown variance σ^2. We are interested in the variance σ^2. The minimum variance unbiased estimator is

$$W_1 = \frac{1}{N-1} \sum_{i=1}^{N} (X_i - \bar{X})^2.$$

The maximum likelihood estimator is

$$W_2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2 = \frac{N-1}{N} \cdot W_1.$$

As the sample gets large, the two estimators get close to each other. They are both consistent and have the same large sample distributions.

$$\sqrt{N} \cdot (W_1 - \sigma^2) \xrightarrow{d} \mathcal{N}(0, 2 \cdot \sigma^4),$$

and

$$\sqrt{N} \cdot (W_2 - \sigma^2) \xrightarrow{d} \mathcal{N}(0, 2 \cdot \sigma^4),$$

□

Proof of Result 1:
For each value of θ, we can apply a law of large numbers so that

$$\frac{1}{N}L(\theta) = \frac{1}{N}\sum_{i=1}^{N}\ln f_X(X_i;\theta) \xrightarrow{p} E[\ln f_X(X;\theta)].$$

In addition we know from Jensen' inequality that

$$\theta^* = \text{argmax}E[\ln f_X(X;\theta)].$$

To get the result that

$$\text{argmax}\frac{1}{N}L(\theta) = \text{argmax}E\left[\frac{1}{N}L(\theta)\right] = \theta^*,$$

we need that the convergence is not just <u>pointwise</u>, but <u>uniform</u> in θ, that is,

$$\sup_{\theta}\left|\frac{1}{N}L(\theta) - E\left[\frac{1}{N}L(\theta)\right]\right| \xrightarrow{p} 0.$$

This implies that the convergence to the limit is not much weaker for some values of θ than for others. It requires stronger regularity conditions than pointwise convergence. (Sufficient

but not necessary is that $\ln f_X(x;\theta) \leq k(x)$, with $k(X) < \infty$.) In large samples at the maximum likelihood estimator the derivative of the log likelihood function must be equal to zero:

$$\frac{\partial L}{\partial \theta}(\hat{\theta}) = 0.$$

Now expand the derivative of the log likelihood function around the true value of theta:

$$0 = \frac{\partial L}{\partial \theta}(\hat{\theta}) = \frac{\partial L}{\partial \theta}(\theta^*) + \frac{\partial^2 L}{\partial \theta^2}(\tilde{\theta}) \cdot (\hat{\theta} - \theta^*),$$

for some $\tilde{\theta}$ between θ^* and $\hat{\theta}$. In large samples $\hat{\theta} \to \theta^*$, and therefore $\tilde{\theta} \to \theta^*$. Rearranging this gives

$$\hat{\theta} - \theta = \left[\frac{\partial^2 L}{\partial \theta^2}(\tilde{\theta})\right]^{-1} \cdot \frac{\partial L}{\partial \theta}(\theta^*),$$

or

$$\sqrt{N} \cdot (\hat{\theta} - \theta) = \left[\frac{\partial^2 L}{\partial \theta^2}(\tilde{\theta}) \Big/ N\right]^{-1} \cdot \left[\frac{\partial L}{\partial \theta}(\theta^*) \Big/ \sqrt{N}\right].$$

In large samples

$$-\frac{\partial^2 L}{\partial \theta^2}(\tilde{\theta}) \Big/ N \sim -\frac{1}{N} \sum_{i=1}^{N} \frac{\partial^2 \ln f_X}{\partial \theta^2}(x_i;\tilde{\theta}) \xrightarrow{p} \mathcal{I}(\theta^*),$$

converges in probability to the information matrix $\mathcal{I}(\theta^*)$. The second part,

$$\frac{\partial L}{\partial \theta}(\theta^*) \Big/ \sqrt{N} = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{\partial \ln f_X}{\partial \theta}(x_i;\theta^*) \mathcal{N}(0, \mathcal{I}(\theta^*)^{-1}),$$

because it satisfies a central limit theorem with variance equal to the information matrix. This completes the argument.

Finally, let us consider the case where there is more than one parameter. Let $\theta = (\theta_0, \theta_1)$. The extension of Result 1 to this case is **Result 2**

Let $X_1, \ldots, X_n$ be a random sample from $f_X(x; \theta_0^*, \theta_1^*)$. Let $\hat{\theta}$ be the maximum likelihood estimator:

$$(\hat{\theta}_0, \hat{\theta}_1) = \text{argmax}_{\theta_{0,}\theta_1} \sum_{i=1}^{N} \ln f_X(x_i; \theta_0, \theta_1).$$

Then $(\hat{\theta}_0, \hat{\theta}_1)$ is <u>consistent</u> for (θ_0^*, θ_1^*):

$$\begin{pmatrix} \hat{\theta}_0 \\ \hat{\theta}_1 \end{pmatrix} \xrightarrow{p} \begin{pmatrix} \theta_0^* \\ \theta_1^* \end{pmatrix},$$

and $(\hat{\theta}_0, \hat{\theta}_1)$ has asymptotically a normal distribution:

$$\sqrt{N} \begin{pmatrix} \hat{\theta}_0 - \theta_0^* \\ \hat{\theta}_1 - \theta_1^* \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \mathcal{I}(\theta_0^*, \theta_1^*)^{-1} \right),$$

where $\mathcal{I}(\theta_0^*, \theta_1^*)$ is the single observation information matrix:

$$\mathcal{I}(\theta_0^*, \theta_1^*) = E \begin{bmatrix} \frac{\partial \ln f_X}{\partial \theta_0}(X; \theta_0^*, \theta_1^*) \cdot \frac{\partial \ln f_X}{\partial \theta_0}(X; \theta_0^*, \theta_1^*) & \frac{\partial \ln f_X}{\partial \theta_0}(X; \theta_0^*, \theta_1^*) \cdot \frac{\partial \ln f_X}{\partial \theta_1}(X; \theta_0^*, \theta_1^*) \\ \frac{\partial \ln f_X}{\partial \theta_1}(X; \theta_0^*, \theta_1^*) \cdot \frac{\partial \ln f_X}{\partial \theta_0}(X; \theta_0^*, \theta_1^*) & \frac{\partial \ln f_X}{\partial \theta_1}(X; \theta_0^*, \theta_1^*) \cdot \frac{\partial \ln f_X}{\partial \theta_1}(X; \theta_0^*, \theta_1^*) \end{bmatrix}$$

$$= -E \begin{bmatrix} \frac{\partial^2 \ln f_X}{\partial \theta_0 \partial \theta_0}(X; \theta_0^*, \theta_1^*) & \frac{\partial^2 \ln f_X}{\partial \theta_0 \partial \theta_1}(X; \theta_0^*, \theta_1^*) \\ \frac{\partial^2 \ln f_X}{\partial \theta_1 \partial \theta_0}(X; \theta_0^*, \theta_1^*) & \frac{\partial^2 \ln f_X}{\partial \theta_1 \partial \theta_1}(X; \theta_0^*, \theta_1^*) \end{bmatrix}.$$

□