Stats 135, Fall 2019
Lecture 11, Monday, 9/23/2019

CLASS ANNOUNCEMENTS: Quiz 2 this Friday will focus on §8.5, 8.7.:

- Calculate MLE
- calculate the asymptotic variance of MLE (FI)
- 95% of MLE of parameter θ

There won't be any sufficiency on the quiz, which is something we'll start today.

1 Review

Last time, we established that $\hat{\theta}_{ML}$ has good properties:

- equivariance $\hat{g(\theta)} = g(\hat{\theta})$ (true for MOM if g is continuous)
- consistency $\hat{\theta} \xrightarrow{p} \theta$ (true for MOM)
- asymptotic: unbiased, normal, efficient

Recall that the Cramer Rao inequality says:

2 Mean Square Error (MSE) of an estimator

The MSE is used to measure how good our estimator is.

Definition: $MSE(\hat{\theta})$ **-**

The Mean Square Error of a parameter is:

$$\begin{aligned} MSE(\hat{\theta}) &= \mathbb{E}_\theta\left[(\hat{\theta}-\theta)^2\right] \\ &= \iiint \cdots \int \left[\hat{\theta}(x_1,\ldots,x_n)-\theta\right]^2 f(x_1|\theta)f(x_2|\theta)\cdots f(x_n|\theta)\; dx_1 \cdots\; dx_n \end{aligned}$$

We can think of the MSE as the average distance of $\hat{\theta}$ from θ.

When we calculate this, it'll be a function of θ, which we won't know. The main theorem that is helpful is:

Theorem 2.1.

$$MSE(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) + \left[\mathrm{Bias}(\hat{\theta})\right]^2,$$

where $\mathrm{Bias}(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$.

Proof.

$$MSE(\hat{\theta}) = \mathbb{E}\left[(\hat{\theta} - \mathbb{E}(\hat{\theta}) + \mathbb{E}(\hat{\theta}) - \theta)^2\right],$$

where Adam Lucas jokes that we then follow our algebraic heart and foil this and complete this for homework (to save time). □

The important picture to have in our heads is that we'll make two estimators that have the same MSE. Let one be an unbiased estimator for θ, so that it has high variance and low bias. We can also have the opposite situation: low variance (pointy at the center) but high bias (θ far off-center). Lucas notes there's a trade-off between variance and bias for the same value of MSE.

Example: These two estimators may have the same MSE. Let $X_1, \dots, X_n \overset{iid}{\sim} \text{Bernoulli}(\theta)$, so that:

$$\mathbb{E}(X_1) = \theta, \text{Var}(X_1) = \theta(1-\theta).$$

Now we may ask: between $\tilde{\theta} = \overline{X}$ and $\hat{\theta} = X_1$, which has the smaller MSE? Notice that $\mathbb{E}(\tilde{\theta}) = \mathbb{E}(\overline{X}) = \theta$, and $\mathbb{E}(\hat{\theta}) = \mathbb{E}(X_1) = \theta$, so $\tilde{\theta}, \hat{\theta}$ are both unbiased.
Now we compare their variances:

$$\text{Var}(\tilde{\theta}) = \text{Var}(\overline{X}) = \frac{\theta(1-\theta)}{n}$$
$$\text{Var}(\hat{\theta}) = \text{Var}(X_1) = \theta(1-\theta),$$

so their MSEs are:

$$MSE(\tilde{\theta}) = \text{Var}(\overline{X}) = \frac{\theta(1-\theta)}{n}$$
$$MSE(\hat{\theta}) = \text{Var}(X_1) = \theta(1-\theta),$$

so it implies:

$$MSE(\tilde{\theta}) \leq MSE(\hat{\theta}).$$

Conclusion: Among all unbiased estimators, if MSE is the most important factor to us, then for large sample size, the MLE has the smallest possible MSE because it is efficient (that is, it achieves the Cramer Rao Lower Bound).

Note that the CR inequality only applies for unbiased estimators, and if we allow biased estimators, we might get an even smaller MSE than what we would find from unbiased estimators. That is, if we care to minimize Mean Square Error, we may want to consider biased estimators.

3 §8.8: Sufficiency

Lucas wants to motivate this section on sufficiency. There are two primary motivations for sufficient estimators.
(a) For one, once we collect all of our data, we can form a **sufficient** statistic and then throw away our data (we need not keep the large storage of data). That is, we only need the sufficient statistic to estimate the parameter θ.
(b) We can make an estimator $\hat{\theta}$ better (for example, lower MSE) by taking the conditional expectation $\tilde{\theta}(T)$ given a sufficient statistic T:

$$\tilde{\theta}(T) = \mathbb{E}\left[\hat{\theta}(X_1, \dots, X_n) \mid T\right],$$

called the Rao-Blackwell theorem.
Before we talk about a sufficient statistic, we define a statistic.

Definition: Statistic -

We say that a **statistic** $T(X_1, \ldots, X_n)$ is a function of our data **only**.

For example, a statistic could be the average $\overline{X}$, or the third element X_3, or the minimum $\min(X_1, \ldots, X_n)$, or an ordered statistic $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$. However, for example, $\overline{X} - \mu$ is NOT a statistic, because it involves the parameter μ.

Definition: Sufficient Statistic -

We say that T is a **sufficient** statistic for θ if the conditional distribution of

$$X_1, \ldots, X_n \mid T$$

dependent on T no longer depends on θ.

Lucas notes this is quite abstract. For example, take:

$$f(X_1, \ldots, X_n \mid T = t, \theta) = f(X_1, \ldots, X_n \mid T = t),$$

where essentially that $T = t$ contains all the data, so conditioning on θ is the same as not conditioning on θ. That is, on the RHS, the conditional density does not depend on θ.
First, recall Bayes' Rule, which gives:

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A, B)}{\mathbb{P}(B)},$$

and for densities, Bayes' Rule gives:

$$f(X_1, \ldots, X_n \mid T = t, \theta) = \frac{\overbrace{f(X_1 = x, \ldots, X_n = x_n, T = t}^{} \mid \theta)}{f(T = t \mid \theta)},$$

where Lucas notes that the overbraced portion is truly overkill because T is a function of the $X_1, \ldots, X_n$. In other words, determining all the $X_i = x_i$ makes it so $T = t$ does not matter. Here's a simple example:

Example: Let $X_1, X_2 \overset{iid}{\sim} \text{Bernoulli}(\theta)$. Then

$$\begin{aligned} f(X_1, X_2 \mid \theta) &= \mathbb{P}(X_1 = x_1 \mid \theta) \cdot \mathbb{P}(X_2 = x_2 \mid \theta) \\ &= \underbrace{\theta}_{x_1 + x_2} \underbrace{(1 - \theta)}_{2 - x_1 - x_2}. \end{aligned}$$

Now if $T := X_1 + X_2$, then

$$f(t \mid \theta) = \mathbb{P}(X_1 + X_2 = t \mid \theta) = \sum_{(x_1, x_2): x_1 + x_2 = t} \mathbb{P}(x_1, x_2),$$

Then we can make a table for this example. The question is if $T = X_1 + X_2$ is a sufficient statistic. Then the right-most table should not depend on θ.

(x_1, x_2)	$T = X_1 + X_2$	$f(x_1, x_2 \mid T = t, \theta) = \frac{f(x_1, x_2 \mid \theta)}{f(t \mid \theta)}$
$(0,0)$	$t = 0$	$\frac{f(0,0\mid\theta)}{f(x_1+x_2=0\mid\theta)} = \frac{(1-\theta)(1-\theta)}{(1-\theta)(1-\theta)} = 1$
$(1,0)$	$t = 1$	$\frac{f(1,0\mid\theta)}{f(x_1+x_2=1\mid\theta)} = \frac{\theta(1-\theta)}{\theta(1-\theta)+(1-\theta)\theta} = \frac{1}{2}$
$(0,1)$	$t = 1$	$\frac{f(0,1\mid\theta)}{f(x_1+x_2=1\mid\theta)} = \frac{(1-\theta)\theta}{\theta(1-\theta)+(1-\theta)\theta} = \frac{1}{2}$
$(1,1)$	$t = 2$	1

and because these conditionals do not depend on θ, this is an example of a sufficient statistic.

Example: Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Poisson}(\theta)$. Then recall:

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

and so

$$f(X_1, \ldots, X_n \mid \theta) = \frac{e^{-n\theta}\theta^{\sum_{i=1}^n X_i}}{(X_1!)(X_2!)\cdots(X_n!)}$$

We will show that $T := X_1 + \cdots + X_n$ is a sufficient statistic. Lucas notes that instead of having to worry about all the data we can have, we worry only about the sums they can have. This is a reduction in storage.
Now we recall that the sum of n independent poisson distributions is poisson, so:

$$T \sim \text{Poisson(n}\theta),$$

so

$$f(t|\theta) = \frac{e^{-n\theta}(n\theta)^t}{t!}.$$

Now we write:

$$\begin{aligned} f(X_1, \ldots, X_n \mid T = t, \theta) &= \frac{f(X_1, \ldots, X_n \mid \theta)}{f(t|\theta)} \\ &= \frac{\frac{e^{n\theta}\theta^{\sum X_i}}{(X_1!)\cdots(X_n!)}}{\frac{e^{-n\theta}n^t\theta^t}{\cdots}}, \end{aligned}$$

which proves this definition of T is sufficient.

Lecture ends here.