Stats 135, Fall 2019
Lecture 10, Friday, 9/20/2019

1 Review

Last time, we covered §8.5.2. We found the key result that the MLE has very nice asymptotic properties. That is, for $X_1, \ldots, X_n$ iid with density $f(x \mid \theta)$, we have:

$$\hat{\theta}_{ML} \sim N\left(\theta, \frac{1}{nI(\theta)}\right),$$

for large n, where $I(\theta)$ denotes Fisher information. In other words, the MLE is approximately normal and is unbiased.
Lucas shows some log-likelihood distribution plots, where histograms for more informative $X \sim N(\mu, 1)$ would be more spread out. When the variance is small, it is very easy to see the location of the max (μ). Of course, all these trajectories are different because they are samples from the population.
We look at the slope (derivative) of the log-likelihood, which we call the "score". We look at it at the true value.

Definition: Fisher information -

Fisher information, $I(\theta)$ is the amount of information that a single observation X has to estimate the max of the log likelihood, $l(\theta)$. That is,

$$I(\theta) := \text{Var}(l'(\theta))$$

Topics Today:

- Alternative definition and examples of F.I. (Fisher information)
- Computing the CI using large sample theory and showing that $(1-\alpha)100\%$ CI is

$$\hat{\theta}_{ML} \pm z\left(\frac{\alpha}{2}\right)\sqrt{\frac{1}{nI(\theta)}}$$

- §8.7, Cramer Rao (CR) inequality

2 Definition of Fisher Information

We define the Fisher Information to be the variance of the score at the true parameter value.

$$I(\theta_0) = \text{Var}\left(\frac{\partial}{\partial\theta}\log f(x \mid \theta)\big|_{\theta=\theta_0}\right)$$

Now because the expectation of the score function at θ_0 is zero, we can write the Fisher information as:

$$I(\theta_0) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log\left(f(x \mid \theta)\right)\big|_{\theta=\theta_0}\right)^2\right]$$

because $\text{Var}(A) = \mathbb{E}(A^2) - (\mathbb{E}(A))^2$, and if $\log f(x \mid \theta)$ is twice-differentiable (see Rice, lemma A on page 276), then

$$I(\theta_0) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log\left(f(x \mid \theta)\right)\big|_{\theta=\theta_0}\right]^2$$

That is, the Fisher information is the average curvature of the log-likelihood over all values of $x \in X$ (all our sample trajectories). Lucas notes that basically, the Fisher information has two equivalent interpretations:

1. Variance of score function at the true parameter θ_0.
2. negative of the average curvature of the log-likelihood θ_0 over all $x \in X$.

Lucas says that for our purposes, we will refer ('go-to') the second definition.

3 Cramér–Rao Inequality (Cramér–Rao Lower Bound, CRLB)

Let $X_1, \ldots, X_n$ be iid random variables with density $f(x \mid \theta)$. Now two things are true.
(1) The MLE is asymptotically normal, unbiased, with variance:

$$\hat{\theta}_{ML} \approx N\left(\theta, \frac{1}{nI(\theta)}\right)$$

for large n.
(2) The Cramer Rao inequality holds. Let the ML estimator, $\hat{\theta}_{ML}$, be unbiased. Then

$$\text{Var}\left(\hat{\theta}_{ML}\right) \geq \frac{1}{nI(\theta)}.$$

We won't prove this in this class (Lucas jokes this is deferred for our next graduate class in Statistics).

Definition: Efficient -

An unbiased estimator where variance achieves the CR lower bound is called **efficient.**

Now from point (1) above, we see that the ML estimator $\hat{\theta}_{ML}$ is asymptotically efficient.

Example: Take $X_1, \ldots, X_n$ iid with Binomial(1, p), where $0 < p < 1$. We wish to find $I(p) := -\mathbb{E}[l''(p)]$.
Recall that

$$f(x \mid p) = p^x(1-p)^{1-x},$$

so taking logs of both sides yields:

$$l(p) = x\log p + (1-x)\log(1-p),$$

and now taking successive derivatives gives:

$$l'(p) = \frac{x}{p} - \frac{1-x}{1-p}$$
$$l''(p) = \frac{-x}{p^2} - \frac{1-x}{(1-p)^2}$$

To continue, we need to take the expectation:

$$\begin{aligned}\mathbb{E}[l''(p)] &= \frac{-\mathbb{E}(x)}{p^2} - \frac{(1-\mathbb{E}(x))}{(1-p)^2}, \text{ where } \mathbb{E}(x) = p \\ &= \frac{-p}{p^2} - \frac{(1-p)}{(1-p)^2} \\ &= -\frac{1}{p} - \frac{1}{1-p} \\ &= \boxed{\frac{-1}{p(1-p)}}\end{aligned}$$

Hence:

$$\boxed{I(p) = -\mathbb{E}(l''(p)) = \frac{1}{p(1-p)}}$$

Now because we want to find the SE of our ML estimator of p, $\hat{p}_{ML}$, we need to find $\frac{1}{nI(p)}$. We assume that n is large and use the asymptotic property. We have:

$$nI(p) = \frac{n}{p(1-p)} \implies \frac{1}{nI(p)} = \boxed{\frac{p(1-p)}{n}},$$

and hence

$$\text{SE of } (\hat{p}_{ML}) = \sqrt{\frac{1}{nI(p)}} = \sqrt{\frac{p(1-p)}{n}}$$

Lucas notes that this used the asymptotic property. Now we try to get this sort of result directly without this assumption that n is large and compare. The Cramer-Rao inequality states that what we found above is a lower bound of what we will directly compute. Hence what we will find must be greater than or equal to our above expression.

3.1 Direct Computation:

To find $\hat{p}_{ML}$ directly, again consider $X_1, \ldots, X_n$ iid Binomial(1, p).
In other words,

$$\text{lik}(p) = f(X_1, \ldots, X_n \mid p) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i},$$

and

$$l(p) = \log \text{lik}(p) = \cdots$$

Lucas skips the details and mentions that it's not surprising that we simply get:

$$\hat{p}_{ML} = \overline{X},$$

which is generally reasonable and Lucas skips proof (or provides supplementally in notes).
Now we want to find SE of $(\hat{p}_{ML})$. Take:

$$\text{Var}(\overline{X}) = \frac{p(1-p)}{n},$$

which implies

$$\text{SE of } (\hat{p}_{ML}) = \sqrt{\frac{p(1-p)}{n}},$$

which is precisely the Cramer-Rao lower bound. We showed the same exact expression as above with the assumption (MLE property).
Hence the estimator of $\hat{P}_{ML}$ is approximately:

$$\text{SE of } (\hat{p}_{ML}) = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{\overline{X}(1-\overline{X})}{n}}.$$

4 Confidence Intervals

Now the approximate $(1-\alpha)100\%$ confidence interval for p is

$$\begin{aligned}
&\hat{p}_{ML} \pm z\left(\frac{\alpha}{2}\right)\sqrt{\frac{1}{nI(p)}} \\
&\approx \overline{X} \pm z\left(\frac{\alpha}{2}\right)\sqrt{\frac{\overline{X}(1-\overline{X})}{n}}
\end{aligned}$$

A question arises in the audience as to when we would use this estimator or the bootstrap method.
Bootstrapping is to use our sample as opposed to our population. Here, we consider that we are using the equivariance property to use $\overline{X}$ as an estimator for p.

5 Information in a Random Sample

Lucas skips through this very quickly but provides the algebraic derivations later.
The Fisher Information for the entire n sample is

$$I_n(\theta) = nI(\theta),$$

which can be shown very simply by taking the variance of a sum of iid variables.
Hence the FI in an iid random sample is simply n times the FI in a single observation. So for large n,

$$\hat{\theta} \sim N\left(\theta, \frac{1}{I_n(\theta)}\right),$$

where $I_n(\theta) = nI(\theta)$.

6 Example

Let $X_1, \ldots, X_n$ be a random iid sample with density $f(x \mid \theta) = \theta x^{\theta-1}$, for $0 < x < 1$, and $\theta > 0$.
(a) Find $\hat{\theta}_{ML}$.
(b) Find $I_n(\theta)$.
(c) What distribution is $\hat{\theta}_{ML}$ approaching as $n \to \infty$?
(d) Find an approximate $(1 - \alpha)100\%$ confidence interval for θ when n is large.