EE 503

Lecture Notes Part 14

Christopher Wayne Walker, Ph.D.

14.0 Statistics: Maximum Likelihood Estimator and the Cramer-Rao Lower Bound

The maximum likelihood estimator is by far the most popular estimator. In this approach we choose the value of the parameter θ that maximizes the likelihood function as given below.

The *likelihood function* is defined as

$$L(\theta|x) = L(\theta_1, \ldots, \theta_k) = \prod_{i=1}^{n} f(x_i|\theta_1, \ldots, \theta_k).$$

Here, $x = \underline{x} = (x_1, \ldots, x_n)$.

Example: Suppose $X = (X_1, \ldots, X_n)$ where each X_i is Bernoulli (0,1) with parameter p, with p unknown. Then

$$L(p|x) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^y(1-p)^{n-y}, \quad \text{where } y = \sum_{i=1}^{n} x_i.$$

We compute

$$\log L(p|x) = y \log p + (n-y)\log(1-p).$$

We thus solve

$$\frac{d}{dp} = \frac{y}{p} + \frac{y-n}{1-p} = 0$$

to get

$$\hat{p} = \frac{y}{n}.$$

Example: Suppose $X = (X_1, \ldots, X_n)$ where each X_i is uniform $(0, \theta)$, with θ unknown. Then

$$f(x) = \begin{cases} \theta^{-n}, & 0 < x_{(1)} < x_{(n)} < \theta, \\ 0, & \text{elsewhere.} \end{cases}$$

We see that $f(x)$ increases as θ decreases so to maximize $f(x)$ we make $\hat{\theta}$ as small as possible. Hence,

$$\hat{\theta} = X_{(n)},$$

that is, we make our estimate of θ to be the largest observed value of the data. Note that we do this even though $X_{(n)}$ can never achieve the true value of θ given our sample space. But for any $\epsilon > 0$ if we let $\hat{\theta} = X_{(n)} + \epsilon$ then it is always possible that $\theta = X_{(n)} + \epsilon/2$ and hence we did not choose the smallest possible $\hat{\theta}$.

Example: Suppose $X = (X_1, \ldots, X_n)$ where each X_i is uniform $(a, a+\theta)$, with both a and θ unknown. Then

$$f(x) = \begin{cases} \theta^{-n}, & a < x_{(1)} < x_{(n)} < a+\theta, \\ 0, & \text{elsewhere.} \end{cases}$$

We see that $f(x)$ increases as θ decreases so to maximize $f(x)$ we make $\hat{\theta}$ as small as possible. To do this note

$$X_{(n)} - X_{(1)} < (a+\theta) - a = \theta.$$

But we also require $a < X_{(1)}$. So to make $\hat{\theta}$ as small as possible we must make a as large as possible since we need $X_{(n)} < a + \theta$. We see if we make a smaller than needed we need to make θ larger in order to satisfy this last inequality. Hence, we choose

$$\begin{aligned} \hat{a} &= X_{(1)} \\ \hat{\theta} &= X_{(n)} - X_{(1)}. \end{aligned}$$

Example: Suppose $X = (X_1, \ldots, X_n)$ where each X_i is uniform $(\theta, \theta+1)$, with θ unknown. Then

$$f(x) = \begin{cases} 1, & \theta < x_{(1)} < x_{(n)} < \theta+1, \\ 0, & \text{elsewhere.} \end{cases}$$

We observe

$$X_{(n)} - 1 < \theta < X_{(1)}.$$

So we choose

$$\hat{\theta} \in \left(X_{(n)} - 1, X_{(1)}\right).$$

We now provide the Cramer-Rao lower bound (CRLB) for any variance of any estimator(not just the MLE).

Theorem (Cramer-Rao Inequality). Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a sample with pdf $f(\mathbf{x}|\theta)$ and let $W(\mathbf{X}) = W(X_1, \ldots, X_n)$ be any estimator satisfying

$$\begin{aligned} \frac{d}{d\theta} E_\theta W(\mathbf{X}) &= \int \frac{\partial}{\partial\theta} \left[W(\mathbf{x} f(\mathbf{x}|\theta) \right] d\mathbf{x} \\ \text{and} \quad Var_\theta W(\mathbf{X}) &< \infty. \end{aligned}$$

Then

$$Var_\theta W(\mathbf{X}) \geq \frac{\left(\frac{d}{d\theta} E_\theta W(\mathbf{X}) \right)^2}{E_\theta \left(\left(\frac{\partial}{\partial\theta} \log f(\mathbf{X}|\theta) \right)^2 \right)}.$$

Proof: To be supplied.

Corollary. If $X_1, \ldots, X_n$ are i.i.d. then

$$Var_\theta W(\mathbf{X}) \geq \frac{\left(\frac{d}{d\theta} E_\theta W(\mathbf{X}) \right)^2}{n E_\theta \left(\left(\frac{\partial}{\partial\theta} \log f(X|\theta) \right)^2 \right)}.$$

If $f(x|\theta)$ satisfies

$$\frac{d}{d\theta} E_\theta \left(\frac{\partial}{\partial\theta} \log f(X|\theta) \right) = \int \frac{\partial}{\partial\theta} \left[\left(\frac{\partial}{\partial\theta} \log f(X|\theta) \right) f(x|\theta) \right] dx$$

then

$$E_\theta \left(\left(\frac{\partial}{\partial\theta} \log f(X|\theta) \right)^2 \right) = -E_\theta \left(\frac{\partial^2}{\partial\theta^2} \log f(X|\theta) \right).$$

This last result holds for so called exponential families of distribution. If this latter case holds and we also have that $W(\mathbf{X})$ is unbiased for θ then for i.i.d. we have

$$Var_\theta W(\mathbf{X}) \geq \frac{1}{-n E_\theta \left(\frac{\partial^2}{\partial\theta^2} \log f(X|\theta) \right)}.$$

Example: Consider $\mathbf{X} = (X_1, \ldots, X_n)$, an i.i.d. sample where each X_i is from the normal distribution with mean μ and variance σ^2. In this case the CRLB for $\theta = \mu$ is found using (note the normal is a member of the exponential family)

$$
\begin{aligned}
Var_\mu W(\mathbf{X}) &\geq \frac{1}{-nE_\mu\left(\frac{\partial^2}{\partial\mu^2}\log f(X|\mu)\right)} \\
&= \frac{\sigma^2}{n}.
\end{aligned}
$$

Suppose we estimate μ using

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Then

$$Var\left(\bar{X}\right) = \frac{\sigma^2}{n}$$

and we wee that the CRLB is actually achieved so this estimator is the best for μ in terms of minimizing variance.