208711 Statistical Theory 1 **Fall 2019**

Lecture 21: October 29

Lecturer: Donlapark Pornnopparath

We recall Lehmann-Scheffé theorem:

If T is a complete sufficient statistic and $\delta(X)$ is unbiased, then $\mathbb{E}[\delta(X)|T]$ is UMVU.

If we can take $X = T$ i.e. an unbiased statistic based on T exists, then $\mathbb{E}[\delta(T)|T] = \delta(T)$ and we have the following corollary:

Corollary 21.1. If T is a complete sufficient statistic and there exists a function δ so that $\delta(T)$ is an unbiased estimator for $g(\theta)$, then $\delta(T)$ is UMVU.

From Lehmann-Scheffé and its corollary, we can come up with a couple of ways to find a UMVU estimator.

1. Rao-Blackwellization: Find an unbiased estimator $\delta(X)$ and a complete sufficient statistic T and then compute $\mathbb{E}[\delta(X)|T]$.
2. Solve for, or find $\delta(T)$ such that $\mathbb{E}[\delta(T)] = g(\theta)$. Usually, we start with $\delta(T)$ as a constant multiple of T.

We have showed how to apply the first method in the previous lectures. Here, we give an example of the second method.

Example 21.2. Let $X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2)$ be iid. Note that $(\overline{X}, S^2)$ is a complete sufficient statistic. From this, we can find an UMVU estimator for

1. μ: Since $\delta(\overline{X}, S^2) = \overline{X}$ is an unbiased estimator of μ, so it is a UMVU for μ.
2. σ^2: The same as above, using $\delta(\overline{X}, S^2) = \frac{S^2}{n-1}$.
3. σ: Since we can write
$$S^2 = \sigma^2 \left(Z_1^2 + Z_2^2 + \ldots + Z_n^2\right) = \sigma^2 \chi^2_{n-1}$$
where $Z_1, Z_2, \ldots, Z_n \sim N(0,1)$ are iid, so we can take the square root:
$$S = \sigma\sqrt{\chi^2_{n-1}}$$
$$\mathbb{E}[S] = \sigma\mathbb{E}\left[\sqrt{\chi^2_{n-1}}\right]$$
$$\mathbb{E}\left[\frac{S}{\mathbb{E}\left[\sqrt{\chi^2_{n-1}}\right]}\right] = \sigma.$$
Thus, $\delta(S^2) = S/\mathbb{E}\left[\sqrt{\chi^2_{n-1}}\right]$ is a UMVU for σ.

◇

21.1 Maximum Likelihood Estimator

Given a sample $X_1, X_2, \dots, X_n$ with distribution given by a pdf $f(x|\theta)$. The likelihood function is defined by

$$\mathcal{L}(\theta|\boldsymbol{x}) = f(x_1, x_2, \dots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta).$$

In context of random sampling, we try to draw samples from a distrisbution with known or unknown parameter. Now we try to "reverse" the process i.e. inferring an unknown parameter θ from the observations, using the principle that the value of θ is the one that the observed sample is most likely to occur.

Definition 21.3. A maximum likelihood estimator (MLE) is given by

$$\hat{\theta} = \arg\max_{\theta\in\Theta} \mathcal{L}(\theta|\boldsymbol{x}).$$

Given that $\mathcal{L}(\theta|\boldsymbol{x})$ is second-order differentiable, a *local* extremum can be obtained by solving the system of equations

$$\frac{\partial}{\partial\theta_i}\mathcal{L}(\theta|\boldsymbol{x}) = 0 \qquad i = 0, 1, 2, \dots, k.$$

We can use the Hessian $H(\theta)$ test to check if the solution is the minimum or maximum. In one dimensional case, it is a maximum if $\mathcal{L}''(\theta|\boldsymbol{x}) < 0$. However, the test fails if $\det(H(\theta)) = 0$, and it does not take into account the values at the boundaries.

In most cases, it is easier to find the maximum of $\log \mathcal{L}(\theta|\boldsymbol{x})$ since it is easier to work with a summation instead of a product.

Example 21.4. Let $X_1, X_2, \dots, X_n \sim \text{Bernoulli}(p)$. The likelihood function is

$$\mathcal{L}(p|\boldsymbol{x}) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^y(1-p)^{n-y},$$

◇

where $y = \sum_i x_i$. Taking the logarithm, we obtain

$$\log \mathcal{L}(p|\boldsymbol{x}) = y \log p + (n-y)\log(1-p) \tag{21.1}$$

Thus, by setting its derivative to zero, we have

$$\frac{y}{\hat{p}} = \frac{n-y}{1-\hat{p}}$$
$$y\left(\frac{1}{\hat{p}} - 1\right) = n - y$$
$$\hat{p} = \frac{y}{n}.$$

To check if this is the local minimum, we can use the fact that (21.1) is a sum of two convex function, or that $\partial/\partial p \log \mathcal{L}(p|\boldsymbol{x}) = -y/p^2 - (n-y)/(1-p)^2 < 0$. Thus the MLE is $\hat{p} = y/n = \sum_i X_i/n$.

Example 21.5. Let $X_1, X_2, \dots, X_n \sim N(\mu, \sigma^2)$. The likelihood function is

$$\mathcal{L}(\mu, \sigma^2|\boldsymbol{x}) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\sum_i (x_i-\mu)^2/(2\sigma^2)}.$$

Thus the log-likelihood function is

$$\log \mathcal{L}(\mu, \sigma^2|\boldsymbol{x}) = -\frac{1}{2\sigma^2}\sum_i (x_i - \mu)^2 - n\log\sigma - \frac{n}{2}\log 2\pi.$$

By setting its derivative with respect to μ and σ^2 equal to zero, we obtain

$$\frac{1}{\sigma^2}\sum_i (x_i - \hat{\mu}) = 0$$
$$\frac{1}{\hat{\sigma}^3}\sum (x_i - \mu)^2 - \frac{n}{\hat{\sigma}} = 0,$$

which yields $\hat{\mu} = \bar{x}$ and $\hat{\sigma}^2 = n^{-1}\sum_i (x - \bar{x})^2$. ◇

The previous example shows that not any MLE is unbiased, since $\mathbb{E}[\hat{\sigma}^2] \neq \sigma^2$. Another property that separate between MLE and unbiased estimators is the *equivariant property* under any function: Let $F : \Theta \to \Theta$ be a function, which might not be one-to-one. For any $\eta \in F(\Theta)$, we want to pick $\hat{\theta}$ from $F^{-1}(\eta)$ so that we can estimate η using $\eta = F(\hat{\theta})$. An appropriate choice is the one among $\hat{\theta}$ that maximizes the likelihood function. Formally, we define

$$\mathcal{L}_F(\eta|\boldsymbol{x}) = \sup_{\theta \in F^{-1}(\eta)} \mathcal{L}(\theta|\boldsymbol{x}).,$$

and we say that the maximum to $\mathcal{L}_F(\theta|\boldsymbol{x})$ is the MLE for $\eta = F(\theta)$.

Theorem 21.6. If $\hat{\theta}$ is MLE for θ, then for any function $F : \Theta \to \Theta$, $F(\hat{\theta})$ is MLE for $F(\theta)$.

Proof: Let $\hat{\theta}$ be the MLE of θ and $\hat{\eta} = F(\hat{\theta})$. Then, for any other $\eta_1 \in F(\Theta)$ and any $\theta_1 \in F^{-1}(\eta_1)$, it follows that

$$\begin{aligned}\sup_{\theta \in F^{-1}(\hat{\eta})} \mathcal{L}(\theta|\boldsymbol{x}) &= \mathcal{L}(\hat{\theta}|\boldsymbol{x}) \\ &\geq \mathcal{L}(\theta_1|\boldsymbol{x}).\end{aligned}$$

Since this is true for any θ_1 in the preimage of η_1,

$$\mathcal{L}_F(\hat{\eta}|\boldsymbol{x}) = \sup_{\theta \in F^{-1}(\eta)} \mathcal{L}(\theta|\boldsymbol{x}) \geq \sup_{\theta_1 \in F^{-1}(\eta_1)} \mathcal{L}(\theta_1|\boldsymbol{x})$$

for any $\eta_1 \in F(\Theta)$. Therefore, $\hat{\eta} = F(\hat{\theta})$ is a MLE for $F(\theta)$. ■

Example 21.7. From Example 21.5, $\bar{x}$ is the MLE estimate of μ, and so $(\bar{x})^2$ is the MLE estimate for μ^2. However, it is a biased estimator since $\mathbb{E}[(\bar{x})^2] = \sigma^2/n + \mu^2$. ◇