208711 Statistical Theory 1 **Fall 2019**

Lecture 20: October 25

Lecturer: Donlapark Pornnopparath

20.1 Decision Theory

Basic setup
A sample: $\boldsymbol{X} \sim \mathcal{P} = \{P_\theta : \theta \in \Theta\}$
Goal: find an *estimator* i.e. a statistic δ such that $\delta(X)$ is close to $g(\theta)$
Loss function: $L(\theta, \delta(X)) \geq 0$ measures the closeness between $\delta(X)$ and $g(\theta)$.
Since $L(\theta, \delta(X))$ is random and be large/small if we are unlucky/lucky, the quantity that we should be interested in is the *risk function*, which is the mean loss:

$$R(\theta, \delta) = \mathbb{E}_\theta[L(\theta, \delta(X))].$$

Example 20.1. Consider 100 independent coin tosses with chance of being head θ and let X be the number of heads. Then $X \sim \text{Bin}(100, \theta)$ and a natural estimator is $\delta(X) = X/100$. Define a loss function by

$$L(\theta, \delta(X)) = (\theta - \delta(X))^2.$$

Then the risk function is

$$R(\theta, \delta) = \mathbb{E}_\theta[(\theta - X/100)^2] = \frac{1}{100^2}\text{Var}[X] = \frac{\theta(1-\theta)}{100}.$$

$\diamond$

We now relate the task of finding a suitable estimator to sufficient statistics. It turns out that the connection can be made by imposing an extra condition on the loss function.

Definition 20.2. A set $\mathcal{C}$ is convex if any line connecting any two points $x, y \in \mathcal{C}$ is contained in $\mathcal{C}$. That is, for any $t \in (0,1)$, $tx + (1-t)y \in \mathcal{C}$.

Definition 20.3. Let $\mathcal{C} \subseteq \mathbb{R}^n$ be a convex set. A function $f : \mathcal{C} \to \mathbb{R}$ is convex if for any $x, y \in \mathcal{C}$ and any $t \in (0,1)$,

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y). \tag{20.1}$$

We say that f is strictly convex if the inequality is strict.

Thus a function f is convex if for any $x, y \in \mathcal{C}$, the graph of f between x and y lies below the segment joining $(x, f(x))$ and $(y, f(y))$. In the case $p = 1$, we can test the convexity of any twice-differentiable function f by checking if f'' is nonnegative everywhere on the interior of $\mathcal{C}$.

Example 20.4. $L(\theta, d) = (\theta - d)^2$ is convex in d since $L''(\theta, d) = 2 > 0$. $\diamond$

The following theorem says that the weighted average in (20.1) can be replaced by the expected value.

> **Theorem 20.5** (Jensen's inequality)**.** Let X be a random variable and $f : \mathcal{C} \to \mathbb{R}$ be a convex function and $E[X] < \infty$, then
> $$f(\mathbb{E}X) \leq \mathbb{E}[f(X)]. \tag{20.2}$$
> If f is strictly convex, then the inequality is strict unless X is almost surely constant.

Proof: Fix $t \in \mathcal{C}$. There exists a constant c such that

$$f(t) + c(x - c) \leq f(x)$$

for all $x \in \mathcal{C}$ (i.e. c is a subgradient of f at t). By taking $t = E[X]$ and $x = X$ we have that

$$f(E[X]) + c(X - E[X]) \leq f(X),$$

which yields (20.2) by applying the expected value on both side of the inequality. ■ Everything is tied up together in the following theorem, which tells us that any estimator can be upgraded by conditioning on a sufficient statistic.

> **Theorem 20.6** (Rao–Blackwell)**.** Let $\boldsymbol{X}$ be a random sample from a distribution with pdf $f(x|\theta)$ and T be a sufficient statistic for θ. Let δ be an estimator of $g(\theta)$ and $L_\theta : d \mapsto L(\theta, d)$ be a convex loss function. Then by defining $\eta(T) = \mathbb{E}[\delta(X)|T]$, we have
> $$R(\theta, \eta) \leq R(\theta, \delta), \tag{20.3}$$
> and the inequality is strict if L_θ is strictly convex, unless $\delta(X) = \eta(T)$ with probability one.

Proof: By conditioning X on T, Jensen's inequality yields

$$\begin{aligned} L_\theta(\mathbb{E}[\delta(X)|T]) &\leq \mathbb{E}[L_\theta(\delta(X)|T] \\ L_\theta(\eta(T)) &\leq \mathbb{E}[L_\theta(\delta(X))|T]. \end{aligned}$$

Taking the expected value on both size, we obtain

$$\begin{aligned} \mathbb{E}_\theta[L_\theta(\eta(T))] &\leq \mathbb{E}_\theta[\mathbb{E}[L_\theta(\delta(X)|T]] \\ R(\theta, \eta) &\leq R(\theta, \delta), \end{aligned}$$

as desired. The statement about strict inequality follows from that of the Jensen's inequality. ■

Example 20.7. Let $\boldsymbol{X} \sim \text{Pois}(\lambda)$. We know that $T = X_1 + X_2 + \ldots + X_n$ is a sufficient statistic. Consider an estimator $\delta(\boldsymbol{X}) = X_1$. If we have observed that $\sum_i x_i = t$, then

$$\eta(T) = \mathbb{E}[X_1|T] = \mathbb{E}\left[X_1|\sum_i X_i = t\right] = \mathbb{E}\left[X_j|\sum_i X_i = t\right]$$

for any other j. Therefore,

$$n\mathbb{E}\left[X_i|\sum_i X_i = t\right] = \sum_j \mathbb{E}\left[X_j|\sum_i X_i = t\right] = \mathbb{E}\left[\sum_j X_j|\sum_i X_i = t\right] = t.$$

Thus the *Rao–Blackwellization* of X_1 is $(X_1 + X_2 + \ldots + X_n)/n$. ◇

20.2 Minimum variance unbiased estimators

Definition 20.8. An estimator δ is unbiased for $g(\theta)$ if

$$\mathbb{E}_\theta[\delta(X)] = g(\theta) \qquad \text{for all } \theta \in \Theta.$$

If an estimator exists, g is called *U-estimable.*

If δ is an unbiased estimator and the loss function is $L(\theta, \delta(X)) = (\theta - \delta(X))^2$, then the risk of δ is

$$R(\theta, \delta) = \mathbb{E}[(\delta(X) - \theta)^2] = \text{Var}[\delta(X)].$$

There can be many unbiased estimators. For example, $\boldsymbol{X} \sim N(\mu, \sigma^2)$ has both X_1 and $\overline{X}$ as unbiased estimators of μ. It is natural that the one that we pick should minimize the variance.

Definition 20.9. An unbiased estimator δ is uniformly minimum variance unbiased (UMVU) if for any other unbiased estimator δ^*,

$$\text{Var}_\theta(\delta) \leq \text{Var}_\theta(\delta^*) \quad \text{for all } \theta.$$

Note that UMVU does not need to exist. However, given that g is U-estimable and the family of distributions has a complete sufficient statistic, a UMVU exists.

Theorem 20.10 (Lehmann–Scheffé)**.** Let T be a complete sufficient statistic from a family of distribution with pdf $f(x|\theta)$. Then any U-estimable function g accepts an essentially unique (i.e. unique up to a set of probability zero) unbiased estimator based on T that is UMVU.

Proof: Let $\delta(X)$ be an unbiased estimator of $g(\theta)$ and define

$$\eta(T) = \mathbb{E}[\delta(X)|T]$$

By smoothing, we see that $\eta(T)$ is also unbiased:

$$\mathbb{E}[\eta(T)] = \mathbb{E}[\mathbb{E}[\delta(X)|T]] = \mathbb{E}[\delta(X)] = g(\theta).$$

For any other unbiased estimator $\eta^*(T)$, we have

$$\mathbb{E}[\eta(T) - \eta^*(T)] = 0,$$

so by Rao–Blackwell theorem, $\eta(T) = \eta^*(T)$ a.e.
To show the variance minimality, let δ^* be another unbiased estimator. Then $\eta^*(T) = \mathbb{E}[\delta^*(X)|T]$ is an unbiased estimator. By Rao–Blackwell theorem with squared error loss as the risk function,

$$\text{Var}[\eta(T)] = \text{Var}[\eta^*(T)] \leq \text{Var}[\delta^*] \quad \text{for all } \theta \in \Theta.$$

Therefore, $\eta(T)$ has minimum variance. ■

Example 20.11. Let $X_1, X_2, \ldots, X_n$ be iid variables from Bernoulli(p). We have already showed in previous lectures that

$$T(\boldsymbol{X}) = X_1 + X_2 + \ldots + X_n \sim \text{Bin}(p, n)$$

is a complete sufficient statistic for p. Suppose that we want to find an unbiased estimator of $g(p) = p^2$.

One possible unbiased choice is $\delta(X) = X_1 X_2$. The UMVU estimator is

$$\begin{aligned}
\mathbb{E}[X_1 X_2 | T = t] = \mathbb{P}(X_1 = 1, X_2 = 1 | T = t) &= \frac{\mathbb{P}(X_1 = 1, X_2 = 1, T = t)}{\mathbb{P}(T = t)} \\
&= \frac{\mathbb{P}(X_1 = 1)\mathbb{P}(X_2 = 1)\mathbb{P}\big(\sum_{i=3}^n X_i = n - 2\big)}{\mathbb{P}(T = t)} \\
&= \frac{p^2 \binom{n-2}{t-2} p^{t-2} (1-p)^{n-t}}{\binom{n}{t} p^t (1-p)^{n-t}} \\
&= \frac{t(t-1)}{n(n-1)}.
\end{aligned}$$

Thus $T(T-1)/(n^2 - n)$ is the UMVU estimator of p^2. $\diamond$