LECTURE 5A NOTES

1. Optimal unbiased estimators. We turn our attention to finding optimal unbiased estimators; i.e. unbiased estimators $\delta(\mathbf{x})$ such that

$$\mathrm{Risk}_\delta(\theta) \le \mathrm{Risk}_{\delta'}(\theta) \text{ at all } \theta \in \Theta$$

for any other unbiased estimator $\delta'(\mathbf{x})$. We call such estimators *uniformly minimum risk unbiased* (UMRU) estimators. By the Rao-Blackwell theorem, as long as the loss function is convex (in $\delta(\mathbf{x})$), we only need to consider estimators that are functions of sufficient statistics. In the rest of this section, we assume the loss function is convex.

Definition 1.1. *A statistic* $\mathbf{t} := \phi(\mathbf{x})$ *is* complete *for a parametric model* $\mathcal{F}$ *if* $\mathbf{E}_\theta[g(\mathbf{t})] = 0$ *for all* $\theta \in \Theta$ *implies* $g(\mathbf{t}) \overset{a.s}{=} 0$ *for all* $\theta \in \Theta$.

By considering $\mathbf{E}_\theta[g(\mathbf{t})] = 0$ as an inner product

$$\langle g(\phi(x)) \cdot f_\theta(x)\rangle = \langle g(t) \cdot f'_\theta(t)\rangle,$$

where $f'_\theta(t)$ is the density of $\mathbf{t}$, we see that the completeness of $\mathbf{t}$ for $\mathcal{F}$ is akin to $g(\phi(x))$ being in the subspace spanned by the densities in $\mathcal{F}$. Thus completeness is a property of a set of densities, and some sources define complete parametric models.

Example 1.2. *Returning to the coin tossing example, let's show that* $\mathbf{t} = \sum_{i\in[n]} \mathbf{x}_i$ *is a complete statistic. Let* g *be a function such that* $\mathbf{E}_p\big[g(\mathbf{t})\big] = 0$ *for any* $p \in (0,1)$*. We expand* $\mathbf{E}_p\big[g(\mathbf{t})\big]$ *to obtain*

$$\begin{aligned}\mathbf{E}_p\big[g(\mathbf{t})\big] &= \sum_{t=0}^{n} g(t)\binom{n}{t}p^t(1-p)^{n-t} \\ &= (1-p)^n \sum_{t=0}^{n} g(t)\binom{n}{t}\left(\frac{p}{1-p}\right)^t.\end{aligned}$$

Since $1-p$ *is non-zero for any* $p \in (0,1)$*, we must have*

$$\sum_{t=0}^{n} g(t)\binom{n}{t}\left(\frac{p}{1-p}\right)^t = 0.$$

For a polynomial to vanish at all $\frac{p}{1-p}$*, the coefficients must be all zero. Since* $\binom{n}{t}$ *is non-zero, we must have* $g(t) = 0$ *for all possible* t*.*

The term completeness is slightly misleading. It suggests a complete statistic is "comprehensive", when it is in fact the opposite. The constant statistic $\mathbf{t} = 1$ is complete:

$$0 = \mathbf{E}_\theta\big[g(\mathbf{t})\big] = g(1)$$

implies $g(\mathbf{t}) \overset{a.s}{=} 0$. The intuition that complete statistics are in fact "incomplete" is confirmed by Bahadur's theorem, which shows a connection between completeness and optimal data reduction.

Theorem 1.3 (Bahadur's theorem). *If $\mathbf{t}$ is a complete and sufficient statistic for a parametric model, then it is a minimal sufficient statistic for the parametric model.*

Proof. Let $\mathbf{t}'$ be a minimal sufficient statistic for $\mathcal{F}$. To show $\mathbf{t}$ is also a minimal sufficient statistic, it suffices to show $\mathbf{t} = g(\mathbf{t}')$ for some function g. Consider $g(\mathbf{t}') = \mathbf{E}_\theta\big[\mathbf{t} \mid \mathbf{t}'\big]$. By the tower property,

$$\begin{aligned}\mathbf{E}_\theta\big[g(\mathbf{t}')\big] &= \mathbf{E}_\theta\big[\mathbf{E}_\theta\big[g(\mathbf{t}') \mid \mathbf{t}\big]\big] = \mathbf{E}_\theta\big[\mathbf{E}_\theta\big[\mathbf{E}_\theta\big[\mathbf{t} \mid \mathbf{t}'\big] \mid \mathbf{t}\big]\big] \\ &= \mathbf{E}_\theta\big[\mathbf{E}_\theta\big[\mathbf{t} \mid \mathbf{t}\big]\big] = \mathbf{E}_\theta\big[\mathbf{t}\big]\end{aligned} \tag{1.1}$$

Since $\mathbf{t}$ is, by assumption, a sufficient statistic, $\mathbf{t}' = h(\mathbf{t})$ for some function g. Thus $\mathbf{E}_\theta\big[g(\mathbf{t}')\big]$ is a function of $\mathbf{t}$. By the completeness of $\mathbf{t}$, (1.1) implies $g(\mathbf{t}') \overset{a.s}{=} \mathbf{t}$. □

A classical result by Erich Lehmann and Henry Scheffé shows that complete and sufficient statistics begets UMRU estimators.

Theorem 1.4 (Lehmann-Scheffé). *If $\mathbf{t}$ is a complete and sufficient statistic and $\mathbf{E}_\theta\big[\delta(\mathbf{t})\big]$ is an unbiased estimator of θ, then $\delta(\mathbf{t})$ is*

1. *the unique function of θ that is an unbiased estimator of $\mathbf{t}$.*
2. *a UMRU estimator under any convex loss function.*
3. *the unique UMRU under any strictly convex loss function.*

Proof. The estimator $\delta(\mathbf{x})$ is unique by the completeness of $\mathbf{t}$. If there is another unbiased estimator $\delta'(\mathbf{x})$ that is also a function of $\mathbf{t}$, then

$$\mathbf{E}_\theta\big[\delta(\mathbf{t}) - \delta'(\mathbf{t})\big] = 0,$$

which, together with the completeness of $\mathbf{t}$, shows that $\delta(\mathbf{t}) \overset{a.s.}{=} \delta'(\mathbf{t})$.

For any other unbiased estimator $\delta'(\mathbf{x})$ (not necessarily a function of $\mathbf{t}$), the Rao-Blackwellized estimator $\delta'_{\mathrm{RB}}(\mathbf{t}) := \mathbf{E}_\theta\big[\delta'(\mathbf{x}) \mid \mathbf{t}\big]$

1. remains unbiased:

$$\mathbf{E}_\theta\big[\delta'_{\mathrm{RB}}(\mathbf{t})\big] = \mathbf{E}_\theta\big[\mathbf{E}_\theta\big[\delta'(\mathbf{x}) \mid \mathbf{t}\big]\big] = \mathbf{E}_\theta\big[\delta'(\mathbf{x})\big] = \theta.$$

2. has smaller risk: $\mathrm{Risk}_{\delta'_{\mathrm{RB}}}(\theta) \le \mathrm{Risk}_{\delta'}(\theta)$.

Since $\delta'_{\mathrm{RB}}(\mathbf{t})$ is a function of $\mathbf{t}$ and $\delta(\mathbf{x})$ is the unique function of θ that is an unbiased estimator of $\mathbf{t}$, $\mathbf{E}_\theta\big[\delta'(\mathbf{x}) \mid \mathbf{t}\big] \overset{a.s.}{=} \delta(\mathbf{t})$. Thus

$$\mathrm{Risk}_{\delta'_{\mathrm{RB}}}(\theta) \le \mathrm{Risk}_{\delta'}(\theta),$$

which shows that $\delta(\mathbf{t})$ is a UMRU estimator.

If the loss function is strictly convex, then the Rao-Blackwellized estimator has strictly smaller risk:

$$\mathrm{Risk}_{\delta'_{\mathrm{RB}}}(\theta) < \mathrm{Risk}_{\delta'}(\theta).$$

Since $\delta'_{\mathrm{RB}}(\mathbf{t}) \overset{a.s.}{=} \delta(\mathbf{t})$, $\delta(\mathbf{t})$ is the unique UMRU estimator. □

The Lehmann-Scheffé theorem suggests Rao-Blackwellization as an approach to obtaining UMRU estimators. Going back to the coin tossing example, it is not hard to check that $\frac{1}{n}\sum_{i\in[n]} \mathbf{x}_i$ is the UMVU estimator of p. We consider a less trivial example: the UMVU estimator of $p(1-p)$.

At first blush, there is no obvious function of $\mathbf{t}$ that is an unbiased estimator of $p(1-p)$. Thus we start with a simple unbiased estimator

$$\delta(\mathbf{x}) = \mathbf{x}_1 - \mathbf{x}_1\mathbf{x}_2$$

and Rao-Blackwellize. We know that $\mathbf{t} = \sum_{i\in[n]} \mathbf{x}_i$ is a complete and sufficient statistic for the *i.i.d.* Bernoulli model. By the Lehmann-Scheffé theorem,

$$\mathbf{E}_p\big[\delta(\mathbf{x}) \mid \mathbf{t}\big] = \mathbf{E}_p\big[\mathbf{x}_1 - \mathbf{x}_1\mathbf{x}_2 \mid \mathbf{t}\big] = \mathbf{E}_p\big[\mathbf{x}_1 \mid \mathbf{t}\big] - \mathbf{E}_p\big[\mathbf{x}_1\mathbf{x}_2 \mid \mathbf{t}\big].$$

is the unique UMVU estimator. The first term is

$$\begin{aligned}\mathbf{E}_p\big[\mathbf{x}_1 \mid \mathbf{t} = t\big] &= \mathbf{P}_p\big(\mathbf{x}_1 = 1 \mid \mathbf{t} = t\big) \\ &= \frac{\mathbf{P}_p\big(\mathbf{x}_1 = 1, \sum_{i=2}^n \mathbf{x}_i = t-1\big)}{\mathbf{P}_p\big(\mathbf{t} = t\big)} \\ &= \frac{\mathbf{P}_p\big(\mathbf{x}_1 = 1\big)\mathbf{P}_p\big(\sum_{i=2}^n \mathbf{x}_i = t-1\big)}{\mathbf{P}_p\big(\mathbf{t} = t\big)} \\ &= \frac{p\binom{n-1}{t-1}p^{t-1}(1-p)^{n-t}}{\binom{n}{t}p^t(1-p)^{n-t}},\end{aligned}$$

which simplifies to $\frac{t}{n}$. A similar calculation shows that $\mathbf{E}_p\big[\mathbf{x}_1\mathbf{x}_2 \mid \mathbf{t}\big]$ simplifies to $\frac{t(t-1)}{n(n-1)}$. Thus the UMVU estimator of variance is

$$\delta(\mathbf{x}) = \frac{\mathbf{t}}{n} - \frac{\mathbf{t}(\mathbf{t}-1)}{n(n-1)} = \frac{\mathbf{t}}{n}\left(1 - \frac{\mathbf{t}-1}{n-1}\right).$$

We observe that the UMVU estimator is not the MLE, which, by equivariance of the MLE, is $\frac{\mathbf{t}}{n}\left(1-\frac{\mathbf{t}}{n}\right)$.

We wrap up our discussion of unbiased estimators by taking a step back and asking whether the set of unbiased estimators is rich enough. Although unbiased estimators are desirable, they are usually not even the minimum MSE estimators. It is often possible to trade a small increase in bias for a larger decrease in variance, resulting in a smaller MSE.

EXAMPLE 1.5. *Let* $\{\mathbf{x}_i\}_{i\in[n]}$ *are* i.i.d. $\mathcal{N}(\mu,\sigma^2)$ *random variables, it is possible to show that the MLE of* σ^2

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i\in[n]}(\mathbf{x}_i - \bar{\mathbf{x}}_n) = \frac{n-1}{n}\mathbf{s}_n^2,$$

despite its bias, has smaller MSE than the unbiased estimator of σ^2

$$\mathbf{s}_n^2 = \frac{1}{n-1}\sum_{i\in[n]}(\mathbf{x}_i - \bar{\mathbf{x}}_n).$$

We remark that just because $\hat{\sigma}^2$ *has smaller MSE than* $\mathbf{s}_n^2$ *does not discount* $\mathbf{s}_n^2$ *as a practical estimator of* σ^2*. We know* $\hat{\sigma}^2$ *tends to underestimate* σ^2*, which is problematic when testing hypothesis.*

YUEKAI SUN
BERKELEY, CALIFORNIA
NOVEMBER 16, 2015