Lecture 23: Sufficiency and Rao-Blackwell theorem, unbiasedness and invariance

Suppose that we have a sufficient statistic $T(X)$ for $P \in \mathcal{P}$.
Intuitively, our decision rule should be a function of T.
This is not true in general, but the following result indicates that this is true if randomized decision rules are allowed.

Proposition 2.2. Suppose that $\mathcal{A}$ is a subset of $\mathcal{R}^k$. Let $T(X)$ be a sufficient statistic for $P \in \mathcal{P}$ and let δ_0 be a decision rule. Then

$$\delta_1(t, A) = E[\delta_0(X, A)|T = t],$$

which is a randomized decision rule depending only on T, is equivalent to δ_0 if $R_{\delta_0}(P) < \infty$ for any $P \in \mathcal{P}$.
Proof. Note that δ_1 is a decision rule since δ_1 does not depend on the unknown P by the sufficiency of T. Then

$$\begin{aligned} R_{\delta_1}(P) &= E\left\{\int_{\mathcal{A}} L(P, a) d\delta_1(X, a)\right\} \\ &= E\left\{E\left[\int_{\mathcal{A}} L(P, a) d\delta_0(X, a)\Big|T\right]\right\} \\ &= E\left\{\int_{\mathcal{A}} L(P, a) d\delta_0(X, a)\right\} \\ &= R_{\delta_0}(P), \end{aligned}$$

where the proof of the second equality is left to the reader.

Note that Proposition 2.2 does not imply that δ_0 is inadmissible.
If δ_0 is a nonrandomized rule,

$$\delta_1(t, A) = E[I_A(\delta_0(X))|T = t] = P(\delta_0(X) \in A|T = t)$$

is still a randomized rule, unless $\delta_0(X) = h(T(X))$ a.s. P for some Borel function h (Exercise 75).
Hence, Proposition 2.2 does not apply to situations where randomized rules are not allowed.

The following result tells us when nonrandomized rules are all we need and when decision rules that are not functions of sufficient statistics are inadmissible.

Theorem 2.5. Suppose that $\mathcal{A}$ is a convex subset of $\mathcal{R}^k$ and that for any $P \in \mathcal{P}$, $L(P, a)$ is a convex function of a.
(i) Let δ be a randomized rule satisfying $\int_{\mathcal{A}} \|a\| d\delta(x, a) < \infty$ for any $x \in \mathcal{X}$ and let $T_1(x) = \int_{\mathcal{A}} a d\delta(x, a)$. Then $L(P, T_1(x)) \leq L(P, \delta, x)$ (or $L(P, T_1(x)) < L(P, \delta, x)$ if L is strictly convex in a) for any $x \in \mathcal{X}$ and $P \in \mathcal{P}$.
(ii) (Rao-Blackwell theorem). Let T be a sufficient statistic for $P \in \mathcal{P}$, $T_0 \in \mathcal{R}^k$ be a nonrandomized rule satisfying $E\|T_0\| < \infty$, and $T_1 = E[T_0(X)|T]$. Then $R_{T_1}(P) \leq R_{T_0}(P)$

for any $P \in \mathcal{P}$. If L is strictly convex in a and T_0 is not a function of T, then T_0 is inadmissible.

The proof of Theorem 2.5 is an application of Jensen's inequality and is left to the reader.

The concept of admissibility helps us to eliminate some decision rules.
However, usually there are still too many rules left after the elimination of some rules according to admissibility and sufficiency.
Although one is typically interested in a $\Im$-optimal rule, frequently it does not exist, if $\Im$ is either too large or too small.

Example 2.22. Let $X_1, ..., X_n$ be i.i.d. random variables from a population $P \in \mathcal{P}$ that is the family of populations having finite mean μ and variance σ^2.
Consider the estimation of μ ($\mathcal{A} = \mathcal{R}$) under the squared error loss.
It can be shown that if we let $\Im$ be the class of all possible estimators, then there is no $\Im$-optimal rule (exercise).
Next, let $\Im_1$ be the class of all linear functions in $X = (X_1, ..., X_n)$, i.e., $T(X) = \sum_{i=1}^n c_i X_i$ with known $c_i \in \mathcal{R}$, $i = 1, ..., n$.
Then

$$R_T(P) = \mu^2 \left(\sum_{i=1}^n c_i - 1 \right)^2 + \sigma^2 \sum_{i=1}^n c_i^2. \tag{1}$$

We now show that there does not exist $T_* = \sum_{i=1}^n c_i^* X_i$ such that $R_{T_*}(P) \leq R_T(P)$ for any $P \in \mathcal{P}$ and $T \in \Im_1$.
If there is such a T_*, then $(c_1^*, ..., c_n^*)$ is a minimum of the function of $(c_1, ..., c_n)$ on the right-hand side of (1).
Then $c_1^*, ..., c_n^*$ must be the same and equal to $\mu^2/(\sigma^2 + n\mu^2)$, which depends on P.
Hence T_* is not a statistic.
This shows that there is no $\Im_1$-optimal rule.
Consider now a subclass $\Im_2 \subset \Im_1$ with c_i's satisfying $\sum_{i=1}^n c_i = 1$.
From (1), $R_T(P) = \sigma^2 \sum_{i=1}^n c_i^2$ if $T \in \Im_2$.
Minimizing $\sigma^2 \sum_{i=1}^n c_i^2$ subject to $\sum_{i=1}^n c_i = 1$ leads to an optimal solution of $c_i = n^{-1}$.
Thus, the sample mean $\bar{X}$ is $\Im_2$-optimal.
There may not be any optimal rule if we consider a small class of decision rules.
For example, if $\Im_3$ contains all the rules in $\Im_2$ except $\bar{X}$, then one can show that there is no $\Im_3$-optimal rule.

Example 2.23. Assume that the sample X has the binomial distribution $Bi(\theta, n)$ with an unknown $\theta \in (0, 1)$ and a fixed integer $n > 1$.
Consider the hypothesis testing problem described in Example 2.20 with $H_0 : \theta \in (0, \theta_0]$ versus $H_1 : \theta \in (\theta_0, 1)$, where $\theta_0 \in (0, 1)$ is a fixed value.
Suppose that we are only interested in the following class of nonrandomized decision rules: $\Im = \{T_j : j = 0, 1, ..., n - 1\}$, where $T_j(X) = I_{\{j+1,...,n\}}(X)$.
From Example 2.20, the risk function for T_j under the 0-1 loss is

$$R_{T_j}(\theta) = P(X > j) I_{(0,\theta_0]}(\theta) + P(X \leq j) I_{(\theta_0,1)}(\theta).$$

For any integers k and j, $0 \leq k < j \leq n-1$,

$$R_{T_j}(\theta) - R_{T_k}(\theta) = \begin{cases} -P(k < X \leq j) < 0 & 0 < \theta \leq \theta_0 \\ P(k < X \leq j) > 0 & \theta_0 < \theta < 1. \end{cases}$$

Hence, neither T_j nor T_k is better than the other.
This shows that every T_j is $\Im$-admissible and, thus, there is no $\Im$-optimal rule.

In view of the fact that an optimal rule often does not exist, statisticians adopt the following two approaches to choose a decision rule.
The first approach is to define a class $\Im$ of decision rules that have some desirable properties (statistical and/or nonstatistical) and then try to find the best rule in $\Im$.
In Example 2.22, for instance, any estimator T in $\Im_2$ has the property that T is linear in X and $E[T(X)] = \mu$.
In a general estimation problem, we can use the following concept.

Definition 2.8 (Unbiasedness). In an estimation problem, the *bias* of an estimator $T(X)$ of a real-valued parameter ϑ of the unknown population is defined to be $b_T(P) = E[T(X)] - \vartheta$ (which is denoted by $b_T(\theta)$ when P is in a parametric family indexed by θ). An estimator $T(X)$ is said to be *unbiased* for ϑ if and only if $b_T(P) = 0$ for any $P \in \mathcal{P}$.

Thus, $\Im_2$ in Example 2.22 is the class of unbiased estimators linear in X.
In Chapter 3, we discuss how to find a $\Im$-optimal estimator when $\Im$ is the class of unbiased estimators or unbiased estimators linear in X.

Another class of decision rules can be defined after we introduce the concept of *invariance.*

Definition 2.9 Let X be a sample from $P \in \mathcal{P}$.
(i) A class $\mathcal{G}$ of one-to-one transformations of X is called a *group* if and only if $g_i \in \mathcal{G}$ implies $g_1 \circ g_2 \in \mathcal{G}$ and $g_i^{-1} \in \mathcal{G}$.
(ii) We say that $\mathcal{P}$ is *invariant* under $\mathcal{G}$ if and only if $\bar{g}(P_X) = P_{g(X)}$ is a one-to-one transformation from $\mathcal{P}$ onto $\mathcal{P}$ for each $g \in \mathcal{G}$.
(iii) A decision problem is said to be *invariant* if and only if $\mathcal{P}$ is invariant under $\mathcal{G}$ and the loss $L(P, a)$ is invariant in the sense that, for every $g \in \mathcal{G}$ and every $a \in \mathcal{A}$, there exists a unique $g(a) \in \mathcal{A}$ such that $L(P_X, a) = L\left(P_{g(X)}, g(a)\right)$. (Note that $g(X)$ and $g(a)$ are different functions in general.)
(iv) A decision rule $T(x)$ is said to be *invariant* if and only if, for every $g \in \mathcal{G}$ and every $x \in \mathcal{X}$, $T(g(x)) = g(T(x))$.

Invariance means that our decision is not affected by one-to-one transformations of data.
In a problem where the distribution of X is in a location-scale family $\mathcal{P}$ on $\mathcal{R}^k$, we often consider location-scale transformations of data X of the form $g(X) = AX + c$, where $c \in \mathcal{C} \subset \mathcal{R}^k$ and $A \in \mathcal{T}$, a class of invertible $k \times k$ matrices.
In §4.2 and §6.3, we discuss the problem of finding a $\Im$-optimal rule when $\Im$ is a class of invariant decision rules.