Lecture 21: Sufficiency and Rao-Blackwell theorem, and two common approaches to derive decision rules

**Sufficiency**

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 \( \mathbb{R}^k \). Let \( T(X) \) be a sufficient statistic for \( P \in \mathcal{P} \) and let \( \delta_0 \) 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 \( \delta_0 \) if \( R_{\delta_0}(P) < \infty \) for any \( P \in \mathcal{P} \).
Proof

Note that $\delta_1$ is a decision rule since $\delta_1$ does not depend on the unknown $P$ by the sufficiency of $T$. Then

$$R_{\delta_1}(P) = E \left\{ \int_{\mathcal{A}} L(P, a) d\delta_1(X, a) \right\}$$

$$= E \left\{ \int_{\mathcal{A}} L(P, a) dE \left[ \delta_0(X, a) \left| T \right. \right] \right\}$$

$$= E \left\{ E \left[ \int_{\mathcal{A}} L(P, a) d\delta_0(X, a) \left| T \right. \right] \right\}$$

$$= E \left\{ \int_{\mathcal{A}} L(P, a) d\delta_0(X, a) \right\}$$

$$= R_{\delta_0}(P),$$

where the proof of the third equality is left to the reader.
Remarks

- Note that Proposition 2.2 does not imply that $\delta_0$ is inadmissible.
- If $\delta_0$ 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.

When can we ignore randomized rules?

Randomized rules involve an additional randomization. It may not be easy to understand and/or interpret. 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 $\mathbb{R}^k$ and that for any $P \in \mathcal{P}$, $L(P, a)$ is a convex function of $a$.

(i) Let $\delta$ 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 $P \in \mathcal{P}$, $T_0 \in \mathbb{R}^k$ 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.

Proof

The proof of Theorem 2.5 is an application of Jensen’s inequality and is left to the reader.
How to find a decision rule?

The concept of admissibility and sufficiency 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 $\mathcal{I}$-optimal rule, frequently it does not exist, if $\mathcal{I}$ is either too large or too small.

Example 2.22

Let $X_1, \ldots, X_n$ be i.i.d. random variables from a population $P \in \mathcal{P}$ that is the family of populations having finite mean $\mu$ and variance $\sigma^2$.

Consider the estimation of $\mu$ ($\mathcal{A} = \mathbb{R}$) under the squared error loss. It can be shown that if we let $\mathcal{I}$ be the class of all possible estimators, then there is no $\mathcal{I}$-optimal rule (exercise).

Next, let $\mathcal{I}_1$ be the class of all linear functions in $X = (X_1, \ldots, X_n)$, i.e., $T(X) = \sum_{i=1}^{n} c_i X_i$ with known $c_i \in \mathbb{R}$, $i = 1, \ldots, n$. 
Example 2.22 (continued)

Then

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

(1)

We now show that there is no \( S_1 \)-optimal rule, i.e., 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 \( T \in S_1 \).

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 \( P \), i.e., \( T_* \) is not a statistic.

Consider now a subclass \( S_2 \subset S_1 \) 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 \( T \in S_2 \).

Minimizing \( \sigma^2 \sum_{i=1}^{n} c_i^2 \) subject to \( \sum_{i=1}^{n} c_i = 1 \) leads to \( c_i = n^{-1} \).

Thus, the sample mean \( \bar{X} \) is \( S_2 \)-optimal.

There may not be any optimal rule if we consider a small class of rules. For example, if \( S_3 \) contains all the rules in \( S_2 \) except \( \bar{X} \), then one can show that there is no \( S_3 \)-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:

$$H_0 : \theta \in (0, \theta_0] \quad \text{versus} \quad H_1 : \theta \in (\theta_0, 1),$$

where $\theta_0 \in (0, 1)$ is a fixed value.

Suppose that we are only interested in the following class of nonrandomized decision rules: $\mathcal{S} = \{ T_j : j = 0, 1, ..., n - 1 \}$, 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 $\mathcal{S}$-admissible and, thus, there is no $\mathcal{S}$-optimal rule.
Approaches

In view of the fact that an optimal rule often does not exist, statisticians adopt two common approaches to choose a decision rule.

The first approach is to define a class $\mathcal{S}$ of decision rules that have some desirable properties (statistical and/or nonstatistical) and then try to find the best rule in $\mathcal{S}$.

In Example 2.22, for instance, any estimator $T$ in $\mathcal{S}_2$ 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 $\vartheta$ of the unknown population is defined to be

$$b_T(P) = E[T(X)] - \vartheta$$

(denoted by $b_T(\theta)$ when $P$ is in a parametric family indexed by $\theta$). An estimator $T(X)$ is said to be unbiased for $\vartheta$ iff $b_T(P) = 0$ for any $P \in \mathcal{P}$. 
Remarks

- $\mathcal{S}_2$ in Example 2.22 is the class of unbiased estimators linear in $X$.
- In Chapter 3, we discuss how to find a $\mathcal{S}$-optimal estimator when $\mathcal{S}$ is the class of unbiased estimators or unbiased estimators linear in $X$.

Another property we may consider is *invariance*

- Consider a class of transformations (such as unit changing)
- Consider rules that are not affected by transformation (invariance)
- Try to find the best rule within the class of invariant rule
- Details are omitted (see textbook)

Approaches

The second approach to finding a good decision rule is to consider some characteristic $R_T$ of $R_T(P)$, for a given decision rule $T$, and then minimize $R_T$ over $T \in \mathcal{S}$.

The following are two popular ways to carry out this idea.
The first method: the Bayes rule

Consider an average of $R_T(P)$ over $P \in \mathcal{P}$:

$$r_T(\Pi) = \int_{\mathcal{P}} R_T(P) d\Pi(P),$$

where $\Pi$ is a known probability measure on $(\mathcal{P}, \mathcal{F}_P)$ with an appropriate $\sigma$-field $\mathcal{F}_P$.

$r_T(\Pi)$ is called the Bayes risk of $T$ w.r.t. $\Pi$.

If $T_\ast \in \mathcal{S}$ and $r_{T_\ast}(\Pi) \leq r_T(\Pi)$ for any $T \in \mathcal{S}$, then $T_\ast$ is called a $\mathcal{S}$-Bayes rule (or Bayes rule when $\mathcal{S}$ contains all possible rules) w.r.t. $\Pi$.

The second method: the minimax rule

Consider the worst situation, i.e., $\sup_{P \in \mathcal{P}} R_T(P)$.

If $T_\ast \in \mathcal{S}$ and

$$\sup_{P \in \mathcal{P}} R_{T_\ast}(P) \leq \sup_{P \in \mathcal{P}} R_T(P)$$

for any $T \in \mathcal{S}$, then $T_\ast$ is called a $\mathcal{S}$-minimax rule (or minimax rule when $\mathcal{S}$ contains all possible rules).
Example 2.25

Consider the estimation of $\theta \in \mathbb{R}$ under loss $L(\theta, a) = (\theta - a)^2$ and

$$r_T(\Pi) = \int_{\mathbb{R}} E[\theta - T(X)]^2 d\Pi(\theta),$$

which is equivalent to $E[\tilde{\theta} - T(X)]^2$, where $\tilde{\theta}$ is random and distributed as $\Pi$ and, given $\tilde{\theta} = \theta$, the conditional distribution of $X$ is $P_\theta$.

Then, the problem can be viewed as a prediction problem for $\tilde{\theta}$ using functions of $X$.

Using the result in Example 1.22, the best predictor is $E(\tilde{\theta}|X)$, which is the $\mathcal{S}$-Bayes rule w.r.t. $\Pi$ with $\mathcal{S}$ being the class of rules $T(X)$ satisfying $E[T(X)]^2 < \infty$ for any $\theta$.

We usually try to find a Bayes rule or a minimax rule in a parametric problem where $P = P_\theta$ for a $\theta \in \mathbb{R}^k$.

A minimax rule in general may be difficult to obtain. Bayes and minimax rules are discussed in Chapter 4.