Lecture 22: Decision rules, loss, and risk

Statistical decision theory

X: a sample from a population $P \in \mathcal{P}$
Decision: an action we take after observing X
$\mathcal{A}$: the set of allowable actions
$(\mathcal{A}, \mathcal{F}_{\mathcal{A}})$: the action space
$\mathcal{X}$: the range of X
Decision rule: a measurable function (a statistic) T from $(\mathcal{X}, \mathcal{F}_{\mathcal{X}})$ to $(\mathcal{A}, \mathcal{F}_{\mathcal{A}})$
If X is observed, then we take the action $T(X) \in \mathcal{A}$

Performance criterion: loss function $L(P, a)$ from $\mathcal{P} \times \mathcal{A}$ to $[0, \infty)$ and is Borel for each P
If $X = x$ is observed and our decision rule is T, then our "loss" is $L(P, T(x))$
It is difficult to compare $L(P, T_1(X))$ and $L(P, T_2(X))$ for two decision rules, T_1 and T_2, since both of them are random.

Risk: Average (expected) loss defined as

$$R_T(P) = E[L(P, T(X))] = \int_{\mathcal{X}} L(P, T(x)) dP_X(x).$$

If $\mathcal{P}$ is a parametric family indexed by θ, the loss and risk are denoted by $L(\theta, a)$ and $R_T(\theta)$

For decision rules T_1 and T_2, T_1 is *as good as* T_2 if and only if

$$R_{T_1}(P) \leq R_{T_2}(P) \quad \text{for any } P \in \mathcal{P},$$

and is *better* than T_2 if, in addition, $R_{T_1}(P) < R_{T_2}(P)$ for at least one $P \in \mathcal{P}$.

Two decision rules T_1 and T_2 are *equivalent* if and only if $R_{T_1}(P) = R_{T_2}(P)$ for all $P \in \mathcal{P}$.

Optimal rule: If T_* is as good as any other rule in $\Im$, a class of allowable decision rules, then T_* is $\Im$-*optimal* (or optimal if $\Im$ contains all possible rules).

Sometimes it is useful to consider *randomized decision rules*.
Randomized decision rule: a function δ on $\mathcal{X} \times \mathcal{F}_{\mathcal{A}}$ such that, for every $A \in \mathcal{F}_{\mathcal{A}}$, $\delta(\cdot, A)$ is a Borel function and, for every $x \in \mathcal{X}$, $\delta(x, \cdot)$ is a probability measure on $(\mathcal{A}, \mathcal{F}_{\mathcal{A}})$.
If $X = x$ is observed, our have a distribution of actions: $\delta(x, \cdot)$.
A nonrandomized decision rule T previously discussed can be viewed as a special randomized decision rule with $\delta(x, \{a\}) = I_{\{a\}}(T(x))$, $a \in \mathcal{A}$, $x \in \mathcal{X}$.

To choose an action in $\mathcal{A}$ when a randomized rule δ is used, we need to simulate a pseudo-random element of $\mathcal{A}$ according to $\delta(x, \cdot)$.
Thus, an alternative way to describe a randomized rule is to specify the method of simulating the action from $\mathcal{A}$ for each $x \in \mathcal{X}$.
For example, a randomized rule can be a discrete distribution $\delta(x, \cdot)$ assigning probability $p_j(x)$ to a nonrandomized decision rule $T_j(x)$, $j = 1, 2, \ldots$, in which case the rule δ can be

equivalently defined as a rule taking value $T_j(x)$ with probability $p_j(x)$, i.e.,

$$T(X) = \begin{cases} T_1(X) & \text{with probability } p_1(X) \\ \cdots & \cdots \\ T_k(X) & \text{with probability } p_k(X) \end{cases}$$

The loss function for a randomized rule δ is defined as

$$L(P, \delta, x) = \int_{\mathcal{A}} L(P, a) d\delta(x, a),$$

which reduces to the same loss function we discussed when δ is a nonrandomized rule.
The risk of a randomized rule δ is then

$$R_\delta(P) = E[L(P, \delta, X)] = \int_{\mathcal{X}} \int_{\mathcal{A}} L(P, a) d\delta(x, a) dP_X(x).$$

For $T(X)$ defined above,

$$L(P, T, x) = \sum_{j=1}^{k} L(P, T_j(x)) p_j(x)$$

and

$$R_T(P) = \sum_{j=1}^{k} E[L(P, T_j(X)) p_j(X)]$$

Example 2.19. Let $X = (X_1, ..., X_n)$ be a vector of iid measurements for a parameter $\theta \in \mathcal{R}$.
Action space: $(\mathcal{A}, \mathcal{F}_{\mathcal{A}}) = (\mathcal{R}, \mathcal{B})$.
A common loss function in this problem is the *squared error loss* $L(P, a) = (\theta - a)^2$, $a \in \mathcal{A}$.
Let $T(X) = \bar{X}$, the sample mean.
The loss for $\bar{X}$ is $(\bar{X} - \theta)^2$.
If the population has mean μ and variance $\sigma^2 < \infty$, then

$$\begin{aligned} R_{\bar{X}}(P) &= E(\theta - \bar{X})^2 \\ &= (\theta - E\bar{X})^2 + E(E\bar{X} - \bar{X})^2 \\ &= (\theta - E\bar{X})^2 + \mathrm{Var}(\bar{X}) \\ &= (\mu - \theta)^2 + \tfrac{\sigma^2}{n}. \end{aligned}$$

If θ is in fact the mean of the population, then

$$R_{\bar{X}}(P) = \tfrac{\sigma^2}{n},$$

is an increasing function of the population variance σ^2 and a decreasing function of the sample size n.
Consider another decision rule $T_1(X) = (X_{(1)} + X_{(n)})/2$.

$R_{T_1}(P)$ does not have a simple explicit form if there is no further assumption on the population P.
Suppose that $P \in \mathcal{P}$. Then, for some $\mathcal{P}$, $\bar{X}$ (or T_1) is better than T_1 (or $\bar{X}$) (exercise), whereas for some $\mathcal{P}$, neither $\bar{X}$ nor T_1 is better than the other.
Consider a randomized rule:

$$T_2(X) = \begin{cases} \bar{X} & \text{with probability } p(X) \\ T_1(X) & \text{with probability } 1 - p(X) \end{cases}$$

The loss for $T_2(X)$ is

$$(\bar{X} - \theta)^2 p(X) + [T_1(X) - \theta]^2 [1 - p(X)]$$

and the risk of T_2 is

$$R_{T_2}(P) = E\{(\bar{X} - \theta)^2 p(X) + [T_1(X) - \theta]^2 [1 - p(X)]\}$$

In particular, if $p(X) = 0.5$, then

$$R_{T_2}(P) = \frac{R_{\bar{X}}(P) + R_{T_1}(P)}{2}.$$

The problem in Example 2.19 is a special case of a general problem called *estimation*.
In an estimation problem, a decision rule T is called an *estimator*.
The following example describes another type of important problem called *hypothesis testing*.

Example 2.20. Let $\mathcal{P}$ be a family of distributions, $\mathcal{P}_0 \subset \mathcal{P}$, and $\mathcal{P}_1 = \{P \in \mathcal{P} : P \notin \mathcal{P}_0\}$.
A hypothesis testing problem can be formulated as that of deciding which of the following two statements is true:

$$H_0: \ P \in \mathcal{P}_0 \qquad \text{versus} \qquad H_1: \ P \in \mathcal{P}_1. \tag{1}$$

Here, H_0 is called the *null hypothesis* and H_1 is called the *alternative hypothesis*.
The action space for this problem contains only two elements, i.e., $\mathcal{A} = \{0, 1\}$, where 0 is the action of accepting H_0 and 1 is the action of rejecting H_0.
A decision rule is called a *test*.
Since a test $T(X)$ is a function from $\mathcal{X}$ to $\{0, 1\}$, $T(X)$ must have the form $I_C(X)$, where $C \in \mathcal{F}_{\mathcal{X}}$ is called the *rejection region* or *critical region* for testing H_0 versus H_1.
0-1 loss: $L(P, a) = 0$ if a correct decision is made and 1 if an incorrect decision is made, i.e., $L(P, j) = 0$ for $P \in \mathcal{P}_j$ and $L(P, j) = 1$ otherwise, $j = 0, 1$.
Under this loss, the risk is

$$R_T(P) = \begin{cases} P(T(X) = 1) = P(X \in C) & P \in \mathcal{P}_0 \\ P(T(X) = 0) = P(X \notin C) & P \in \mathcal{P}_1. \end{cases}$$

See Figure 2.2 on page 127 for an example of a graph of $R_T(\theta)$ for some T and P in a parametric family.

The 0-1 loss implies that the loss for two types of incorrect decisions (accepting H_0 when $P \in \mathcal{P}_1$ and rejecting H_0 when $P \in \mathcal{P}_0$) are the same.
In some cases, one might assume unequal losses: $L(P, j) = 0$ for $P \in \mathcal{P}_j$, $L(P, 0) = c_0$ when $P \in \mathcal{P}_1$, and $L(P, 1) = c_1$ when $P \in \mathcal{P}_0$.

Admissibility

Definition 2.7. Let $\Im$ be a class of decision rules (randomized or nonrandomized). A decision rule $T \in \Im$ is called $\Im$-*admissible* (or admissible when $\Im$ contains all possible rules) if and only if there does not exist any $S \in \Im$ that is better than T (in terms of the risk).

If a decision rule T is inadmissible, then there exists a rule better than T.
Thus, T should not be used in principle.
However, an admissible decision rule is not necessarily good.
For example, in an estimation problem a silly estimator $T(X) \equiv$ a constant may be admissible.

If T_* is $\Im$-optimal, then it is $\Im$-admissible.
If T_* is $\Im$-optimal and T_0 is $\Im$-admissible, then T_0 is also $\Im$-optimal and is equivalent to T_*.
If there are two $\Im$-admissible rules that are not equivalent, then there does not exist any $\Im$-optimal rule.