Basic elements

- 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

Loss function \( L(P, a) \): a function from \( \mathcal{P} \times \mathcal{A} \) to \([0, \infty)\).
- \( L(P, a) \) is Borel for each \( P \)
- If \( X = x \) is observed and our decision rule is \( T \), then our “loss” is \( L(P, T(x)) \)
Risk

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. The average (expected) loss is defined as

$$R_T(P) = E[L(P, T(X))] = \int_X L(P, T(x))dP_X(x).$$

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

Comparisons

- For decision rules $T_1$ and $T_2$, $T_1$ is as good as $T_2$ iff
  $$R_{T_1}(P) \leq R_{T_2}(P) \quad \text{for any } P \in \mathcal{P},$$
  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 iff
  $$R_{T_1}(P) = R_{T_2}(P) \quad \text{for all } P \in \mathcal{P}.$$
Optimal rule
If $T_*$ is as good as any other rule in $\mathcal{G}$, a class of allowable decision rules, then $T_*$ is $\mathcal{G}$-optimal (or optimal if $\mathcal{G}$ contains all possible rules).

Randomized decision rules
A function $\delta$ on $X \times \mathcal{F}_A$ such that, for every $A \in \mathcal{F}_A$, $\delta(\cdot, A)$ is a Borel function and, for every $x \in X$, $\delta(x, \cdot)$ is a probability measure on $(A, \mathcal{F}_A)$.

- If $X = x$ is observed, we 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 $\delta(x, \{a\}) = I_{\{a\}}(T(x))$, $a \in A$, $x \in X$.
- To choose an action in $A$ when a randomized rule $\delta$ is used, we need to simulate a pseudorandom element of $A$ according to $\delta(x, \cdot)$.
- Thus, an alternative way to describe a randomized rule is to specify the method of simulating the action from $A$ for each $x \in X$. 
Randomized decision rules (cont.)

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 \( \delta \) can be equivalently defined as a rule taking value \( T_j(x) \) with probability \( p_j(x) \), i.e.,

\[
T(X) = \begin{cases} 
T_1(X) & \text{with probability } p_1(X) \\
\vdots & \vdots \\
T_k(X) & \text{with probability } p_k(X)
\end{cases}
\]

The loss function for a randomized rule \( \delta \) is defined as

\[
L(P, \delta, x) = \int_A L(P, a) d\delta(x, a),
\]

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

\[
R_\delta(P) = E[L(P, \delta, X)] = \int_X \int_A L(P, a) d\delta(x, a) dP_X(x).
\]
Randomized decision rules (cont.)

For

\[ T(X) = \begin{cases} 
T_1(X) & \text{with probability } p_1(X) \\
\vdots & \vdots \\
T_k(X) & \text{with probability } p_k(X) 
\end{cases} \]

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

and

\[ R_T(P) = \sum_{j=1}^{k} E[L(P, T_j(X)) p_j(X)] \]
Randomized decision rules (cont.)

For

\[ T(X) = \begin{cases} 
T_1(X) & \text{with probability } p_1(X) \\
\vdots & \vdots \\
T_k(X) & \text{with probability } p_k(X) 
\end{cases} \]

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

and

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

Example 2.19

Let \( X = (X_1, \ldots, X_n) \) be a vector of iid measurements for a parameter \( \theta \in \mathcal{R}. \)

We want to estimate \( \theta. \)

Action space: \( (\mathcal{A}, \mathcal{F}_A) = (\mathcal{R}, \mathcal{B}). \)

A common loss function in this problem is the \textit{squared error loss} \( L(P, a) = (\theta - a)^2, \ a \in \mathcal{A}. \)
Example 2.19 (continued)

Let $T(X) = \bar{X}$, the sample mean.
The loss for $\bar{X}$ is $(\bar{X} - \theta)^2$.
If the population has mean $\mu$ and variance $\sigma^2 < \infty$, then

$$R_{\bar{X}}(P) = E(\theta - \bar{X})^2$$
$$= (\theta - E\bar{X})^2 + E(E\bar{X} - \bar{X})^2$$
$$= (\theta - E\bar{X})^2 + \text{Var}(\bar{X})$$
$$= (\mu - \theta)^2 + \frac{\sigma^2}{n}.$$

If $\theta$ is in fact the mean of the population, then

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

is an increasing function of the population variance $\sigma^2$ and a decreasing function of the sample size $n$. 
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:

$$H_0 : P \in \mathcal{P}_0 \quad \text{versus} \quad H_1 : P \in \mathcal{P}_1.$$ 

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 $H_0$. 
Example 2.20 (continued)

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 $C \in \mathcal{F}_X$ 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

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

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 $P \in \mathcal{P}_0$. 
Definition 2.7 (Admissibility)

Let $\mathcal{S}$ be a class of decision rules (randomized or nonrandomized). A decision rule $T \in \mathcal{S}$ is called $\mathcal{S}$-admissible (or admissible when $\mathcal{S}$ contains all possible rules) iff there does not exist any $S \in \mathcal{S}$ that is better than $T$ (in terms of the risk).

Remarks

- If a decision rule $T$ is inadmissible, then there exists a rule better than $T$ and $T$ 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 $T_\ast$ is $\mathcal{S}$-optimal, then it is $\mathcal{S}$-admissible.
- If $T_\ast$ is $\mathcal{S}$-optimal and $T_0$ is $\mathcal{S}$-admissible, then $T_0$ is also $\mathcal{S}$-optimal and is equivalent to $T_\ast$.
- If there are two $\mathcal{S}$-admissible rules that are not equivalent, then there does not exist any $\mathcal{S}$-optimal rule.
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 $\delta$ be a randomized rule satisfying $\int_{\mathcal{A}} \|a\| d\delta(x, a) < \infty$ for any $x \in \mathcal{X}$ and let $T_1(x) = \int_{\mathcal{A}} ad\delta(x, a)$. 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.

Remark
Under the conditions of Theorem 2.5, we can concentrate on non-randomized rules. The proof of Theorem 2.5 is an application of Jensen’s inequality.
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 $P \in \mathcal{P}$.

Definition 2.9 (Invariance)

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* iff $g_i \in \mathcal{G}$ 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 $\mathcal{G}$ iff $\bar{g}(P_X) = P_{g(X)}$ is a one-to-one transformation from $\mathcal{P}$ onto $\mathcal{P}$ for each $g \in \mathcal{G}$. 
Definition 2.9 (continued)

(iii) A decision problem is said to be invariant iff \( \mathcal{P} \) 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 \( L(P, a) = L(P_{g(X)}, g(a)) \).
(Note that \( g(X) \) and \( g(a) \) are different functions in general.)

(iv) A decision rule \( T(x) \) is said to be invariant iff, for every \( g \in \mathcal{G} \) and every \( x \in \mathcal{X} \), \( T(g(x)) = g(T(x)) \).

Remarks

▶ 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.
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. $\Pi$.

Minimix 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).