Consulting Time and Venue:
Fridays (2:00 - 5:00pm), Math Building 2304

Text Book:

Reference Books:

Evaluation:
1: Mid-term paper and presentation (30%, ReferencePapersList.pdf).
2: Final-term examination (70%, one A4 size help sheet is allowed).
Chapter 1: Basic Concepts

1. Introduction

2. Basic Elements

3. Expected Loss, Decision Rules, and Risk
   - Bayesian Expected Loss
   - Frequentist Risk

4. Randomized Decision Rules

5. Decision Principles
   - The Conditional Bayes Decision Principle
   - Frequentist Decision Principles

6. Sufficient Statistics

7. Convexity
• **Statistical decision theory**: to make decisions in the presence of statistical knowledge which sheds light on some of the uncertainties (denoted by $\theta$)

• **Classical statistics**: to use sample information to make inferences about $\theta$.

• **Decision theory**: to combine the sample information with other relevant aspects to make the decision

• Two typical types of relevant *nonsample information*:
  - A knowledge of the possible consequences of the decisions (loss for various values of $\theta$: loss function)
  - Information arising from sources other than the statistical investigation, generally comes from past experience (prior information)

• **Bayesian analysis**: a statistical approach which formally seeks to utilize prior information
Notations and Terminologies

- $\theta$: the unknown quantity, *state of nature (parameter)*
- $\Theta$: the set of all possible states of nature (*parameter space*)
- $a$: actions (decisions)
- $A$: the set of all possible actions
- $L(\theta, a)$: the loss function, $\theta$ is the true state of nature, $a$ is the action taken, $(L(\theta, a) \geq -K > -\infty)$
Notations and Terminologies (cont.)

- $X = (X_1, X_2, \cdots, X_n)$: the outcome, $X_i$s are independent observations from a common distribution
- $x$: a particular realization of $X$
- $\mathcal{H}$: the set of possible outcomes (the sample space, e.g., $\mathbb{R}^n$)
- $P_{\theta}(A)$: or $P_{\theta}(X \in A)$: the probability of the event $A$ ($A \subseteq \mathcal{H}$) when $\theta$ is the true state of nature.

For example, $X$ will be assumed to be either a continuous or a discrete random variable with density $f(x|\theta)$, let $F^X(x|\theta)$ be the cumulative distribution function (c.d.f), then $P_{\theta}(A) = \int_A dF^X(x|\theta)$, the expectation of function $h(x)$, $E_{\theta}[h(X)]$ is defined to be $\int_{\mathcal{H}} h(x)dF^X(x|\theta)$
It is useful and natural to state prior beliefs in terms of probabilities of various possible values of $\theta$ being true.

Let $\pi(\theta)$ represent a prior density of $\theta$ (continuous or discrete), if $A \subseteq \Theta$,

$$P(\theta \in A) = \int_A dF^{\pi}(\theta) = \begin{cases} \int_A \pi(\theta) d\theta & (\text{continuous}) \\ \sum_{\theta \in A} \pi(\theta) & (\text{discrete}) \end{cases}$$
An Example:

Example 1

A drug company needs to decide on whether or not to market the drug, how much to market, what price to charge, etc. Two main factors affecting its decision are the proportion of people for which the drug will prove effective $\theta_1$, and the proportion of the market the drug will capture ($\theta_2$).

Assume it is desired to estimate $\theta_2$. Clearly, $\Theta = [0, 1]$ and $\mathcal{A} = [0, 1]$. 
An Example (cont.):

- The loss function might be

\[ L(\theta_2, a) = \begin{cases} 
\theta_2 - a & \text{if } \theta_2 - a \geq 0 \\
2(a - \theta_2) & \text{if } \theta_2 - a \leq 0 
\end{cases} \]

- The sample information about \( \theta_2 \) is obtained from a sample survey. For example, assume \( n \) people are interviewed, and the number \( X \) who would buy the drug is observed. It might be reasonable to assume

\[ f(x|\theta_2) = \binom{n}{x} \theta_2^x (1 - \theta_2)^{n-x}. \]

- There could well be considerable prior information about \( \theta_2 \), arising from previous introductions of new similar drugs into the market. Assume the prior information could be modeled by giving \( \theta_2 \) a \( \mathcal{U}(0.1, 0.2) \) prior density, i.e., letting

\[ \pi(\theta_2) = 10I_{(0.1, 0.2)}(\theta_2) \]
It is important to note that:

- Loss function and prior information could be very vague or even nonunique.
- The goal in statistical inference is to provide a “summary” of the statistical evidence which a wide variety of future “users” of this evidence can easily incorporate into their own decision-making processes.
The reasons for involvement of losses and prior information in inference:

- Reports from statistical inferences should (ideally) be constructed so that they can be easily utilized in individual decision making.
- The investigator may very well possess such information; he will often be very informed about the uses to which his inferences are likely to be put, and may have considerable prior knowledge about the situation. It is then almost imperative that he present such information in his analysis.
- The choice of an inference can be viewed as a decision problem.
To summarize: decision theory can be useful even when such incorporation is proscribed. This is because many standard inference criteria can be formally reproduced as decision-theoretic criteria with respect to certain formal loss functions.
Chapter 1

1. Introduction

2. Basic Elements

3. Expected Loss, Decision Rules, and Risk
   - Bayesian Expected Loss
   - Frequentist Risk

4. Randomized Decision Rules

5. Decision Principles
   - The Conditional Bayes Decision Principle
   - Frequentist Decision Principles

6. Sufficient Statistics

7. Convexity
Definition 2 (Bayesian Expected Loss)

If $\pi^*(\theta)$ is the believed probability distribution of $\theta$ at the time of decision making, the Bayesian expected loss of an action $a$ is

$$\rho(\pi^*, a) = \mathbb{E}^{\pi^*} L(\theta, a) = \int_{\Theta} L(\theta, a) dF^{\pi^*}(\theta).$$
Example 3 (Example 1 cont.)

Assume no data is obtained, so that the believed distribution of $\theta_2$ is simply $\pi(\theta_2) = 10I_{(0.1,0.2)}(\theta_2)$. Then

$$\rho(\pi, a) = \int_0^1 L(\theta_2, a)\pi(\theta_2)d\theta_2$$

$$= \int_0^a 2(a - \theta_2)10I_{(0.1,0.2)}(\theta_2)d\theta_2$$

$$+ \int_a^1 (\theta_2 - a)10I_{(0.1,0.2)}(\theta_2)d\theta_2$$

$$= \begin{cases} 
0.15 - a & \text{if } a \leq 0.1 \\
15a^2 - 4a + 0.3 & \text{if } 0.1 \leq a \leq 0.2 \\
2a - 0.3 & \text{if } a \geq 0.2 
\end{cases}$$
Chapter 1

1. Introduction

2. Basic Elements

3. Expected Loss, Decision Rules, and Risk
   - Bayesian Expected Loss
   - Frequentist Risk

4. Randomized Decision Rules

5. Decision Principles
   - The Conditional Bayes Decision Principle
   - Frequentist Decision Principles

6. Sufficient Statistics

7. Convexity
Nonrandomized decision rule

Definition 4 (Nonrandomized Decision Rule)
A (nonrandomized) decision rule $\delta(x)$ is a function from $\mathcal{H}$ into $\mathcal{A}$. If $X = x$ is the observed value of the sample information, then $\delta(x)$ is the action that will be taken. (For a no-data problem, a decision rule is simply an action.) Two decision rules, $\delta_1$ and $\delta_2$, are considered equivalent if $P_{\theta}(\delta_1(X) = \delta_2(X)) = 1$ for all $\theta$.

Example 5 (Example 1 cont.)
For the situation of Example 1, $\delta(x) = x/n$ is the standard decision rule (estimator) for estimating $\theta_2$. 
Definition 6 (Risk Function For Nonrandomized Decision Rules)

The risk function of a decision rule $\delta(x)$ is defined by

$$R(\theta, \delta) = E_\theta^X [L(\theta, \delta(X))] = \int_{\mathcal{H}} L(\theta, \delta(x)) dF^X(x|\theta).$$

For a no-data problem, $R(\theta, \delta) \equiv L(\theta, \delta)$.

To a frequentist, it is desirable to use a decision rule $\delta$ which has small $R(\theta, \delta)$. However, whereas the Bayesian expected loss of an action was a single number, the risk is a function on $\Theta$, and since $\theta$ is unknown we have a problem in saying what “small” means.
Definition 7
A decision rule $\delta_1$, is $R$ – better than a decision rule $\delta_2$ if $R(\theta, \delta_1) \leq R(\theta, \delta_2)$ for all $\theta \in \Theta$, with strict inequality for some $\theta$. A rule $\delta_1$ is $R$ – equivalent to $\delta_2$ if $R(\theta, \delta_1) = R(\theta, \delta_2)$ for all $\theta$.

Definition 8 (Admissibility)
A decision rule $\delta$ is *admissible* if there exists no $R$ – better decision rule. A decision rule $\delta$ is *inadmissible* if there does exist an $R$ – better decision rule.
Example 9

Assume $X$ is $\mathcal{N}(\theta, 1)$ and that it is desired to estimate $\theta$ under loss $L(\theta, a) = (\theta - a)^2$. (This loss is called *squared-error loss.*) Consider the decision rules $\delta_c(x) = cx$. Clearly the risk function $R(\theta, \delta_x) = E_X L(\theta, \delta_c(X)) = E_X (\theta - cX)^2 = c^2 + (1 - c)^2 \theta^2$. Hence, $\delta_1$ is $R$–better than $\delta_c$ for $c > 1$, the rules $\delta_c$ are inadmissible for $c > 1$. 
Definition 10 (Bayes Risk)

The Bayes risk of a decision rule $\delta$, with respect to a prior distribution $\pi$ on $\Theta$ is defined as $r(\pi, \delta) = \mathbb{E}^\pi[R(\theta, \delta)]$.

Example 11 (Example 9 cont.)

In Example 9, suppose that $\pi(\theta)$ is a $N(0, \tau^2)$ density. Then, for the decision rule $\delta_c$,

$$r(\pi, \delta_c) = \mathbb{E}^\pi[R(\theta, \delta_c)] = \mathbb{E}^\pi[c^2 + (1 - c)^2 \theta^2] = c^2 + (1 - c)^2 \tau^2.$$
Example 12 (Matching Pennies)

- Persons A and B are to simultaneously uncover a penny (they can secretly turn the penny to heads or tails)
- If the two coin match, A wins $1 from B; otherwise, B wins $1 from A
- Two available actions for A: $a_1$-choose heads, or $a_2$-choose tails
- Two possible states of nature: $\theta_1$-B’s coin is a head, and $\theta_2$-B’s coin is a tail
- Both $a_1$ and $a_2$ are admissible actions (Why?)

If the game is to be played a number of times, a way of preventing ultimate defeat is to choose $a_1$ and $a_2$ with probabilities $p$ and $1 - p$ respectively.
Definition 13 (Randomized Decision Rule)

A *randomized decision rule* \( \delta^*(x, \cdot) \) is, for each \( x \), a probability distribution on \( A \), with the interpretation that if \( x \) is observed, \( \delta^*(x, A) \) is the probability that an action in \( A \) (a subset of \( A \)) will be chosen. In no-data problems, a randomized decision rule, also called a *randomized action*, denoted by \( \delta^*(\cdot) \) is a probability distribution on \( A \).

**Remark:** Nonrandomized decision rules will be considered a special case of randomized rules, in that they correspond to the randomized rules which, for each \( x \), choose a specific action with probability 1. If \( \delta(x) \) is a nonrandomized decision rule, let \( < \delta > \) denote the equivalent randomized rule given by

\[
< \delta > (x, A) = I_A(\delta(x)) = \begin{cases} 
1 & \text{if } \delta(x) \in A, \\
0 & \text{if } \delta(x) \notin A
\end{cases}
\]


Example 14

In Example 12, the randomized action is defined by \( \delta^*(a_1) = p, \delta^*(a_2) = 1 - p \). It is equivalent to be denoted by \( \delta^* = p < a_1 > + (1 - p) < a_2 > \).

Definition 15 (Risk Function for Randomized Decision Rules)

The loss function \( L(\theta, \delta^*(x, \cdot)) \) of the randomized rule \( \delta^* \) is defined to be \( L(\theta, \delta^*(x, \cdot)) = \mathbb{E}^{\delta^*(x, \cdot)}[L(\theta, a)] \), where the expectation is taken over \( a \) (a random variable with distribution \( \delta^*(x, \cdot) \)). The risk function of \( \delta^* \) is \( R(\theta, \delta^*) = \mathbb{E}^X_{\theta} [L(\theta, \delta^*(X, \cdot))] \).

Definition 16

Let \( \mathcal{D}^* \) be the set of all randomized decision rules \( \delta^* \) for which \( R(\theta, \delta^*) < +\infty \) for all \( \theta \). A decision rule will be said to be admissible if there exists no \( R \)-better randomized decision rules in \( \mathcal{D}^* \).
Chapter 1

1 Introduction

2 Basic Elements

3 Expected Loss, Decision Rules, and Risk
   - Bayesian Expected Loss
   - Frequentist Risk

4 Randomized Decision Rules

5 Decision Principles
   - The Conditional Bayes Decision Principle
   - Frequentist Decision Principles

6 Sufficient Statistics

7 Convexity
Theorem 17 (The Conditional Bayes Principle)

Choose an action $a \in A$ which minimize $\rho(\pi^*, a)$ (def 2) (assume the minimum is attained). Such an action will be called a Bayes action and will be denoted by $a^{\pi^*}$.

Example 18 (Example 3 cont.)

In Example 3, when $a = 2/15$, $\rho(\pi, a)$ reaches its minimum $1/30$. Hence the Bayes estimator of $\theta_2$ is $2/15$ assuming no data was available.
Chapter 1

1 Introduction

2 Basic Elements

3 Expected Loss, Decision Rules, and Risk
   - Bayesian Expected Loss
   - Frequentist Risk

4 Randomized Decision Rules

5 Decision Principles
   - The Conditional Bayes Decision Principle
   - Frequentist Decision Principles

6 Sufficient Statistics

7 Convexity
**Theorem 19 (The Bayes Risk Principle)**

A decision rule $\delta_1$ is preferred to a rule $\delta_2$ if $r(\pi, \delta_1) < r(\pi, \delta_2)$. A decision rule which minimizes $r(\pi, \delta)$ is optimal; it is called a Bayes rule and will be denoted by $\delta^\pi$. The quantity $r(\pi) = r(\pi, \delta^\pi)$ is then called the Bayes risk for $\pi$.

**Example 20 (Example 11 cont.)**

In Example 11, $r(\pi, \delta_c)$ is minimized when $c = c_0 \overset{\text{def}}{=} \tau^2/(1 + \tau^2)$, hence $\delta_{c_0}$ is the Bayes rule (or Bayes estimator).

**Remark:** In a no-data problem, the Bayes risk principle will give the same answer as the conditional Bayes decision principle.
Theorem 21 (The Minimax Principle)

A decision rule \( \delta_1^* \) is preferred to a rule \( \delta_2^* \) if \( \sup_\theta R(\theta, \delta_1^*) < \sup_\theta R(\theta, \delta_2^*) \). A rule \( \delta^*_M \) is a minimax decision rule if it minimizes \( R(\theta, \delta^*) \) among all randomized rules in \( \mathcal{D}^* \), i.e,

\[
\sup_\theta R(\theta, \delta^*_M) = \inf_{\delta^* \in \mathcal{D}^*} \sup_\theta R(\theta, \delta^*)(\text{minimax value})
\]

Remark: minimax action (minimax decision rule in no-data problems), minimax nonrandomized rule, minimax nonrandomized action.

Example 22 (Example 9 cont.)

In Example 9, for decision rules \( \delta_c \),

\[
\sup_\theta R(\theta, \delta_c) = \begin{cases} 
1 & \text{if } c = 1 \\
\infty & \text{if } c \neq 1
\end{cases}
\]

The minimax rule is \( \delta_1 \).
Theorem 23 (The Invariance Principle)

If two decision problems have the same formal structure (e.g., \( X, Y = g(X) \)), then the same decision rule should be used in each problem. I.e., let \( \delta, \delta^* \) be the decision rules used in the \( X \) and \( Y \) problems, then \( \tilde{g}(\delta(x)) = \delta^*(y) \). If a decision problem is invariant under a group \( \phi \) of transformations, a (nonrandomized) decision rule \( \delta(x) \) is invariant under \( \phi \) if for all \( x \in \mathcal{H}, g \in \phi, \delta(g(x)) = \tilde{g}(\delta(x)) \).

Remark: For the comparisons of the above principles, please refer to Sec 1.6.
**Definition 24 (Sufficient Statistics)**

Let $X$ be a random variable whose distribution depends on the unknown parameter $\theta$, but is otherwise known. A function $T$ of $X$ is said to be a **sufficient statistic** for $\theta$ if the conditional distribution of $X$, given $T(X) = t$, is independent of $\theta$ (with probability one).

**Definition 25**

If $T(X)$ is a statistic with range $\mathcal{F}$ (i.e. $\mathcal{F} = T(x) : x \in \mathcal{H}$), the partition of $\mathcal{H}$ induced by $T$ is the collection of all sets of the form $\mathcal{H}_t = \{x \in \mathcal{H} : T(x) = t\}$ for $t \in \mathcal{F}$.

**Definition 26**

A **sufficient partition** of $\mathcal{H}$ is a partition induced by a sufficient statistic $T$. 
Theorem 27

Assume that $T$ is a sufficient statistic for $\theta$, and let $\delta_0^*(x, \cdot)$ be any randomized rule in $D^*$. Then (subject to measurability conditions) there exists a randomized rule $\delta_1^*(t, \cdot)$, depending only on $T(x)$, which is $R$-equivalent to $\delta_0^*$. 
Definition 28
A set $Q \subseteq \mathbb{R}^m$ is convex if for any two points $x$ and $y$ in $\Omega$, the point $[\alpha x + (1 - \alpha)y]$ is in $\Omega$ for $0 \leq \alpha \leq 1$. (i.e, the line segment joining $x$ and $y$ is a subset of $\Omega$.)

Definition 29
If $\{x^1, x^2, \cdots \}$ is a sequence of points in $\mathbb{R}^m$, and $0 \leq \alpha_i \leq 1$ are numbers such that $\sum_{i=1}^{\infty} \alpha_i = 1$, then $\sum_{i=1}^{\infty} \alpha_i x^i$ (providing it is finite) is called a convex combination of the $\{x^i\}$. The convex hull of a set $\Omega$ is the set of all points which are convex combinations of points in $\Omega$.

Definition 30
A real valued function $g(x)$ defined on a convex set $\Omega$ is convex if $g(\alpha x + (1 - \alpha)y) \leq \alpha g(x) + (1 - \alpha)g(y)$ for all $x \in \Omega$, and $0 < \alpha < 1$. If the inequality is strict for $x \neq y$, then $g$ is strictly convex. If $g(\alpha x + (1 - \alpha)y) \geq \alpha g(x) + (1 - \alpha)g(y)$ then $g$ is concave. If the inequality is strict for $x \neq y$, then $g$ is strictly concave.
Lemma 31

Let $g(x)$ be a function defined on an open convex subset $\Omega$ of $\mathbb{R}^m$ for which all second-order partial derivatives $g^{(i,j)}(x) = \frac{\partial^2}{\partial x_i \partial x_j} g(x)$ exist and are finite. Then $g$ is convex if and only if the matrix $G = (g^{(i,j)}(x))$ is nonnegative definite for $x \in \Omega$. Likewise, $g$ is concave if $-G$ is nonnegative definite. If $G$ is positive (negative) definite, then $g$ is strictly convex (strictly concave).

Lemma 32

Let $X$ be an $m$-variate random vector such that $E(|X|) < \infty$ and $P(X \in \Omega) = 1$, where $\Omega$ is a convex subset of $\mathbb{R}^m$. Then $E(X) \in \Omega$. 
**Theorem 33 (Jensen’s Inequality)**

Let $g(x)$ be a convex real-valued function defined on a convex subset $\Omega$ of $\mathbb{R}^m$, and let $X$ be an $m$-variate random vector for which $E(|X|) < \infty$. Suppose also that $P(X \in \Omega) = 1$. Then $g(E(X)) \leq E[g(X)]$, with strict inequality if $g$ is strictly convex and $X$ is not concentrated at a point.

**Theorem 34**

Assume that $A$ is a convex subset of $\mathbb{R}^m$, and that for each $\theta \in \Theta$ the loss function $L(\theta, a)$ is a convex function of $a$. Let $\delta^*$ be a randomized decision rule in $D^*$ for which $E^{\delta^*(x, \cdot)}[|a|] < \infty$ for all $x \in \mathcal{H}$. Then (subject to measurability conditions) the nonrandomized rule $\delta(x) = E^{\delta^*(x, \cdot)}[a]$ has $L(\theta, \delta(x)) \leq L(\theta, \delta^*(x, \cdot))$ for all $x$ and $\theta$. 
Theorem 35 (Rao-Blackwell Theorem)

Assume that \( \mathcal{A} \) is a convex subset of \( \mathbb{R}^m \) and that \( L(\theta, a) \) is a convex function of \( a \) for \( \theta \in \Theta \). Suppose also that \( T \) is a sufficient statistic for \( \theta \), and that \( \delta^0(x) \) is a nonrandomized decision rule in \( \mathcal{D} \). Then the decision rule, based on \( T(x) = t \), defined by \( \delta^1(t) = E^X|t[\delta^0(X)] \), is \( R \)-equivalent to or \( R \)-better than \( \delta^0 \), provided the expectation exists.

Proof.

By the definition of a sufficient statistic, the expectation above does not depend on \( \theta \), so that \( \delta^1 \) is an obtainable decision rule. By Jensen’s inequality

\[
L(\theta, \delta^1(t)) = L(\theta, E^X|t[\delta^0(X)]) \leq E^X|t[L(\theta, \delta^0(X))].
\]

Hence

\[
R(\theta, \delta^1) = E_\theta^T[L(\theta, \delta^1(T))] \\
\leq E_\theta^T[E^X|^T[L(\theta, \delta^0(X))]] \\
= E^X[L(\theta, \delta^0(X))] = R(\theta, \delta^0).
\]