# Statistical Decision Theory and Bayesian Analysis
# Chapter1 Basic Concepts

9139511 蕭奕融

1

## Outline

- Introduction
- Basic Elements
- Bayesian Expected Loss
- Frequentist Risk
- Randomized Decision Rules
- Decision Principles
- Misuse of Classical Inference Procedures

2

## Introduction

- Decision theory is concerned with **the problem of making decisions**.
- Statistical decision theory is concerned ith the making of decisions in the presence of statistical knowledge which sheds light on some of the **uncertainties** involved in the decision problem (will be presented by $\theta$ )

3

## Introduction(cont.)

- Classical statistics is directed towards the use of **sample information** in making inference.
- In decision theory, an attempt is made to combine **the sample information** with **loss function** and **prior information**.

4

## Basic Elements

- The unknown quantity $\theta$ which affects the decision process is commonly called the **state of nature**.
- The symbol $\Theta$ will be used to denote **the set of all possible states of nature** and it is also called **the parameter** or **parameter space**.
- Decisions are more commonly called **actions** in the literature. Particular actions will be denoted by **a**, while the set of all possible actions under consideration will be denoted $\mathcal{A}$

5

## Basic Elements (cont.)

- Loss function $L(\theta, a)$ is defined for all $(\theta, a) \in \Theta \times \mathcal{A}$ . For technical convenience, only loss functions satisfying $L(\theta, a) \geq -K > -\infty$
- Outcome (Observations) will be denoted X (that's a vector), the possible outcomes is the sample space, and will be denoted $\aleph$
- Denote $P_\theta(A)$ the probability of the event A.

6

## Bayesian Expected Loss

- Definition 1.

  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

  $$\rho(\pi^*, a) = E^{\pi^*} L(\theta, a) = \int_{\Theta} L(\theta, a) dF^{\pi^*}(\theta)$$



## Frequentist Risk

- Definition 2.

  A decision rule $\delta(x)$ is a function from $\aleph$ into $\mathcal{A}$. If X=x is the observed value of the sample information, then $\delta(x)$ is the action that will be taken. Two decision rules, $\delta_1(x)$ and $\delta_2(x)$, are considered equivalent if $P_\theta(\delta_1(X) = \delta_2(X)) = 1$ for all $\theta$



## Frequentist Risk(cont.)

- Definition 3.

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

  $$R(\theta, \delta) = E_\theta^X[L(\theta, \delta(X))] = \int_{\aleph} L(\theta, \delta(x)) dF^X(x \mid \theta)$$

  example

- Definition 4.

  A decision rule $\delta_1$ is **R-better** than a decision rule $\delta_2$ if $R(\theta, \delta_1) \le R(\theta, \delta_2)$ 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 $\theta \in \Theta$



## Frequentist Risk(cont.)

- Definition 5.

  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.

| | $a_1$ | $a_2$ | $a_3$ |
|---|---|---|---|
| $\theta_1$ | 1 | 3 | 4 |
| $\theta_2$ | -1 | 5 | 5 |
| $\theta_3$ | 0 | -1 | -1 |



## Frequentist Risk(cont.)

$D$ ={all decision rules $\delta$: $R(\theta, \delta) < \infty$ for all $\theta \in \Theta$}.

- Definition 6.

  The Bayes risk of a decision rule $\delta$, with respect to a prior distribution $\pi$ on $\Theta$, is defined as

  $$r(\pi, \delta) = E^\pi[R(\theta, \delta)]$$



## Randomized Decision Rules

- Definition 7.

  A **randomized decision rule** is $\delta^*(x, \bullet)$, for each x, a probability distribution on $\mathcal{A}$, with the interpretation that if x is observed, $\delta^*(x, A)$ is the probability that an action in A will be chosen.

## Randomized Decision Rules( cont.)

- Definition 8.

  The **loss function** $L(\theta,\delta^*(x,\bullet))$ of the randomized rule $\delta^*$ is defined to be

  $$L(\theta,\delta^*(x,\bullet)) = E^{\delta^*(x,\bullet)}[L(\theta,a)]$$

  The **risk function** of $\delta^*$ will then be defined to be

  $$R(\theta,\delta^*) = E_\theta^X L(\theta,\delta^*(x,\bullet))$$

13

## Randomized Decision Rules(cont.)

- Matching Pennies

| | $a_1$ | $a_2$ |
|---|---|---|
| $\theta_1$ | -1 | 1 |
| $\theta_2$ | 1 | -1 |

- Definition 9.

  Let $D^*$ be the set of all randomized decision rules $\delta^*$ for which $R(\theta,\delta) < \infty$ for all $\theta \in \Theta$ . The concepts introduced in **Definitions 4 and 5** will hence forth be considered to apply to all randomized rules in $D^*$ .

14

## Decision Principles

- **The Conditional Bayes Principle**

  Choose an action $a \in A$ which minimizes $\rho(\pi^*,a)$. Such an action will be called a **Bayes action** and will be denoted $a^{\pi^*}$.
- **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 $\delta^\pi$.The quantity $r(\pi) = r(\pi,\delta^\pi)$ is then called the **Bayes risk** for $\pi$ .

15

## Decision Principles(cont.)

- **The Minimax Principle**

  A decision rule $\delta_1^*$ is preferred to a rule $\delta_2^*$ if

  $$\sup_{\theta\in\Theta} R(\theta,\delta_1^*) < \sup_{\theta\in\Theta} R(\theta,\delta_2^*)$$

- Definition 10.

  A rule $\delta^{*M}$ is a **minimax decision rule** if it minimizes $\sup_{\theta\in\Theta} R(\theta,\delta^*)$ among all randomized rules in $D^*$,if

  $$\sup_{\theta\in\Theta} R(\theta,\delta^{*M}) < \inf_{\delta^*\in D^*} \sup_{\theta\in\Theta} R(\theta,\delta^*)$$

16

## Decision Principles(cont.)

- **The Invariance Principle**

  The invariance principle basically states that if two problems have **identical formal structured**,then the same decision rule should be used in each problem.That leads to a restriction to so-called **"invariant" decision rules**.

17

## Misuse of Classical Inference Procedures

- For a **large enough sample size**, the classical test will be virtually certain to reject,because the point null hypothesis is **almost certainly not exactly true**, and that this will always be confirmed by a large enough sample.
- In the other hand, a **"statistically significant"** difference between the true parameter and the null hypothesis can be **an unimportant difference practically**.

Example:

$X_1, X_2, \dots, X_n \sim N(\theta,1)$ $\qquad \dfrac{\overline{X}-\theta_0}{1/\sqrt{n}}$

18

## Misuse of Classical Inference Procedures

- **The Frequentist Perspective:** it's risk funtion has not eliminated dependence on $\theta$

  $R(\theta,\delta)=E_\theta^X[L(\theta,\delta(X))]=\int_{\aleph} L(\theta,\delta(x))dF^X(x\mid\theta)$

- **The Conditional Perspective :**

At level $\alpha=0.01$ , rejection region: R= $\{X:1,2\}$

$H_0:\theta=0$

$H_1:\theta=1$

| | | x | |
|---|---|---|---|
| | 1 | 2 | 3 |
| $f(x\mid 0)$ | 0.005 | 0.005 | 0.99 |
| $f(x\mid 1)$ | 0.0051 | 0.9849 | 0.01 |



## Misuse of Classical Inference Procedures

- The standard frequentist ,observing x=1, would report that the decision is H1 and that the test had **error probabilities** of 0.01.
- But the likelihood ratio is very close to 1. To a conditionalist that means the data does very little to distinguish $\theta=0$ and $\theta=1$.

$$\lambda(X)=\frac{\sup_{\theta\in\Theta_0} L(\theta\mid X)}{\sup_{\theta\in\Theta} L(\theta\mid X)}=\frac{0.005}{0.0051}$$



## The Likelihood Principle

- **The Likelihood Principle:** In making inferences or decisions about $\theta$ after x is observed, all relevant experimental information is contained in the likelihood function for the observed x. Furthermore **two likelihood functions contain the same information** $\theta$ **about** if they are proportional to each other.

$$L_1(\theta\mid X)\propto L_2(\theta\mid X)$$



## Sufficient Statistics

- **Definition:** 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 $\theta$ .
- **Theorem:** Assume that T is a sufficient statistic for $\theta$ ,and let $\delta_0^*(t,\bullet)$ be any randomized rule in $D^*$ . Then there exists a randomized rule $\delta_1^*(t,\bullet)$, depending only on T(x), which is R-equivalent to $\delta_0^*$ .



## Convexity

- Theorem: Assume that $A$ is a convex subset of $R^m$ ,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 $x\in\chi$ . Then the nonrandomized rule

$$\delta(x)=E^{\delta^*(x,\cdot)}[a]$$

has $L(\theta,\delta(x))\le L(\theta,\delta^*(x))$ for all x and $\theta$



## Conclusion

- Baysian analysis combine the sample information with loss function and prior information.
- It can improve the defects of classical statistic.

## example

- Consider the situation of a drug company deciding whether or not to market a new pain reliever. Two of the many factors affecting its decision are the proportion of people for for which the drug will prove **effective** $\theta_2$,and the **proportion** of the market the drug will capture $\theta_2$.both $\theta_1$ and $\theta_2$ will be generally unknown.
- Loss function:

$$L(\theta_2, a) = \{ \begin{matrix} \theta_2 - a & ,\theta_2 - a > 0 \\ 2(a - \theta_2), & \theta_2 - a \le 0 \end{matrix}$$

25

## example

- Assume n people are interviewed , and the number X who would buy the drug is observed. It might be reasonable to assume that X is $\mathrm{B}(n, \theta_2)$

$$f(x \mid \theta_2) = \binom{n}{x} \theta_2^x (1 - \theta_2)^{n-x}$$

- By historical record, assume prior distribution

$$\pi(\theta_2) = 10 I_{(0.1, 0.2)}(\theta_2)$$

26

## example

- **Bayesian expected loss of an action**

$$\rho(\pi^*, a) = E^{\pi^*} L(\theta, a) = \int_0^1 L(\theta_2, a)\pi(\theta_2) d\theta_2$$

$$= \begin{cases} 0.15 - a & , a \le 0.1 \\ 15a^2 - 4a + 0.3, & 0.1 \le a \le 0.2 \\ 2a - 0.3 & a \ge 0.2 \end{cases}$$

- **Frequentist Risk**

$$\delta(x) = x / n$$

$$R(\theta, \delta(x)) = E_\theta^X [L(\theta, \delta(X))]$$

27