

Decision theory and Bayesian Inference

決策理論

- Set forth a unifying framework for theory of statistic. ← including estimation & testing
- Deal in a systematic way with problems that are not amenable to analysis by traditional methods such as estimation and testing e.g. classification

Example 10.1 (Sampling Inspection, 2nd Ed., TBp. 572)

- A lot of N items, n ($n < N$) of which are sampled randomly and determined to be either defective or nondefective. n, N : known.

unknown p : proportion of the N items that are defective (parameter) → $0 \leq p \leq 1$

– \hat{p} : proportion of the n items that are defective (observed data, the distribution of \hat{p} depends on p) X : # of observed defeated items → $\hat{p} = X/n$
 r.v. \sim hypergeometric/binomial L.r.v.

- For any lot, the manufacturer has two possible actions: depending on p .
 ◯ sell the lot, for a price $\$M$ with a guarantee that if $p > p_0$ the manufacturer will pay a $\$P$ penalty, ← return policy unknown when sell it. a fixed known value.
 ◯ junk it at a cost $\$C$.

- The loss function is

i.e., minimize loss

	State of Nature	Sell	Junk
<u>unknown when taking action</u>	$p \leq p_0$	$-\$M$	$\$C$
	$p > p_0$	$\$P$	$\$C$

- **Question:** For best profit, how to make a decision (sell or junk) based on the observed data \hat{p} ?

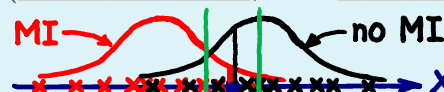
→ can have more than 2 classes

Example 10.2 (Classification, 2nd Ed., TBp. 572)

← cf. hypotheses testing problem

- On the basis of several physiological measurements, a decision must be made concerning whether a patient has suffered a myocardial infarction (MI) and should be admitted to intensive care

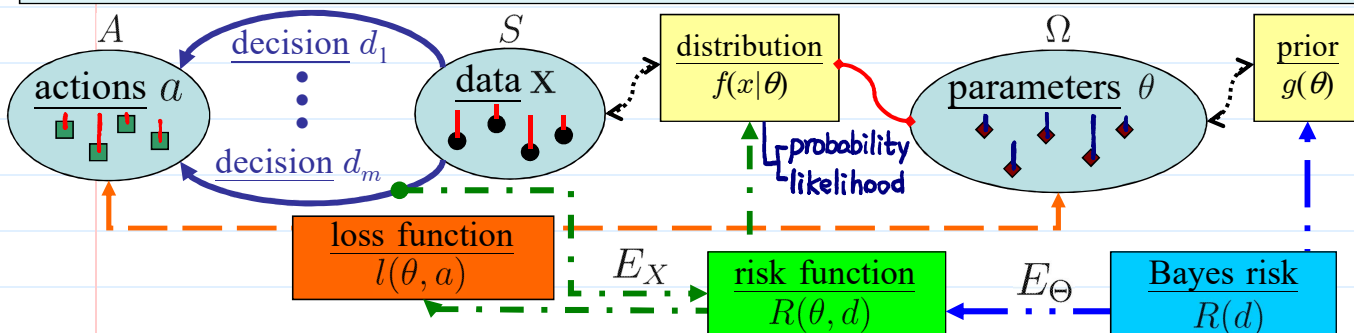
– patient status: {MI, no MI} (parameter) ← unknown

decide X : physiological measurements (observed data, the distribution of X depends on parameter) e.g. 
 ◯ actions: {admit, not admit}

- loss function:

	patient status	admin	not admit
<u>unknown</u>	MI	correct	subjective
	no MI	economic costs	correct

- **Question:** How to make a decision (admin or not admin) based on the observed data X so that the loss can be minimized?



Summary and Definitions (Decision Theory, 2nd Ed., TBp.571)

- Making decisions in the face of uncertainty.

– a : an **action**
 – A : **action space**, the set of all possible actions

$a \in A$

- The decision maker chooses an action based on
 - observation of random variables, or **data**, \mathbf{X} .
 - $\mathbf{X} \in S$ (S : **sample space**, set of all possible data values).
 - d : **statistical decision function** (decision rule), a map from S to A .

a transformation of \mathbf{X}

- The probability distribution of \mathbf{X} depends on
 - a **parameter** θ , called the **state of nature**
 - $\theta \in \Omega$ (Ω : **parameter space**, the set of all possible values of θ)

- A **loss function**, $l(\theta, a)$, is a real function defined on $\Omega \times A$.

- Under a decision function d , by taking the action $a = d(\mathbf{X})$, the decision maker incurs a **loss**

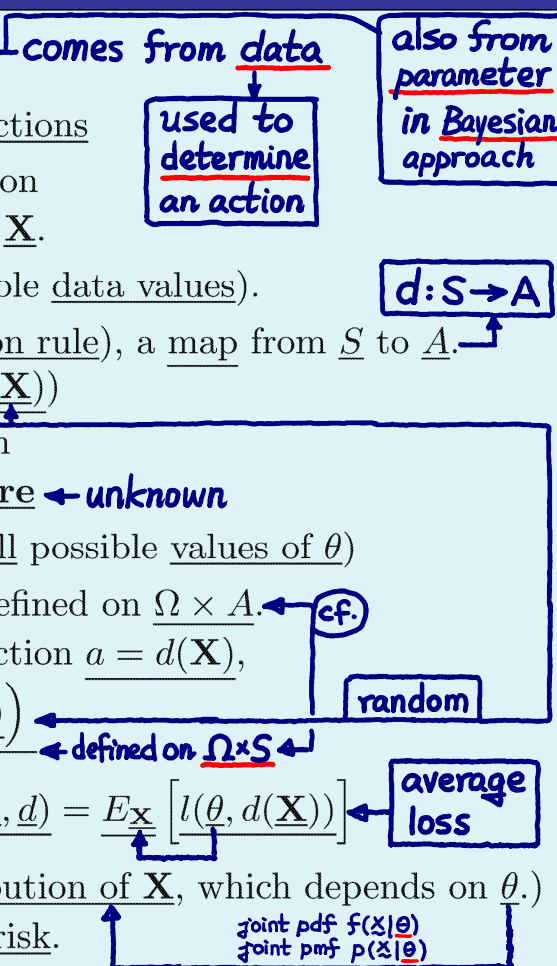
- The expected loss of a decision d is called **risk function**:

$$l(\theta, d(\mathbf{X}))$$

$$R(\theta, d) = E_{\mathbf{X}} [l(\theta, d(\mathbf{X}))]$$

(Note: the expectation is taken w.r.t. distribution of \mathbf{X} , which depends on θ .)

- Good decision function d should have small risk.



Example 10.3 (Estimation, 2nd Ed., TBp. 573)

- $\mathbf{X} = (X_1, X_2, \dots, X_n)$: a random sample, distribution of \mathbf{X} depends on θ .
- Goal: estimate $\nu(\theta)$
- $d(\mathbf{X})$: an estimator of $\nu(\theta)$ (Note. $A = \Omega$, and $d(\mathbf{X}) : S \rightarrow \Omega$)
- quadratic loss function:

$$l(\theta, d(\mathbf{X})) = [\nu(\theta) - d(\mathbf{X})]^2$$

- The risk function is then

$$R(\theta, d) = E_{\mathbf{X}} [\nu(\theta) - d(\mathbf{X})]^2$$

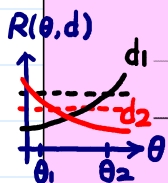
which is the mean squared error.

Recall. UMVUE:
 minimizes MSE
 among d : unbiased

Question 10.1

- A good decision function d is one that has a small risk smallest risk $\forall \theta \in \Omega$

$$R(\theta, d) = E_{\mathbf{X}} [l(\theta, d(\mathbf{X}))]$$



– Difficulty 1: $R(\theta, d)$ depends on θ , which is unknown.

– Difficulty 2: There might be two decisions, d_1 and d_2 , and two values of θ , θ_1 and θ_2 , such that

$$R(\theta_1, d_1) < R(\theta_1, d_2)$$

$$R(\theta_2, d_1) > R(\theta_2, d_2)$$

Hint: Transform a function of θ into a single value

Question: how to confront the difficulties to make a good choice of d ?

Definition 10.1 (Minimax Rule, 2nd Ed., TBp.573)

1. For a given decision function d , consider the worst that the risk could be

a function of $\theta \rightarrow$ a single value $\rightarrow \max_{\theta \in \Omega} R(\theta, d) \leftarrow$ largest average loss of d

2. Choose a decision function d^* that minimizes this maximum risk

Why not taking average?
or weighed average?

$$\min_d \left[\max_{\theta \in \Omega} R(\theta, d) \right]$$

a decision function.

3. Such a decision function d^* , if exists, is called a minimax rule.

Notes (Minimax Rule)

- weakness: very **conservative**, places all emphasis on guarding against the worst possible case, which may not be very likely to occur.
- A rigorous statement is to change $\max \rightarrow \sup$, $\min \rightarrow \inf$

Θ (parameter) : unknown
Fixed value
random variable

1. Consider Ex10.1 in LNp.1. What if there are many lots? Each lot has its own p , and these p 's can be different. \Rightarrow The p of a randomly chosen lot is a random variable.

2. Extra information. From past experience, we may know what value of p is more possible to appear. How to include this information in analysis?

not observe when taking action

Definition 10.2 (Prior Distribution, Bayes Risk, Bayes Rule, 2nd Ed., TBp.574)

1. Assign a probability distribution, called prior distribution, to θ .
 Θ : a random element of Ω drawn according to the prior distribution

Bayesian \leftarrow SF
Frequentist \leftarrow

2. The Bayes risk of a decision function d is

a function of $\theta \rightarrow$ a single value of $d \rightarrow B(d) = E_{\Theta} [R(\Theta, d)] \leftarrow$ average risk, with weight from

where the expectation is taken with respect to the prior distribution of Θ .

3. Bayes rule: a decision function d^{**} that minimizes the Bayes risk $B(d)$.

Notes (Bayes risk)

- Bayes risk can be interpreted as the average of the risks with respect to the prior distribution of θ . \rightarrow check the graph in LNp.2.
- Bayes risk is a function of decision d only, not depending on θ .

Example 10.4 (steel section of firm stratum, 2nd Ed., TBp. 574-575)

• Benjamin and Cornell (1970): As part of the foundation of a building, a steel section is to be driven down to a firm stratum below ground.

• Two possible states of nature (parameters):

θ_1 : depth of firm stratum is 40 ft; θ_2 : depth of firm stratum is 50 ft

• Two possible actions:

a_1 : select a 40-ft section; a_2 : select a 50-ft section

• loss function $l(\theta, a)$:

	a_1	a_2
θ_1	\$0	\$100
θ_2	\$400	\$0

unknown \rightarrow

- Data:

X = depth measured by a sonic test, which has probability distribution:

	X	θ_1 (<u>40 ft</u>)	θ_2 (<u>50 ft</u>)
$S \leftarrow$	$x_1 = \underline{40}$	<u>0.6</u>	0.1
	$x_2 = \underline{45}$	0.3	0.2
	$x_3 = \underline{50}$	0.1	<u>0.7</u>

Sum = 1

- Consider the following four decision rules:

		x_1 (<u>=40</u>)	x_2 (<u>=45</u>)	x_3 (<u>=50</u>)
<u>always 40-ft section</u> →	d_1	a_1	a_1	a_1
do not use the information in data	d_2	a_1	a_2	a_2
	d_3	a_1	a_1	a_2
<u>always 50-ft section</u> →	d_4	a_2	a_2	a_2

$d: S \rightarrow A$
 \parallel
 $\{a_1, a_2\}$

- Risk functions: For $j = 1, 2$, and i -th decision function d_i , $i = 1, \dots, 4$,

$$R(\theta_j, d_i) = E_X [l(\theta_j, d_i(X))] = \sum_{k=1}^3 l(\theta_j, d_i(x_k)) P(X = x_k | \theta = \theta_j).$$

- Thus
 - $R(\theta_1, d_1) = 0 \times 0.6 + 0 \times 0.3 + 0 \times 0.1 = 0$
 - $R(\theta_1, d_2) = 0 \times 0.6 + 100 \times 0.3 + 100 \times 0.1 = 40$
 - $R(\theta_1, d_3) = 0 \times 0.6 + 0 \times 0.3 + 100 \times 0.1 = 10$
 - $R(\theta_1, d_4) = 100 \times 0.6 + 100 \times 0.3 + 100 \times 0.1 = 100$

Similarly, $R(\theta_2, d_1) = 400$, $R(\theta_2, d_2) = 40$, $R(\theta_2, d_3) = 120$, $R(\theta_2, d_4) = 0$.

- Minimax rule: d_2 is the minimax rule since

$$\begin{aligned} \max_j R(\theta_j, d_1) &= 400, & \max_j R(\theta_j, d_2) &= 40, \\ \max_j R(\theta_j, d_3) &= 120, & \max_j R(\theta_j, d_4) &= 100 \end{aligned}$$

- Bayes rules: θ (fixed unknown constant) → Θ (random variable)

may come from past experience

- prior distribution: $g(\theta_1) = 0.8$, $g(\theta_2) = 0.2$.
- Bayes risk: $B(d) = E_{\Theta} [R(\Theta, d)] = R(\theta_1, d)g(\theta_1) + R(\theta_2, d)g(\theta_2)$
- Thus,

becomes r.v.

$$\begin{aligned} B(d_1) &= 0 \times 0.8 + 400 \times 0.2 = 80 \\ B(d_2) &= 40 \times 0.8 + 40 \times 0.2 = 40 \\ B(d_3) &= 10 \times 0.8 + 120 \times 0.2 = 32 \\ B(d_4) &= 100 \times 0.8 + 0 \times 0.2 = 80 \end{aligned}$$

Q: What if $\begin{cases} g(\theta_1) = 0.2 \\ g(\theta_2) = 0.8 \end{cases}$?

- Thus d_3 is the Bayes rule corresponding to that prior. ← reasonable?

Example 10.5 (sampling inspection, 2nd Ed., TBp. 576-577)

- A manufacturer produces items in lots of 21. One item is selected at random and tested to determine whether or not it is defective.
- Two possible actions: (1) sell the remaining 20 items at \$1 per item with a double-your-money-back guarantee on each item, or (2) junk the whole lot at a cost of \$1.

m : # of defeated items in the remaining 20 items
 unknown →

	<u>sell</u>	<u>junk</u>
<u>loss</u>	$-20 + 2m$	1

- parameter: $k = \text{number of defectives in a lot of } 21 \rightarrow k = 0, 1, 2, \dots, 21$
- data: $X = \begin{cases} 1, & \text{if the tested item is good} \\ 0, & \text{if the tested item is defective} \end{cases} \Rightarrow m = \begin{cases} k-1, & \text{if } X=0, \\ k, & \text{if } X=1. \end{cases}$
- For a given value of k , the distribution of X is Bernoulli($1 - k/21$):

$$P(X = 0 | k) = k/21, \quad P(X = 1 | k) = 1 - (k/21).$$

- Consider the following two decisions: $d: S \rightarrow A$
 - d_1 : sell if tested item is good, junk if defective; d_2 : sell in either case

The loss are

$$l(k, d_1(X)) = \begin{cases} -20 + 2 \times k, & \text{if } X = 1, \rightarrow m=k \\ 1, & \text{if } X = 0. \end{cases}$$

$$l(k, d_2(X)) = \begin{cases} -20 + 2 \times k, & \text{if } X = 1, \rightarrow m=k \\ -20 + 2 \times (k - 1), & \text{if } X = 0. \rightarrow m=k-1 \end{cases}$$

- The risk functions are $R(k, d_i) = E_X[l(k, d_i(X))]$.

$$R(k, d_1) = (-20 + 2k)[1 - (k/21)] + 1 \times (k/21) = -20 + 3k - (2k^2/21) \leftarrow \text{quadratic polynomial}$$

$$R(k, d_2) = (-20 + 2k)[1 - (k/21)] + [-20 + 2(k - 1)](k/21) = -20 + (40/21)k \leftarrow \text{linear polynomial}$$

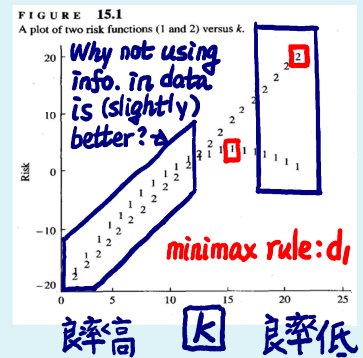


Figure 15.1: the two risk functions. d_1 is the minimax rule.

- Suppose the prior distribution of K is Binomial($21, p$). Then $\leftarrow \text{probability of generating a defective item. assume known}$

$$E(K) = 21p, \quad Var(K) = 21p(1 - p), \quad E(K^2) = 21p(1 - p) + (21p)^2$$

- The Bayes risks of d_1 and d_2 are

$$B(d_1) = -20 + 3 \times E(K) - (2/21)E(K^2) = -20 + 3 \times 21p - (2/21)[21p(1 - p) + (21p)^2] = -20 + 61p - 40p^2 \leftarrow \text{quadratic polynomial}$$

$$B(d_2) = -20 + (40/21) \times 21p = 40p - 20 \leftarrow \text{linear polynomial}$$

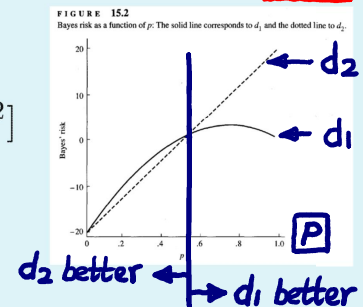


Figure 15.2: Bayes risks versus p , d_2 has a smaller Bayes risk as long as $p \leq 0.5$. (If the product is fairly reliable, may prefer d_2 .)

Reading: textbook (2nd ed.), 15.1, 15.2, 15.2.1 5/26

Posterior Analysis --- A simple method for finding Bayes rule

Definition 10.3 (Posterior Distribution and Posterior Risk, 2nd Ed., TBP.578-579)

- In Bayesian procedures, we have

Θ : a random variable with a pdf/pmf $g_\Theta(\theta)$

$g_\Theta(\theta)$: prior distribution of Θ \leftarrow 事前分配

$f_{X|\Theta}(x|\theta)$: pdf/pmf of X , conditional on the value θ of Θ

In estimation (CH8) and testing (CH9), the joint pdf/pmf of X (data)

Frequentist approach

not observed
Conditioned on $\Theta = \theta$
random variable Θ
 \rightarrow a fixed value θ