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)

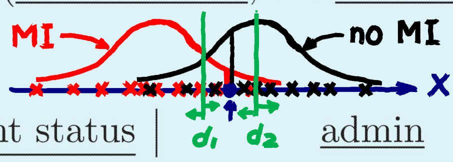
- A lot of N items, n ($n < N$) of which are sampled randomly and determined to be either defective or nondefective. n, N : known
  - unknown → p: proportion of the N items that are defective (parameter) ← $0 \le p \le 1$
  - known → $\hat{p}$: proportion of the n items that are defective (observed data, the distribution of $\hat{p}$ depends on p) X: # of observed defeated items ⇒ $\hat{p} = X/n$ ← r.v.; r.v. ~ hypergeometric/binomial depending on p.
- For any lot, the manufacturer has two possible actions: decide
  - **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) (p_0: a fixed known value)
  - **junk** it at a cost $\$C$.
- The loss function is (i.e., minimize loss)

State of Nature (unknown when taking action)	Sell	Junk
$p \le p_0$	$-\$M$	$\$C$
$p > p_0$	$\$P$	$\$C$

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

Example 10.2 (Classification, 2nd Ed., TBp. 572) ← cf. → hypotheses testing problem

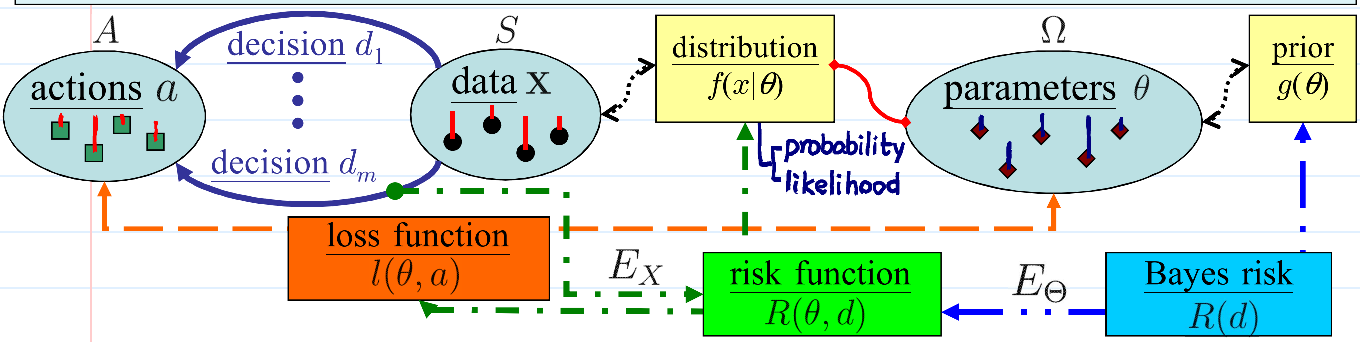
- On the basis of several physiological measurements, a decision must be made concerning whether a patient has suffered a myocardial infarction (MI) and should be admitted to intensive care
  - patient status: {MI, no MI} (parameter) ← unknown
  - X: physiological measurements (observed data, the distribution of X depends on parameter) (r.v.) e.g.



- decide → actions: {admit, not admit}
- loss function:

patient status (unknown)	admin	not admit
MI	correct	subjective
no MI	economic costs	correct

- Question: How to make a decision (admin or not admit) 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. ← comes from data → used to determine an action
  - a: an **action** — $a \in A$
  - A: **action space**, the set of all possible actions
- 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. — $d: S \to A$

    (Note: $d(\mathbf{X})$ is *random*, given as $a = d(\mathbf{X})$) — a transformation of X
- The probability distribution of $\mathbf{X}$ depends on
  - a **parameter** θ, called the **state of nature** ← unknown
  - $\theta \in \Omega$ (Ω: **parameter space**, the set of all possible values of θ)

also from parameter in Bayesian approach

- 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** $\quad l\left(\theta, d(\mathbf{X})\right)$ ← random

  random ($\because X$ is random)
- The expected loss of a decision d is called **risk function**: $\quad R(\theta, d) = E_{\mathbf{X}}\left[l(\theta, d(\mathbf{X}))\right]$ ← average loss

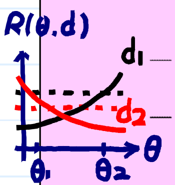
  (Note: the expectation is taken w.r.t. distribution of $\mathbf{X}$, which depends on θ.) — joint pdf $f(\underset{\sim}{x}|\theta)$, joint pmf $p(\underset{\sim}{x}|\theta)$
- Good decision function d should have small risk.

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

- $\mathbf{X} = (X_1, X_2, \ldots, X_n)$: a random sample, distribution of $\mathbf{X}$ depends on θ. ← unknown parameter
- Goal: estimate $\nu(\theta)$ — decision function
- $d(\mathbf{X})$: an estimator of $\nu(\theta)$ (Note. $A = \Omega$, and $d(\mathbf{X}) : S \to \Omega$) — $d: S \to A$
- quadratic loss function:
$$l(\theta, d(\mathbf{X})) = \left[\nu(\theta) - d(\mathbf{X})\right]^2.$$
- The risk function is then
$$R(\theta, d) = E_{\mathbf{X}}\left[\nu(\theta) - d(\mathbf{X})\right]^2$$
  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 $\forall \theta \in \Omega$
$$R(\theta, d) = E_{\mathbf{X}}[l(\theta, d(\mathbf{X}))]$$
  on any $\theta \in \Omega$



  - 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

$$\boxed{\text{a function of } \theta \to \text{a single value}} \longrightarrow \max_{\theta\in\Omega} R(\theta, d) \longleftarrow \text{largest average loss of } d$$

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

Why not taking average? or weighed average?

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

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

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 → sup, min → 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. ⇒ 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 possibe to appear. How to include this information in analysis?

not observe when taking action

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

1. Assign a probability distribution, called **prior distribution**, to θ. ← Bayesian cf. Frequentist

   $\boldsymbol{\Theta}$: a random element of Ω drawn according to the prior distribution

2. The **Bayes risk** of a decision function d is

$$\boxed{\begin{array}{l}\text{a function of } \theta \\ \to \text{a single value of } d\end{array}} \to B(d) = E_{\boldsymbol{\Theta}}\left[R(\boldsymbol{\Theta}, d)\right], \longleftarrow \text{average risk, with weight from}$$

   where the expectation is taken with respect to the prior distribution of $\boldsymbol{\Theta}$.

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. → 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 → θ_1, θ_2

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

X	θ_1 (40 ft)	θ_2 (50 ft)
$x_1 = 40$	0.6	0.1
$x_2 = 45$	0.3	0.2
$x_3 = 50$	0.1	0.7

$S \Leftarrow \{x_1, x_2, x_3\}$ sum = 1

- Consider the following four decision rules:

	x_1 (=40)	x_2 (=45)	x_3 (=50)
always 40-ft section → d_1	a_1	a_1	a_1
d_2	a_1	a_2	a_2
d_3	a_1	a_1	a_2
always 50-ft section → d_4	a_2	a_2	a_2

do not use the information in data

$d: S \to A = \{a_1, a_2\}$

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

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

- Thus

$$\begin{aligned}
R(\theta_1, d_1) &= 0 \times 0.6 + 0 \times 0.3 + 0 \times 0.1 = 0\\
R(\theta_1, d_2) &= 0 \times 0.6 + 100 \times 0.3 + 100 \times 0.1 = 40\\
R(\theta_1, d_3) &= 0 \times 0.6 + 0 \times 0.3 + 100 \times 0.1 = 10\\
R(\theta_1, d_4) &= 100 \times 0.6 + 100 \times 0.3 + 100 \times 0.1 = 100
\end{aligned}$$

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

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

- Bayes rules: θ (fixed unknown constant) → Θ (random variable)
  - prior distribution: $g(\theta_1) = 0.8$, $g(\theta_2) = 0.2$. ← may come from past experience
  - Bayes risk: $B(d) = E_\Theta[R(\Theta, d)] = R(\theta_1, d)g(\theta_1) + R(\theta_2, d)g(\theta_2)$
  - Thus, (Θ becomes r.v.)

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

cf. Q: What if $g(\theta_1) = 0.2$, $g(\theta_2) = 0.8$?

  - 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)

	sell	junk
loss	$-20+2m$	1

- parameter: k = number of defectives in a lot of 21 → $k = 0, 1, 2, \dots, 21$
- 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$(1 - k/21)$:

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

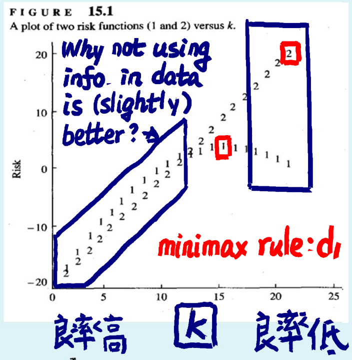
- Consider the following two decisions: $d: S \to A$
  - d_1: sell if tested item is good, junk if defective;
  - d_2: sell in either case ← not use the information in data
- The loss are

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

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

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

$$\begin{aligned} R(k, d_1) &= (-20 + 2k)[1 - (k/21)] + 1 \times (k/21) \\ &= -20 + 3k - (2k^2/21) \leftarrow \text{quadratic polynomial} \\ R(k, d_2) &= (-20 + 2k)[1 - (k/21)] \\ &\quad + [-20 + 2(k-1)](k/21) \\ &= -20 + (40/21)k \leftarrow \text{linear polynomial} \end{aligned}$$



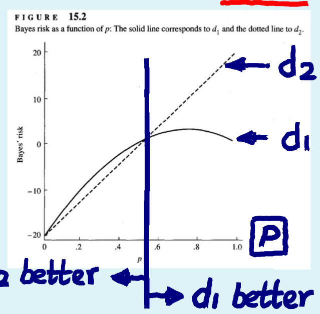
- **Figure 15.1**: the two risk functions. d_1 is the minimax rule.

- Suppose the prior distribution of K ($\boxed{\text{r.v.}}$) is Binomial$(21, p)$. Then, (p: 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 ($\boxed{\theta \text{ is gone}}$)

$$\begin{aligned} B(d_1) &= -20 + 3 \times E(K) - (2/21)E(K^2) \\ &= -20 + 3 \times 21p - (2/21)[21p(1-p) + (21p)^2] \\ &= -20 + 61p - 40p^2 \leftarrow \text{quadratic polynomial} \\ B(d_2) &= -20 + (40/21) \times 21p \\ &= 40p - 20 \leftarrow \text{linear polynomial} \end{aligned}$$



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

❖ **Reading**: textbook (2nd ed.), 15.1, 15.2, 15.2.1 $\boxed{5/20}$

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

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

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

- In Bayesian procedures, we have
  - Θ: a random variable with a pdf/pmf $g_\Theta(\theta)$ (θ a fixed value)
  - $g_\Theta(\theta)$: prior distribution of Θ ← 事前分配
  - $f_{X|\Theta}(x|\theta)$: pdf/pmf of X, conditional on the value θ of Θ ← not observed; Conditioned on $\Theta = \theta$: random variable Θ → a fixed value θ