Mathematical Statistics

Guo-Liang TIAN

Department of Mathematics,
Southern University of Science and Technology,
Shenzhen, Guangdong Province, P.R. China

Xuejun JIANG

Department of Mathematics,
Southern University of Science and Technology,
Shenzhen, Guangdong Province, P.R. China

Yin LIU

School of Statistics and Mathematics,
Zhongnan University of Economics and Law,
Wuhan, Hubei Province, P.R. China

Science Press, Beijing, P.R. China

1 March 2018

To Yanli, Margaret and Adam

To Zhengzhuo

To Wei and Qinyi

Contents

Chapter 1

Probability and Distributions

1.1 Probability

1.1.1 Permutation, combination and binomial coefficients

1• Several notions

1.1• Factorial

— We represent the product $n(n-1)(n-2)\cdots 3\cdot 2\cdot 1$ by the symbol $n!$, which is read "n factorial."

1.2• Permutation

— The number of permutations of n distinct objects taken r at a time is

$$ {}_nP_r = n(n-1)\cdots(n-r+1) = \frac{n!}{(n-r)!}, \quad r = 0, 1, 2, \ldots, n. $$

1.3• Combination

— The number of combinations of n distinct objects taken r at a time is

$$ \binom{n}{r} = \frac{n(n-1)\cdots(n-r+1)}{r!} = \frac{n!}{r!(n-r)!}, \quad r = 0, 1, 2, \ldots, n. $$

1.4• Binomial coefficient

— The binomial coefficient of the term $x^r y^{n-r}$ in the expansion of

$$ (x+y)^n = \sum_{r=0}^{n} \binom{n}{r} x^r y^{n-r} $$

is $\binom{n}{r}$, where n is a positive integer and r is a non-negative integer less than or equal to n.

1.5• Multinomial coefficient

— The number of ways in which a set of n distinct objects can be partitioned into k subsets with n_1 objects in the first subset, n_2 objects in the second subset, . . . ,and n_k objects in the k-th subset is

$$\binom{n}{n_1, \ldots, n_k} = \frac{n!}{n_1! \cdots n_k!},$$

which is the multinomial coefficient of the term $x_1^{n_1} \cdots x_k^{n_k}$ in the expansion of $(x_1 + \cdots + x_k)^n$, where $n_1 + \cdots + n_k = n$.

2• Some useful formulae

- $\binom{n}{r} = \binom{n}{n-r}$.
- $\binom{n}{r} = \binom{n-1}{r} + \binom{n-1}{r-1}$.
- When $n = x$ is not a positive integer or zero, the generalized binomial coefficient is defined by

$$\binom{x}{r} = \frac{x(x-1)\cdots(x-r+1)}{r!}.$$

- $\binom{-x}{r} = (-1)^r \binom{x+r-1}{r}$ for any $x > 0$.
- $\binom{n}{r_1, \ldots, r_k} = \binom{n}{r_1}\binom{n-r_1}{r_2} \cdots \binom{n-r_1-\cdots-r_{k-1}}{r_k}$.

Example 1.1 (Some important identities). By equating the coefficients of x^n in the expressions on both sides of the equation

$$(1+x)^{a+b} = (1+x)^a (1+x)^b, \tag{1.1}$$

we have

$$\binom{a}{0}\binom{b}{n}+\binom{a}{1}\binom{b}{n-1}+\cdots+\binom{a}{n}\binom{b}{0}=\binom{a+b}{n}. \tag{1.2}$$

Especially, in (1.2) let $a=b=n$, we obtain

$$\binom{n}{0}\binom{n}{n}+\binom{n}{1}\binom{n}{n-1}+\cdots+\binom{n}{n}\binom{n}{0}=\binom{2n}{n}. \tag{1.3}$$

Note that $\binom{n}{r}=\binom{n}{n-r}$, then (1.3) becomes

$$\binom{n}{0}^2+\binom{n}{1}^2+\cdots+\binom{n}{n}^2=\binom{2n}{n}. \tag{1.4}$$

If we compare the coefficients of x^{a-r} on both sides of (1.1), we have

$$\begin{aligned}\binom{a+b}{a-r} &= \sum_{i+j=a-r}\binom{a}{i}\binom{b}{j}=\sum_{k=0}^{a-r}\binom{a}{a-r-k}\binom{b}{k}\\ &= \sum_{k=0}^{a-r}\binom{a}{r+k}\binom{b}{k}.\end{aligned} \tag{1.5}$$

Furthermore, by differentiating the identity $(1+x)^n=\sum_{i=0}^n\binom{n}{i}x^i$ on both sides with respect to x, and setting $x=1$, we obtain

$$n2^{n-1}=\sum_{i=1}^{n} i\binom{n}{i}. \tag{1.6}$$

Similarly, by differentiating $(1-x)^n=\sum_{i=0}^n(-1)^i\binom{n}{i}x^i$ on both sides with respect to x, and substituting $x=1$, we have $0=\sum_{i=1}^n(-1)^i i\binom{n}{i}$. ‖

1.1.2 Sample space

3• Outcomes of an experiment

- An experiment is a process of observation or measurement.
- The results obtained from an experiment are called the *outcomes* of the experiment.

- The set of all possible outcomes of an experiment is called the *sample space* denoted by $\mathbb{S}$.
- Each outcome in a sample space is called an *element* or a *sample point*.
- An *event* is a subset of a sample space.

4• Sample space

- According to the number of elements they contain, sample spaces can be classified into *discrete* sample space and *continuous* sample space.

4.1• Discrete sample space

— A sample space is discrete, if the number of elements is finite or countable.

4.2• Continuous sample space

— A sample space is continuous, if the sample space consists of a continuum.

— For example, a set of real numbers includes both the rational numbers and the irrational numbers.

1.1.3 Events

5• Complement, union and intersection

- Given two events $\mathbb{A} \subset \mathbb{S}$ and $\mathbb{B} \subset \mathbb{S}$, we define three events as follows:

$$\begin{aligned} \mathbb{A}' &\triangleq \{e\colon\ e \in \mathbb{S} \quad \text{but} \quad e \notin \mathbb{A}\}, \\ \mathbb{A} \cup \mathbb{B} &\triangleq \{e\colon\ e \in \mathbb{A} \quad \text{or} \quad e \in \mathbb{B}\}, \quad \text{and} \\ \mathbb{A} \cap \mathbb{B} &\triangleq \{e\colon\ e \in \mathbb{A} \quad \text{and} \quad e \in \mathbb{B}\}. \end{aligned}$$

  They are called the *complement* of $\mathbb{A}$, the *union* of $\mathbb{A}$ and $\mathbb{B}$, and the *intersection* of $\mathbb{A}$ and $\mathbb{B}$, respectively.

- Let $\emptyset$ denote the empty set. If $\mathbb{A} \cap \mathbb{B} = \emptyset$, then $\mathbb{A}$ and $\mathbb{B}$ are *mutually exclusive*.
- If $\mathbb{A} \subset \mathbb{B}$, then $\mathbb{A}$ is contained in $\mathbb{B}$ or $\mathbb{A}$ is a subset of $\mathbb{B}$.
- Figures 1.1 and 1.2 illustrate these concepts.

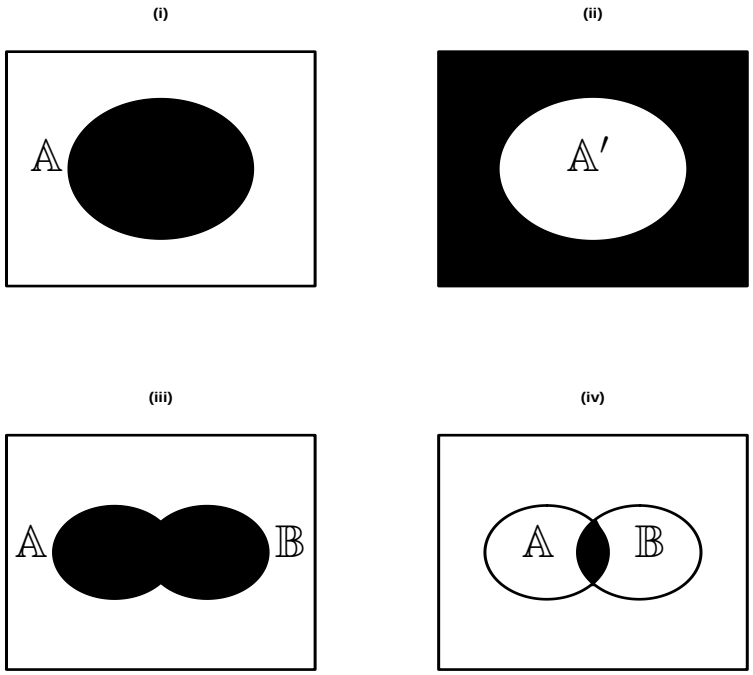

Figure 1.1 Venn diagrams. (i) $\mathbb{A}$; (ii) $\mathbb{A}'$; (iii) $\mathbb{A} \cup \mathbb{B}$; (iv) $\mathbb{A} \cap \mathbb{B}$.

1.1.4 Properties of probability

6• Definition of probability

Definition 1.1 (Probability of a set). Let $\mathbb{A}$ be a subset of the sample space $\mathbb{S}$, then $\Pr(\mathbb{A})$ is said to be the probability of $\mathbb{A}$ if

(1) $\Pr(\mathbb{A}) \geqslant 0$ and $\Pr(\mathbb{S}) = 1$;

(2) If $\mathbb{A}_1, \mathbb{A}_2, \ldots$ is a sequence of mutually exclusive events of $\mathbb{S}$, then

$$\Pr\left(\bigcup_{i=1}^{\infty} \mathbb{A}_i\right) = \sum_{i=1}^{\infty} \Pr(\mathbb{A}_i). \tag{1.7}$$

‖

7• Some properties of probability

Property 1.1 $\Pr(\emptyset) = 0$. ‖

Proof. Since $\emptyset = \cup_{i=1}^{\infty} \emptyset$, from (1.7), we have $\Pr(\emptyset) = \sum_{i=1}^{\infty} \Pr(\emptyset)$ and thus $\Pr(\emptyset) = 0$. □

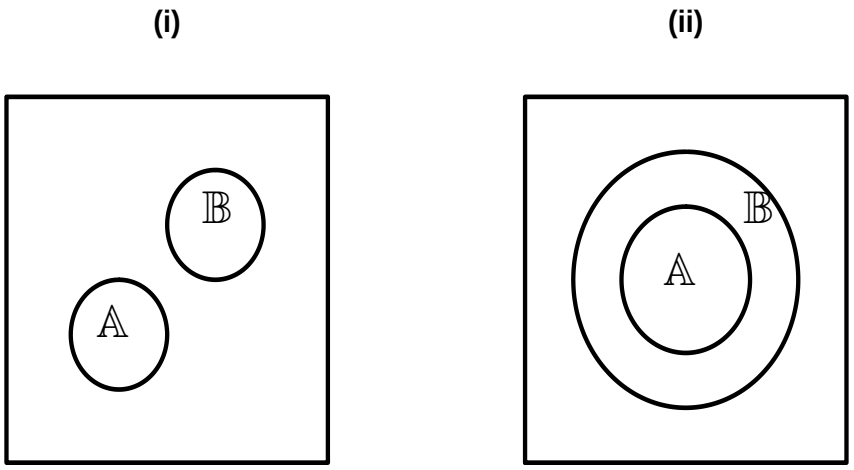

Figure 1.2 Diagrams showing special relationships among events. (i) $\mathbb{A}$ and $\mathbb{B}$ are mutually exclusive; (ii) $\mathbb{A}$ is contained in $\mathbb{B}$.

Property 1.2 If $\mathbb{A}_1, \ldots, \mathbb{A}_n$ are mutually exclusive, then

$$\Pr\left(\bigcup_{i=1}^{n} \mathbb{A}_i\right) = \sum_{i=1}^{n} \Pr(\mathbb{A}_i). \qquad \|$$

Proof. This is trivial, since $\cup_{i=1}^{n} \mathbb{A}_i = \mathbb{A}_1 \cup \cdots \cup \mathbb{A}_n \cup \emptyset \cup \emptyset \cup \cdots$. □

Property 1.3 $\Pr(\mathbb{A}') = 1 - \Pr(\mathbb{A})$. ‖

Property 1.4 If $\mathbb{A} \subseteq \mathbb{B}$, then $\Pr(\mathbb{A}) \leqslant \Pr(\mathbb{B})$. ‖

Proof. Note that $\mathbb{B} = \mathbb{A} \cup (\mathbb{A}' \cap \mathbb{B})$. Since $\mathbb{A}$ and $\mathbb{A}' \cap \mathbb{B}$ are mutually exclusive, then $\Pr(\mathbb{B}) = \Pr(\mathbb{A}) + \Pr(\mathbb{A}' \cap \mathbb{B}) \geqslant \Pr(\mathbb{A})$. □

Property 1.5 $0 \leqslant \Pr(\mathbb{A}) \leqslant 1$. ‖

Property 1.6 $\Pr(\mathbb{A} \cup \mathbb{B}) = \Pr(\mathbb{A}) + \Pr(\mathbb{B}) - \Pr(\mathbb{A} \cap \mathbb{B})$. ‖

Proof. We first partition $\mathbb{A} \cup \mathbb{B}$ into three mutually exclusive events $\mathbb{E}_1$, $\mathbb{E}_2$ and $\mathbb{E}_3$ as shown in Figure 1.3, then

$$\begin{aligned} \Pr(\mathbb{A} \cup \mathbb{B}) &= \Pr(\mathbb{E}_1 \cup \mathbb{E}_2 \cup \mathbb{E}_3) \\ &= \Pr(\mathbb{E}_1) + \Pr(\mathbb{E}_2) + \Pr(\mathbb{E}_3) \\ &= \Pr(\mathbb{E}_1 \cup \mathbb{E}_3) + \Pr(\mathbb{E}_2 \cup \mathbb{E}_3) - \Pr(\mathbb{E}_3) \\ &= \Pr(\mathbb{A}) + \Pr(\mathbb{B}) - \Pr(\mathbb{A} \cap \mathbb{B}). \end{aligned}$$

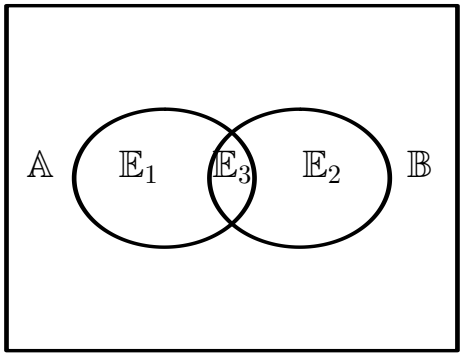

Figure 1.3 A partition of $\mathbb{A} \cup \mathbb{B}$. □

Property 1.7 $\Pr(\mathbb{A}\cup\mathbb{B}\cup\mathbb{C}) = \Pr(\mathbb{A})+\Pr(\mathbb{B})+\Pr(\mathbb{C})-\Pr(\mathbb{A}\cap\mathbb{B})-\Pr(\mathbb{B}\cap\mathbb{C})-\Pr(\mathbb{A}\cap\mathbb{C})+\Pr(\mathbb{A}\cap\mathbb{B}\cap\mathbb{C})$. In general, we have

$$\begin{aligned}\Pr\left(\bigcup_{i=1}^{n}\mathbb{A}_i\right) &= \sum_{i=1}^{n}\Pr(\mathbb{A}_i) - \sum_{i<j}\Pr(\mathbb{A}_i\cap\mathbb{A}_j)\\ &\quad + \sum_{i<j<k}\Pr(\mathbb{A}_i\cap\mathbb{A}_j\cap\mathbb{A}_k) - \cdots\\ &\quad + (-1)^{n+1}\Pr(\mathbb{A}_1\cap\mathbb{A}_2\cap\cdots\cap\mathbb{A}_n).\end{aligned}$$ ‖

1.2 Conditional Probability

8• Definition of conditional probability

Definition 1.2 (Conditional probability of two sets). If $\mathbb{A}$ and $\mathbb{B}$ are two events in the sample space $\mathbb{S}$, the conditional probability of $\mathbb{B}$ given $\mathbb{A}$ is defined by

$$\Pr(\mathbb{B}|\mathbb{A}) = \frac{\Pr(\mathbb{A}\cap\mathbb{B})}{\Pr(\mathbb{A})}, \tag{1.8}$$

where $\Pr(\mathbb{A}) > 0$. ‖

8.1• Equivalent formulae

— From (1.8), we immediately obtain

$$\Pr(\mathbb{A}\cap\mathbb{B}) = \Pr(\mathbb{A})\times\Pr(\mathbb{B}|\mathbb{A}), \quad \Pr(\mathbb{A})>0, \quad \text{and} \tag{1.9}$$

$$\Pr(\mathbb{A}\cap\mathbb{B}) = \Pr(\mathbb{B})\times\Pr(\mathbb{A}|\mathbb{B}), \quad \Pr(\mathbb{B})>0. \tag{1.10}$$

Example 1.2 (Car dealer service data). A research report of the services under warranty provided by 50 new-car dealers in a certain city is summarized in Table 1.1. Let $\mathbb{G}$ denote the selection of a dealer who provides good service under warranty, and let $\mathbb{T}$ denote the selection of a dealer who has been in business 10 years or more. The aim is to find $\Pr(\mathbb{G})$ and $\Pr(\mathbb{G}|\mathbb{T})$.

Table 1.1 Car dealer service data

Time in business	Service attitude		
	Good service ($\mathbb{G}$)	Poor service ($\mathbb{G}'$)	Total
$\geqslant$ 10 years ($\mathbb{T}$)	16	4	20
$<$ 10 years ($\mathbb{T}'$)	10	20	30
Total	26	24	50

Solution. Note that

$$\Pr(\mathbb{G}) = \frac{\#\text{ of favourable outcomes}}{\#\text{ of possible outcomes}} = \frac{26}{50} = \frac{13}{25}.$$

From (1.8), we have

$$\Pr(\mathbb{G}|\mathbb{T}) = \frac{\Pr(\mathbb{G} \cap \mathbb{T})}{\Pr(\mathbb{T})} = \frac{16/50}{20/50} = \frac{4}{5}. \qquad \|$$

9• Definition of independency of two events

Definition 1.3 (Independency of two events). Two events $\mathbb{A}$ and $\mathbb{B}$ are said to be *independent*, denoted by $\mathbb{A} \perp\!\!\!\perp \mathbb{B}$, if

$$\Pr(\mathbb{A} \cap \mathbb{B}) = \Pr(\mathbb{A}) \times \Pr(\mathbb{B}). \qquad \|$$

Theorem 1.1 (Independency). Let $\mathbb{A} \perp\!\!\!\perp \mathbb{B}$, then $\mathbb{A} \perp\!\!\!\perp \mathbb{B}'$ and $\mathbb{A}' \perp\!\!\!\perp \mathbb{B}'$. $\|$

Proof. Since $\mathbb{A} = (\mathbb{A} \cap \mathbb{B}) \cup (\mathbb{A} \cap \mathbb{B}')$, where $\mathbb{A} \cap \mathbb{B}$ and $\mathbb{A} \cap \mathbb{B}'$ are mutually exclusive, we obtain

$$\begin{aligned} \Pr(\mathbb{A} \cap \mathbb{B}') &= \Pr(\mathbb{A}) - \Pr(\mathbb{A} \cap \mathbb{B}) \\ &= \Pr(\mathbb{A}) - \Pr(\mathbb{A}) \times \Pr(\mathbb{B}) \\ &= \Pr(\mathbb{A}) \times \Pr(\mathbb{B}'), \end{aligned}$$

which indicates $\mathbb{A} \perp\!\!\!\perp \mathbb{B}'$. Using it again, we have $\mathbb{A}' \perp\!\!\!\perp \mathbb{B}'$. □

10• Mutual independency and pairwise independency

Definition 1.4 (Mutual independency). Events $\mathbb{A}_1, \ldots, \mathbb{A}_n$ are said to be mutually independent, if the probability of the intersection of any $2, 3, \ldots,$ or n of these events equals the product of their respective probabilities. ‖

10.1• Pairwise independency

— For $n = 3$, $\mathbb{A}_1, \mathbb{A}_2$ and $\mathbb{A}_3$ are mutually independent *if and only if* (iff)

$$\mathbb{A}_1 \perp\!\!\!\perp \mathbb{A}_2, \quad \mathbb{A}_1 \perp\!\!\!\perp \mathbb{A}_3, \quad \mathbb{A}_2 \perp\!\!\!\perp \mathbb{A}_3 \tag{1.11}$$

and

$$\Pr(\mathbb{A}_1 \cap \mathbb{A}_2 \cap \mathbb{A}_3) = \Pr(\mathbb{A}_1) \times \Pr(\mathbb{A}_2) \times \Pr(\mathbb{A}_3). \tag{1.12}$$

— Note that $\mathbb{A}_1, \mathbb{A}_2$ and $\mathbb{A}_3$ are called *pairwise* independent if (1.11) holds.

1.3 Bayes Theorem

11• Partition and Bayes formula

Definition 1.5 (Partition). A partition of the sample space $\mathbb{S}$ is a collection of mutually exclusive sets $\mathbb{B}_1, \ldots, \mathbb{B}_n$ such that $\mathbb{S} = \cup_{i=1}^{n} \mathbb{B}_i$. ‖

Theorem 1.2 (Bayes formula). Let $\mathbb{B}_1, \ldots, \mathbb{B}_n$ be a partition of the sample space $\mathbb{S}$ and $\mathbb{A}$ be an event, then

(1) *Law of total probability*:

$$\Pr(\mathbb{A}) = \sum_{i=1}^{n} \Pr(\mathbb{A}|\mathbb{B}_i) \Pr(\mathbb{B}_i). \tag{1.13}$$

(2) *Bayes formula*:

$$\Pr(\mathbb{B}_j|\mathbb{A}) = \frac{\Pr(\mathbb{A}|\mathbb{B}_j) \Pr(\mathbb{B}_j)}{\sum_{i=1}^{n} \Pr(\mathbb{A}|\mathbb{B}_i) \Pr(\mathbb{B}_i)} \quad \text{for } j = 1, \ldots, n. \tag{1.14}$$

‖

Example 1.3 (Insurance data). An insurance company has three types of customers: high risk, medium risk, and low risk. Twenty percent of its customers are of high risk, 30% are of medium risk, and 50% are of low risk. The probability that a customer has at least one accident in the current year is 0.25 for high risk, 0.16 for medium risk, and 0.10 for low risk.

Let the events that a customer is high, medium, and low risk be $\mathbb{H}$, $\mathbb{M}$ and $\mathbb{L}$, respectively. Let the event that a customer has at least one accident in the current year be $\mathbb{A}$. Find $\Pr(\mathbb{A})$ and $\Pr(\mathbb{H}|\mathbb{A})$.

Solution. By the law of total probability, we have

$$\begin{aligned}\Pr(\mathbb{A}) &= \Pr(\mathbb{A}|\mathbb{H})\Pr(\mathbb{H}) + \Pr(\mathbb{A}|\mathbb{M})\Pr(\mathbb{M}) + \Pr(\mathbb{A}|\mathbb{L})\Pr(\mathbb{L}) \\ &= 0.25 \times 0.20 + 0.16 \times 0.30 + 0.10 \times 0.50 \\ &= 0.148.\end{aligned}$$

By Bayes formula, we obtain

$$\Pr(\mathbb{H}|\mathbb{A}) = \frac{\Pr(\mathbb{A}|\mathbb{H})\Pr(\mathbb{H})}{\Pr(\mathbb{A})} = \frac{0.25 \times 0.20}{0.148} = 0.3378. \quad \|$$

1.4 Probability Distributions

12• Discrete and continuous random variables

Definition 1.6 (Random variable). A *random variable* (r.v.) is a function from a sample space $\mathbb{S}$ into the real numbers. An r.v. is *discrete* if it takes values in a finite or countable set. An r.v. is *continuous* if it takes values over some interval. $\|$

Definition 1.7 (Probability mass function). If X is a discrete r.v., the function defined by

$$p(x) = \Pr(X = x)$$

for each x within the range of X is called the *probability mass function* (pmf) of X. $\|$

13• Basic properties of the pmf $p(x)$

- $p(x) \geqslant 0$.
- $\sum_x p(x) = 1$.

Definition 1.8 (Probability density function). Let X be a continuous r.v.. A non-negative function $f(x)$ is call the *probability density function* (pdf) of X, if

$$\Pr(\mathbb{A}) = \int_{\mathbb{A}} f(x)\,\mathrm{d}x$$

for an arbitrary set $\mathbb{A}$ in the range of X. In particular, if the range of X is the real line (or one-dimensional Euclidean space) $\mathbb{R} = (-\infty, \infty)$, then

$$\Pr(a \leqslant X \leqslant b) = \int_a^b f(x)\,\mathrm{d}x. \qquad \|$$

14• Basic properties of the pdf $f(x)$

- $f(x) \geqslant 0$.
- $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$.

Definition 1.9 (Cumulative density function). The *cumulative distribution function* (cdf) of an r.v. X is defined by

$$F(x) = \Pr(X \leqslant x) = \begin{cases} \sum_{t \leqslant x} p(t), & \text{if } X \text{ is discrete,} \\ \int_{-\infty}^{x} f(t)\,\mathrm{d}t, & \text{if } X \text{ is continuous.} \end{cases} \qquad \|$$

15• Basic properties of the cdf $F(x)$

- $F(-\infty) = 0$ and $F(\infty) = 1$.
- If $a \leqslant b$, then $F(a) \leqslant F(b)$.
- If the range of a discrete r.v. X consists of the ordered values $x_1 < x_2 < \cdots < x_n$, then $p(x_1) = F(x_1)$ and $p(x_i) = F(x_i) - F(x_{i-1})$, $i = 2, 3, \ldots, n$.
- $f(x) = F'(x) \mathrel{\hat{=}} \mathrm{d}F(x)/\mathrm{d}x$.
- $\Pr(a < X < b) = \Pr(a \leqslant X < b) = \Pr(a < X \leqslant b) = \Pr(a \leqslant X \leqslant b) = F(b) - F(a) = \int_a^b f(x)\,\mathrm{d}x$.

Example 1.4 (Geometric distribution). If X has the pmf $p(x) = 0.5^x$ for $x = 1, 2, \ldots$, please determine the cdf $F(x)$.

Solution. Let $S_n = \alpha + \alpha^2 + \cdots + \alpha^n$. First, we prove that

$$S_n = \begin{cases} \dfrac{\alpha(1-\alpha^n)}{1-\alpha}, & \text{if } \alpha \neq 1, \\ n, & \text{if } \alpha = 1. \end{cases} \tag{1.15}$$

$$\lim_{n\to\infty} S_n = \frac{\alpha}{1-\alpha}, \qquad \text{if } \alpha \in (0,1). \tag{1.16}$$

In fact, $\alpha S_n = \alpha^2 + \cdots + \alpha^n + \alpha^{n+1}$, then

$$S_n - \alpha S_n = \alpha - \alpha^{n+1}.$$

Thus, we immediately obtain (1.15) and (1.16).

Let $\alpha = 0.5$ in (1.16), we have $\sum_{x=1}^{\infty} p(x) = 1$, i.e., $p(x)$ is really a pmf. Furthermore, let $\alpha = 0.5$ in (1.15), we obtain

$$F(n) = \Pr(X \leqslant n) = \sum_{x=1}^{n} 0.5^x = 1 - 0.5^n,$$

$$F(x) = \begin{cases} 0, & \text{if } x < 1, \\ 1 - 0.5^n, & \text{if } n \leqslant x < n+1, \quad n = 1, 2, \ldots. \end{cases}$$

Obviously, $F(\infty) = \lim_{n\to\infty}(1 - 0.5^n) = 1$. ‖

Example 1.5 (Exponential distribution). Let X have the pdf $f(x) = \lambda\, \mathrm{e}^{-3x}$ for $x > 0$, where α is the normalizing constant.

1) Find the λ.

2) Find the cdf.

3) Evaluate $\Pr(0.5 \leqslant X < 1)$.

Solution. 1) Since $1 = \int_0^\infty f(x)\,\mathrm{d}x = \int_0^\infty \lambda\, \mathrm{e}^{-3x}\,\mathrm{d}x = \frac{\lambda}{3}$, we obtain $\lambda = 3$.

2) According to the relationship between the cdf and pdf, we have

$$F(x) = \int_0^x f(t)\,\mathrm{d}t = \int_0^x 3\,\mathrm{e}^{-3t}\,\mathrm{d}t = 1 - \mathrm{e}^{-3x}.$$

3) Therefore, we obtain

$$\Pr(0.5 \leqslant X < 1) = \int_{0.5}^{1} f(x)\,\mathrm{d}x = F(1) - F(0.5) = \mathrm{e}^{-1.5} - \mathrm{e}^{-3}. \qquad \|$$

1.5 Bivariate Distributions

1.5.1 Joint distribution

Definition 1.10 (Bivariate pmf). If X and Y are two discrete r.v.'s, the function defined by

$$p(x, y) = \Pr(X = x,\ Y = y)$$

for each pair of values (x, y) within the range of X and Y is called the joint pmf of X and Y. ‖

16• Basic properties of the joint pmf $p(x, y)$

- $p(x, y) \geqslant 0$.
- $\sum_x \sum_y p(x, y) = 1$.

Definition 1.11 (Bivariate pdf). A bivariate function $f(x, y)$ is called a joint pdf of the continuous r.v.'s X and Y if

$$\Pr\{(X, Y) \in \mathbb{A}\} = \int \int_{\mathbb{A}} f(x, y)\, \mathrm{d}x\, \mathrm{d}y$$

for a region $\mathbb{A}$ in the domain of (X, Y). ‖

17• Basic properties of the joint pdf $f(x, y)$

- $f(x, y) \geqslant 0$.
- $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, \mathrm{d}x\, \mathrm{d}y = 1$.

Definition 1.12 (Bivariate cdf). The joint distribution (or joint cdf) of r.v.'s (X, Y) is defined by

$$\begin{aligned} F(x, y) &= \Pr(X \leqslant x,\ Y \leqslant y) \\ &= \begin{cases} \displaystyle\sum_{s \leqslant x} \sum_{t \leqslant y} p(s, t), & \text{if } X \text{ and } Y \text{ are discrete,} \\ \displaystyle\int_{-\infty}^{y} \int_{-\infty}^{x} f(s, t)\, \mathrm{d}s\, \mathrm{d}t, & \text{if } X \text{ and } Y \text{ are continuous.} \end{cases} \end{aligned}$$

For the continuous case, the joint pdf and cdf have the following relationship:

$$f(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y} = \frac{\partial^2 F(x, y)}{\partial y \partial x}.$$ ‖

1.5.2 Marginal and conditional distributions

Definition 1.13 (Marginal pmfs and conditional pmfs). Let $p(x, y)$ be the joint pmf of discrete r.v.'s (X, Y). The *marginal* pmfs of X and Y are defined by

$$p(x) = \sum_y p(x, y) \quad \text{and} \quad p(y) = \sum_x p(x, y),$$

respectively. The *conditional* pmfs of X given $Y = y$ and Y given $X = x$ are defined by

$$p(x|y) = \frac{p(x, y)}{p(y)}, \quad p(y) \neq 0 \quad \text{and} \quad p(y|x) = \frac{p(x, y)}{p(x)}, \quad p(x) \neq 0,$$

respectively. ‖

Definition 1.14 (Marginal pdfs and conditional pdfs). Let $f(x, y)$ be the joint pdf of continuous r.v.'s (X, Y). The *marginal* pdfs of X and Y are defined by

$$f(x) = \int_{-\infty}^{\infty} f(x, y)\, \mathrm{d}y \quad \text{and} \quad f(y) = \int_{-\infty}^{\infty} f(x, y)\, \mathrm{d}x,$$

respectively. The *conditional* pdfs of X given $Y = y$ and Y given $X = x$ are defined by

$$f(x|y) = \frac{f(x, y)}{f(y)}, \quad f(y) \neq 0 \quad \text{and} \quad f(y|x) = \frac{f(x, y)}{f(x)}, \quad f(x) \neq 0,$$

respectively. ‖

1.5.3 Independency of two random variables

Definition 1.15 (Independency of two r.v.'s). Let $f(x, y)$ denote the joint pdf of r.v.'s (X, Y), and $f(x)$ and $f(y)$ be their marginal pdfs. The r.v.'s X and Y are said to be *independent*, denoted by $X \perp\!\!\!\perp Y$, if

$$f(x, y) = f(x) \times f(y), \quad \forall\, (x, y) \in \mathcal{S}_{(X,Y)}, \quad \text{or} \tag{1.17}$$

$$F(x, y) = F(x) \times F(y), \quad \forall\, (x, y) \in \mathcal{S}_{(X,Y)}. \tag{1.18}$$

where $\mathcal{S}_{(X,Y)} \hat{=} \{(x, y)\colon f(x, y) > 0\}$ denotes the joint *support* of (X, Y). ‖

Example 1.6 (Uniform distribution in a circle). Let

$$f(x,y) = \begin{cases} \dfrac{1}{\pi r^2}, & \text{if } x^2 + y^2 \leqslant r^2, \\ 0, & \text{otherwise.} \end{cases}$$

1) Find the marginal density $f(x)$.

2) Find the conditional density $f(y|x)$.

3) Evaluate $\Pr(Y \geqslant 0.5r | X = 0.5r)$.

3) Are X and Y independent?

Solution. Let $I_{\mathbb{S}}(z)$ denote the *indicator function*, i.e., $I_{\mathbb{S}}(z) = 1$ if $z \in \mathbb{S}$ and $I_{\mathbb{S}}(z) = 0$ if $z \notin \mathbb{S}$. Note that the joint support of (X,Y) is $\mathcal{S}_{(X,Y)} = \{(x,y)\colon\ x^2 + y^2 \leqslant r^2\}$, we can rewrite the joint pdf as

$$f(x,y) = \frac{1}{\pi r^2} \cdot I_{\mathcal{S}_{(X,Y)}}(x,y).$$

1) Since the marginal support of X is $\mathcal{S}_X = \{x\colon\ |x| \leqslant r\}$, then the marginal density of X is given by

$$f(x) = \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}y = \int_{-\sqrt{r^2-x^2}}^{\sqrt{r^2-x^2}} f(x,y)\,\mathrm{d}y = \frac{2\sqrt{r^2-x^2}}{\pi r^2} \cdot I_{\mathcal{S}_X}(x).$$

By symmetry, we have $\mathcal{S}_Y = \{y\colon\ |y| \leqslant r\}$ and

$$f(y) = \frac{2\sqrt{r^2-y^2}}{\pi r^2} \cdot I_{\mathcal{S}_Y}(y).$$

2) Now, the conditional support $\mathcal{S}_{(Y|X=x)} = \{y\colon\ |y| \leqslant \sqrt{r^2-x^2}\,\}$. Thus, the conditional density is

$$f(y|x) = \frac{f(x,y)}{f(x)} = \frac{1}{2\sqrt{r^2-x^2}} \cdot I_{\mathcal{S}_{(Y|X=x)}}(y).$$

3) When $X = 0.5r$, we have $\mathcal{S}_{(Y|X=0.5r)} = \{y\colon\ |y| \leqslant \sqrt{3}\,r/2\}$. Hence,

$$\begin{aligned} \Pr(Y \geqslant 0.5r | X = 0.5r) &= \int_{0.5r}^{\sqrt{3}\,r/2} f(y|0.5r)\,\mathrm{d}y = \int_{0.5r}^{\sqrt{3}\,r/2} \frac{1}{\sqrt{3}\,r}\,\mathrm{d}y \\ &= \frac{1}{\sqrt{3}\,r}\left(\frac{\sqrt{3}\,r}{2} - \frac{r}{2}\right) = \frac{3-\sqrt{3}}{6}. \end{aligned}$$

4) Since $f(x,y) \neq f(x) \times f(y)$ for $(x,y) \in \mathcal{S}_{(X,Y)}$, X and Y are not independent. ‖

1.6 Expectation, Variance and Moments

1.6.1 Moments

18• The general case

- Let X be a discrete (or continuous) r.v. with pmf $p(x)$ (or pdf $f(x)$).
- Let $g(\cdot)$ be an arbitrary function, then $g(X)$ itself is also a random variable.
- The expectation of $g(X)$ is defined by

$$E\{g(X)\} = \begin{cases} \sum_x g(x)p(x), & \text{if } X \text{ is discrete,} \\ \int_{-\infty}^{\infty} g(x)f(x)\,\mathrm{d}x, & \text{if } X \text{ is continuous,} \end{cases}$$

  provided that $E\{g(X)\}$ exists.

19• Mean or expectation

- The expectation of X is defined as

$$\mu = E(X) = \begin{cases} \sum_x xp(x), & \text{if } X \text{ is discrete,} \\ \int_{-\infty}^{\infty} xf(x)\,\mathrm{d}x, & \text{if } X \text{ is continuous.} \end{cases}$$

- It is a measure of the *central location* of the pdf of X.

19.1• Some basic properties of expectation

— $E(c) = c$ for a constant c.

— $E\{cg(X)\} = cE\{g(X)\}$ for a constant c.

— $E\{\sum c_i g_i(X)\} = \sum c_i E\{g_i(X)\}$.

— $E\{g_1(X)\} \leqslant E\{g_2(X)\}$ if $g_1(x) \leqslant g_2(x) \quad \forall x$.

— If $X_1 \perp\!\!\!\perp X_2$, then $E(X_1X_2) = E(X_1)E(X_2)$.

20• Variance and standard deviation

- Let $\mu = E(X)$, then

$$\sigma^2 = \mathrm{Var}(X) = E(X-\mu)^2 = E(X^2) - \mu^2,$$

is a measure of the *dispersion* of the pdf of X.

20.1• Some basic properties of variance

— $\mathrm{Var}(c) = 0$ for a constant c.

— $\mathrm{Var}\{cg(X)\} = c^2\mathrm{Var}\{g(X)\}$ for a constant c.

— $\mathrm{Var}(\sum c_i X_i) = \sum c_i^2 \mathrm{Var}(X_i) + 2\sum_{i<j} c_i c_j \mathrm{Cov}(X_i, X_j)$.

20.2• Standard deviation

— $\sigma = \sqrt{\mathrm{Var}(X)}$.

21• Covariance

- $\mathrm{Cov}(X_1, X_2) = E\{(X_1 - \mu_1)(X_2 - \mu_2)\}$, where $\mu_i = E(X_i)$, $i = 1, 2$.

21.1• Some basic properties of covariance

— $\mathrm{Cov}(X_1, X_2) = E(X_1X_2) - E(X_1)E(X_2)$.

— If $X_1 \perp\!\!\!\perp X_2$, then $\mathrm{Cov}(X_1, X_2) = 0$.

22• Moments and central moments

- The r-th moment of the r.v. X is defined by $\mu'_r = E(X^r)$, which is also called the r-th *raw moment* or the r-th *moment about the origin*.
- The r-th central moment of the r.v. X is defined by $\mu_r = E(X-\mu)^r$, which is also called the r-th *central moment* or the r-th *moment about the mean*.
- It is easy to show that μ_r and μ'_r have the following relationship:

$$\mu_r = \sum_{i=0}^{r} (-1)^i \binom{r}{i} \mu'_{r-i}\mu^i. \tag{1.19}$$

23• SKEWNESS AND KURTOSIS

- The third central moment $\mu_3 = E(X-\mu)^3$ is a measure of *asymmetry* of the pdf of X.
- μ_3/σ^3 is called the *coefficient of skewness*.
- The fourth central moment $\mu_4 = E(X-\mu)^4$ is a measure of kurtosis, which is the *degree of flatness* of a density near its center.
- μ_4/σ^4, called the *coefficient of kurtosis*, is sometimes used to indicate that a density is more peaked around its center than the normal density.

24• QUANTILE AND MEDIAN

- The q-th quantile of an r.v. X with cdf $F(\cdot)$, denoted by ξ_q, is defined as the smallest real number ξ satisfying $F(\xi) = \Pr(X \leqslant \xi) \geqslant q$.
- If X is continuous, then the q-th quantile of X is defined as the smallest real number ξ satisfying $F(\xi) = \Pr(X \leqslant \xi) = q$.
- The 0.5-th quantile $\xi_{0.5}$ is defined as the median of X, denoted by $\mathrm{med}(X)$.
- Alternatively, the median of X satisfies

$$\Pr\{X \leqslant \mathrm{med}(X)\} \geqslant 0.5 \quad \text{and} \quad \Pr\{X \geqslant \mathrm{med}(X)\} \geqslant 0.5.$$

- If X is a continuous r.v. with pdf $f(x)$, then the median of X satisfies

$$\int_{-\infty}^{\mathrm{med}(X)} f(x)\,\mathrm{d}x = 0.5 = \int_{\mathrm{med}(X)}^{\infty} f(x)\,\mathrm{d}x.$$

1.6.2 Some probability inequalities

25• A MORE GENERAL INEQUALITY

Theorem 1.3 (The general case). If X is an r.v. and $g(\cdot)$ is a non-negative function defined on the real line $\mathbb{R}$, then

$$\Pr\{g(X) \geqslant c\} \leqslant \frac{E\{g(X)\}}{c} \tag{1.20}$$

for any $c > 0$. ‖

Proof. We only consider the case that X is a continuous r.v. . Let the pdf of X be $f(x)$ and define $\mathbb{D} = \{x\colon\ g(x) \geqslant c\}$, we have

$$\begin{aligned} E\{g(X)\} &= \int_{-\infty}^{\infty} g(x) \cdot f(x)\,\mathrm{d}x \\ &= \int_{\mathbb{D}} g(x) \cdot f(x)\,\mathrm{d}x + \int_{\mathbb{D}'} g(x) \cdot f(x)\,\mathrm{d}x \\ &\geqslant c \int_{\mathbb{D}} f(x)\,\mathrm{d}x \\ &= c\,\Pr\{g(X) \geqslant c\}, \end{aligned}$$

which means (1.20). □

25.1• Chebyshev inequality

— Especially, in (1.20), by setting $g(X) = (X-\mu)^2$ and replacing c with $c^2\sigma^2$, we obtain the well–known Chebyshev inequality as follows.

— Let X be an r.v. and c be a positive constant, then

$$\Pr(|X-\mu| \geqslant c\,\sigma) \leqslant \frac{1}{c^2}, \tag{1.21}$$

where $\mu = E(X)$ and $\sigma^2 = \mathrm{Var}(X)$.

Example 1.7 (Three–point distribution). Let X be a discrete r.v. with pmf $\Pr(X=-1) = 1/(2c^2)$, $\Pr(X=0) = 1-1/c^2$, and $\Pr(X=1) = 1/(2c^2)$, where $c > 1$. Then $E(X) = 0$, $\mathrm{Var}(X) = 1/c^2$, and

$$\Pr(|X-\mu| \geqslant c\,\sigma) = \Pr(|X| \geqslant 1) = \Pr(X = 1 \text{ or } -1) = 1/c^2.$$

Thus equality holds in (1.21). ‖

25.2• Markov inequality

— Especially, in (1.20), by setting $g(X) = |X|^r$ and replacing c with c^r, we obtain Markov inequality as follows.

— Let X be an r.v. and c, r be two positive constants, then

$$\Pr(|X| \geqslant c) \leqslant \frac{E(|X|^r)}{c^r}, \tag{1.22}$$

provided that $E(|X|^r)$ exists.

26• Jensen's inequality

26.1• Convex function

— A continuous function $g(\cdot)$ defined on a subset $\mathbb{S}$ of the real line $\mathbb{R}$ is said to be *convex* if for each point $x_0 \in \mathbb{R}$, there exists a line which goes through the point $(x_0, g(x_0))$ and lies on or under the curve of the function $g(\cdot)$.

— A twice continuously differentiable function $g(x)$ is convex (or strictly convex) iff its second derivative $g''(x) \geqslant 0$ (or $g''(x) > 0$) for all $x \in \mathbb{S}$.

— A function $h(x)$ is *concave* (or strictly concave) iff $-h(x)$ is convex (or strictly convex).

Theorem 1.4 (Jensen's inequality). Let $g(\cdot)$ be a convex function. If X is an r.v. taking values in the domain of $g(\cdot)$, then

$$E\{g(X)\} \geqslant g(E(X)), \tag{1.23}$$

provided that both expectations $E(X)$ and $E\{g(X)\}$ exist. ‖

Proof. Since $g(x)$ is continuous and convex, there exists a straight line, say $\ell(x) = a + bx$, satisfying

$$\ell(x) = a + bx \leqslant g(x)$$

and

$$g(x_0) = g(E(X)) = \ell(E(X))$$

for any point $x_0 = E(X)$. Note that $E\{\ell(X)\} = E(a + bX) = a + bE(X) = \ell(E(X))$, we have

$$g(E(X)) = \ell(E(X)) = E\{\ell(X)\} \leqslant E\{g(X)\},$$

implying (1.23). □

Theorem 1.5 (Cauchy–Schwarz inequality). If two random variables X and Y have finite second moments, then

$$\{E(XY)\}^2 \leqslant E(X^2)E(Y^2), \tag{1.24}$$

with equality iff $\Pr(Y = cX) = 1$ for some constant c. ‖

Proof. Let

$$h(t) = E(tX - Y)^2 = E(X^2)t^2 - 2E(XY)t + E(Y^2),$$

then $h(t) \geqslant 0$ and it is a quadratic function of t.

If $h(t) > 0$, then the roots of $h(t)$ are not real; so

$$4\{E(XY)\}^2 - 4E(X^2)E(Y^2) < 0,$$

which means $\{E(XY)\}^2 < E(X^2)E(Y^2)$.

If $h(t) = 0$ for some t, say c, then $E(cX - Y)^2 = 0$, which implies $\Pr(Y = cX) = 1$. □

1.6.3 Conditional expectation

Definition 1.16 (Conditional expectation). Let X and Y be two r.v.'s and $p(x|y)$ (or $f(x|y)$) be the conditional pmf (or pdf) of X given $Y = y$, then the conditional expectation of $g(X)$ given $Y = y$ is

$$E\{g(X)|Y = y\} = \begin{cases} \sum_x g(x)p(x|y), & \text{if } X \text{ is discrete,} \\ \int_{-\infty}^{\infty} g(x)f(x|y)\,\mathrm{d}x, & \text{if } X \text{ is continuous.} \end{cases} \quad \|$$

27• Basic properties of $E\{g(X)|Y = y\}$

- $E\{g_1(X) + g_2(X)|Y = y\} = E\{g_1(X)|Y = y\} + E\{g_2(X)|Y = y\}$.
- $E\{g_1(X)g_2(Y)|Y = y\} = g_2(y)E\{g_1(X)|Y = y\}$.

28• Calculating $E(X)$ and $\mathrm{Var}(X)$ via $E(X|Y)$ and $\mathrm{Var}(X|Y)$

28.1• $E\{g(X)|Y\}$ is a random variable

— Note that $E\{g(X)|Y\}$ is a function of the r.v. Y, thus we can write

$$E\{g(X)|Y\} \hat{=} h(Y),$$

which is also a random variable.

— That is, for any y in the range of Y, we have $E\{g(X)|Y = y\} = h(y)$.

28.2• Two general formulae

— For the continuous case, let $f(y)$ be the marginal density of Y, then

$$\begin{aligned} E[E\{g(X)|Y\}] &= E\{h(Y)\} = \int_{-\infty}^{\infty} h(y)f(y)\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty} E\{g(X)|Y=y\}f(y)\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{\infty} g(x)f(x|y)\,\mathrm{d}x \right\} f(y)\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)f(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ &= E\{g(X)\}. \end{aligned} \tag{1.25}$$

— Furthermore, we have

$$\begin{aligned} E[\mathrm{Var}\{g(X)|Y\}] &= E[E\{g(X)g(X)|Y\} - E^2\{g(X)|Y\}] \\ &\overset{(1.25)}{=} E\{g(X)\}^2 - E\{h(Y)\}^2 \\ &= E\{g(X)\}^2 - [\mathrm{Var}\{h(Y)\} + E^2\{h(Y)\}] \\ &\overset{(1.25)}{=} E\{g(X)\}^2 - \mathrm{Var}\{h(Y)\} - E^2\{g(X)\} \\ &= \mathrm{Var}\{g(X)\} - \mathrm{Var}[E\{g(X)|Y\}]. \end{aligned}$$

That is,

$$\mathrm{Var}\{g(X)\} = E[\mathrm{Var}\{g(X)|Y\}] + \mathrm{Var}[E\{g(X)|Y\}]. \tag{1.26}$$

28.3• Two important formulae

— In particular, letting $g(X) = X$ in (1.25) and (1.26), we obtain

$$E(X) = E\{E(X|Y)\} = \int E(X|Y=y)\,f(y)\,\mathrm{d}y \quad \text{and} \tag{1.27}$$

$$\mathrm{Var}(X) = E\{\mathrm{Var}(X|Y)\} + \mathrm{Var}\{E(X|Y)\}. \tag{1.28}$$

1.6.4 Compound random variables

29• A COMPOUND RANDOM VARIABLE

- Let $X_1, X_2, \ldots$ be a sequence of *independent and identically distributed* (i.i.d.) r.v.'s that are independent of the non-negative integer-valued r.v. N.
- The random variable
$$S_N = \sum_{i=1}^{N} X_i \tag{1.29}$$
is called a *compound* random variable.

Theorem 1.6 (Expectation and variance of S_N). Let S_N be defined by (1.29), then

$$E(S_N) = E(N)E(X) \quad \text{and} \tag{1.30}$$
$$\text{Var}(S_N) = E(N)\text{Var}(X) + \text{Var}(N)\{E(X)\}^2. \tag{1.31}$$

where the r.v. X has the same distribution with X_1. ‖

Proof. From (1.27), we have

$$E(S_N) = E\{E(S_N|N)\} = E\{NE(X)\} = E(N)E(X).$$

From (1.28), we obtain

$$\begin{aligned}\text{Var}(S_N) &= E\{\text{Var}(S_N|N)\} + \text{Var}\{E(S_N|N)\} \\ &= E\{N\text{Var}(X)\} + \text{Var}\{NE(X)\} \\ &= E(N)\text{Var}(X) + \text{Var}(N)\{E(X)\}^2.\end{aligned}$$

which means (1.31). □

1.6.5 Calculation of (conditional) probability via (conditional) expectation

30• EQUIVALENCE BETWEEN PROBABILITY AND EXPECTATION

- Let $\mathbb{A}$ be an event and Y be a random variable. The aim is to find probability $\Pr(\mathbb{A})$ and the conditional probability $\Pr(\mathbb{A}|Y)$.

- To do this, we first introduce an indicator r.v. associated with the event $\mathbb{A}$ as follows:

$$X = \begin{cases} 1, & \text{if the event } \mathbb{A} \text{ occurs with probability } \Pr(\mathbb{A}), \\ 0, & \text{if the event } \mathbb{A} \text{ does not occurs with probability } 1 - \Pr(\mathbb{A}). \end{cases}$$

- It is easy to see that X is a two-point discrete random variable.
- Hence, we have

$$E(X) = 1 \times \Pr(\mathbb{A}) + 0 \times \{1 - \Pr(\mathbb{A})\} = \Pr(\mathbb{A}) \tag{1.32}$$

and $E(X|Y) = \Pr(\mathbb{A}|Y)$.

- Moreover, we obtain

$$\begin{aligned} \Pr(\mathbb{A}) = E(X) = E\{E(X|Y)\} = E\{\Pr(\mathbb{A}|Y)\} \\ = \begin{cases} \sum_y \Pr(\mathbb{A}|Y = y)p(y), & \text{If } Y \text{ is discrete,} \\ \int \Pr(\mathbb{A}|Y = y)f(y)\,\mathrm{d}y, & \text{If } Y \text{ is continuous.} \end{cases} \end{aligned} \tag{1.33}$$

1.7 Moment Generating Function

31• Definition of mgf

Definition 1.17 (mgf). For an r.v. X, if $E(\mathrm{e}^{tX})$ exists for any $t \in (-\varepsilon, \varepsilon)$ for some $\varepsilon > 0$, then

$$M_X(t) = E(\mathrm{e}^{tX})$$

is called the *moment generating function* (mgf) of X. ‖

31.1• The relationship between mgf and the n-th moment

— By using Maclaurin's expansion, we have

$$M_X(t) = E\left\{\sum_{n=0}^{\infty} \frac{(tX)^n}{n!}\right\} = \sum_{n=0}^{\infty} \frac{t^n}{n!}\, E(X^n) = \sum_{n=0}^{\infty} \frac{t^n}{n!}\, \mu'_n.$$

— Thus, we obtain

$$\mu'_n = E(X^n) = \left.\frac{\mathrm{d}^n M_X(t)}{\mathrm{d}t^n}\right|_{t=0}. \tag{1.34}$$

31.2• Some basic properties of mgf

— $M_{a+bX}(t) = \mathrm{e}^{at} M_X(bt)$.

— If $X \perp\!\!\!\perp Y$, then $M_{X+Y}(t) = M_X(t) M_Y(t)$.

— Jensen's inequality provides a lower bound: $M_X(t) \geqslant \exp\{tE(X)\}$.

Example 1.8 (Normal distribution). Assume that $X \sim N(\mu, \sigma^2)$, then

$$M_X(t) = \exp(\mu t + 0.5\sigma^2 t^2). \tag{1.35}$$

Solution. If $Z \sim N(0,1)$, then the mgf of Z is

$$\begin{aligned} M_Z(t) &= \int_{-\infty}^{\infty} \mathrm{e}^{tz} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-0.5z^2}\, \mathrm{d}z \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\{-0.5(z^2 - 2tz + t^2) + 0.5t^2\}\, \mathrm{d}z \\ &= \mathrm{e}^{0.5t^2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\{-0.5(z-t)^2\}\, \mathrm{d}z = \mathrm{e}^{0.5t^2}. \end{aligned} \tag{1.36}$$

Since $X = \mu + \sigma Z$, we obtain

$$M_X(t) = M_{\mu+\sigma Z}(t) = \mathrm{e}^{\mu t} M_Z(\sigma t) = \mathrm{e}^{\mu t + 0.5\sigma^2 t^2}. \qquad \|$$

Table 1.2 Discrete probability distributions

Name	pmf $p(x)$	Parameter	mgf	$E(X)$	$\mathrm{Var}(X)$
Binomial	$\binom{n}{x} p^x q^{n-x}$, $x = 0, 1, \ldots, n$	$0 \leqslant p \leqslant 1$	$(p\,\mathrm{e}^t + q)^n$	np	npq
Poisson	$\lambda^x \mathrm{e}^{-\lambda}/x!$, $x = 0, 1, 2, \ldots$	$\lambda > 0$	$\exp\{\lambda(\mathrm{e}^t - 1)\}$	λ	λ
Geometric	pq^{x-1}, $x = 1, 2, \ldots$	$0 \leqslant p \leqslant 1$	$p\,\mathrm{e}^t/(1 - q\,\mathrm{e}^t)$	$1/p$	q/p^2
Negative binomial	$\binom{x-1}{r-1} p^r q^{x-r}$, $x = r, r+1, \ldots$	$0 \leqslant p \leqslant 1$	$\{p\,\mathrm{e}^t/(1 - q\,\mathrm{e}^t)\}^r$	r/p	rq/p^2

NOTE: $q \hat{=} 1 - p$.

成功了r次

Table 1.3 Continuous probability distributions

Name	pdf $f(x)$	Parameter	mgf	$E(X)$	$\mathrm{Var}(X)$
Uniform	$1/(b-a)$, $a \leqslant x \leqslant b$	$a < b$	$\frac{\mathrm{e}^{tb}-\mathrm{e}^{ta}}{t(b-a)}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Beta	$\frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}$, $0 \leqslant x \leqslant 1$	$a > 0$ $b > 0$		$\frac{a}{a+b}$	$\frac{ab}{c^2(c+1)}$,
Exponential	$\lambda\,\mathrm{e}^{-\lambda x}$, $x \geqslant 0$	$\lambda > 0$	$\lambda/(\lambda - t)$ $t < \lambda$	$1/\lambda$	$1/\lambda^2$
Gamma	$\frac{\lambda^n}{\Gamma(n)} x^{n-1}\,\mathrm{e}^{-\lambda x}$, $x \geqslant 0$	$\lambda > 0$ $n \in \mathbb{N}$	$\{\lambda/(\lambda - t)\}^n$ $t < \lambda$	n/λ	n/λ^2
Normal	$\frac{\exp[-(x-\mu)^2/2\sigma^2]}{\sqrt{2\pi}\,\sigma}$, $x \in \mathbb{R}$	$\mu \in \mathbb{R}$ $\sigma^2 > 0$	$\mathrm{e}^{\mu t + 0.5\sigma^2 t^2}$	μ	σ^2

NOTE: $c \hat{=} a + b$ and $\mathbb{N} \hat{=} \{1, 2, 3, \ldots\}$.

Theorem 1.7 (Alternative to probability distribution). Two r.v.'s have the same mgf iff their distributions are identical. In other words, for all values of t, $M_X(t) = M_Y(t)$ iff $F_X(x) = F_Y(x)$ for all values of x (or equivalently $X \overset{\mathrm{d}}{=} Y$). ‖

31.3• A remark to Theorem 1.5

— The above statement is not equivalent to the statement "if two distributions have the same moments, then they are identical at all points."

— For some cases, the moments exist and yet the mgf does not, because

$$\sum_{n=0}^{\infty} \frac{t^n}{n!}\,\mu'_n = \lim_{n\to\infty} \sum_{i=0}^{n} \frac{t^i}{i!}\,\mu'_i$$

may not exist.

32• Relation to other functions

32.1• Characteristic function

— The mgf may not exist for some distributions.

— In general, we may apply the *characteristic function* (cf)

$$\varphi_X(t) = E(\,\mathrm{e}^{\mathrm{i}tX}) = M_{\mathrm{i}X}(t) = M_X(\mathrm{i}t),$$

which always exists, where $\mathrm{i}^2 = -1$.

32.2• Probability generating function

— If X is a discrete r.v. taking values in the non-negative integers $\{0, 1, \ldots\}$, then the *probability generating function* of X is defined as

$$G(z) = E(z^X) = \sum_{x=0}^{\infty} p(x) z^x,$$

where $p(\cdot)$ is the pmf of X.

— This immediately implies that $G(\,\mathrm{e}^t) = M_X(t)$.

1.8 Beta and Gamma Distributions

1.8.1 Beta distribution

33• Definition

Definition 1.18 (Beta density). An r.v. X is said to follow a beta distribution with parameters $a > 0$ and $b > 0$, if it has the pdf

$$\mathrm{Beta}(x|a,b) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}, \quad 0 \leqslant x \leqslant 1, \tag{1.37}$$

where $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ is the beta function and $\Gamma(\cdot)$ is the gamma function defined by (1.39). We will write $X \sim \mathrm{Beta}(a,b)$. ‖

33.1• Densities

— Beta densities with various parameters are shown in Figure 1.4.

— Because $\int_0^1 \mathrm{Beta}(x|a,b)\,\mathrm{d}x = 1$, we have the following identity:

$$\int_0^1 x^{a-1}(1-x)^{b-1}\,\mathrm{d}x = B(a,b).$$

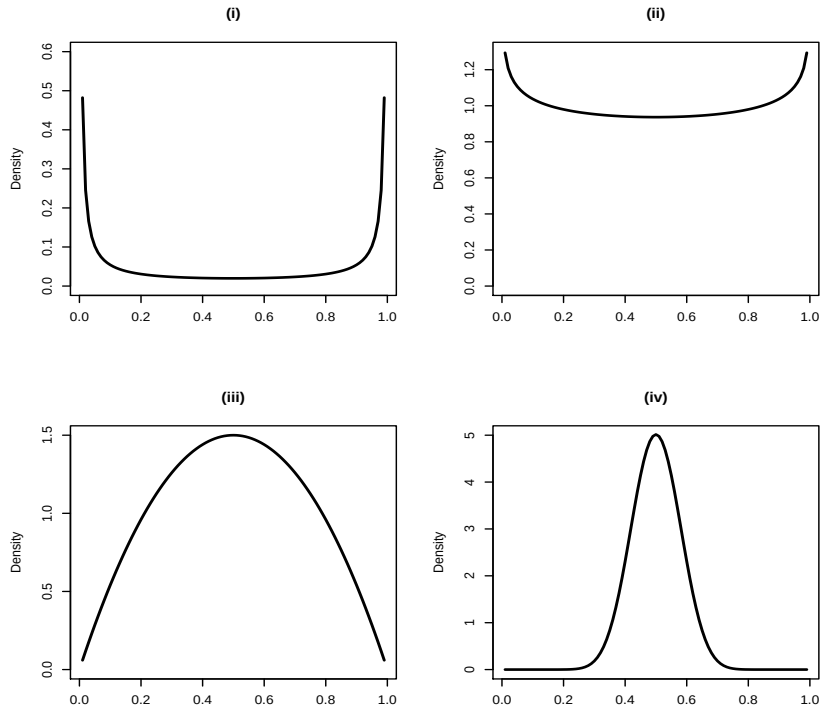

Figure 1.4 Plots of the densities of $X \sim \text{Beta}(a, a)$ with various parameter values. (i) $a = 0.01$; (ii) $a = 0.9$; (iii) $a = 2$; (iv) $a = 20$.

33.2• Moments

— Let $X \sim \text{Beta}(a, b)$, then the r-th moment of X is given by

$$E(X^r) = \frac{B(a+r, b)}{B(a, b)} = \frac{\Gamma(a+r)}{\Gamma(a)} \cdot \frac{\Gamma(a+b)}{\Gamma(a+b+r)}. \tag{1.38}$$

— Especially, we obtain

$$\begin{aligned} E(X) &= \frac{a}{a+b}, \\ E(X^2) &= \frac{a(a+1)}{(a+b)(a+b+1)}, \quad \text{and} \\ \text{Var}(X) &= \frac{ab}{(a+b)^2(a+b+1)}. \end{aligned}$$

33.3• Basic properties on $X \sim \text{Beta}(a, b)$

— $\text{Beta}(1, 1) = U(0, 1)$.

— $1 - X \sim \text{Beta}(b, a)$.

— The k-th order statistic (cf. Section 2.4) from a sample of n i.i.d. $U(0, 1)$ follows $\text{Beta}(k,\ n - k + 1)$.

33.4• Usefulness

— The beta distribution is the conjugate prior for the binomial likelihood.

— A non-informative distribution is obtained as $a, b \to 0$.

— The built-in `R` function, `rbeta`(N, a, b), can be used to generate N i.i.d. samples from $\text{Beta}(a, b)$.

1.8.2 Gamma distribution

34• The gamma function

- First of all, we briefly review the gamma function

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} \, \mathrm{e}^{-x} \, \mathrm{d}x, \tag{1.39}$$

  which is well defined for $\alpha > 0$.

34.1• Basic properties

— $\Gamma(1) = 1$.

— $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$.

— $\Gamma(1/2) = \sqrt{\pi}$.

— For a positive integer n, $\Gamma(n + 1) = n!$.

35• Definition of gamma distribution

Definition 1.19 (Gamma density). An r.v. X has a gamma distribution with shape parameter $\alpha > 0$ and rate parameter (1/rate is called scale parameter) $\beta > 0$, if its density function is given by

$$\text{Gamma}(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \, \mathrm{e}^{-\beta x}, \quad x \geqslant 0. \tag{1.40}$$

It is denoted by $X \sim \text{Gamma}(\alpha, \beta)$. ‖

35.1• Densities

— Gamma densities with various parameters are shown in Figure 1.5.

— By (1.39), we have

$$\int_0^\infty x^{\alpha-1}\,\mathrm{e}^{-\beta x}\,\mathrm{d}x = \frac{\Gamma(\alpha)}{\beta^\alpha}, \tag{1.41}$$

which implies $\int_0^\infty \text{Gamma}(x|\alpha,\beta)\,\mathrm{d}x = 1$.

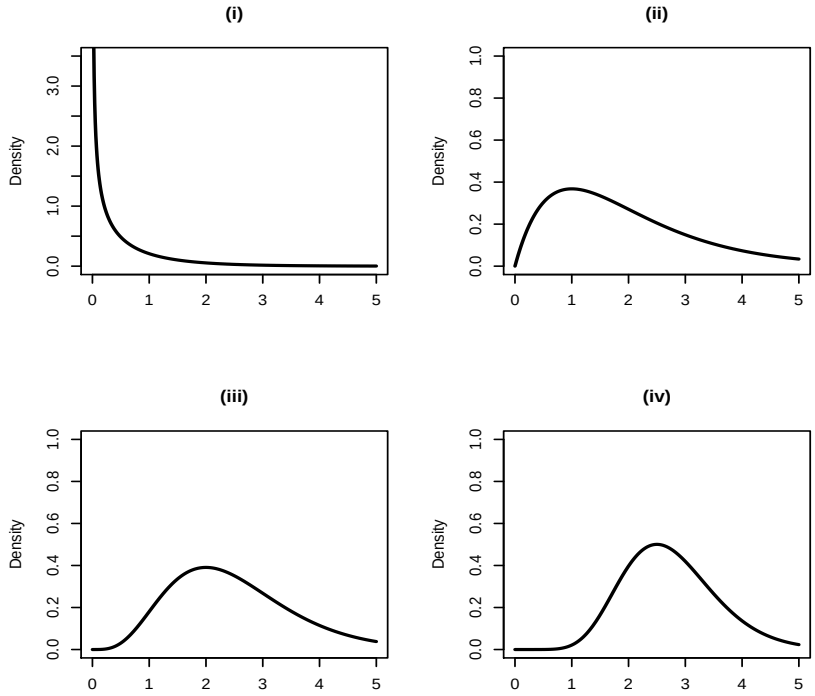

Figure 1.5 Plots of the densities of $X \sim \text{Gamma}(\alpha,\beta)$ with various parameter values. (i) $\alpha = 0.5, \beta = 1$; (ii) $\alpha = 2, \beta = 1$; (iii) $\alpha = 5, \beta = 2$; (iv) $\alpha = 11, \beta = 4$.

35.2• Moment generating function

— The mgf of X is given by

$$\begin{aligned} M_X(t) &= \frac{\beta^\alpha}{\Gamma(\alpha)}\int_0^\infty \mathrm{e}^{tx}x^{\alpha-1}\,\mathrm{e}^{-\beta x}\,\mathrm{d}x = \frac{\beta^\alpha}{\Gamma(\alpha)}\int_0^\infty x^{\alpha-1}\,\mathrm{e}^{-(\beta-t)x}\,\mathrm{d}x \\ &\overset{(1.41)}{=} \frac{\beta^\alpha}{\Gamma(\alpha)}\cdot\frac{\Gamma(\alpha)}{(\beta-t)^\alpha} = \left(\frac{\beta}{\beta-t}\right)^\alpha, \quad t<\beta. \end{aligned} \tag{1.42}$$

35.3• Moments

— Since

$$\begin{aligned} M'_X(t) &= \frac{\mathrm{d}M_X(t)}{\mathrm{d}t} = \alpha\beta^\alpha(\beta-t)^{-\alpha-1} \quad \text{and} \\ M''_X(t) &= \frac{\mathrm{d}^2M_X(t)}{\mathrm{d}t^2} = \alpha(\alpha+1)\beta^\alpha(\beta-t)^{-\alpha-2}, \end{aligned}$$

we obtain

$$\begin{aligned} E(X) &= M'_X(0) = \frac{\alpha}{\beta}, \\ E(X^2) &= M''_X(0) = \frac{\alpha(\alpha+1)}{\beta^2} \quad \text{and} \\ \mathrm{Var}(X) &= E(X^2) - \{E(X)\}^2 = \frac{\alpha}{\beta^2}. \end{aligned}$$

35.4• Other properties of $X \sim \mathrm{Gamma}(\alpha, \beta)$

— $\mathrm{Gamma}(1, \beta) = \mathrm{Exponential}(\beta)$, the exponential distribution with rate parameter β.

— $\mathrm{Gamma}(n/2, 1/2) = \chi^2(n)$, the chi-squared distribution with n degrees of freedom.

— If $X \sim \mathrm{Gamma}(\alpha, \beta)$ and $c > 0$, then $Y = cX \sim \mathrm{Gamma}(\alpha, \beta/c)$.

— If $\{X_i\}_{i=1}^n \overset{\text{ind}}{\sim} \mathrm{Gamma}(\alpha_i, \beta)$, then $\sum_{i=1}^n X_i \sim \mathrm{Gamma}(\sum_{i=1}^n \alpha_i, \beta)$.

— If $\{X_i\}_{i=1}^n \overset{\text{iid}}{\sim} \mathrm{Exponential}(\beta)$, then $\sum_{i=1}^n X_i \sim \mathrm{Gamma}(n, \beta)$.

36• A useful formula

- Let $X \sim \mathrm{Gamma}(\alpha, \beta)$, then

$$E\{\log(X)\} = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} - \log(\beta). \tag{1.43}$$

Proof. Let $Y = \beta X$, then $Y \sim \mathrm{Gamma}(\alpha, 1)$ and

$$E\{\log(Y)\} = \log(\beta) + E\{\log(X)\}.$$

Differentiating both sides of the following identity

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1}\,\mathrm{e}^{-y}\,\mathrm{d}y$$

with respect to α, we obtain

$$\Gamma'(\alpha) = \int_0^\infty y^{\alpha-1}\log(y)\cdot\,\mathrm{e}^{-y}\,\mathrm{d}y, \qquad [\because (a^x)' = a^x\log(a)]$$

or

$$\frac{\Gamma'(\alpha)}{\Gamma(\alpha)} = \int_0^\infty \log(y)\cdot\frac{1}{\Gamma(\alpha)}y^{\alpha-1}\,\mathrm{e}^{-y}\,\mathrm{d}y = E\{\log(Y)\},$$

which indicates (1.43). □

1.9 Bivariate Normal Distribution

1.9.1 Univariate normal distribution

37• Definition

- It is well known that X is normally distributed with mean μ and variance σ^2, denoted by $X \sim N(\mu, \sigma^2)$, if its pdf is

$$N(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \quad -\infty < x < \infty. \tag{1.44}$$

37.1• Densities

— Normal densities with various parameters are shown in Figure 1.6.

— From (1.44), we have

$$\int_{-\infty}^{+\infty}\exp\left(-\frac{z^2}{2}\right)\mathrm{d}z = \sqrt{2\pi}.$$

37.2• Basic properties of $X \sim N(\mu, \sigma^2)$

— If $\mu = 0$ and $\sigma = 1$, then $E(X^{2r+1}) = 0$ and $E(X^{2r}) = (2r)!/(2^r r!)$.

Proof. On the one hand, if $g(\cdot)$ is an odd function defined in $(-\infty, \infty)$, then $\int_{-\infty}^{\infty} g(x)\,\mathrm{d}x = 0$. We immediately obtain the first formula. On the other hand, if $g(\cdot)$ is an even function defined in $(-\infty, \infty)$, then

$$\int_{-\infty}^{\infty} g(x)\,\mathrm{d}x = 2\int_{0}^{\infty} g(x)\,\mathrm{d}x.$$

By combining this fact with the identity (1.41), we can obtain the second formula. □

— $a + bX \sim N(a + b\mu, b^2\sigma^2)$.

— In particular, $(X - \mu)/\sigma \sim N(0, 1)$.

— $E(X) = \mu,\ \mathrm{Var}(X) = \sigma^2$.

— $M_X(t) = \exp(\mu t + 0.5\sigma^2 t^2)$.

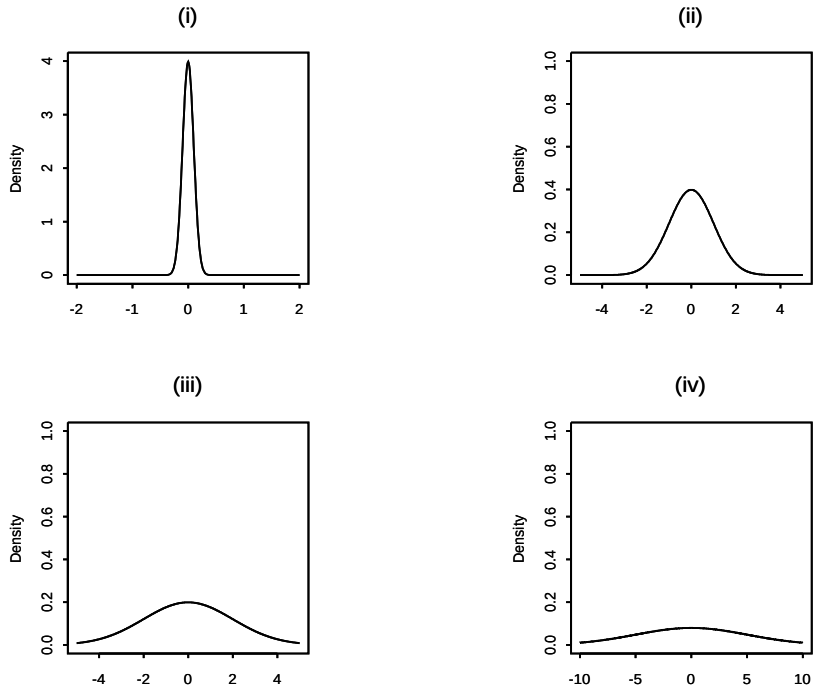

Figure 1.6 Plots of the densities of $X \sim N(0, \sigma^2)$ with various variances. (i) $\sigma = 0.1$; (ii) $\sigma = 1$; (iii) $\sigma = 2$; (iv) $\sigma = 5$.

1.9.2 Correlation coefficient

38• Definition of correlation coefficient

- To introduce the bivariate normal distribution, first of all, we introduce the concept of *correlation coefficient*.
- Given two r.v.'s X_1 and X_2 with $E(X_1) = \mu_1, E(X_2) = \mu_2, \mathrm{Var}(X_1) = \sigma_1^2$ and $\mathrm{Var}(X_2) = \sigma_2^2$, the covariance of X_1 and X_2 is

$$\mathrm{Cov}(X_1, X_2) = E(X_1X_2) - E(X_1)E(X_2).$$

- The correlation coefficient of X_1 and X_2 is defined by

$$\rho = \mathrm{Corr}(X_1, X_2) = \frac{\mathrm{Cov}(X_1, X_2)}{\sigma_1\sigma_2}. \tag{1.45}$$

38.1• Using the Cauchy–Schwarz inequality to prove $|\rho| \leqslant 1$

— The correlation coefficient ρ defined by (1.45) can be rewritten as

$$\rho = \frac{E(X_1 - \mu_1)(X_2 - \mu_2)}{\sqrt{E(X_1 - \mu_1)^2 \cdot E(X_2 - \mu_2)^2}}.$$

— In (1.24), let $X = X_1 - \mu_1$ and $Y = X_2 - \mu_2$, we obtain

$$\frac{\{E(X_1 - \mu_1)(X_2 - \mu_2)\}^2}{E(X_1 - \mu_1)^2 \cdot E(X_2 - \mu_2)^2} \leqslant 1,$$

i.e., $\rho^2 \leqslant 1$ so that $-1 \leqslant \rho \leqslant 1$. □

1.9.3 Joint density

39• Multivariate case

- A random vector $\mathbf{x} = (X_1, \ldots, X_d)^\top$ is said to follow a d-dimensional normal distribution, if its joint pdf is given by

$$N_d(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(\sqrt{2\pi})^d |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \right\}, \tag{1.46}$$

for $\boldsymbol{x} \in \mathbb{R}^d$, where the mean vector $\boldsymbol{\mu} \in \mathbb{R}^d$ and the covariance matrix $\boldsymbol{\Sigma}$ is positive definite, denoted by $\boldsymbol{\Sigma} > 0$ (which is equivalent to $\boldsymbol{y}^\top \boldsymbol{\Sigma} \boldsymbol{y} > 0$ for any non-zero vector $\boldsymbol{y}$).

- We will write $\mathbf{x} \sim N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ or $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

39.1• Two-dimensional case

— Especially, when $d = 2$ and

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix},$$

we use $(X_1, X_2)^\top \sim N_2(\mu_1, \mu_2; \sigma_1^2, \sigma_2^2, \rho)$ to represent the bivariate normal distribution.

— Its joint pdf is then given by

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{ -\frac{Q(x_1, x_2)}{2(1-\rho^2)} \right\}, \; x_1, x_2 \in \mathbb{R}, \quad (1.47)$$

where

$$Q(x_1, x_2) = \left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2.$$

39.2• Marginal and conditional densities

— Note that for any joint pdf $f(x_1, x_2)$, we have $f(x_1, x_2) = f(x_1) \times f(x_2|x_1)$.

— From (1.47), we can write

$$\begin{aligned}
&\frac{1}{2(1-\rho^2)} Q(x_1, x_2) \\
= \; &\frac{1}{2(1-\rho^2)} \left\{ \left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 \right\} \\
= \; &\frac{(x_1-\mu_1)^2}{2\sigma_1^2}\left(1 + \frac{\rho^2}{1-\rho^2}\right) + \frac{1}{2(1-\rho^2)}\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 \\
&+ \frac{1}{2(1-\rho^2)}\left\{ -2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) \right\} \\
= \; &\frac{(x_1-\mu_1)^2}{2\sigma_1^2} + \frac{(x_2-\mu_2)^2}{2\sigma_2^2(1-\rho^2)} - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{2\sigma_1\sigma_2(1-\rho)^2} + \frac{\rho^2(x_1-\mu_1)^2}{2\sigma_1^2(1-\rho^2)} \\
= \; &\frac{(x_1-\mu_1)^2}{2\sigma_1^2} + \frac{\{x_2-\mu_2-\rho\frac{\sigma_2}{\sigma_1}(x_1-\mu_1)\}^2}{2\sigma_2^2(1-\rho^2)}.
\end{aligned}$$

— Thus, we can decompose (1.47) into

$$\begin{aligned} f(x_1, x_2) &= \frac{1}{\sqrt{2\pi}\,\sigma_1} \exp\left\{ -\frac{(x_1-\mu_1)^2}{2\sigma_1^2} \right\} \\ &\quad \times \frac{1}{\sqrt{2\pi}\,\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{\{x_2-\mu_2-\rho\frac{\sigma_2}{\sigma_1}(x_1-\mu_1)\}^2}{2\sigma_2^2(1-\rho^2)} \right], \end{aligned}$$

which indicates

$$X_1 \sim N(\mu_1,\ \sigma_1^2) \quad \text{and} \tag{1.48}$$

$$X_2|(X_1 = x_1) \sim N\left(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x_1-\mu_1),\ \sigma_2^2(1-\rho^2)\right). \tag{1.49}$$

— By symmetry, we also have

$$X_2 \sim N(\mu_2,\ \sigma_2^2) \quad \text{and} \tag{1.50}$$

$$X_1|(X_2 = x_2) \sim N\left(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(x_2-\mu_2),\ \sigma_1^2(1-\rho^2)\right). \tag{1.51}$$

39.3• Basic properties of $(X_1, X_2)^\top \sim N_2(\mu_1, \mu_2; \sigma_1^2, \sigma_2^2, \rho)$

— $E(X_i) = \mu_i$, $\mathrm{Var}(X_i) = \sigma_i^2$ for $i = 1, 2$.

— $E(X_1|X_2 = x_2) = \mu_1 + \rho\frac{\sigma_1}{\sigma_2}(x_2-\mu_2), \mathrm{Var}(X_1|X_2 = x_2) = \sigma_1^2(1-\rho^2)$,

$E(X_2|X_1 = x_1) = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x_1-\mu_1), \mathrm{Var}(X_2|X_1 = x_1) = \sigma_2^2(1-\rho^2)$.

— $\mathrm{Cov}(X_1, X_2) = \rho\sigma_1\sigma_2$, $\mathrm{Corr}(X_1, X_2) = \rho$.

Proof. It is clear that

$$\begin{aligned} \mathrm{Cov}(X_1, X_2) &= E\{(X_1-\mu_1)(X_2-\mu_2)\} \\ &\overset{(1.25)}{=} E\Big[E\{(X_1-\mu_1)(X_2-\mu_2)|X_2\}\Big] \\ &= E\Big[(X_2-\mu_2)E\{(X_1-\mu_1)|X_2\}\Big] \\ &= E\Big\{(X_2-\mu_2)\cdot\rho\frac{\sigma_1}{\sigma_2}(X_2-\mu_2)\Big\} \\ &= \rho\frac{\sigma_1}{\sigma_2}\cdot\mathrm{Var}(X_2) \\ &= \rho\sigma_1\sigma_2, \end{aligned}$$

so that $\mathrm{Corr}(X_1, X_2) = \rho$. □

— $M_{(X_1,\,X_2)}(t_1, t_2) = \exp\Big\{\mu_1 t_1 + \mu_2 t_2 + 0.5(\sigma_1^2 t_1^2 + \sigma_2^2 t_2^2 + 2\rho\sigma_1\sigma_2 t_1 t_2)\Big\}$.

Proof. Let $t^* = \rho\frac{\sigma_1}{\sigma_2} t_1 + t_2$, we have

$$\begin{aligned}
& M_{(X_1,\,X_2)}(t_1, t_2) \\
= \;& E(\,\mathrm{e}^{t_1 X_1 + t_2 X_2}) \\
\overset{(1.25)}{=} \;& E\Big\{E(\,\mathrm{e}^{t_1 X_1 + t_2 X_2}|X_2)\Big\} \\
= \;& E\Big\{\,\mathrm{e}^{t_2 X_2} E(\,\mathrm{e}^{t_1 X_1}|X_2)\Big\} \\
\overset{(1.51)}{=} \;& E\Big[\,\mathrm{e}^{t_2 X_2} \cdot \,\mathrm{e}^{\{\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(X_2 - \mu_2)\}t_1 + 0.5\sigma_1^2(1-\rho^2)t_1^2}\Big] \\
= \;& \mathrm{e}^{\mu_1 t_1 - \rho\frac{\sigma_1}{\sigma_2}\mu_2 t_1 + 0.5\sigma_1^2(1-\rho^2)t_1^2} \cdot E(\,\mathrm{e}^{t^* X_2}) \\
= \;& \mathrm{e}^{\mu_1 t_1 - \rho\frac{\sigma_1}{\sigma_2}\mu_2 t_1 + 0.5\sigma_1^2(1-\rho^2)t_1^2} \cdot \,\mathrm{e}^{\mu_2 t^* + 0.5\sigma_2^2 t^{*2}} \\
= \;& \exp\Big\{\mu_1 t_1 + \mu_2 t_2 + 0.5(\sigma_1^2 t_1^2 + \sigma_2^2 t_2^2 + 2\rho\sigma_1\sigma_2 t_1 t_2)\Big\},
\end{aligned}$$

which completes the proof. □

— $X_1 \perp\!\!\!\perp X_2$ iff $\rho = 0$.

— $a_1 X_1 + a_2 X_2 \sim N\Big(a_1\mu_1 + a_2\mu_2,\; a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + 2a_1 a_2\rho\sigma_1\sigma_2\Big)$.

1.9.4 Stochastic representation of random variables or random vectors

40• Definition of stochastic representation

- Let X and $Y_1, \ldots, Y_n$ be r.v.'s and $g(\cdot)$ a function.
- If X and $g(Y_1, \ldots, Y_n)$ have the same distribution, denoted by

$$X \overset{\mathrm{d}}{=} g(Y_1, \ldots, Y_n), \tag{1.52}$$

we say (1.52) is a one-to-many *stochastic representation* (SR) of the r.v. X.

- The symbol '$\stackrel{\mathrm{d}}{=}$' means that the r.v.'s on both sides of the equality have the same distribution.

40.1• Several examples on the operator '$\stackrel{\mathrm{d}}{=}$'

— If $X \sim N(0,1)$, then $X \stackrel{\mathrm{d}}{=} -X$.

— If $U \sim U(0,1)$, then $U \stackrel{\mathrm{d}}{=} 1-U$.

— If $X \sim \text{Exponential}(\beta)$ and $U \sim U(0,1)$, then $X \stackrel{\mathrm{d}}{=} -\log(U)/\beta$.

— If $X_1, \ldots, X_n \stackrel{\text{iid}}{\sim} \text{Exponential}(\beta)$ and $Y \sim \text{Gamma}(n,\beta)$, then

$$Y \stackrel{\mathrm{d}}{=} \sum_{i=1}^{n} X_i.$$

41• An SR of a multivariate normal random vector

- Let $Y_1, \ldots, Y_d \stackrel{\text{iid}}{\sim} N(0,1)$ or

$$\mathbf{y} \stackrel{\wedge}{=} \begin{pmatrix} Y_1 \\ \vdots \\ Y_d \end{pmatrix} \sim N_d(\mathbf{0}, \mathbf{I}_d).$$

- If

$$\begin{pmatrix} X_1 \\ \vdots \\ X_d \end{pmatrix} \stackrel{\mathrm{d}}{=} \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_d \end{pmatrix} + \mathbf{A} \begin{pmatrix} Y_1 \\ \vdots \\ Y_d \end{pmatrix}, \tag{1.53}$$

  where $\mathbf{A}$ is a $d \times d$ matrix, then $\mathbf{x} = (X_1, \ldots, X_d)^\top \sim N_d(\boldsymbol{\mu}, \mathbf{A}\mathbf{A}^\top)$, where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)^\top$ and $\mathbf{A}\mathbf{A}^\top$ is not necessarily positive definite.

41.1• An advantage of using the SR (1.53) to define a multivariate normal distribution

— If $\mathbf{A}\mathbf{A}^\top$ is positive definite (i.e., $\mathbf{A}\mathbf{A}^\top > 0$), then $(\mathbf{A}\mathbf{A}^\top)^{-1}$ exists, the joint density of $\mathbf{x}$ exists and it is given by (1.46) with $\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}^\top$.

— It is clear that $\mathbf{A}\mathbf{A}^\top$ is always positive semi-definite (i.e., $\mathbf{A}\mathbf{A}^\top \geqslant 0$).

— If the determinant of $\mathbf{A}\mathbf{A}^\top$ is zero (i.e., $|\mathbf{A}\mathbf{A}^\top| = 0$), then the joint density of $\mathbf{x}$ does not exist but the distribution still exists.

— To define a multivariate normal distribution, we can see that using the SR (1.53) is better than using the joint density (1.46).

41.2• Other advantages starting from the SR (1.53)

— $E(\mathbf{x}) = \boldsymbol{\mu}$ and $\mathrm{Var}(\mathbf{x}) = \mathbf{A}\mathbf{A}^\top$.

— $M_{\mathbf{x}}(\boldsymbol{t}) = E(\,\mathrm{e}^{\boldsymbol{t}^\top \mathbf{x}}) = \exp(\boldsymbol{t}^\top \boldsymbol{\mu} + 0.5\,\boldsymbol{t}^\top \mathbf{A}\mathbf{A}^\top \boldsymbol{t})$.

Proof. Let $\boldsymbol{s} = \mathbf{A}^\top \boldsymbol{t}$, we have

$$\begin{aligned}
M_{\mathbf{x}}(\boldsymbol{t}) &= E(\,\mathrm{e}^{\boldsymbol{t}^\top \mathbf{x}}) \\
&= E[\exp\{\boldsymbol{t}^\top(\boldsymbol{\mu} + \mathbf{A}\mathbf{y})\}] \\
&= \exp(\boldsymbol{t}^\top \boldsymbol{\mu}) \cdot E\{\exp(\boldsymbol{s}^\top \mathbf{y})\} \\
&= \exp(\boldsymbol{t}^\top \boldsymbol{\mu}) \cdot E\{\exp(s_1 Y_1 + \cdots + s_d Y_d)\} \\
&= \exp(\boldsymbol{t}^\top \boldsymbol{\mu}) \cdot \prod_{i=1}^{d} E\{\exp(s_i Y_i)\} \\
&= \exp(\boldsymbol{t}^\top \boldsymbol{\mu}) \cdot \prod_{i=1}^{d} \exp(0.5\, s_i^2) \\
&= \exp(\boldsymbol{t}^\top \boldsymbol{\mu}) \cdot \exp(0.5\, \boldsymbol{s}^\top \boldsymbol{s}) \\
&= \exp(\boldsymbol{t}^\top \boldsymbol{\mu} + 0.5\, \boldsymbol{t}^\top \mathbf{A}\mathbf{A}^\top \boldsymbol{t}),
\end{aligned}$$

which completes the proof. □

— Partition $\mathbf{x}$ into two parts

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}^{(1)} \\ \mathbf{x}^{(2)} \end{pmatrix}, \quad \text{where} \quad \mathbf{x}^{(1)} = \begin{pmatrix} X_1 \\ \vdots \\ X_r \end{pmatrix} \quad \text{and} \quad \mathbf{x}^{(2)} = \begin{pmatrix} X_{r+1} \\ \vdots \\ X_d \end{pmatrix}.$$

We can partition $\boldsymbol{\mu}$ and $\mathbf{y}$ in the same fashion. From the SR (1.53), we obtain

$$\mathbf{x}^{(k)} \overset{\mathrm{d}}{=} \boldsymbol{\mu}^{(k)} + \mathbf{A}_k \mathbf{y}, \quad k = 1, 2,$$

which indicates that $\mathbf{x}^{(k)} \sim N(\boldsymbol{\mu}^{(k)}, \mathbf{A}_k\mathbf{A}_k^\top)$ for $k = 1, 2$, where

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}^{(1)} \\ \mathbf{A}^{(2)} \end{pmatrix}$$

with $\mathbf{A}^{(1)}$: $r \times d$ and $\mathbf{A}^{(2)}$: $(d - r) \times d$.

1.10 Inverse Bayes Formulae

1.10.1 Three inverse Bayes formulae

42• The issue

- Let two conditional densities $f_{(X|Y)}(x|y)$ and $f_{(Y|X)}(y|x)$ be known.
- Next, suppose that the assumption of $\mathcal{S}_{(X,Y)} = \mathcal{S}_X \times \mathcal{S}_Y$ is valid, where $\mathcal{S}_X$, $\mathcal{S}_Y$ and $\mathcal{S}_{(X,Y)}$ denote the marginal supports of X, Y and the joint support of (X, Y), respectively.
- What are the marginal density $f_X(x)$ and hence the joint density of (X, Y)?

42.1• What is the definition of support?

— Let $f_X(x)$ is the pdf of the r.v. X, then $\mathcal{S}_X = \{x\colon\ f_X(x) > 0\}$ is called the *support* of X.

— For example, if $X \sim U(0, 1)$, then the pdf of X is

$$f_X(x) = \begin{cases} 1, & \text{if } 0 < x < 1, \\ 0, & \text{otherwise.} \end{cases}$$

Hence, $\mathcal{S}_X = (0, 1)$.

— Let $f_{(X,Y)}(x, y)$ denote the joint density of (X, Y), then

$$\mathcal{S}_{(X,Y)} = \{(x, y)\colon\ f_{(X,Y)}(x, y) > 0\}$$

is called the *joint support* of (X, Y), see Definition 1.15 on page 14.

42.2• The product space and non-product space

— If $\mathcal{S}_{(X,Y)} = \mathcal{S}_X \times \mathcal{S}_Y$, we say $\mathcal{S}_{(X,Y)}$ is a *product space*; otherwise, it is called a *non-product space*.

— For example, let $(X,Y) \sim N_2(\mathbf{0},\mathbf{I})$, then $\mathcal{S}_{(X,Y)} = \mathbb{R}^2 = \mathbb{R}\times\mathbb{R} = \mathcal{S}_X \times \mathcal{S}_Y$; that is $\mathcal{S}_{(X,Y)}$ is a product space.

— If $(X,Y) \sim U(\mathbb{B}_2)$, where $\mathbb{B}_2 = \{(x,y)\colon x^2+y^2 \leqslant 1\}$, then $\mathcal{S}_{(X,Y)} = \mathbb{B}_2 \neq [-1,1]\times[-1,1] = \mathcal{S}_X \times \mathcal{S}_Y$; that is, $\mathcal{S}_{(X,Y)}$ is a non-product space.

42.3• The point-wise formula

— We always have the following identity:

$$f_{(X|Y)}(x|y) f_Y(y) = f_{(Y|X)}(y|x) f_X(x), \quad (x,y) \in \mathcal{S}_{(X,Y)}. \tag{1.54}$$

— From (1.54), by division, we obtain

$$f_Y(y) = \frac{f_{(Y|X)}(y|x)}{f_{(X|Y)}(x|y)} \cdot f_X(x), \qquad x \in \mathcal{S}_X, \;\; y \in \mathcal{S}_Y. \tag{1.55}$$

— Integrating this identity with respect to y on support $\mathcal{S}_Y$, i.e.,

$$\int_{\mathcal{S}_Y} f_Y(y)\,\mathrm{d}y = \int_{\mathcal{S}_Y} \frac{f_{(Y|X)}(y|x)}{f_{(X|Y)}(x|y)} \cdot f_X(x)\,\mathrm{d}y = f_X(x) \cdot \int_{\mathcal{S}_Y} \frac{f_{(Y|X)}(y|x)}{f_{(X|Y)}(x|y)}\,\mathrm{d}y,$$

we immediately have the following *point-wise formula*:

$$f_X(x) = \left\{ \int_{\mathcal{S}_Y} \frac{f_{(Y|X)}(y|x)}{f_{(X|Y)}(x|y)}\,\mathrm{d}y \right\}^{-1}, \quad \text{for any } x \in \mathcal{S}_X. \tag{1.56}$$

42.4• The function-wise formula

— Now substituting (1.56) into (1.55), we obtain the dual form of *inverse Bayes formula* (IBF) for $f_Y(y)$ and hence by symmetry we obtain the *function-wise formula* of $f_X(x)$:

$$f_X(x) = \left\{ \int_{\mathcal{S}_X} \frac{f_{(X|Y)}(x|y_0)}{f_{(Y|X)}(y_0|x)}\,\mathrm{d}x \right\}^{-1} \frac{f_{(X|Y)}(x|y_0)}{f_{(Y|X)}(y_0|x)}, \tag{1.57}$$

for all $x \in \mathcal{S}_X$ and an arbitrarily fixed $y_0 \in \mathcal{S}_Y$.

42.5• The sampling-wise formula

— By dropping the normalizing constant in (1.57), we obtain the so-called *sampling-wise formula*:

$$f_X(x) \propto \frac{f_{(X|Y)}(x|y_0)}{f_{(Y|X)}(y_0|x)}, \tag{1.58}$$

for all $x \in \mathcal{S}_X$ and an arbitrarily fixed $y_0 \in \mathcal{S}_Y$.

42.6• Discrete versions of (1.56) and (1.58)

— When both X and Y are discrete random variables, we have

$$\Pr(X = x) = \left\{ \sum_{y \in \mathcal{S}_Y} \frac{\Pr(Y = y|X = x)}{\Pr(X = x|Y = y)} \right\}^{-1}, \tag{1.59}$$

for any $x \in \mathcal{S}_X$, which is called the discrete version of the point-wise formula.

— The discrete version of the sampling-wise formula is

$$\Pr(X = x) \propto \frac{\Pr(X = x|Y = y_0)}{\Pr(Y = y_0|X = x)}, \tag{1.60}$$

for all $x \in \mathcal{S}_X$ and an arbitrarily fixed $y_0 \in \mathcal{S}_Y$.

1.10.2 Understanding the IBF

43• Why do they have the name of IBF?

- To answer this question, we first introduce Bayes formula or Bayes Theorem.
- From (1.55), we obtain

$$f_{(Y|X)}(y|x) = \frac{f_{(X|Y)}(x|y) f_Y(y)}{f_X(x)} = \frac{f_{(X|Y)}(x|y) f_Y(y)}{\int f_{(X|Y)}(x|y) f_Y(y)\, \mathrm{d}y}.$$

- Replacing y by θ, the above identity becomes

$$p(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{f(x)} = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)\, \mathrm{d}\theta}, \tag{1.61}$$

which is called *Bayes formula*, where the parameter θ is treated as a random variable in Bayesian statistics, $\pi(\theta)$ is the prior density of θ, $p(\theta|x)$ is the posterior density of θ after the x was observed.

- The formula (1.61) states that given $f(x|\theta)$ and $\pi(\theta)$, the posterior density $p(\theta|x)$ can be determined uniquely.
- While (1.56)–(1.58) state that given both conditional densities, the marginal density can also be determined uniquely. This is why the name of IBF is taken.

44• Differences among the three IBF

- If we could obtain the *closed-form expression* for the integral in (1.56) or (1.57) (e.g., see Example 1.9 below), the three formulae (1.56)–(1.58) should have the same name.
- However, in practice, usually, it is not so.
- We need to numerically evaluate the integral by using the Monte Carlo method.

44.1• What does the "point-wise" mean?

— To evaluate the integral in (1.56), we must *fix the value of* x, say $x = x_0 = 0.5$, because the integration is with respect to the variable y.

— In other words, given a point $x = x_0$, (1.56) can only be used to calculate the value of the density $f_X(x)$ evaluated at $x = x_0$, i.e., $f_X(x_0)$.

— For example, given $\{x_i\}_{i=1}^n$, say $n = 100$, we can calculate the values of $\{f_X(x_i)\}_{i=1}^n$ by performing n integrations.

— That is why (1.56) is called the point-wise formula.

44.2• What does the "function-wise" mean?

— The normalizing constant in (1.57) is $f_Y(y_0)$, which can be obtained by performing one integration like in (1.56).

— In other words, by only performing one integration, we can obtain the expression of the marginal density $f_X(x)$.

— That is why (1.57) is called the function-wise formula.

44.3• What does the "sampling-wise" mean?

— A density $f(x) = c \cdot g(x)$ can be factorized into two parts: c and $g(x)$, where $c = 1/\int g(x)\,\mathrm{d}x$ is called the *normalizing constant* and $g(x)$ is call the *kernel* of $f(x)$.

— If we would like to generate random samples from $f_X(x)$, there are many methods.

— For some methods (e.g., the acceptance–rejection algorithm, the grid method, the sampling/important resampling algorithm), it is not necessary to know the value of the normalizing constant.

— And it only needs to know the kernel. In (1.58), we only know the kernel of $f_X(x)$.

— That is why (1.58) is called the sampling-wise formula.

44.4• Remarks

— Often in practice, we know $f_{(X|Y)}(x|y)$ only up to a normalizing constant.

— In other words, $f_{(X|Y)}(x|y) = c(y) \cdot g(x|y)$, where $c(y)$ is unknown and $g(x|y)$ is completely known, then the function-wise IBF (1.57) and sampling-wise IBF (1.58) still hold if we replace $f_{(X|Y)}(x|y_0)$ by $g(x|y_0)$.

1.10.3 Two examples

Example 1.9 (Bivariate normal distribution). Assume that

$$\begin{aligned} X|(Y=y) &\sim N(\mu_1 + \rho(y-\mu_2),\, 1-\rho^2) \quad \text{and} \\ Y|(X=x) &\sim N(\mu_2 + \rho(x-\mu_1),\, 1-\rho^2). \end{aligned}$$

Find the marginal distribution of X and the joint distribution of (X, Y).

Solution. Since $\mathcal{S}_{(X,Y)} = \mathcal{S}_X \times \mathcal{S}_Y = \mathbb{R}^2$, from (1.56), we obtain

$$\{f_X(x)\}^{-1} = \sqrt{2\pi}\exp\{(x-\mu_1)^2/2\},$$

which means $X \sim N(\mu_1, 1)$. Therefore, the joint distribution of (X, Y) exists and is bivariate normal with means μ_1 and μ_2, unit variances and correlation coefficient ρ.

Alternative solution. When using (1.56), we need to evaluate an integral. In contrast, using (1.58), the integration can be avoided. In fact, let the arbitrary y_0 be μ_2, then

$$f_X(x) \propto \exp\{-(x-\mu_1)^2/2\}. \qquad \|$$

Example 1.10 (Bivariate discrete distribution). Let X be a discrete random variable with probability mass function (pmf) $p_i = \Pr(X = x_i)$ for $i = 1, 2, 3$ and Y be a discrete random variable with pmf $q_j = \Pr(Y = y_j)$ for $j = 1, 2, 3$. Given two conditional distribution matrices

$$\mathbf{A} = (a_{ij}) = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = \begin{pmatrix} 1/6 & 0 & 3/14 \\ 0 & 1/4 & 4/14 \\ 5/6 & 3/4 & 7/14 \end{pmatrix}$$

and

$$\mathbf{B} = (b_{ij}) = \begin{pmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{pmatrix} = \begin{pmatrix} 1/4 & 0 & 3/4 \\ 0 & 1/3 & 2/3 \\ 5/18 & 6/18 & 7/18 \end{pmatrix},$$

where the (i, j) element of $\mathbf{A}$ is $a_{ij} = \Pr\{X = x_i | Y = y_j\}$ and the (i, j) element of $\mathbf{B}$ is $b_{ij} = \Pr\{Y = y_j | X = x_i\}$.

1) Find the marginal distributions of X and Y.

2) Find the joint distribution of (X, Y).

Solution. 1) The support of X and Y are $\mathcal{S}_X = \{x_1, x_2, x_3\}$ and $\mathcal{S}_Y = \{y_1, y_2, y_3\}$. By using (1.60) with $y_0 = y_3$, the X-marginal is given by

$$\begin{aligned} p_1 &\hat{=} \Pr(X = x_1) = f_X(x_1) \\ &\propto \frac{f_{(X|Y)}(x_1|y_0)}{f_{(Y|X)}(y_0|x_1)} = \frac{\Pr(X = x_1|Y = y_3)}{\Pr(Y = y_3|X = x_1)} \\ &= \frac{a_{13}}{b_{13}} = \frac{3/14}{3/4} = \frac{4}{14}, \\ p_2 &\hat{=} \Pr(X = x_2) = f_X(x_2) \\ &\propto \frac{f_{(X|Y)}(x_2|y_0)}{f_{(Y|X)}(y_0|x_2)} = \frac{\Pr(X = x_2|Y = y_3)}{\Pr(Y = y_3|X = x_2)} \\ &= \frac{a_{23}}{b_{23}} = \frac{4/14}{2/3} = \frac{6}{14}, \end{aligned}$$

$$\begin{aligned} p_3 &\hat{=} \Pr(X = x_3) = f_X(x_3) \\ &\propto \frac{f_{(X|Y)}(x_3|y_0)}{f_{(Y|X)}(y_0|x_3)} = \frac{\Pr(X = x_3|Y = y_3)}{\Pr(Y = y_3|X = x_3)} \\ &= \frac{a_{33}}{b_{33}} = \frac{7/14}{7/18} = \frac{18}{14}. \end{aligned}$$

Note that $p_1 + p_2 + p_3 = 1$, we obtain

$$\begin{aligned} p_1 &= \frac{4/14}{4/14 + 6/14 + 18/14} = \frac{4}{4+6+18} = \frac{4}{28} = \frac{2}{14}, \\ p_2 &= \frac{6/14}{4/14 + 6/14 + 18/14} = \frac{6}{4+6+18} = \frac{6}{28} = \frac{3}{14}, \\ p_3 &= \frac{18/14}{4/14 + 6/14 + 18/14} = \frac{18}{4+6+18} = \frac{18}{28} = \frac{9}{14}, \end{aligned}$$

which are summarized into

X	x_1	x_2	x_3
$p_i = \Pr(X = x_i)$	2/14	3/14	9/14

Similarly, letting $x_0 = x_3$ in (1.60) yields the following Y-marginal

Y	y_1	y_2	y_3
$q_j = \Pr(Y = y_j)$	3/14	4/14	7/14

2) The joint distribution of (X, Y) is given by

$$\mathbf{P} = \begin{pmatrix} 1/28 & 0 & 3/28 \\ 0 & 2/28 & 4/28 \\ 5/28 & 6/28 & 7/28 \end{pmatrix}. \qquad \|$$

1.11 Categorical Distribution

45• Finite discrete distribution

- Let the discrete r.v. X be defined as follows:

X	x_1	$\cdots$	x_i	$\cdots$	x_d
$\Pr(X = x_i)$	p_1	$\cdots$	p_i	$\cdots$	p_d

  where the probabilities $p_i > 0$ and $\sum_{i=1}^{d} p_i = 1$.

- When d is finite, we say X follows a finite discrete distribution, denoted by $X \sim \text{FDiscrete}_d(\boldsymbol{x}, \boldsymbol{p})$, where $\boldsymbol{x} = (x_1, \ldots, x_d)^\top$ and $\boldsymbol{p} = (p_1, \ldots, p_d)^\top \in \mathbb{T}_d \hat{=} \{(p_1, \ldots, p_d)^\top \colon p_i > 0,\ \sum_{i=1}^d p_i = 1\}$.

45.1• Basic features of $X \sim \text{FDiscrete}_d(\boldsymbol{x}, \boldsymbol{p})$

— X is an r.v., not a random vector.

— The support of X is $\{x_1, \ldots, x_d\}$.

— $\{x_i\}_{i=1}^d$ are *numeric* or real numbers.

— $E(X) = \sum_{i=1}^d x_i p_i$ is meaningful.

— Bernoulli, binomial, and hypergeometric distributions are special cases.

46• Categorical random variable

- Define a categorical random variable Y with d-category as follows:

Y	A_1	$\cdots$	A_i	$\cdots$	A_d
$\Pr(Y = A_i)$	p_1	$\cdots$	p_i	$\cdots$	p_d

  where $\{A_i\}_{i=1}^d$ denote *characters, labels* or *symbols*.

46.1• Some examples

— Blood types can be classified as A, B, AB, and O.

— USA citizen can be classified as White, African-American, Asian or Pacific Islander, and Native America.

— The color can be classified as red, green, blue, and others.

46.2• A key feature

— Obviously, $E(Y)$ is *meaningless*.

— Therefore, we cannot define above distribution of Y as a categorial distribution.

47• Categorical distribution

- However, we can define a *one-to-one mapping* between a random vector $\mathbf{y} = (Y_1, \ldots, Y_d)^\top$ and the above categorical random variable Y.
- For the purpose of illustration, let $d = 4$, we define

$$\begin{aligned}
\mathbf{y} &= (1,0,0,0)^\top = \boldsymbol{e}_4^{(1)} \leftrightarrow Y = A_1 \text{ (if the blood type is A)},\\
\mathbf{y} &= (0,1,0,0)^\top = \boldsymbol{e}_4^{(2)} \leftrightarrow Y = A_2 \text{ (if the blood type is B)},\\
\mathbf{y} &= (0,0,1,0)^\top = \boldsymbol{e}_4^{(3)} \leftrightarrow Y = A_3 \text{ (if the blood type is AB)},\\
\mathbf{y} &= (0,0,0,1)^\top = \boldsymbol{e}_4^{(4)} \leftrightarrow Y = A_4 \text{ (if the blood type is O)}.
\end{aligned}$$

- Therefore, we can define the categorical distribution as follows:

$\mathbf{y}$	$\boldsymbol{e}_d^{(1)}$	$\cdots$	$\boldsymbol{e}_d^{(i)}$	$\cdots$	$\boldsymbol{e}_d^{(d)}$
$\Pr\{Y = \boldsymbol{e}_d^{(i)}\}$	p_1	$\cdots$	p_i	$\cdots$	p_d

  and denote it by $\mathbf{y} = (Y_1, \ldots, Y_d)^\top \sim \text{Categorical}_d(\boldsymbol{p})$, which is a special case of the multinomial distribution with the first parameter being 1, i.e., $\mathbf{y} \sim \text{Multinomial}_d(1; \boldsymbol{p})$.
- The pmf of $\mathbf{y}$ is given by

$$\Pr(\mathbf{y} = \boldsymbol{y}) = \binom{1}{y_1, \ldots, y_d} \prod_{i=1}^{d} p_i^{y_i} = \prod_{i=1}^{d} p_i^{y_i}, \tag{1.62}$$

  where $\boldsymbol{y} = (y_1, \ldots, y_d)^\top$, $y_i = 0$ or 1 and $\sum_{i=1}^{d} y_i = 1$. In other words, only one component of $\boldsymbol{y}$ is 1 and others are zeros.

47.1• Basic features of $\mathbf{y} \sim \text{Categorical}_d(\boldsymbol{p})$

— $\mathbf{y}$ is a random vector.

— The support of $\mathbf{y}$ is $\{\boldsymbol{e}_d^{(1)}, \ldots, \boldsymbol{e}_d^{(d)}\}$.

— $\{\boldsymbol{e}_d^{(1)}, \ldots, \boldsymbol{e}_d^{(d)}\}$ are unit/base vectors.

— $E(\mathbf{y}) = \boldsymbol{p}$ is meaningful.

— It reduces to Bernoulli distribution when $d = 2$.

— There is a one-to-one mapping between $\mathbf{y} \sim \text{Categorical}_d(\boldsymbol{p})$ and $X \sim \text{FDiscrete}_d(\boldsymbol{x}, \boldsymbol{p})$.

— If $\mathbf{y}_1, \ldots, \mathbf{y}_n \overset{\text{iid}}{\sim} \text{Categorical}_d(\boldsymbol{p})$, then

$$\sum_{j=1}^{n} \mathbf{y}_j \sim \text{Multinomial}_d(n; \boldsymbol{p}). \tag{1.63}$$

1.12 Zero-inflated Poisson Distribution

48• Definition

- Let $Z \sim \text{Bernoulli}(1-\phi)$, $X \sim \text{Poisson}(\lambda)$ and $Z \perp\!\!\!\perp X$.
- Let $Y \overset{\text{d}}{=} ZX$, we say Y follows the *zero-inflated Poisson* (ZIP) distribution, denoted by $Y \sim \text{ZIP}(\phi, \lambda)$.

48.1• The pmf of $Y \sim \text{ZIP}(\phi, \lambda)$

— The r.v. $Y \sim \text{ZIP}(\phi, \lambda)$ has the following stochastic representation:

$$Y \overset{\text{d}}{=} ZX = \begin{cases} 0, & \text{with probability } \phi, \\ X, & \text{with probability } 1-\phi \end{cases} \tag{1.64}$$

with support $\mathcal{S}_Y = \{0, 1, 2, \ldots\}$.

— Since

$$\{Y = 0\} \Leftrightarrow \{Z = 0\} \quad \text{or} \quad \{Z = 1, X = 0\}$$

and for $y > 0$,

$$\{Y = y\} \Leftrightarrow \{Z = 1, X = y\},$$

we obtain

$$\begin{aligned} \Pr(Y = 0) &= \Pr(Z = 0) + \Pr(Z = 1, X = 0) = \phi + (1-\phi)\,\mathrm{e}^{-\lambda}, \\ \Pr(Y = y) &= \Pr(Z = 1, X = y) = (1-\phi)\,\Pr(X = y) \\ &= (1-\phi)\frac{\lambda^y\,\mathrm{e}^{-\lambda}}{y!}, \qquad y > 0. \end{aligned}$$

— The pmf of Y is given by

$$\begin{aligned} & f(y|\phi,\lambda) = \Pr(Y=y) \\ = & \begin{cases} \phi + (1-\phi)\,\mathrm{e}^{-\lambda}, & \text{if } y=0, \\ (1-\phi)\dfrac{\lambda^y\,\mathrm{e}^{-\lambda}}{y!}, & \text{if } y \geqslant 1 \end{cases} \\ = & [\phi + (1-\phi)\,\mathrm{e}^{-\lambda}]I(y=0) + \left[(1-\phi)\frac{\lambda^y\,\mathrm{e}^{-\lambda}}{y!}\right]I(y>0) \qquad (1.65) \\ = & \phi \cdot I(y=0) + (1-\phi)\Pr(X=y), \end{aligned}$$

where the $\phi \in [0,1)$ denotes the unknown proportion for incorporating extra-zeros than those permitted by the original Poisson distribution.

48.2• Comments

— The ZIP distribution may be viewed as a *mixture* of a degenerate distribution with all mass at zero (denoted by $\xi \sim$ Degenerate(0)) and a Poisson(λ) distribution.

— In particular, when $\phi = 0$, the ZIP distribution is reduced to the original Poisson distribution.

— The ZIP distribution can be used to model count data with *extra zeros*.

48.3• Examples of count data with extra zeros

— Number of insurance claims within a population for a certain type of risk would be zero-inflated by those people who have not taken out insurance against the risk and thus are unable to claim.

— Number of workdays missed due to sickness of a dependent in a 4-week period.

— Number of papers published per year by a researcher.

49• Basic properties

49.1• Expectation and variance of Y

— From (1.64), we immediately obtain

$$\begin{aligned} E(Y) &= E(ZX) &= E(Z)E(X) &= (1-\phi)\lambda, \\ E(Y^2) &= E(Z^2X^2) &= E(Z)E(X^2) &= (1-\phi)(\lambda^2+\lambda), \end{aligned}$$

and $\mathrm{Var}(Y) = E(Y^2) - [E(Y)]^2$.

49.2• The mgf of Y

— We have

$$\begin{aligned} M_Y(t) &= E(\mathrm{e}^{tY}) = E(\mathrm{e}^{tZX}) \\ &= \phi + (1-\phi)E(\mathrm{e}^{tX}) \\ &= \phi + (1-\phi)\exp[\lambda(\mathrm{e}^t - 1)]. \end{aligned}$$

49.3• Show that $Y|(Z=z) \sim$ Poisson(λz)

— It is obvious. □

49.4• The conditional distribution of $Z|Y$

— The r.v. Z is a Bernoulli variable, only taking the value 0 or 1.

— Note that

$$\begin{aligned} \Pr(Z=1|Y=y) &= \frac{\Pr(Z=1,Y=y)}{\Pr(Y=y)} = \frac{\Pr(Z=1,X=y)}{f(y|\phi,\lambda)} \\ &= \frac{(1-\phi)\,\mathrm{e}^{-\lambda}\lambda^y/y!}{f(y|\phi,\lambda)} \;\hat{=}\; p_y, \end{aligned}$$

where $f(y|\phi,\lambda)$ is given by (1.65), then we have

$$p_0 = \frac{(1-\phi)\,\mathrm{e}^{-\lambda}}{\phi + (1-\phi)\,\mathrm{e}^{-\lambda}} \qquad \text{and} \qquad p_y = 1 \text{ for } y > 0. \tag{1.66}$$

— Therefore,

$$Z|(Y=y) \sim \begin{cases} \text{Bernoulli}\,(p_0), & \text{if } y = 0, \\ \text{Degenerate}(1), & \text{if } y > 0. \end{cases} \tag{1.67}$$

49.5• The conditional distribution of $X|Y$

— We first find the conditional distribution of $X|(Y = y = 0)$.

— Note that

$$\begin{aligned}
& \Pr(X = x|Y = 0) \\
= \quad & \frac{\Pr(X = x, Y = 0)}{\Pr(Y = 0)} \\
= \quad & \frac{\Pr(X = 0, Y = 0)}{f(0|\phi, \lambda)} I_{(x=0)} + \frac{\Pr(X = x, Z = 0)}{f(0|\phi, \lambda)} I_{(x>0)} \\
= \quad & \frac{\Pr(X = 0)}{f(0|\phi, \lambda)} I_{(x=0)} + \frac{\phi \Pr(X = x)}{f(0|\phi, \lambda)} I_{(x>0)} \quad [\because \{X = 0\} \subseteq \{Y = 0\}] \\
= \quad & \frac{\mathrm{e}^{-\lambda}}{\phi + (1-\phi)\,\mathrm{e}^{-\lambda}} I_{(x=0)} + \frac{\phi}{\phi + (1-\phi)\,\mathrm{e}^{-\lambda}} \cdot \frac{\mathrm{e}^{-\lambda}\lambda^x}{x!} I_{(x>0)} \\
\overset{(1.66)}{=} \quad & [p_0 + (1-p_0)\,\mathrm{e}^{-\lambda}] I_{(x=0)} + \left[(1-p_0)\frac{\mathrm{e}^{-\lambda}\lambda^x}{x!}\right] I_{(x>0)}. \qquad (1.68)
\end{aligned}$$

— By comparing (1.68) with (1.65), we have

$$X|(Y = 0) \sim \mathrm{ZIP}(p_0, \lambda). \qquad (1.69)$$

— We then find the conditional distribution of $X|(Y = y > 0)$.

— Note that

$$\begin{aligned}
& \Pr(X = x|Y = y) \\
= \quad & \frac{\Pr(X = x, Y = y)}{\Pr(Y = y)} \qquad [\because y > 0 \Rightarrow x = y > 0 \;\&\; Z = 1] \\
= \quad & \frac{\Pr(X = y, Z = 1)}{f(y|\phi, \lambda)} \\
\overset{(1.65)}{=} \quad & \frac{(1-\phi)\Pr(X = y)}{(1-\phi)\,\mathrm{e}^{-\lambda}\lambda^y/y!} = 1,
\end{aligned}$$

implying that $X|(Y = y > 0) \sim \mathrm{Degenerate}(y)$.

Exercise 1

1.1 Let $X \sim \text{Binomial}(n, p)$, $Y \sim \text{Poisson}(\lambda)$, and $X \perp\!\!\!\perp Y$.

(a) Find the mgf of X.

(b) Use the formula (1.34) to find the variance of X.

(c) Find the distribution of $X + Y$.

1.2 The joint pmf of X and Y is given by

(X, Y)	$(1,1)$	$(1,2)$	$(1,3)$	$(1,4)$	$(2,2)$
Probability	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{2}{16}$
(X, Y)	$(2,3)$	$(2,4)$	$(3,3)$	$(3,4)$	$(4,4)$
Probability	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{3}{16}$	$\frac{1}{16}$	$\frac{4}{16}$

(a) Find the marginal distribution of X.

(b) Find the pmf of $X + Y$.

1.3 Let two conditional distributions be exponential restricted to the interval $[0, b)$; that is,

$$f_{(X|Y)}(x|y) = \frac{y \exp(-yx)}{1 - \exp(-by)}, \quad 0 \leqslant x < b < +\infty,$$

$$f_{(Y|X)}(y|x) = \frac{x \exp(-xy)}{1 - \exp(-bx)}, \quad 0 \leqslant y < b < +\infty.$$

(a) Find the marginal distribution of X.

(b) If $b = +\infty$, please discuss the existence of $f_X(x)$.

1.4 Let X be a discrete r.v. with pmf $p_i = \Pr(X = x_i)$ for $i = 1, 2, 3$ and Y be another discrete r.v. with pmf $q_j = \Pr(Y = y_j)$ for $j = 1, 2, 3, 4$. Given two conditional distribution matrices

$$\mathbf{A} = \begin{pmatrix} 1/7 & 1/4 & 3/7 & 1/7 \\ 2/7 & 1/2 & 1/7 & 2/7 \\ 4/7 & 1/4 & 3/7 & 4/7 \end{pmatrix}$$

and

$$\mathbf{B} = \begin{pmatrix} 1/6 & 1/6 & 1/2 & 1/6 \\ 2/7 & 2/7 & 1/7 & 2/7 \\ 1/3 & 1/12 & 1/4 & 1/3 \end{pmatrix},$$

where the (i, j) element of $\mathbf{A}$ is $a_{ij} = \Pr(X = x_i|Y = y_j)$ and the (i, j) element of $\mathbf{B}$ is $b_{ij} = \Pr(Y = y_j|X = x_i)$.

(a) Find the marginal distributions of X and Y.

(b) Find the joint distribution of (X, Y).

1.5 Let X be a continuous r.v. with pdf $f(x)$. If m is the unique median of the distribution of X and b is a real constant.

(a) Show that

$$E(|X - b|) = E(|X - m|) + 2\int_m^b (b - x)f(x)\,\mathrm{d}x,$$

provided that the expectation exists.

(b) Find the value of b such that $E(|X - b|)$ is minimized.

1.6 Let X be a r.v. having the following cdf

$$F(x) = \begin{cases} 0, & x < 0, \\ 2x^2, & 0 \leqslant x < 1/2, \\ 1 - 2(1 - x)^2, & 1/2 \leqslant x < 3/4, \\ 1, & 3/4 \leqslant x. \end{cases}$$

(a) Calculate $\Pr(1/4 < X < 5/8)$.

(b) Find the variance of X.

1.7 Let $\mathbf{x} = (X_1, \ldots, X_d)^\top$ be a random vector, the joint mgf of $\mathbf{x}$ is defined by $M_{\mathbf{x}}(\boldsymbol{t}) = E[\exp(t_1X_1 + \cdots + t_dX_d)]$, where $\boldsymbol{t} = (t_1, \ldots, t_d)^\top$ be a real vector. For either the discrete case or the continuous case, prove that

(a) For a fixed i $(i = 1, \ldots, d)$, the partial derivative of the joint mgf with respect to t_i evaluated at $t_1 = \cdots = t_d = 0$ is $E(X_i)$.

(b) For two fixed i, j $(i \neq j,\ i, j = 1, \ldots, d)$, the second derivative of the joint mgf with respect to t_i and t_j evaluated at $t_1 = \cdots = t_d = 0$ is $E(X_iX_j)$.

(c) If two r.v.'s have the following joint density

$$f(x, y) = \begin{cases} \exp(-x - y), & \text{for } x > 0,\ y > 0, \\ 0, & \text{elsewhere,} \end{cases}$$

find the joint mgf and use it to derive $E(XY)$, $E(X)$, $E(Y)$ and $\mathrm{Cov}(X, Y)$.

1.8 A discrete r.v. X is said to follow a *zero-truncated Poisson* (ZTP) distribution if its pmf is

$$\Pr(X = x) = c \cdot \frac{\lambda^x \, \mathrm{e}^{-\lambda}}{x!}, \quad x = 1, 2, \ldots,$$

where $\lambda > 0$ is an unknown parameter, and c is the normalizing constant such that $\sum_{x=1}^{\infty} \Pr(X = x) = 1$. We will write $X \sim \mathrm{ZTP}(\lambda)$.

A. Let $X \sim \mathrm{ZTP}(\lambda)$.

(a) Find the normalizing constant c.

(b) Find $E(X)$, $E(X^2)$ and $\mathrm{Var}(X)$.

(c) Find the moment generating function of X.

B. Let $X_1 \sim \mathrm{ZTP}(\lambda_1)$, $X_2 \sim \mathrm{ZTP}(\lambda_2)$, and they are independent.

(d) Find the distribution of $X_1 + X_2$.

(e) Find the conditional distribution of $X_1 | (X_1 + X_2 = x)$, where $x \geqslant 2$ and x is an integer.

1.9 Let $U \sim \mathrm{Poisson}(\lambda_0)$, $V \sim \mathrm{Poisson}(\lambda)$, $W \sim \mathrm{Poisson}(\beta\lambda)$, $Z \sim \mathrm{Bernoulli}(1 - \phi)$, and U, V, W, Z are mutually independent. Define

$$X = U + V \quad \text{and} \quad Y = Z(U + W).$$

(a) Find $\mathrm{Var}(X)$, $\mathrm{Var}(Y)$ and $\mathrm{Cov}(X, Y)$.

(b) Find the joint distribution of X and Y.

Chapter 2

Sampling Distributions

2.1 Distribution of the Function of Random Variables

1• Aim of this section

- Given a set of r.v.'s $X_1, \ldots, X_n$ with the cdf $F(x_1, \ldots, x_n)$ or the pdf $f(x_1, \ldots, x_n)$, we seek the distribution of $Y = h(X_1, \ldots, X_n)$ for some function $h(\cdot)$.
- In this section, we will introduce three commonly used methods.

1.1• Three techniques

— Cumulative distribution function technique.

— Transformation technique.

— Moment generating function technique.

2.1.1 Cumulative distribution function technique

2• The continuous case

- A set of r.v.'s $X_1, \ldots, X_n$ can define a new r.v. $Y = h(X_1, \ldots, X_n)$ via the function $h(\cdot)$.
- The distribution of Y can be determined by the transformation $h(\cdot)$ together with the joint distribution of $X_1, \ldots, X_n$.

2.1• The procedure of cdf

— If $X_1, \ldots, X_n$ are continuous r.v.'s, then the cdf of Y can be determined by integrating $f(x_1, \ldots, x_n)$ over the domain

$$\mathbb{D} = \{(x_1, \ldots, x_n)\colon\ h(x_1, \ldots, x_n) \leqslant y\};$$

that is

$$\begin{aligned} G(y) &= \Pr(Y \leqslant y) \\ &= \Pr\{h(X_1, \ldots, X_n) \leqslant y\} \\ &= \int_{\mathbb{D}} f(x_1, \ldots, x_n)\, \mathrm{d}x_1 \cdots \mathrm{d}x_n. \end{aligned}$$

— Then by differentiating it with respect to y, we obtain the density of Y as $g(y) = G'(y)$.

Example 2.1 (Beta distribution). Suppose that $X \sim \text{Beta}(2, 2)$, then its pdf is $f(x) = 6x(1-x)$, $0 \leqslant x \leqslant 1$. Find the pdf of $Y = X^3$.

Solution. The distribution function of Y for $0 \leqslant y \leqslant 1$ is

$$\begin{aligned} G(y) &= \Pr(X^3 \leqslant y) \\ &= \Pr(X \leqslant y^{1/3}) \\ &= \int_0^{y^{1/3}} 6x(1-x)\, \mathrm{d}x \\ &= 3y^{2/3} - 2y. \end{aligned}$$

Then, the pdf of Y is $g(y) = 2y^{-1/3} - 2$, $0 \leqslant y \leqslant 1$.

The corresponding densities and distribution functions of $X \sim \text{Beta}(2, 2)$ and $Y = X^3$ are shown in Figure 2.1. ‖

Example 2.2 (Bivariate exponential distribution). Let

$$(X_1, X_2) \sim f(x_1, x_2) = 6\exp(-3x_1 - 2x_2), \quad x_1 \geqslant 0, \quad x_2 \geqslant 0.$$

Find the pdf of $Y = X_1 + X_2$.

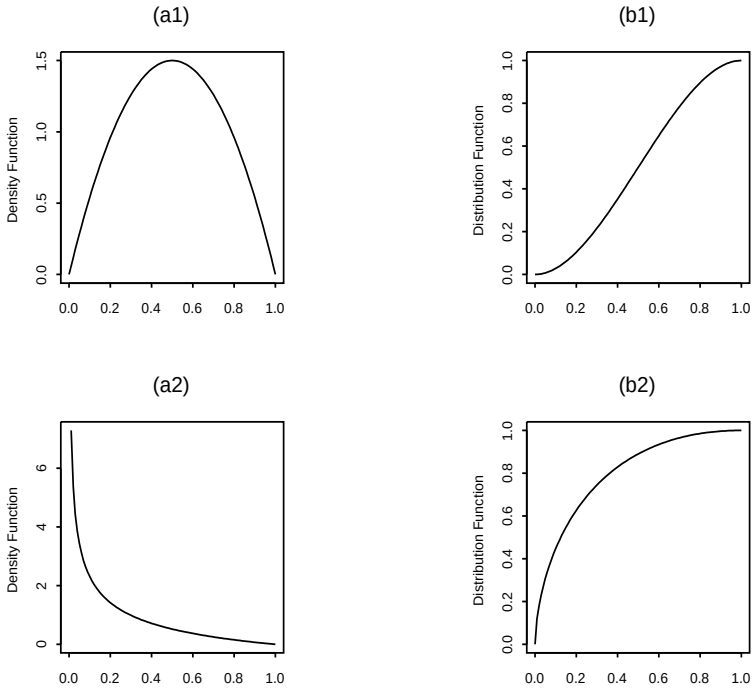

Figure 2.1 The densities and distribution functions of $X \sim \text{Beta}(2,2)$ and $Y = X^3$. (a1) The density $f(x)$ of X; (b1) The cdf $F(x)$ of X; (a2) The density $g(y)$ of Y; (b2) The cdf $G(y)$ of Y.

Solution. The cdf of Y is

$$\begin{aligned} G(y) &= \int\int_{\mathbb{D}} 6\exp(-3x_1 - 2x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2 \\ &= \int_0^y \left\{ \int_0^{y-x_2} 6\exp(-3x_1 - 2x_2)\,\mathrm{d}x_1 \right\} \mathrm{d}x_2 \\ &= \int_0^y 2\,\mathrm{e}^{-2x_2}\{1 - \mathrm{e}^{-3(y-x_2)}\}\,\mathrm{d}x_2 \\ &= 1 + 2\,\mathrm{e}^{-3y} - 3\,\mathrm{e}^{-2y}, \quad y \geqslant 0, \end{aligned}$$

where $\mathbb{D} = \{(x_1, x_2)\colon\ x_1 \geqslant 0, x_2 \geqslant 0, x_1 + x_2 \leqslant y\}$ with $y \geqslant 0$ denotes the integration region. Figure 2.2 gives an illustration for the $\mathbb{D}$.

Therefore, the density of Y is

$$g(y) = 6(\,\mathrm{e}^{-2y} - \,\mathrm{e}^{-3y}), \quad y \geqslant 0.$$

Figure 2.3 shows the pdf $g(y)$ and the cdf $G(y)$ of $Y = X_1 + X_2$.

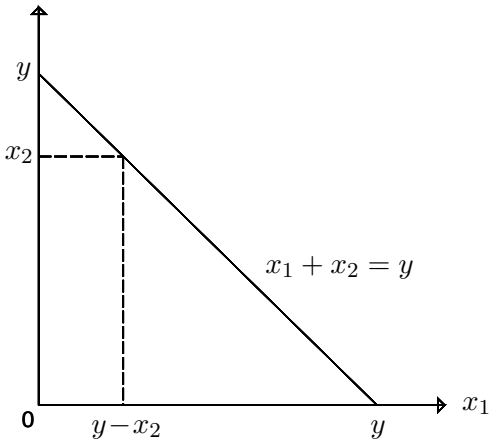

Figure 2.2 The integration region $\mathbb{D} = \{(x_1, x_2)\colon\ x_1 \geqslant 0, x_2 \geqslant 0, x_1 + x_2 \leqslant y\}$.

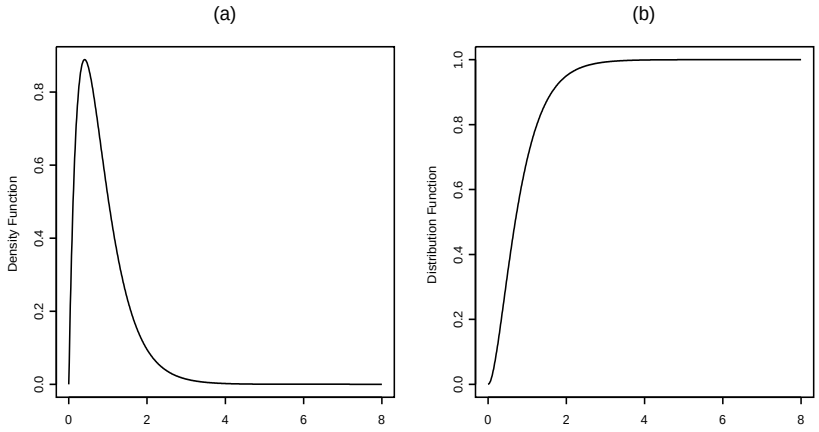

Figure 2.3 The density function and distribution function of $Y = X_1 + X_2$. (a) The pdf $g(y)$ of Y; (b) The cdf $G(y)$ of Y. ‖

3• The discrete case

- For the purpose of illustration, first we let $n = 1$.
- If X is a discrete r.v. taking values $\{x_i\}$ with probabilities $\{p_i\}$, then the distribution of $Y = h(X)$ is determined directly by the laws of probability.

- It may be that several values of X give rise to the same value of Y.
- The probability that Y takes on a given value, say y_j, is

$$\Pr(Y = y_j) = \sum_{\{i:\, h(x_i)=y_j\}} p_i.$$

Example 2.3 (Finite discrete distribution). Suppose that X takes the values of 0, 1, 2, 3, 4, 5 with the corresponding probabilities p_0, p_1, p_2, p_3, p_4 and p_5. Find the pmf of $Y = h(X) = (X-2)^2$.

Solution. From the following table

X	0	1	2	3	4	5
$p_i = \Pr(X = x_i)$	p_0	p_1	p_2	p_3	p_4	p_5
$Y = (X-2)^2$	4	1	0	1	4	9

we note that Y can take on values 0, 1, 4 and 9; then

$$\begin{aligned} \Pr(Y=0) &= p_2, & \Pr(Y=1) &= p_1 + p_3, \\ \Pr(Y=4) &= p_0 + p_4, & \Pr(Y=9) &= p_5. \end{aligned}$$

‖

Example 2.4 (Joint discrete distribution). Let (X_1, X_2, X_3) have a joint discrete distribution given by

(X_1, X_2, X_3)	$(0,0,0)$	$(0,0,1)$	$(0,1,1)$	$(1,0,1)$	$(1,1,0)$	$(1,1,1)$
Probability	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$

Find the pmf of $Y = h(X_1, X_2, X_3) = X_1 + X_2 + X_3$.

Solution. We note that Y can take on values 0, 1, 2 and 3; then

$$\begin{aligned} \Pr(Y=0) &= \frac{1}{8}, \\ \Pr(Y=1) &= \frac{3}{8}, \\ \Pr(Y=2) &= \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{3}{8}, \\ \Pr(Y=3) &= \frac{1}{8}. \end{aligned}$$

‖

Example 2.5 (Poisson distribution). Let $X_i \sim \text{Poisson}(\lambda_i)$, $i = 1, 2$, and $X_1 \perp\!\!\!\perp X_2$, find the pmf of $Y = X_1 + X_2$.

Solution. The pmf of $Y = X_1 + X_2$ is

可以直观上理解

$$\begin{aligned}
\Pr(Y = y) &= \Pr(X_1 + X_2 = y) \\
&= \sum_{x=0}^{y} \Pr(X_1 = x, X_2 = y - x) \\
&= \sum_{x=0}^{y} \Pr(X_1 = x) \cdot \Pr(X_2 = y - x) \\
&= \sum_{x=0}^{y} \frac{\lambda_1^x}{x!} \mathrm{e}^{-\lambda_1} \cdot \frac{\lambda_2^{y-x}}{(y-x)!} \mathrm{e}^{-\lambda_2} \\
&= \frac{1}{y!} \mathrm{e}^{-(\lambda_1+\lambda_2)} \sum_{x=0}^{y} \binom{y}{x} \lambda_1^x \lambda_2^{y-x} \\
&= \frac{(\lambda_1 + \lambda_2)^y}{y!} \mathrm{e}^{-(\lambda_1+\lambda_2)}, \quad y = 0, 1, 2, \ldots.
\end{aligned}$$

卷积适合于两个独立的随机变量的和的分布

离散型卷积公式

Therefore, $X_1 + X_2 \sim \text{Poisson}(\lambda_1 + \lambda_2)$. ‖

2.1.2 Transformation technique

4• Monotone transformation

- Let $f(x)$ and $F(x)$ denote the corresponding pdf and cdf of an r.v. X.
- If $y = h(x)$ is a differentiable and monotone function and the inverse function is $x = h^{-1}(y)$, then the pdf of $Y = h(X)$ is given by

$$g(y) = f(x) \times \left| \frac{\mathrm{d}x}{\mathrm{d}y} \right| = f(h^{-1}(y)) \times \left| \frac{\mathrm{d}h^{-1}(y)}{\mathrm{d}y} \right|. \tag{2.1}$$

  Proof. We first assume that $y = h(x)$ is increasing. Thus, $\mathrm{d}h(x)/\mathrm{d}x \geqslant 0$ and $\mathrm{d}h^{-1}(y)/\mathrm{d}y \geqslant 0$. Since

$$\begin{aligned}
G(y) &= \Pr(Y \leqslant y) = \Pr\{h^{-1}(Y) \leqslant h^{-1}(y)\} \\
&= \Pr\{X \leqslant h^{-1}(y)\} = F(h^{-1}(y)),
\end{aligned}$$

by differentiating, we have

$$\begin{aligned} g(y) &= \frac{\mathrm{d}G(y)}{\mathrm{d}y} \\ &= \frac{\mathrm{d}F(h^{-1}(y))}{\mathrm{d}y} \qquad \text{let } x = h^{-1}(y) \\ &= \left.\frac{\mathrm{d}F(x)}{\mathrm{d}x}\right|_{x=h^{-1}(y)} \times \frac{\mathrm{d}x}{\mathrm{d}y} \\ &= f(h^{-1}(y)) \times \frac{\mathrm{d}h^{-1}(y)}{\mathrm{d}y}. \end{aligned}$$

When $y = h(x)$ is decreasing, the proof is similar. □

Example 2.6 (Pareto distribution). Suppose that X has the Pareto density $f(x) = \theta x^{-\theta-1}$, $x \geqslant 1$, $\theta > 0$, find the pdf of $Y = \log(X)$.

Solution. Because $y = \log(x)$ is increasing with inverse $x = \mathrm{e}^y$, we have

$$\begin{aligned} g(y) &= f(x) \times \left|\frac{\mathrm{d}x}{\mathrm{d}y}\right| \\ &= \theta x^{-\theta-1} \cdot \mathrm{e}^y = \theta\, \mathrm{e}^{-\theta y}, \quad y \geqslant 0. \end{aligned}$$

Thus, Y follows an exponential distribution with mean $1/\theta$. Figure 2.4 shows the density functions of X and Y.

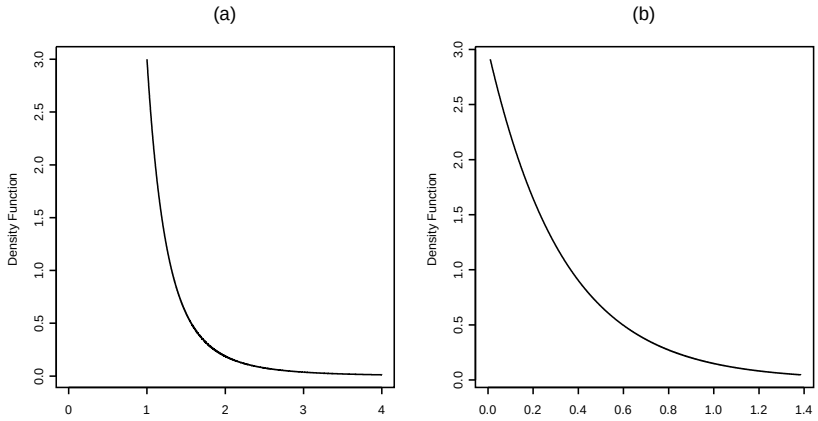

Figure 2.4 (a) The Pareto density $f(x) = \theta x^{-\theta-1} I_{[1,\infty)}(x)$; (b) The density of $Y = \log(X) \sim$ Exponential(θ). ‖

5• Piecewise monotone transformation

- Let $\mathbb{A}_1, \ldots, \mathbb{A}_n$ be a partition of the real line $\mathbb{R} = (-\infty, \infty)$, i.e., they are mutually exclusive and $\cup_{i=1}^n \mathbb{A}_i = \mathbb{R}$.
- If $y = h(x)$ is monotone on each $\mathbb{A}_i$, then $h_i(x) \hat{=} h(x) I_{\mathbb{A}_i}(x)$ has a unique inverse h_i^{-1} on $\mathbb{A}_i$, and the pdf of Y is given by

$$g(y) = \sum_{i=1}^{n} f(h_i^{-1}(y)) \times \left| \frac{\mathrm{d}h_i^{-1}(y)}{\mathrm{d}y} \right|. \tag{2.2}$$

Example 2.7 (Standard normal distribution). Let $X \sim N(0,1)$, find the pdf of $Y = X^2$.

Solution. The function $y = x^2$ is decreasing on $\mathbb{A}_1 = (-\infty, 0]$ and increasing on $\mathbb{A}_2 = (0, \infty)$. For $y \geqslant 0$, the inverse in $\mathbb{A}_1$ is $x = -\sqrt{y}$ and the inverse in $\mathbb{A}_2$ is $x = \sqrt{y}$. We apply (2.2) to get

$$\begin{aligned} g(y) &= \sum_{i=1}^{2} f(h_i^{-1}(y)) \times \left| \frac{\mathrm{d}h_i^{-1}(y)}{\mathrm{d}y} \right| \\ &= \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-y/2} \cdot \frac{y^{-1/2}}{2} + \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-y/2} \cdot \frac{y^{-1/2}}{2} \\ &= \frac{1}{\sqrt{2\pi}} y^{-1/2} \mathrm{e}^{-y/2} \\ &= \frac{(1/2)^{1/2}}{\Gamma(1/2)} y^{-1/2} \mathrm{e}^{-y/2}. \end{aligned}$$

Then, $Y = X^2 \sim \text{Gamma}(1/2, 1/2) = \chi^2(1)$.

Figure 2.5 shows the density functions of the standard normal distribution and the chi-squared distribution with 1 degree of freedom. ‖

6• Bivariate transformation

- Let $(X_1, X_2) \sim f(x_1, x_2)$.
- Let the functions $y_i = h_i(x_1, x_2)$ for $i = 1, 2$ are differentiable and their inverse functions

$$x_i = h_i^{-1}(y_1, y_2) \quad \text{for} \quad i = 1, 2$$

exist.

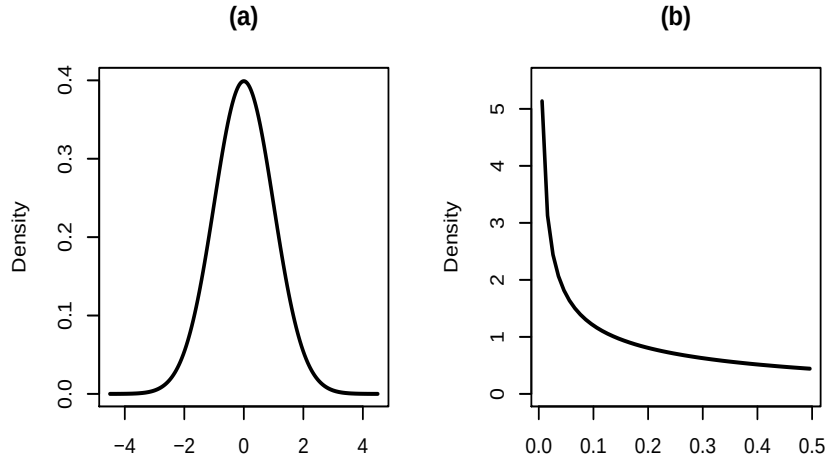

Figure 2.5 (a) The pdf of $X \sim N(0,1)$; (b) The pdf of $Y = X^2 \sim \chi^2(1)$.

- Then, the joint pdf of $Y_1 = h_1(X_1, X_2)$ and $Y_2 = h_2(X_1, X_2)$ is

$$\begin{aligned} g(y_1, y_2) &= f(x_1, x_2) \times |J(x_1, x_2 \to y_1, y_2)| \\ &= f(h_1^{-1}(y_1, y_2), h_2^{-1}(y_1, y_2)) \\ &\quad \times |J(x_1, x_2 \to y_1, y_2)|, \end{aligned} \tag{2.3}$$

  where

$$J(x_1, x_2 \to y_1, y_2) = \left| \frac{\partial(x_1, x_2)}{\partial(y_1, y_2)} \right| = \det \begin{pmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{pmatrix}$$

  denotes the Jacobian determinant of the transformation from (x_1, x_2) to (y_1, y_2).

Example 2.8 (Quotient of two independent normal variables). Let X_1 and X_2 be two independent standard normal random variables. Define

$$Y_1 = X_1 + X_2 \quad \text{and} \quad Y_2 = \frac{X_1}{X_2}.$$

1) Find the joint density of Y_1 and Y_2.

2) Find the marginal density of Y_2.

Solution. 1) From $y_1 = x_1 + x_2$ and $y_2 = x_1/x_2$, we have

$$x_1 = \frac{y_1 y_2}{1+y_2} \quad \text{and} \quad x_2 = \frac{y_1}{1+y_2}.$$

The Jacobian determinant is

$$\begin{aligned} J(x_1, x_2 \to y_1, y_2) &= \left| \frac{\partial(x_1, x_2)}{\partial(y_1, y_2)} \right| \\ &= \det \begin{pmatrix} \dfrac{y_2}{1+y_2} & \dfrac{y_1}{(1+y_2)^2} \\ \dfrac{1}{1+y_2} & -\dfrac{y_1}{(1+y_2)^2} \end{pmatrix} = -\frac{y_1}{(1+y_2)^2} \end{aligned}$$

so that

$$\begin{aligned} g(y_1, y_2) &= f(x_1, x_2) \times |J(x_1, x_2 \to y_1, y_2)| \\ &= \frac{1}{2\pi} \exp\left[-\frac{1}{2} \left\{ \frac{(y_1 y_2)^2}{(1+y_2)^2} + \frac{y_1^2}{(1+y_2)^2} \right\} \right] \times \frac{|y_1|}{(1+y_2)^2} \\ &= \frac{1}{2\pi} \frac{|y_1|}{(1+y_2)^2} \exp\left[-\frac{1}{2} \left\{ \frac{(1+y_2^2) y_1^2}{(1+y_2)^2} \right\} \right]. \end{aligned}$$

2) The marginal density of Y_2 is given by

$$\begin{aligned} h(y_2) &= \int_{-\infty}^{\infty} g(y_1, y_2)\, \mathrm{d}y_1 \\ &= \frac{1}{2\pi} \frac{1}{(1+y_2)^2} \int_{-\infty}^{\infty} |y_1| \exp\left[-\frac{1}{2} \left\{ \frac{(1+y_2^2) y_1^2}{(1+y_2)^2} \right\} \right] \mathrm{d}y_1 \end{aligned}$$

Let

$$u = \frac{1}{2} \frac{(1+y_2^2) y_1^2}{(1+y_2)^2},$$

then $u \geqslant 0$ and

$$\mathrm{d}u = \frac{(1+y_2^2) y_1}{(1+y_2)^2}\, \mathrm{d}y_1,$$

so

$$h(y_2) = \frac{1}{2\pi(1+y_2)^2} \cdot 2 \int_0^{\infty} \mathrm{e}^{-u} \frac{(1+y_2)^2}{(1+y_2^2)}\, \mathrm{d}u = \frac{1}{\pi(1+y_2^2)},$$

which is a Cauchy density. ‖

Example 2.9 (Uniform distribution on the unit square). Let

$$(X_1, X_2)^\top \sim f(x_1, x_2) = 1, \quad 0 \leqslant x_1 \leqslant 1, \quad 0 \leqslant x_2 \leqslant 1,$$

1) Find the joint pdf of $Y = X_1 + X_2$ and $Z = X_2$.

2) Find the marginal density of Y.

Solution. 1) Make the transformation $y = x_1 + x_2$ and $z = x_2$, where

$$(x_1, x_2) \in \mathcal{S}_{(X_1,X_2)} = \{(x_1, x_2)\colon\ 0 \leqslant x_i \leqslant 1,\ i = 1, 2\},$$

then the corresponding inverse transformation is given by $x_1 = y - z$ and $x_2 = z$, where

$$(y, z) \in \mathcal{S}_{(Y,Z)} = \{(y, z)\colon\ z \leqslant y \leqslant z + 1,\ 0 \leqslant z \leqslant 1\}.$$

Figure 2.6 shows the two regions.

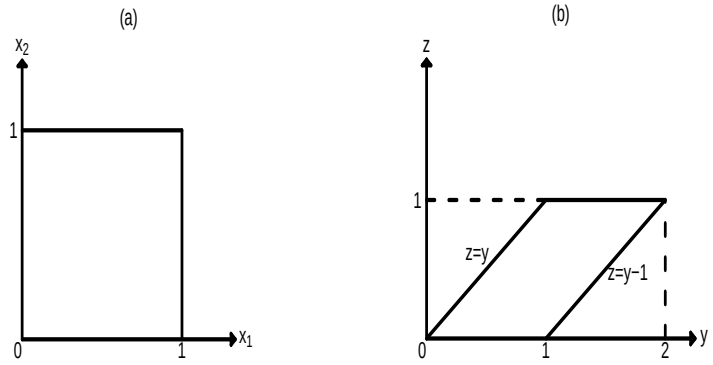

Figure 2.6 (a) $\mathcal{S}_{(X_1,X_2)} = \{(x_1, x_2)\colon\ 0 \leqslant x_i \leqslant 1,\ i = 1, 2\}$; (b) $\mathcal{S}_{(Y,Z)} = \{(y, z)\colon\ z \leqslant y \leqslant z + 1,\ 0 \leqslant z \leqslant 1\}$.

Hence, we have

$$J(x_1, x_2 \to y, z) = \det \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} = 1.$$

By using (2.3), we obtain the joint pdf of (Y, Z) as

$$g(y,z) = f(x_1, x_2) \times |J(x_1, x_2 \to y, z)| = 1 \cdot I_{\mathcal{S}_{(Y,Z)}}(y,z);$$

that is, $(Y,Z)^\top \sim U(\mathcal{S}_{(Y,Z)})$.

2) The marginal density of Y is given by

$$\begin{aligned} g(y) &= \int_{-\infty}^{\infty} g(y,z)\,\mathrm{d}z \\ &= \begin{cases} \int_0^y \mathrm{d}z, & \text{if } 0 \leqslant y \leqslant 1 \\ \int_{y-1}^1 \mathrm{d}z, & \text{if } 1 < y \leqslant 2 \\ 0, & \text{otherwise} \end{cases} \\ &= \begin{cases} y, & \text{if } 0 \leqslant y \leqslant 1 \\ 2-y, & \text{if } 1 < y \leqslant 2 \\ 0, & \text{otherwise} \end{cases} \end{aligned}$$

Figure 2.7 shows this density function. The key point for the transformation technique is to determine the image domain $\mathcal{S}_{(Y,Z)}$.

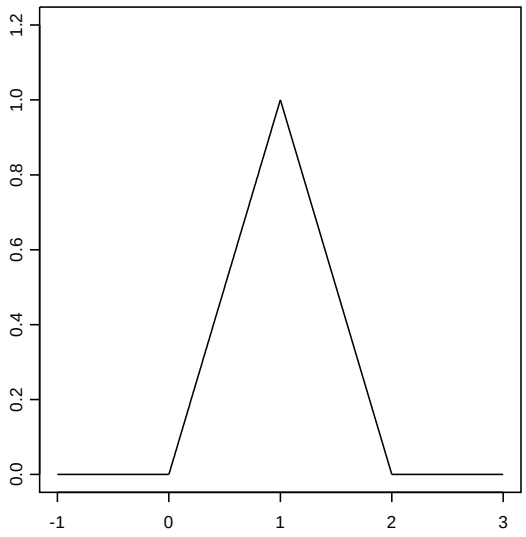

Figure 2.7 The density function of $Y = X_1 + X_2$, where $X_1, X_2 \overset{\text{iid}}{\sim} U[0,1]$. ‖

7• Multivariate transformation

- Let $(X_1, \ldots, X_n)^\top \sim f(x_1, \ldots, x_n)$.
- If the functions $y_i = h_i(x_1, \ldots, x_n)$ for $i = 1, \ldots, n$ are differentiable, then the joint pdf of $Y_i = h_i(X_1, \ldots, X_n)$ for $i = 1, \ldots, n$ is given by

$$g(y_1, \ldots, y_n) = f(x_1, \ldots, x_n) \times |J(x_1, \ldots, x_n \to y_1, \ldots, y_n)|. \quad (2.4)$$

Example 2.10 (Multivariate t-distribution). Let $Z \sim \chi^2(\nu)$, $Z \perp\!\!\!\perp \mathbf{y}$, and $\mathbf{y} = (Y_1, \ldots, Y_d)^\top \sim N_d(\mathbf{0}, \mathbf{\Sigma})$. Define

$$X_i = \mu_i + \frac{Y_i}{\sqrt{Z/\nu}}, \quad i = 1, \ldots, d, \quad (2.5)$$

then $\mathbf{x} = (X_1, \ldots, X_d)^\top$ is said to follow a d-dimensional t-distribution with location parameter vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)^\top \in \mathbb{R}^d$, dispersion matrix $\mathbf{\Sigma} > 0$ and degree of freedom $\nu > 0$, denoted by $\mathbf{x} \sim t_d(\boldsymbol{\mu}, \mathbf{\Sigma}, \nu)$.

1) Find the joint density of $\mathbf{x}$ and Z.

2) Find the joint density of $\mathbf{x}$.

3) Find the marginal density of X_i for $i = 1, \ldots, d$.

4) When $\mathbf{\Sigma} = \mathbf{I}_d$, are X_i and X_j $(i \neq j)$ independent?

Solution. 1) Making the following transformation

$$\begin{cases} x_i &= \mu_i + \dfrac{y_i}{\sqrt{z/\nu}}, \quad i = 1, \ldots, d, \\ z &= z, \end{cases}$$

we have

$$\begin{cases} y_i &= \sqrt{z/\nu}\,(x_i - \mu_i), \quad i = 1, \ldots, d, \\ z &= z, \end{cases}$$

or

$$\begin{cases} \boldsymbol{y} &= (y_1, \ldots, y_d)^\top = \sqrt{z/\nu}\,(\boldsymbol{x} - \boldsymbol{\mu}), \\ z &= z, \end{cases}$$

where $\boldsymbol{x} = (x_1, \ldots, x_d)^\top \in \mathbb{R}^d$ and $z > 0$. The Jacobian determinant is

$$\begin{aligned}
& J(y_1, \ldots, y_d, z \to x_1, \ldots, x_d, z) \\
= \; & \det \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_d} & \frac{\partial y_1}{\partial z} \\ \vdots & \ddots & \vdots & \vdots \\ \frac{\partial y_d}{\partial x_1} & \cdots & \frac{\partial y_d}{\partial x_d} & \frac{\partial y_d}{\partial z} \\ \frac{\partial z}{\partial x_1} & \cdots & \frac{\partial z}{\partial x_d} & \frac{\partial z}{\partial z} \end{pmatrix} \\
= \; & \det \begin{pmatrix} \sqrt{z/\nu} & 0 & \cdots & 0 & 0.5(x_1 - \mu_1)/\sqrt{z/\nu} \\ 0 & \sqrt{z/\nu} & \cdots & 0 & 0.5(x_2 - \mu_2)/\sqrt{z/\nu} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \sqrt{z/\nu} & 0.5(x_d - \mu_d)/\sqrt{z/\nu} \\ 0 & 0 & \cdots & 0 & 1 \end{pmatrix} \\
= \; & (z/\nu)^{d/2}.
\end{aligned}$$

Thus, the joint pdf of $\mathbf{x}$ and Z is

$$\begin{aligned}
& f(x_1, \ldots, x_d, z) \\
= \; & f(y_1, \ldots, y_d, z) \times |J(y_1, \ldots, y_d, z \to x_1, \ldots, x_d, z)| \\
= \; & f(y_1, \ldots, y_d) \times f(z) \times (z/\nu)^{d/2} \\
= \; & N_d(\boldsymbol{y}|\mathbf{0}, \boldsymbol{\Sigma}) \times \chi^2(z|\nu) \times (z/\nu)^{d/2} \\
= \; & \frac{1}{(\sqrt{2\pi})^d |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} \boldsymbol{y}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{y} \right) \times \frac{2^{-\nu/2}}{\Gamma(\nu/2)} z^{\frac{\nu}{2}-1} \mathrm{e}^{-z/2} \times (z/\nu)^{\frac{d}{2}} \\
= \; & c \cdot \exp\left\{ -\frac{z}{2\nu} (\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \right\} \times z^{\frac{\nu+d}{2}-1} \mathrm{e}^{-z/2} \\
= \; & c \cdot z^{\frac{\nu+d}{2}-1} \exp\left[-z \left\{ \frac{1}{2} + \frac{(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})}{2\nu} \right\} \right],
\end{aligned}$$

where $\boldsymbol{x} \in \mathbb{R}^d$, $z > 0$ and

$$c = \frac{2^{-\frac{\nu}{2}}}{(2\pi\nu)^{\frac{d}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}} \Gamma(\frac{\nu}{2})} = \frac{1}{2^{\frac{\nu+d}{2}} \Gamma(\frac{\nu}{2}) (\sqrt{\pi\nu})^d |\boldsymbol{\Sigma}|^{\frac{1}{2}}}.$$

2) By using (1.41) in Chapter 1, we obtain the joint pdf of $\mathbf{x}$ given by

$$\begin{aligned}
& f(x_1,\ldots,x_d) \\
= & \int_0^\infty f(x_1,\ldots,x_d,z)\,\mathrm{d}z \\
= & c\cdot\int_0^\infty z^{\frac{\nu+d}{2}-1}\exp\left[-z\left\{\frac{1}{2}+\frac{(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{2\nu}\right\}\right]\mathrm{d}z \\
\overset{(1.41)}{=} & c\cdot\frac{\Gamma(\frac{\nu+d}{2})}{\left\{\frac{1}{2}+\frac{(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{2\nu}\right\}^{\frac{\nu+d}{2}}} \\
= & \frac{\Gamma(\frac{\nu+d}{2})}{\Gamma(\frac{\nu}{2})(\sqrt{\pi\nu})^d|\boldsymbol{\Sigma}|^{\frac{1}{2}}}\left\{1+\frac{(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{\nu}\right\}^{-\frac{\nu+d}{2}}, \quad \boldsymbol{x}\in\mathbb{R}^d,
\end{aligned}$$

which is the density of d-dimensional t-distribution.

3) In particular, let $d=1$ and denote $\boldsymbol{\Sigma}$ by σ^2. The density of X_1 is

$$\frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}\,\sigma}\left\{1+\frac{(x_1-\mu)^2}{\nu\sigma^2}\right\}^{-\frac{\nu+1}{2}}, \quad x_1\in\mathbb{R},$$

which is the density of the univariate t-distribution with location parameter $\mu\in\mathbb{R}$, dispersion parameter $\sigma^2>0$ and degree of freedom $\nu>0$. We denote it by $X_1\sim t(\mu,\sigma^2,\nu)$.

4) When $d=2$ and $\boldsymbol{\Sigma}=\mathbf{I}_2$, it is easy to show that

$$f_{(X_1,X_2)}(x_1,x_2)\neq f_{X_1}(x_1)\times f_{X_2}(x_2),$$

So X_1 and X_2 are not independent. From (2.5), it is clear that X_i and X_j $(i\neq j)$ share a common r.v. Z, so they are not independent. ‖

2.1.3 Moment generating function technique

8• The procedure of mgf

- Let $Y=\sum_{i=1}^n X_i$.
- If $\{X_i\}_{i=1}^n$ are independent r.v.'s, then

$$M_Y(t)=\prod_{i=1}^n M_{X_i}(t). \tag{2.6}$$

Example 2.11 (Sum of independent binomial r.v.'s with a common p). Let $\{X_i\}_{i=1}^n$ be independent r.v.'s and $X_i \sim \text{Binomial}(m_i, p)$ for $i = 1, \ldots, n$, find the distribution of $Y = \sum_{i=1}^n X_i$. 结果可加

Solution. From (2.6) and Table 1.2, we have

$$M_Y(t) = \prod_{i=1}^n M_{X_i}(t) = \prod_{i=1}^n (p\,\mathrm{e}^t + 1 - p)^{m_i} = (p\,\mathrm{e}^t + 1 - p)^{\sum_{i=1}^n m_i},$$

indicating that $\sum_{i=1}^n X_i \sim \text{Binomial}(\sum_{i=1}^n m_i, p)$. This result means that binomial distribution is additive. ‖

Example 2.12 (Sum of independent Poisson r.v.'s). Let $\{X_i\}_{i=1}^n$ be independent r.v.'s and $X_i \sim \text{Poisson}(\lambda_i)$ for $i = 1, \ldots, n$, find the distribution of $Y = \sum_{i=1}^n X_i$.

Solution. From (2.6) and Table 1.2, we have 泊松流：单位时间里通过的粒子数

$$M_Y(t) = \prod_{i=1}^n M_{X_i}(t) = \prod_{i=1}^n \exp\{\lambda_i(\mathrm{e}^t - 1)\} = \exp\left\{\sum_{i=1}^n \lambda_i(\mathrm{e}^t - 1)\right\},$$

which means $\sum_{i=1}^n X_i \sim \text{Poisson}(\sum_{i=1}^n \lambda_i)$; i.e., Poisson distribution is also additive. This result is a generalization of the result in Example 2.5. ‖

Example 2.13 (Sum of independent chi-squared r.v.'s). Let $\{X_i\}_{i=1}^n$ be independent r.v.'s and $X_i \sim \chi^2(m_i)$ for $i = 1, \ldots, n$, find the distribution of $Y = \sum_{i=1}^n X_i$.

Solution. Note that $\chi^2(m) = \text{Gamma}(\frac{m}{2}, \frac{1}{2})$. From (2.6) and Table 1.3, we have

$$\begin{aligned} M_Y(t) &= \prod_{i=1}^n M_{X_i}(t) \\ &= \prod_{i=1}^n \left(\frac{0.5}{0.5 - t}\right)^{m_i/2} \\ &= \left(\frac{0.5}{0.5 - t}\right)^{\sum_{i=1}^n m_i/2}, \end{aligned}$$

which means $\sum_{i=1}^n X_i \sim \chi^2(\sum_{i=1}^n m_i)$. ‖

是一个随机变量，而E(x)则是一个值、参数

2.2 Statistics, Sample Mean and Sample Variance

9• What is a random sample?

- Let $F(x)$ be the cdf of an r.v. X.
- If $\{X_i\}_{i=1}^n \overset{\text{iid}}{\sim} F(x)$, then $\{X_i\}_{i=1}^n$ is said to be a *random sample* of X, or $\{X_i\}_{i=1}^n$ is a random sample from $F(x)$.

独立性和代表性

10• What is a statistic?

Definition 2.1 (Function of random variables). Let $\{X_i\}_{i=1}^n \overset{\text{iid}}{\sim} F(x)$. An arbitrary function $T(X_1, \ldots, X_n)$ of $\{X_i\}_{i=1}^n$ is called a *statistic*. ‖

二次型？？？

统计量也是 v.r.

统计量，the function of random sample

10.1• The sample mean and sample variance

— For example,

$$\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i \quad \text{and} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2 \tag{2.7}$$

are two statistics.

— They are called the sample mean and sample variance, respectively.

2.2.1 Distribution of the sample mean

11• Basic properties of the sample mean

- Let $\{X_i\}_{i=1}^n \overset{\text{iid}}{\sim} F(x)$ with $E(X_1) = \mu$ and $\text{Var}(X_1) = \sigma^2$.
- For any $F(x)$, we have $E(\bar{X}) = \mu$ and $\text{Var}(\bar{X}) = \sigma^2/n$.
- If $F(x)$ is the cdf of the normal distribution $N(\mu, \sigma^2)$, then

$$\bar{X} \sim N(\mu, \sigma^2/n). \tag{2.8}$$

Proof. In fact, by the mgf technique, we have

$$\begin{aligned} M_{\bar{X}}(t) &= M_{\sum_{i=1}^n X_i/n}(t) = \prod_{i=1}^n M_{X_i/n}(t) = \prod_{i=1}^n M_{X_i}\left(\frac{t}{n}\right) \\ &= \left\{M_{X_1}\left(\frac{t}{n}\right)\right\}^n = \left\{\exp\left(\mu\frac{t}{n} + 0.5\sigma^2\frac{t^2}{n^2}\right)\right\}^n \end{aligned}$$

$$= \exp\left\{\mu t + 0.5\left(\frac{\sigma^2}{n}\right)t^2\right\},$$

indicating that $\bar{X} \sim N(\mu, \sigma^2/n)$. □

2.2.2 Distribution of the sample variance

To prove (2.10) below, we need the following theorem with proof given in Section 2.6.

Theorem 2.1 (Linear combination of normal components). Let $\mathbf{A}_{m\times n}$ and $\mathbf{B}_{r\times n}$ be two scalar matrices and $\mathbf{x} = (X_1, \ldots, X_n)^\top \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then

1) $\mathbf{Ax} \sim N_m(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top)$.

2) $\mathbf{Bx} \sim N_r(\mathbf{B}\boldsymbol{\mu}, \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^\top)$.

3) $\mathbf{Ax} \perp\!\!\!\perp \mathbf{Bx}$ iff $\mathbf{A}\boldsymbol{\Sigma}\mathbf{B}^\top = \mathbf{O}_{m\times r}$. ‖

12• Basic properties of the sample variance

- Let $\{X_i\}_{i=1}^n \overset{\text{iid}}{\sim} F(x)$ with $E(X_1) = \mu$ and $\text{Var}(X_1) = \sigma^2$.
- For any $F(x)$, the sample variance is an unbiased estimator of the variance, i.e.,

$$E(S^2) = \sigma^2. \tag{2.9}$$

Proof. Since

$$(n-1)S^2 = \sum_{i=1}^n [X_i - \mu - (\bar{X} - \mu)]^2 = \sum_{i=1}^n (X_i - \mu)^2 - n(\bar{X} - \mu)^2,$$

taking expectation on both sides, we have

$$(n-1)E(S^2) = n\sigma^2 - n \cdot \frac{\sigma^2}{n},$$

which means (2.9). □

- If $\{X_i\}_{i=1}^n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, then

$$S^2 \perp\!\!\!\perp \bar{X} \quad \text{and} \quad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1). \tag{2.10}$$

前者代表数据波动，后者表示数据中心，所以可以无关

Proof. Define $\mathbf{Q}_n = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$, it is easy to show that

冥等矩阵

$$\mathbf{Q}_n = \mathbf{Q}_n^\top = \mathbf{Q}_n^2 \quad \text{and} \quad \mathbf{Q}_n\mathbf{1}_n = \mathbf{0}_n. \tag{2.11}$$

Let $\mathbf{x} = (X_1, \ldots, X_n)^\top$, then $\mathbf{x} \sim N_n(\mu\mathbf{1}_n, \sigma^2\mathbf{I}_n)$. From the result 1) of Theorem 2.1 and (2.11), we have

$$\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i = \frac{1}{n}\mathbf{1}_n^\top\mathbf{x} \sim N(\mu, \sigma^2/n)$$

and

$$\begin{pmatrix} X_1 - \bar{X} \\ \vdots \\ X_n - \bar{X} \end{pmatrix} = \mathbf{x} - \bar{X}\mathbf{1}_n = \mathbf{x} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top\mathbf{x} = \mathbf{Q}_n\mathbf{x} \sim N_n(\mathbf{0}, \sigma^2\mathbf{Q}_n).$$

Note that $\mathbf{Q}_n \cdot \sigma^2\mathbf{I}_n \cdot \mathbf{1}_n = \mathbf{0}$, by the result 3) of Theorem 2.1, we can conclude that $\mathbf{Q}_n\mathbf{x} \perp\!\!\!\perp \mathbf{1}_n^\top\mathbf{x}$. Since

$$S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n-1}(\mathbf{Q}_n\mathbf{x})^\top\mathbf{Q}_n\mathbf{x}$$

is a function of $\mathbf{Q}_n\mathbf{x}$ and $\bar{X}$ is a function of $\mathbf{1}_n^\top\mathbf{x}$, we have $S^2 \perp\!\!\!\perp \bar{X}$.

Since

$$\begin{aligned} \sum_{i=1}^n (X_i - \mu)^2 &= \sum_{i=1}^n (X_i - \bar{X} + \bar{X} - \mu)^2 \\ &= \sum_{i=1}^n (X_i - \bar{X})^2 + n(\bar{X} - \mu)^2, \end{aligned}$$

we have

$$W \hat{=} \sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma}\right)^2 = \frac{(n-1)S^2}{\sigma^2} + \left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right)^2 \hat{=} U + V,$$

where $W \sim \chi^2(n)$, $V \sim \chi^2(1)$, and $U \perp\!\!\!\perp V$. Then

$$M_W(t) = M_U(t) \cdot M_V(t),$$

or

$$(1-2t)^{-n/2} = M_U(t) \cdot (1-2t)^{-1/2}.$$

Hence

$$M_U(t) = (1-2t)^{-(n-1)/2}.$$

This implies that $U = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$. □

2.3 The t and F Distributions

2.3.1 The t distribution

13• Definition

- Let $Y \sim \chi^2(n)$, $Z \sim N(0,1)$ and $Y \perp\!\!\!\perp Z$.
- The distribution of
$$T = \frac{Z}{\sqrt{Y/n}} \tag{2.12}$$
is called the t distribution with n degrees of freedom and is written as $T \sim t(n)$.

13.1• The name of the t distribution

— The t distribution was introduced originally by W. S. Gosset, who published his scientific writings under the pen name "Student" since the company for which he worked, a brewery, did not permit publication by employees.

— Thus, the t distribution is also known as the *Student t distribution*, or *Student's t distribution*.

Theorem 2.2 (Density of the t distribution). The density of $T \sim t(n)$ is given by
$$f(t) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{\pi n}\,\Gamma(\frac{n}{2})}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}, \quad -\infty < t < \infty. \qquad \|$$

Proof. Let $\phi(z) = \frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{-z^2/2}$ denote the pdf of $Z \sim N(0,1)$ and $g(y)$ denote the pdf of $Y \sim \chi^2(n)$. The cdf of T is
$$\begin{aligned} F(x) &= \Pr(T \leqslant x) = \Pr\left(\frac{Z}{\sqrt{Y/n}} \leqslant x\right) \\ &\overset{(1.33)}{=} \int \Pr\left(\frac{Z}{\sqrt{Y/n}} \leqslant x \,\middle|\, Y = y\right) \cdot g(y)\,\mathrm{d}y \\ &= \int_0^\infty \Pr\left(Z \leqslant x\sqrt{y/n}\right) \cdot g(y)\,\mathrm{d}y \\ &= \int_0^\infty \left\{\int_{-\infty}^{x\sqrt{y/n}} \phi(z)\,\mathrm{d}z\right\} \cdot g(y)\,\mathrm{d}y. \end{aligned}$$

Let $t = \frac{z}{\sqrt{y/n}}$, then $-\infty < t \leqslant x$, $\mathrm{d}z = \sqrt{y/n}\,\mathrm{d}t$, and $F(x)$ becomes

$$\begin{aligned} F(x) &= \int_0^\infty \left\{ \int_{-\infty}^x \phi\Big(t\sqrt{y/n}\Big) \cdot \sqrt{y/n}\,\mathrm{d}t \right\} \cdot g(y)\,\mathrm{d}y \\ &= \int_{-\infty}^x \left\{ \int_0^\infty \phi\Big(t\sqrt{y/n}\Big) \cdot \sqrt{y/n} \cdot g(y)\,\mathrm{d}y \right\} \mathrm{d}t \\ &= \int_{-\infty}^x f(t)\,\mathrm{d}t. \end{aligned}$$

Hence, the density of T is given by

$$\begin{aligned} f(t) &= \int_0^\infty \phi\Big(t\sqrt{y/n}\Big) \cdot \sqrt{y/n} \cdot g(y)\,\mathrm{d}y \\ &= \int_0^\infty \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-t^2 y/(2n)} \cdot \sqrt{y/n} \cdot \frac{(1/2)^{n/2}}{\Gamma(n/2)} y^{\frac{n}{2}-1} \mathrm{e}^{-y/2}\,\mathrm{d}y \\ &= \frac{1}{\sqrt{2\pi n}} \cdot \frac{(1/2)^{n/2}}{\Gamma(n/2)} \cdot \int_0^\infty y^{\frac{n+1}{2}-1} \mathrm{e}^{-y(\frac{1}{2}+\frac{t^2}{2n})}\,\mathrm{d}y \\ &\overset{(1.39)}{=} \frac{(1/2)^{(n+1)/2}}{\sqrt{\pi n}\,\Gamma(n/2)} \cdot \frac{\Gamma(\frac{n+1}{2})}{(\frac{1}{2}+\frac{t^2}{2n})^{\frac{n+1}{2}}} \\ &= \frac{\Gamma(\frac{n+1}{2})}{\sqrt{\pi n}\,\Gamma(\frac{n}{2})} \left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}. \end{aligned}$$

This completes the proof of Theorem 2.2. □

13.2• The usefulness of the t distribution

— The t distribution is an important distribution in statistical inferences on the mean of the normal population.

— Figure 2.8 compares the $t(4)$ density with the standard normal density.

— Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$. From (2.8), we obtain

$$\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \sim N(0,1). \tag{2.13}$$

— By using (2.10), we have

$$T = \frac{\sqrt{n}(\bar{X}-\mu)/\sigma}{\sqrt{\frac{(n-1)S^2}{\sigma^2}\Big/(n-1)}} = \frac{\sqrt{n}(\bar{X}-\mu)}{S} \sim t(n-1). \tag{2.14}$$

sigma 是 S 的中心

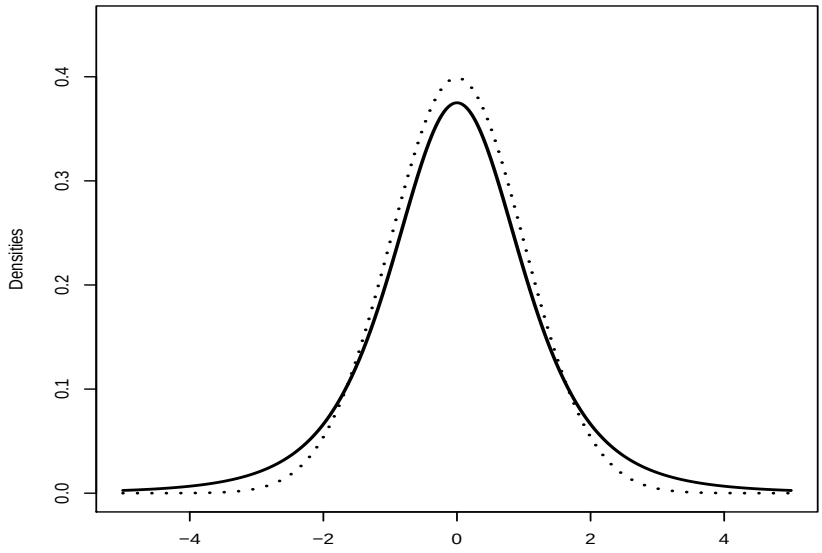

Figure 2.8 The comparison between the $t(4)$ density (solid curve) and the standard normal density (dotted curve).

2.3.2 The F distribution

偏斜的，非对称的

14• Definition

- Let $U \sim \chi^2(m)$, $V \sim \chi^2(n)$ and $U \perp\!\!\!\perp V$.
- The distribution of the r.v.

$$W = \frac{U/m}{V/n} \tag{2.15}$$

  is said to have an F distribution with m and n degrees of freedom. We write $W \sim F(m, n)$.

14.1• The name of the F distribution

— Besides the t distribution, another distribution that plays an important role in connection with sampling from normal populations is the F distribution, named after Sir Ronald A. Fisher, one of the most prominent statisticians of the last century.

— The F distribution is also known as Snedecor's F distribution (after George W. Snedecor) or the Fisher–Snedecor distribution.

Theorem 2.3 (Density of the F distribution). The density of $W \sim F(m,n)$ is given by

$$f(w) = \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \left(\frac{m}{n}\right)^{m/2} w^{m/2-1} \left(1+\frac{m}{n}w\right)^{-\frac{m+n}{2}}, \quad w>0. \qquad \|$$

Proof. Let $h(u)$ and $g(v)$ denote the densities of $U \sim \chi^2(m)$ and $V \sim \chi^2(n)$, respectively. Since $U \perp\!\!\!\perp V$, the cdf of W is

$$\begin{aligned} F(x) &= \Pr(W \leqslant x) = \Pr\left(\frac{U/m}{V/n} \leqslant x\right) \\ &= \int \Pr\left(\frac{U/m}{V/n} \leqslant x \middle| V=v\right) \cdot g(v)\,\mathrm{d}v \quad \text{条件概率} \\ &= \int_0^\infty \Pr\left(U \leqslant xvm/n\right) \cdot g(v)\,\mathrm{d}v \\ &= \int_0^\infty \left\{\int_0^{xvm/n} h(u)\,\mathrm{d}u\right\} \cdot g(v)\,\mathrm{d}v. \end{aligned}$$

Let $w = \frac{u/m}{v/n}$, then $0 \leqslant w \leqslant x$, $\mathrm{d}u = \frac{mv}{n}\,\mathrm{d}w$, and $F(x)$ becomes

$$\begin{aligned} F(x) &= \int_0^\infty \left\{\int_0^x h\Big(\frac{mv}{n}w\Big) \cdot \frac{mv}{n}\,\mathrm{d}w\right\} \cdot g(v)\,\mathrm{d}v \\ &= \int_0^x \left\{\int_0^\infty h\Big(\frac{mv}{n}w\Big) \cdot \frac{mv}{n} \cdot g(v)\,\mathrm{d}v\right\} \mathrm{d}w = \int_0^x f(w)\,\mathrm{d}w. \end{aligned}$$

Hence, the density of W is given by

$$\begin{aligned} f(w) &= \int_0^\infty h\Big(\frac{mv}{n}w\Big) \cdot \frac{mv}{n} \cdot g(v)\,\mathrm{d}v \\ &= \int_0^\infty \frac{(\frac{1}{2})^{m/2}}{\Gamma(\frac{m}{2})} \Big(\frac{mv}{n}w\Big)^{\frac{m}{2}-1} \mathrm{e}^{-\frac{mvw}{2n}} \cdot \frac{mv}{n} \cdot \frac{(\frac{1}{2})^{n/2}}{\Gamma(\frac{n}{2})} v^{\frac{n}{2}-1} \mathrm{e}^{-v/2}\,\mathrm{d}v \\ &= \frac{(\frac{1}{2})^{(m+n)/2}}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \Big(\frac{m}{n}\Big)^{m/2} w^{m/2-1} \cdot \int_0^\infty v^{\frac{m+n}{2}-1} \mathrm{e}^{-v(\frac{1}{2}+\frac{mw}{2n})}\,\mathrm{d}v \\ &\overset{(1.39)}{=} \frac{(\frac{1}{2})^{(m+n)/2}}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \Big(\frac{m}{n}\Big)^{m/2} w^{m/2-1} \cdot \frac{\Gamma(\frac{m+n}{2})}{(\frac{1}{2}+\frac{mw}{2n})^{\frac{m+n}{2}}} \\ &= \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \Big(\frac{m}{n}\Big)^{m/2} w^{m/2-1} \Big(1+\frac{m}{n}w\Big)^{-\frac{m+n}{2}}. \end{aligned}$$

This completes the proof of Theorem 2.3. □

Theorem 2.4 (Ratio of two normal sample variances). If S_1^2 and S_2^2 are the sample variances of independent random samples of size n_1 and n_2 from normal populations $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, respectively, then

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2} \sim F(n_1 - 1, n_2 - 1). \qquad \|$$

Proof. Note that

$$\frac{(n_1-1)S_1^2}{\sigma_1^2} \sim \chi^2(n_1 - 1) \quad \text{and} \quad \frac{(n_2-1)S_2^2}{\sigma_2^2} \sim \chi^2(n_2 - 1)$$

are independent, then

$$F = \frac{\frac{(n_1-1)S_1^2}{\sigma_1^2}/(n_1-1)}{\frac{(n_2-1)S_2^2}{\sigma_2^2}/(n_2-1)} = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1). \qquad \square$$

14.2• The usefulness of the F distribution

— If $X \sim F(m, n)$, then $Y = 1/X \sim F(n, m)$.

— The densities of $F(m, n)$ with various degrees of freedom are shown in Figure 2.9.

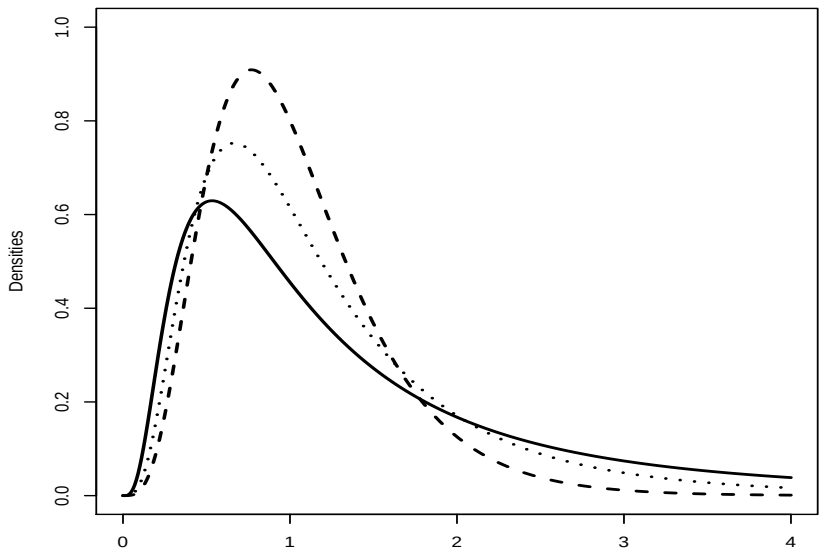

Figure 2.9 Plots of the densities of $W \sim F(m, n)$ with $m = 10$ and $n = 4$ (solid curve), $n = 10$ (dotted curve), $n = 50$ (broken curve).

2.4 Order Statistics

都是些 random sample

15• Definition

- Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F(\cdot)$, and $f(\cdot)$ is the pdf.
- Let
  - $X_{(1)} = \min(X_1, \ldots, X_n)$ be the smallest of $X_1, \ldots, X_n$;
  - $X_{(2)}$ be the second smallest of $X_1, \ldots, X_n$;
  - $\vdots$
  - $X_{(n)} = \max(X_1, \ldots, X_n)$ be the largest of $X_1, \ldots, X_n$.
- Then $X_{(1)}, \ldots, X_{(n)}$ are called the *order statistics* and $X_{(r)}$ is called the r-th *order statistic* for $r = 1, \ldots, n$.
- We use $x_{(1)}, \ldots, x_{(n)}$ to denote the realizations of $X_{(1)}, \ldots, X_{(n)}$.

15.1• An example

— Let $\{x_1, \ldots, x_5\} = \{2, 5, -1, 0, 6\}$, then we have $x_{(1)} = -1$, $x_{(2)} = 0$, $x_{(3)} = 2$, $x_{(4)} = 5$, and $x_{(5)} = 6$.

15.2• Remarks

— The $X_{(r)}$'s are statistics since they are functions of the random sample $X_1, \ldots, X_n$ and are in order.

— Unlike the random sample themselves, the order statistics are clearly *not independent*, because if $X_{(r)} \geqslant x$, then $X_{(r+1)} \geqslant x$.

2.4.1 Distribution of a single order statistic

16• The distribution of the largest order statistic

- Let $G_r(x)$ denote the cdf of the r-th order statistic $X_{(r)}$.
- Then the cdf of the largest order statistic $X_{(n)}$ is

$$\begin{aligned} G_n(x) &= \Pr\{\max(X_1, \ldots, X_n) \leqslant x\} \\ &= \Pr(X_1 \leqslant x, \ldots, X_n \leqslant x) \overset{?}{=} F^n(x). \end{aligned} \tag{2.16}$$

- The pdf of $X_{(n)}$ is

$$g_n(x) = \frac{\mathrm{d}G_n(x)}{\mathrm{d}x} = nf(x)F^{n-1}(x). \tag{2.17}$$

17• The distribution of the smallest order statistic

- Similarly, we have

$$\begin{aligned} G_1(x) &= \Pr(X_{(1)} \leqslant x) \\ &= 1 - \Pr\{\min(X_1, \ldots, X_n) > x\} \\ &= 1 - \Pr(X_1 > x, \ldots, X_n > x) \\ &= 1 - \{1 - F(x)\}^n. \end{aligned} \tag{2.18}$$

- The pdf of $X_{(1)}$ is

$$g_1(x) = \frac{\mathrm{d}G_1(x)}{\mathrm{d}x} = nf(x)\{1 - F(x)\}^{n-1}. \tag{2.19}$$

18• The distribution of the r-th order statistic

18.1• The cdf of $X_{(r)}$

— Let $G_r(x)$ denote the cdf of $X_{(r)}$, then

$$G_r(x) = \frac{1}{B(r, n-r+1)} \int_0^{F(x)} t^{r-1}(1-t)^{n-r}\,\mathrm{d}t. \tag{2.20}$$

Proof. The formulae (2.16) and (2.18) are important special cases of the general result:

$$\begin{aligned} G_r(x) &= \Pr(X_{(r)} \leqslant x) \\ &= \Pr(\text{at least } r \text{ of } X_1, \ldots, X_n \leqslant x) \\ &= \sum_{i=r}^{n} \Pr(\text{exact } i \text{ of } X_1, \ldots, X_n \leqslant x) \\ &= \sum_{i=r}^{n} \binom{n}{i} \Pr(X_1, \ldots, X_i \leqslant x) \cdot \Pr(X_{i+1}, \ldots, X_n > x) \\ &= \sum_{i=r}^{n} \binom{n}{i} F^i(x)\{1 - F(x)\}^{n-i}. \end{aligned} \tag{2.21}$$

biominol 离散

连续

By using the identity

$$\sum_{i=r}^{n} \binom{n}{i} p^i (1-p)^{n-i} = \frac{1}{B(r, n-r+1)} \int_0^p t^{r-1}(1-t)^{n-r}\, \mathrm{d}t \qquad (2.22)$$

for any $p \in [0,1]$, we can rewrite (2.21) into (2.20) and hence complete the proof. □

可以用微分的思路

18.2• Proof of (2.22)

— Let $f(p)$ denote the left-hand side of (2.22), we have

$$\begin{aligned} f'(p) &= \sum_{i=r}^{n} \binom{n}{i} \left\{ i p^{i-1}(1-p)^{n-i} - (n-i)p^i(1-p)^{n-i-1} \right\} \\ &= \sum_{i=r}^{n} \frac{n!}{i!(n-i)!} \left\{ i p^{i-1}(1-p)^{n-i} - (n-i)p^i(1-p)^{n-i-1} \right\} \\ &= \sum_{i=r}^{n} \frac{n! p^{i-1}(1-p)^{n-i}}{(i-1)!(n-i)!} - \sum_{i=r}^{n} \frac{n! p^i (1-p)^{n-i-1}}{i!(n-i-1)!} \\ &= \frac{n!}{(n-r)!(r-1)!} p^{r-1}(1-p)^{n-r} \end{aligned}$$

— Let $g(p)$ denote the right-hand side of (2.22), we obtain

$$\begin{aligned} g'(p) &= \frac{1}{B(r, n-r+1)} p^{r-1}(1-p)^{n-r} \\ &= \frac{n!}{(r-1)!(n-r)!} p^{r-1}(1-p)^{n-r}, \end{aligned}$$

so that $f'(p) = g'(p)$.

— This implies $f(p) = g(p) + c$ for any $p \in [0,1]$, where c is a constant.

— In particular, let $p = 0$, we have

$$c = f(0) - g(0) = 0.$$

Thus $f(p) = g(p)$. □

18.3• The pdf of $X_{(r)}$

— Let $g_r(x)$ denote the pdf of $X_{(r)}$, from (2.20), we obtain

$$\begin{aligned} g_r(x) &= \frac{\mathrm{d}}{\mathrm{d}x}G_r(x) \\ &= \frac{1}{B(r, n-r+1)} \cdot \frac{\mathrm{d}}{\mathrm{d}x}\int_0^{F(x)} t^{r-1}(1-t)^{n-r}\,\mathrm{d}t \\ &= \frac{n!}{(r-1)!(n-r)!} f(x)F^{r-1}(x)\{1-F(x)\}^{n-r}. \end{aligned} \tag{2.23}$$

— In (2.23), we utilized the following formula:

$$\frac{\mathrm{d}}{\mathrm{d}x}\int_0^{A(x)} g(t)\,\mathrm{d}t = \frac{\mathrm{d}}{\mathrm{d}x}\{G(A(x)) - G(0)\} = A'(x)\cdot g(A(x)),$$

where $G'(t) = g(t)$.

中位数

Example 2.14 (Distribution of sample median). In a random sample of size $n = 2m+1$, the *sample median* is $X_{(m+1)}$, whose sampling distribution is

$$\frac{(2m+1)!}{m!m!} f(x)F^m(x)\{1-F(x)\}^m, \quad -\infty < x < \infty.$$

For a random sample of size $n = 2m$, the median is defined as $\frac{1}{2}(X_{(m)} + X_{(m+1)})$. ‖

2.4.2 Joint distribution of more order statistics

19• The general case

- The joint density of $X_{(r_1)}, \ldots, X_{(r_k)}$ $(1 \leqslant r_1 \leqslant \cdots \leqslant r_k \leqslant n;\ 1 \leqslant k \leqslant n)$ is, for $x_1 \leqslant \cdots \leqslant x_k$ (or $x_{(r_1)} \leqslant \cdots \leqslant x_{(r_k)}$),

$$\begin{aligned} & g_{r_1\cdots r_k}(x_1, \ldots, x_k) \\ = \ & n!\left\{\prod_{i=1}^{k} f(x_i)\right\} \cdot \prod_{i=0}^{k} \frac{\{F(x_{i+1}) - F(x_i)\}^{r_{i+1}-r_i-1}}{(r_{i+1}-r_i-1)!}, \end{aligned} \tag{2.24}$$

where $x_0 = -\infty$, $x_{k+1} = +\infty$, $r_0 = 0$ and $r_{k+1} = n+1$.

19.1• Three special cases

— The joint pdf of $X_{(r)}$ and $X_{(s)}$ $(1 \leqslant r < s \leqslant n)$ is, for $x \leqslant y$,

$$\begin{aligned} g_{rs}(x,y) &= \frac{n!}{(r-1)!(s-r-1)!(n-s)!} f(x)f(y) \\ &\quad \times F^{r-1}(x)\{F(y)-F(x)\}^{s-r-1}\{1-F(y)\}^{n-s}. \end{aligned} \tag{2.25}$$

— The joint pdf of $X_{(1)}, \ldots, X_{(r)}$ $(1 \leqslant r \leqslant n)$ is, for $x_1 \leqslant \cdots \leqslant x_r$,

$$g_{1\cdots r}(x_1,\ldots,x_r) = \frac{n!}{(n-r)!} f(x_1)\cdots f(x_r)\{1-F(x_r)\}^{n-r}. \tag{2.26}$$

— The joint pdf of $X_{(1)}, \ldots, X_{(n)}$ is, for $x_1 \leqslant \cdots \leqslant x_n$,

$$g_{1\cdots n}(x_1,\ldots,x_n) = n! f(x_1)\cdots f(x_n). \tag{2.27}$$

Example 2.15 (Distribution of $X_{(s)} - X_{(r)}$ for uniform population). Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} U[0,1]$.

1) Find the distribution of $X_{(r)}$.

2) Find the distribution of $X_{(s)} - X_{(r)}$, where $1 \leqslant r < s \leqslant n$.

Solution. 1) Obviously, the corresponding cdf is

$$F(x) = \begin{cases} 0, & \text{if } x < 0, \\ x, & \text{if } 0 \leqslant x \leqslant 1, \\ 1, & \text{if } x > 1. \end{cases}$$

From (2.23), we have at once

$$g_r(x) = \frac{1}{B(r, n-r+1)} x^{r-1}(1-x)^{n-r}, \quad 0 \leqslant x \leqslant 1.$$

Thus $X_{(r)} \sim \text{Beta}(r, n-r+1)$.

2) From (2.25), the joint density of $X_{(r)}$ and $X_{(s)}$ is

$$g_{rs}(x_{(r)}, x_{(s)}) = c \cdot x_{(r)}^{r-1}\{x_{(s)} - x_{(r)}\}^{s-r-1}\{1-x_{(s)}\}^{n-s},$$

where $0 \leqslant x_{(r)} \leqslant x_{(s)} \leqslant 1$ and

$$c \;\hat{=}\; \frac{n!}{(r-1)!(s-r-1)!(n-s)!}$$

Making the transformation $z = x_{(s)} - x_{(r)}$ and $x = x_{(r)}$, we have

$$\begin{aligned} J(z, x \to x_{(r)}, x_{(s)}) &= \left| \frac{\partial(z, x)}{\partial(x_{(r)}, x_{(s)})} \right| \\ &= \det \begin{pmatrix} -1 & 1 \\ 1 & 0 \end{pmatrix} = -1. \end{aligned}$$

Hence, the joint density of $Z = X_{(s)} - X_{(r)}$ and $X = X_{(r)}$ is

$$\begin{aligned} h(z, x) &= g_{rs}(x_{(r)}, x_{(s)}) / |J(z, x \to x_{(r)}, x_{(s)})| \\ &= c \cdot x^{r-1} z^{s-r-1} (1 - x - z)^{n-s}, \end{aligned}$$

where $0 \leqslant x \leqslant 1$, $0 \leqslant z \leqslant 1$, and $0 \leqslant x + z \leqslant 1$. The marginal density of $Z = X_{(s)} - X_{(r)}$ is given by

$$\begin{aligned} h(z) &= \int_0^{1-z} h(z, x)\, \mathrm{d}x \\ &= c \cdot z^{s-r-1} \int_0^{1-z} x^{r-1} (1 - z - x)^{n-s}\, \mathrm{d}x \\ &= c \cdot z^{s-r-1} (1 - z)^{n-s} \int_0^{1-z} x^{r-1} \left(1 - \frac{x}{1-z}\right)^{n-s} \mathrm{d}x. \end{aligned}$$

Let $w = x/(1 - z)$, note that

$$\begin{aligned} \int_0^{1-z} x^{r-1} \left(1 - \frac{x}{1-z}\right)^{n-s} \mathrm{d}x &= \int_0^1 (1 - z)^r w^{r-1} (1 - w)^{n-s}\, \mathrm{d}x \\ &= (1 - z)^r \cdot B(r, n - s + 1), \end{aligned}$$

we obtain $h(z) \propto z^{s-r-1}(1 - z)^{n-s+r}$, i.e.,

$$X_{(s)} - X_{(r)} \sim \text{Beta}(s - r, n - s + r + 1). \qquad \|$$

2.5 Limit Theorems

2.5.1 Convergency of a sequence of distribution functions

20• A MOTIVATION EXAMPLE 亿分布收敛

- Consider a sequence of i.i.d. r.v.'s $\{Y_i\}_{i=1}^{\infty}$ each having a uniform distribution on the unit interval $(0, 1)$.

- The mgf of $Y_1 \sim U(0,1)$ is

$$M_{Y_1}(t) = \begin{cases} 1, & \text{if } t = 0, \\ (\mathrm{e}^t - 1)/t, & \text{if } t \neq 0. \end{cases} \tag{2.28}$$

- Let $X_n \hat{=} \bar{Y} = \sum_{i=1}^n Y_i/n$. Since $X_1 = Y_1$ and $X_2 = (Y_1 + Y_2)/2 = (X_1 + Y_2)/2$, $\{X_n\}_{n=1}^\infty$ are dependent. The mgf of X_n is

$$M_{X_n}(t) = \begin{cases} 1, & \text{if } t = 0, \\ \{n(\mathrm{e}^{t/n} - 1)/t\}^n \to \mathrm{e}^{t/2} \text{ as } n \to \infty, & \text{if } t \neq 0. \end{cases} \tag{2.29}$$

- Since $\mathrm{e}^{t/2}$ is the mgf of the degenerate r.v. Z with all mass at 0.5; i.e., $\Pr(Z = 0.5) = 1$, we may expect the cdf F_n of X_n has the following limitation distribution

$$F_n(x) \to F_Z(x) = \begin{cases} 0, & x < 0.5, \\ 1, & x \geqslant 0.5. \end{cases}$$

20.1• Proof of (2.28)

— The pdf of $Y_1 \sim U(0,1)$ is $f(y_1) = 1 \cdot I_{(0,1)}(y_1)$.

— The mgf of Y_1 is defined by $M_{Y_1}(t) = E(\mathrm{e}^{tY_1})$.

— If $t = 0$, we have $M_{Y_1}(t) = M_{Y_1}(0) = E(\mathrm{e}^0) = 1$.

— If $t \neq 0$, we obtain

$$M_{Y_1}(t) = \int_0^1 \mathrm{e}^{ty_1}\,\mathrm{d}y_1 = \frac{1}{t}\,\mathrm{e}^{ty_1}\Big|_0^1 = \frac{1}{t}(\mathrm{e}^t - 1),$$

which completes the proof of (2.28). □

20.2• Proof of (2.29)

— We have

$$M_{X_n}(t) = M_{\bar{Y}}(t) = E\left\{\exp\left(\sum_{i=1}^n tY_i/n\right)\right\} = \left\{M_{Y_1}\left(\frac{t}{n}\right)\right\}^n.$$

— If $t = 0$, from the first one of (2.28), we have $M_{X_n}(t) = \{M_{Y_1}(0)\}^n = 1$.

— If $t \neq 0$, from the second formula of (2.28), we have

$$M_{X_n}(t) = \left(\frac{\mathrm{e}^{\frac{t}{n}} - 1}{\frac{t}{n}} \right)^{\frac{n}{t} \cdot t} \to \mathrm{e}^{t/2}, \quad \text{as } n \to \infty, \tag{2.30}$$

which completes the proof of (2.29). □

20.3• Proof of (2.30)

— To prove (2.30), we need to prove that

$$\lim_{x \to 0} \left(\frac{\mathrm{e}^x - 1}{x} \right)^{\frac{1}{x}} = \mathrm{e}^{1/2}. \tag{2.31}$$

Proof. Note that $\mathrm{e}^x = 1 + x + x^2/2! + x^3/3! + \cdots$, we have

$$\frac{\mathrm{e}^x - 1}{x} = 1 + \frac{x}{2} + \frac{x^2}{6} + \cdots. \tag{2.32}$$

Define

$$y = \left(\frac{\mathrm{e}^x - 1}{x} \right)^{\frac{1}{x}},$$

we obtain

$$\log(y) = \frac{1}{x} \log \left(\frac{\mathrm{e}^x - 1}{x} \right) \overset{(2.32)}{=} \frac{\log(1 + x/2 + x^2/6 + \cdots)}{x},$$

so that

$$\lim_{x \to 0} \log(y) = \lim_{x \to 0} \frac{1/2 + x/3 + \cdots}{1 + x/2 + x^2/6 + \cdots} = \frac{1}{2}.$$

Hence,

$$\lim_{x \to 0} y = \lim_{x \to 0} \mathrm{e}^{\log(y)} = \mathrm{e}^{1/2},$$

which completes the proof of (2.31). □

21• Convergence in distribution via cdf

Definition 2.2 (Convergence in distribution). Given a sequence of r.v.'s $\{X_n\}_{n=1}^{\infty}$. Let $F_n(x)$ be the cdf of X_n, if there exists an r.v. X with cdf $F(x)$ such that

$$\lim_{n \to \infty} F_n(x) = F(x)$$

for all points x at which $F(x)$ is continuous, then we say that $\{X_n\}_{n=1}^{\infty}$ converges *in distribution* or *in law* to X and write $X_n \xrightarrow{\mathrm{D}} X$ or $X_n \xrightarrow{\mathrm{L}} X$. ‖

21.1• Remarks on Definition 2.2

— It is possible that $\lim_{n\to\infty} F_n(x_0) \neq F(x_0)$ for such points x_0 at which $F(x)$ is discontinuous. distribution

— $X_n \xrightarrow{\text{L}} X \iff$ as $n \to \infty$, $X_n \overset{\text{d}}{=} X$.

— The procedure for proving $X_n \xrightarrow{\text{L}} X$ is as follows:

Step 1: Find $F_n(x)$.

Step 2: Find $F(x)$.

Step 3: Prove $F_n(x) \to F(x)$ as $n \to \infty$.

Example 2.16 (Uniform distribution). Let $\{Y_i\}_{i=1}^n \overset{\text{iid}}{\sim} U(0,\theta)$ and $X_n = Y_{(n)}$ be the n-th order statistic of $Y_1, \ldots, Y_n$. Show that $X_n \xrightarrow{\text{L}} X$, where X is an r.v. with $\Pr(X = \theta) = 1$.

Solution. The pdf and cdf of $Y \sim U(0,\theta)$ are $g(y) = 1/\theta$, $0 < y < \theta$, and

$$G(y) = \begin{cases} 0, & y < 0, \\ y/\theta, & 0 \leqslant y < \theta, \\ 1, & y \geqslant \theta, \end{cases}$$

respectively. From (2.17), we know that the pdf of X_n is

$$f_n(x) = ng(x)G^{n-1}(x) = nx^{n-1}/\theta^n, \quad 0 < x < \theta.$$

Thus, the cdf of X_n is

$$F_n(x) = \begin{cases} 0, & x < 0, \\ x^n/\theta^n, & 0 \leqslant x < \theta, \\ 1, & x \geqslant \theta, \end{cases} \quad \to \quad F(x) = \begin{cases} 0, & x < \theta, \\ 1, & x \geqslant \theta. \end{cases}$$

Therefore, $X_n \xrightarrow{\text{L}} X$. ‖

Example 2.17 (Degenerate distribution). Let $\{X_n\}_{n=1}^\infty$ be a sequence of r.v's with $\Pr(X_n = 2 + \frac{1}{n}) = 1$. Show that $X_n \xrightarrow{\text{L}} X$, where X is an r.v. with $\Pr(X = 2) = 1$.

Solution. The cdf of X_n is

$$F_n(x) = \begin{cases} 0, & x < 2 + 1/n, \\ 1, & x \geqslant 2 + 1/n, \end{cases}$$

$$\to F(x) = \begin{cases} 0, & x < 2, \\ 1, & x \geqslant 2, \end{cases} \qquad \text{as } n \to \infty.$$

Thus, $\lim_{n\to\infty} F_n(x) = F(x)$ for $x \neq 2$; i.e., all points where $F(x)$ is continuous. Thus $X_n \xrightarrow{\text{L}} X$. ‖

22• Convergence in distribution via mgf

Theorem 2.5 (Equivalent result). Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of r.v's. Assume that the mgf $M_{X_n}(t) = M(t; n)$ of X_n exists for $|t| < h$ for all n, and there exists an r.v. X with mgf $M(t)$ that exists for $|t| < h_1 < h$. If

$$\lim_{n\to\infty} M(t; n) = M(t),$$

then $X_n \xrightarrow{\text{L}} X$. ‖

Example 2.18 (Binomial distribution). Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of r.v.'s and $X_n \sim \text{Binomial}(n, p)$ with $np = \mu$, then $X_n \xrightarrow{\text{L}} X$, where $X \sim \text{Poisson}(\mu)$.

n，p 是变化的

Solution. The mgf of $X_n \sim \text{Binomial}(n, p)$ is

$$M(t; n) = (p\, \mathrm{e}^t + q)^n = \left\{1 + \frac{\mu(\mathrm{e}^t - 1)}{n}\right\}^n$$

$$\to \exp\{\mu(\mathrm{e}^t - 1)\} \quad \text{as } n \to \infty. \tag{2.33}$$

for all real t. Since $\exp\{\mu(\mathrm{e}^t - 1)\}$ is the mgf of Poisson r.v. X, we have $X_n \xrightarrow{\text{L}} X$. ‖

22.1• Proof of (2.33). To prove (2.33), we need to prove that

$$\lim_{n\to\infty} \left(1 + \frac{a}{n}\right)^n = \mathrm{e}^a \quad \text{or} \quad \lim_{x\to 0}(1 + ax)^{\frac{1}{x}} = \mathrm{e}^a. \tag{2.34}$$

— Proof. Define $y = (1+ax)^{\frac{1}{x}}$, we have $\log(y) = (1/x)\log(1+ax)$ so that 洛必达

$$\lim_{x\to 0}\log(y) = \lim_{x\to 0}\frac{\log(1+ax)}{x} = \lim_{x\to 0}\frac{\frac{a}{1+ax}}{1} = a.$$

Therefore, $\lim_{x\to 0} y = \mathrm{e}^a$, which completes the proof of (2.34). □

2.5.2 Convergence in probability 依概率收敛

Definition 2.3 (Weak convergence). A sequence of r.v.'s $\{X_n\}_{n=1}^{\infty}$ is said to *weakly* converge *in probability* to an r.v. X, denoted by $X_n \xrightarrow{\mathrm{P}} X$, if for any $\varepsilon > 0$,

$$\lim_{n\to\infty}\Pr(|X_n - X| \geqslant \varepsilon) = 0. \qquad \|$$

Theorem 2.6 (Markov inequality). Let $E|X|^r < \infty$, $r > 0$, $\varepsilon > 0$. Then

$$\Pr(|X| \geqslant \varepsilon) \leqslant \frac{E|X|^r}{\varepsilon^r}. \tag{2.35}$$

In particular, let $r = 2$, then $\mathrm{Var}(X) < \infty$ and

$$\begin{aligned} \Pr(|X-\mu| \geqslant \varepsilon) &\leqslant \frac{\mathrm{Var}(X)}{\varepsilon^2} \quad \text{or} \\ \Pr(|X-\mu| < \varepsilon) &\geqslant 1 - \frac{\mathrm{Var}(X)}{\varepsilon^2}, \end{aligned} \tag{2.36}$$

where $\mu = E(X)$. $\|$

Proof. If $|x| \geqslant \varepsilon$, then $|x|^r \geqslant \varepsilon^r$; i.e.,

$$1 \leqslant \frac{|x|^r}{\varepsilon^r}.$$

Let $X \sim F(x)$, we have

$$\begin{aligned} \Pr(|X| \geqslant \varepsilon) &= \int_{|x|\geqslant\varepsilon} \mathrm{d}F(x) \\ &\leqslant \int_{|x|\geqslant\varepsilon} \frac{|x|^r}{\varepsilon^r}\,\mathrm{d}F(x) \\ &\leqslant \int_{-\infty}^{\infty} \frac{|x|^r}{\varepsilon^r}\,\mathrm{d}F(x) \\ &= \frac{E|X|^r}{\varepsilon^r}, \end{aligned}$$

which implies (2.35). □

2.5.3 Relationship of four classes of convergency

Definition 2.4 (Strong convergence). A sequence of r.v.'s $\{X_n\}_{n=1}^{\infty}$ is said to *strongly* converge *almost surely* to an r.v. X, denoted by $X_n \overset{\text{a.s.}}{\to} X$, if

$$\Pr\left(\lim_{n\to\infty} X_n = X\right) = 1. \qquad \|$$

Definition 2.5 (Convergence in mean square). A sequence of r.v.'s $\{X_n\}_{n=1}^{\infty}$ is said to converge *in mean square* to an r.v. X, denoted by $X_n \overset{\text{m.s.}}{\to} X$, if

$$\lim_{n\to\infty} E(X_n - X)^2 = 0. \qquad \|$$

The relationship of the four classes of convergency can be summarized by

$$\begin{matrix} X_n \overset{\text{a.s.}}{\to} X \\ X_n \overset{\text{m.s.}}{\to} X \end{matrix} \Longrightarrow X_n \overset{\text{P}}{\to} X \Longrightarrow X_n \overset{\text{L}}{\to} X.$$

Property 2.1 $X_n \overset{\text{P}}{\to} X \Longrightarrow X_n \overset{\text{L}}{\to} X.$ $\|$

Proof. We first prove the following facts: (i) $\forall\, x' < x$, if $X_n \overset{\text{P}}{\to} X$, then

$$\Pr(X_n \geqslant x, X < x') \to 0. \tag{2.37}$$

(ii) $\forall\, x < x''$, if $X_n \overset{\text{P}}{\to} X$, then

$$\Pr(X_n < x, X \geqslant x'') \to 0. \tag{2.38}$$

In fact, $\{X_n \geqslant x, X < x'\} \Longrightarrow X_n - X \geqslant x - x' > 0$, then

$$|X_n - X| = X_n - X \geqslant x - x' > 0.$$

Thus,

$$0 \leqslant \Pr\{X_n \geqslant x, X < x'\} \leqslant \Pr\{|X_n - X| \geqslant x - x'\} \to 0,$$

which implies (2.37). Similarly, we can prove (2.38).

On the one hand, for $x' < x$, since

$$\begin{aligned} \{X < x'\} &= \{X_n < x, X < x'\} + \{X_n \geqslant x, X < x'\} \\ &\subset \{X_n < x\} + \{X_n \geqslant x, X < x'\}, \end{aligned}$$

we have

$$F(x') \leqslant F_n(x) + \Pr\{X_n \geqslant x, X < x'\} \leqslant \underline{\lim}_{n\to\infty} F_n(x).$$

On the other hand, for $x < x''$, since

$$\begin{aligned} \{X \geqslant x''\} &= \{X_n \geqslant x, X \geqslant x''\} + \{X_n < x, X \geqslant x''\} \\ &\subset \{X_n \geqslant x\} + \{X_n < x, X \geqslant x''\}, \end{aligned}$$

we have

$$1 - F(x'') \leqslant \underline{\lim}_{n\to\infty} \Pr\{X_n \geqslant x\} = 1 - \overline{\lim}_{n\to\infty} F_n(x),$$

i.e., $F(x'') \geqslant \overline{\lim}_{n\to\infty} F_n(x)$.

Therefore, for $x' < x < x''$, we have

$$F(x') \leqslant \underline{\lim}_{n\to\infty} F_n(x) \leqslant \overline{\lim}_{n\to\infty} F_n(x) \leqslant F(x'').$$

Let x be a point at which $F(x)$ is continuous. Let $x' \to x$ and $x'' \to x$, then $F(x) = \lim_{n\to\infty} F_n(x)$. □

Property 2.2 $X_n \xrightarrow{\text{L}} c \Longleftrightarrow X_n \xrightarrow{\text{P}} c$, where c is a constant. ‖

Proof. Property 2.1 indicates that we only need to prove " $\Longrightarrow$ ". Note that the cdf of $X = c$ is

$$F_X(x) = \begin{cases} 0, & \text{if } x \leqslant c, \\ 1, & \text{if } x > c, \end{cases}$$

hence, as $n \to \infty$,

$$\begin{aligned} \Pr(|X_n - c| \geqslant \varepsilon) &= \Pr(X_n \geqslant c + \varepsilon) + \Pr(X_n \leqslant c - \varepsilon) \\ &= 1 - F_n(c + \varepsilon) + F_n(c - \varepsilon) \\ &\to 1 - F_X(c + \varepsilon) + F_X(c - \varepsilon) \\ &\to 1 - 1 + 0 = 0, \end{aligned}$$

which completes the proof. □

Property 2.3 $X_n \xrightarrow{\text{m.s.}} X \Longrightarrow X_n \xrightarrow{\text{P}} X$. ‖

Proof. If $X_n \xrightarrow{\text{m.s.}} X$, by using (2.35), then

$$\Pr(|X_n - X| \geqslant \varepsilon) \leqslant \frac{E(X_n - X)^2}{\varepsilon^2} \to 0, \quad \text{as } n \to \infty.$$

This means that $X_n \xrightarrow{\text{P}} X$. □

2.5.4 Law of large number

Theorem 2.7 (Weak law of large number). Assume that $\{X_n\}_{n=1}^{\infty}$ is a sequence of i.i.d. random variables with $E(X_n) = \mu < \infty$. Let $\bar{X}_n = \sum_{i=1}^{n} X_i/n$, then $\bar{X}_n \xrightarrow{\text{P}} \mu$. ‖

Proof. We prove it under an additional assumption $\text{Var}(X_n) = \sigma^2 < \infty$. By using (2.35), we have

$$\Pr(|\bar{X}_n - \mu| \geqslant \varepsilon) \leqslant \frac{\text{Var}(\bar{X}_n)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \to 0, \quad \text{as } n \to \infty.$$

This means that $\bar{X}_n \xrightarrow{\text{P}} \mu$. □

Theorem 2.8 (Strong law of large number). Assume that $\{X_n\}_{n=1}^{\infty}$ is a sequence of i.i.d. random variables with $E(X_n) = \mu < \infty$. Let $\bar{X}_n = \sum_{i=1}^{n} X_i/n$, then $\bar{X}_n \xrightarrow{\text{a.s}} \mu$. ‖

2.5.5 Central limit theorem

23• Proof of the central limit theorem via mgf 任何分布

Theorem 2.9 (Central limit theorem). Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of i.i.d. random variables with common mean μ and common variance $\sigma^2 > 0$. Let $\bar{X}_n = \sum_{i=1}^{n} X_i/n$ and $Y_n = \sqrt{n}(\bar{X}_n - \mu)/\sigma$, then $Y_n \xrightarrow{\text{L}} Z$ as $n \to \infty$, where $Z \sim N(0,1)$. ‖

标准化

算术平均

Proof. Assume that the mgf of X exists for $|t| < h$. Let

$$m(t) = E\{\,\mathrm{e}^{t(X-\mu)}\}.$$

自然界，很多不同因子并不起决定，最后导致正态分布

Then $m(0) = 1$, $m'(0) = E(X-\mu) = 0$, $m''(0) = E(X-\mu)^2 = \sigma^2$. By Maclaurin's expansion,

$$m(t) = m(0) + m'(0)t + \frac{1}{2}m''(\xi)t^2 = 1 + \frac{m''(\xi)}{2}t^2, \quad 0 < \xi < t,$$

where $m''(\xi) \to m''(0) = \sigma^2$ as $t \to 0$. Now

$$\begin{aligned} M(t;n) &= E(\,\mathrm{e}^{tY_n}) \\ &= E[\exp\{t\sqrt{n}(\bar{X}_n - \mu)/\sigma\}] \\ &= E[\exp\{t\textstyle\sum_{i=1}^{n}(X_i - \mu)/(\sqrt{n}\sigma)\}] \end{aligned}$$

二项分布可以近似成正态分布

$$
\begin{aligned}
&= \prod_{i=1}^{n} E[\exp\{t(X_i-\mu)/(\sqrt{n}\sigma)\}] \\
&= \{m(t/(\sqrt{n}\sigma))\}^n \\
&= \left\{1+\frac{m''(\xi(n))}{2}(t/(\sqrt{n}\sigma))^2\right\}^n \qquad \text{和n相关} \\
&= \left\{1+\frac{m''(\xi(n))}{2n\sigma^2}t^2\right\}^n, \quad 0<\xi(n)<t/(\sqrt{n}\sigma) \\
&\to \mathrm{e}^{t^2/2} \quad \text{as } n\to\infty,
\end{aligned}
$$

since $\xi(n)\to 0$ and $m''(\xi(n))\to m''(0)=\sigma^2$. Because $\mathrm{e}^{t^2/2}$ is the mgf of $Z\sim N(0,1)$, this means that $Y_n \xrightarrow{\mathrm{L}} Z$. □

Example 2.19 (Bernoulli distribution). Let $X_1,\ldots,X_n$ be a random sample from Bernoulli(θ). Let $Z_n=\sum_{i=1}^n X_i$, then binomial

$$
\frac{Z_n-n\theta}{\sqrt{n\theta(1-\theta)}} \xrightarrow{\mathrm{L}} N(0,1) \quad \text{as } n\to\infty. \tag{2.39}
$$

Solution. Because $\mu=\theta$ and $\sigma^2=\theta(1-\theta)$, by the central limit theorem, we have

$$
\frac{Z_n-n\theta}{\sqrt{n\theta(1-\theta)}} = \frac{n\bar{X}_n-n\theta}{\sqrt{n\theta(1-\theta)}} = \frac{\sqrt{n}(\bar{X}_n-\mu)}{\sigma} \xrightarrow{\mathrm{L}} Z \quad \text{as } n\to\infty,
$$

where $Z\sim N(0,1)$. ‖

23.1• Remarks on normal approximation

— Since $Z_n\sim$ Binomial(n,θ), we have $E(Z_n)=n\theta$ and $\mathrm{Var}(Z_n)=n\theta(1-\theta)$. Then (2.39) means

$$
\frac{Z_n-E(Z_n)}{\sqrt{\mathrm{Var}(Z_n)}} \xrightarrow{\mathrm{L}} Z\sim N(0,1) \quad \text{as } n\to\infty.
$$

— If n is large, approximately we have

$$
Z_n \dot{\sim} N(n\theta, n\theta(1-\theta)).
$$

That is, Binomial(n,θ) can be approximated by $N(n\theta,n\theta(1-\theta))$.

— If Z_n is a discrete r.v., in using normal approximation, we should use

$$\Pr(Z_n = k) = \Pr(k - 0.5 < Z_n < k + 0.5),$$

连续性矫正

and number 0.5 here is called the continuity correction.

Example 2.20 (Binomial distribution). Let $X \sim \text{Binomial}(10, 0.5)$, directly calculate $\Pr(X = 4)$ and compute $\Pr(X = 4)$ by normal approximation.

Solution. First, we directly compute

$$\Pr(X = 4) = \binom{10}{4} 0.5^4 0.5^6 = 0.2051.$$

Second, we use normal approximation $X \sim N(5, 2.5)$ and obtain

$$\begin{aligned}
\Pr(X = 4) &= \Pr(4 - 0.5 < X < 4 + 0.5) \\
&= \Pr(3.5 < X < 4.5) \\
&= \Pr\left(\frac{3.5 - 5}{\sqrt{2.5}} < \frac{X - 5}{\sqrt{2.5}} < \frac{4.5 - 5}{\sqrt{2.5}}\right) \\
&\doteq \Pr(-0.9487 < Z < -0.3162) \\
&= \Phi(-0.3162) - \Phi(-0.9487) \\
&= \Phi(0.9487) - \Phi(0.3162) \\
&= 0.8286 - 0.6241 = 0.2045.
\end{aligned}$$

分布间区别越大，收敛到正态分布就越慢

The error is $0.2051 - 0.2045 = 0.0006$ and the percentage error is

$$\frac{|0.2051 - 0.2045|}{0.2051} = 0.29\%. \qquad \|$$

2.6 Some Challenging Questions

24• Dependency and correlation

- Let r.v. $X \sim N(0, 1)$ and we define a new random variable $Y = X^2$.
- In Example 2.7, we know that $Y \sim \chi^2(1)$.

24.1• Dependency and correlation between X and Y

— It is clear that X and Y are *dependent* because $Y = X^2$ is uniquely determined when X is given.

— Let $\phi(x)$ be the pdf of $N(0,1)$. Since $x^3\phi(x)$ is an odd function, we have

$$E(XY) = E(X^3) = \int_{-\infty}^{\infty} x^3\phi(x)\,\mathrm{d}x = 0.$$

— Note that $E(X) = 0$, we obtain

协方差

$$\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} = \frac{E(XY) - E(X)E(Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} = 0.$$

— In other words, X and Y are uncorrelated but surely dependent.

24.2• Conditional distributions of $Y|(X = x)$ and $X|(Y = y)$

— The conditional distribution of $Y|(X = x)$ is

$$\Pr(Y = x^2|X = x) = 1;$$

i.e., $Y|(X = x) \sim \mathrm{Degenerate}(x^2)$.

— The conditional distribution of $X|(Y = y > 0)$ is given by

$$\Pr(X = -\sqrt{y}|Y = y) = \Pr(X = \sqrt{y}|Y = y) = 0.5;$$

that is, $X|(Y = y > 0)$ follows a uniform two-point distribution.

— The conditional distribution of $X|(Y = y = 0)$ is

$$\Pr(X = 0|Y = 0) = 1;$$

that is, $X|(Y = y = 0) \sim \mathrm{Degenerate}(0)$.

24.3• The joint cdf of X and Y

— Let $F(x,y)$ denote the cdf of (X,Y), we have

$$\begin{aligned} F(x,y) &= \Pr(X \leqslant x, X^2 \leqslant y) = \Pr(X \leqslant x, -\sqrt{y} \leqslant X \leqslant \sqrt{y}) \\ &= \Pr\{-\sqrt{y} \leqslant X \leqslant \min(x, \sqrt{y})\} \\ &= \Phi(\min\{x, \sqrt{y}\}) - \Phi(-\sqrt{y}), \quad -\infty < x < \infty,\ y > 0, \end{aligned}$$

where $\Phi(\cdot)$ is the cdf of the standard normal distribution.

24.4• Can the identities

?

$$f_{(X,Y)}(x,y) = f_X(x)f_{(Y|X)}(y|x) = f_Y(y)f_{(X|Y)}(x|y) \tag{2.40}$$

be used to derive the joint density function of X and Y?

— No.

24.5• Comment on the existence of $f_{(X,Y)}(x,y)$ in the xy-plane

— The joint pdf of (X,Y) does *not exist* in the xy-plane because the support of (X,Y) is

$$\mathbb{S}_{(X,Y)} = \{(x,y):\ -\infty < x < \infty,\ y = x^2\},$$

which is a curve and the *measure/area of* $\mathbb{S}_{(X,Y)}$ *is zero.*

25• Proof of Theorem 2.1

- In **41.2•** of Chapter 1, it was shown that the mgf of $\mathbf{x} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is

$$M_{\mathbf{x}}(\boldsymbol{t}) = \exp(\boldsymbol{t}^\top\boldsymbol{\mu} + 0.5\boldsymbol{t}^\top\boldsymbol{\Sigma}\boldsymbol{t}). \tag{2.41}$$

25.1• $\mathbf{Ax} \sim N_m(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top)$ and $\mathbf{Bx} \sim N_r(\mathbf{B}\boldsymbol{\mu}, \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^\top)$

— Let $\boldsymbol{s} = (s_1, \ldots, s_m)^\top$ and define

$$\underset{m\times 1}{\mathbf{y}} = \underset{m\times n}{\mathbf{A}}\ \underset{n\times 1}{\mathbf{x}},$$

then the mgf of $\mathbf{y}$ is

$$\begin{aligned}
M_{\mathbf{y}}(\boldsymbol{s}) &= E\{\exp(\boldsymbol{s}^\top\mathbf{y})\} = E\{\exp(\boldsymbol{s}^\top\mathbf{A}\mathbf{x})\} \\
&= E[\exp\{(\mathbf{A}^\top\boldsymbol{s})^\top\mathbf{x}\}] \\
&= M_{\mathbf{x}}(\mathbf{A}^\top\boldsymbol{s}) \qquad [\text{Let } \boldsymbol{t} = \mathbf{A}^\top\boldsymbol{s}] \\
&= M_{\mathbf{x}}(\boldsymbol{t}) \\
&\overset{(2.41)}{=} \exp(\boldsymbol{t}^\top\boldsymbol{\mu} + 0.5\boldsymbol{t}^\top\boldsymbol{\Sigma}\boldsymbol{t}) \\
&= \exp(\boldsymbol{s}^\top\mathbf{A}\boldsymbol{\mu} + 0.5\boldsymbol{s}^\top\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top\boldsymbol{s}) \\
&= \exp\{\boldsymbol{s}^\top(\mathbf{A}\boldsymbol{\mu}) + 0.5\boldsymbol{s}^\top(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top)\boldsymbol{s}\},
\end{aligned}$$

implying $\mathbf{y} \sim N_m(\mathbf{A}\boldsymbol{\mu},\ \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top)$.

— Similarly, we can prove $\mathbf{Bx} \sim N_r(\mathbf{B}\boldsymbol{\mu}, \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^\top)$.

25.2• $\mathbf{Ax} \perp\!\!\!\perp \mathbf{Bx}$ iff $\mathbf{A}\boldsymbol{\Sigma}\mathbf{B}^\top = \mathbf{O}_{m\times r}$

— Define

$$\underset{(m+r)\times 1}{\mathbf{z}} = \begin{pmatrix}\mathbf{Ax}\\ \mathbf{Bx}\end{pmatrix} = \begin{pmatrix}\mathbf{A}\\ \mathbf{B}\end{pmatrix}\mathbf{x} \triangleq \underset{(m+r)\times n}{\mathbf{C}}\ \underset{n\times 1}{\mathbf{x}},$$

then, we have $\mathbf{z} \sim N_{m+r}(\mathbf{C}\boldsymbol{\mu},\ \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^\top)$.

— Note that

$$\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^\top = \begin{pmatrix}\mathbf{A}\\ \mathbf{B}\end{pmatrix}\boldsymbol{\Sigma}(\mathbf{A}^\top\ \mathbf{B}^\top) = \begin{pmatrix}\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top & \mathbf{A}\boldsymbol{\Sigma}\mathbf{B}^\top\\ \mathbf{B}\boldsymbol{\Sigma}\mathbf{A}^\top & \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^\top\end{pmatrix},$$

we can see that $\mathbf{Ax} \perp\!\!\!\perp \mathbf{Bx}$ iff $\mathbf{A}\boldsymbol{\Sigma}\mathbf{B}^\top = \mathbf{O}_{m\times r}$. □

Exercise 2

2.1 Calculate the expectation and variance of the $T \sim t(n)$ via the stochastic representation (SR):

$$T \overset{\mathrm{d}}{=} \frac{Z}{\sqrt{Y/n}},$$

where $Z \sim N(0,1)$, $Y \sim \chi^2(n)$ and $Z \perp\!\!\!\perp Y$.

2.2 Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Beta}(3,2)$. Find the sampling distributions of $X_{(1)} = \min\{X_1, \ldots, X_n\}$ and $X_{(n)} = \max\{X_1, \ldots, X_n\}$.

2.3 Let $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ be the order statistics of a random sample of size n from the exponential distribution with pdf $f(x) = \mathrm{e}^{-x}$, $x > 0$, zero elsewhere.

(a) Show that $Z_1 = nX_{(1)}$, $Z_2 = (n-1)[X_{(2)} - X_{(1)}]$, $Z_3 = (n-2)[X_{(3)} - X_{(2)}], \ldots, Z_n = X_{(n)} - X_{(n-1)}$ are independent and that each Z_i has the exponential distrbution.

(b) Demonstrate that all linear functions of $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$, such as $\sum_{i=1}^n a_i X_{(i)}$, can be expressed as linear functions of independent random variables.

2.4 Let $X_i \sim \text{Gamma}(a_i, 1)$, $i = 1, \ldots, n$, and $X_1, \ldots, X_n$ are mutually independent. Define

$$Y_i = \frac{X_i}{X_1 + \cdots + X_n}, \quad i = 1, \ldots, n-1.$$

(a) Find the joint density of $(Y_1, \ldots, Y_{n-1})$.

(b) Find the density of $X_1 + \cdots + X_n$.

2.5 Let $X \sim \text{Gamma}(p, 1)$, $Y \sim \text{Beta}(q, p-q)$, and $X \perp\!\!\!\perp Y$, where $0 < q < p$. Find the distribution of XY.

2.6 Let $Z \sim \text{Bernoulli}\,(1-\phi)$, $\mathbf{x} = (X_1, \ldots, X_m)^\top$, $X_i \sim \text{Poisson}\,(\lambda_i)$ for $i = 1, \ldots, m$, and $(Z, X_1, \ldots, X_m)$ be mutually independent. Define $\mathbf{y} = (Y_1, \ldots, Y_m)^\top = Z\mathbf{x}$. Find the joint pmf of $\mathbf{y}$.

Chapter 3

Point Estimation

3.1 Maximum Likelihood Estimator

3.1.1 Point estimator and point estimate

1• Difference between point estimator and point estimate

- Let the pdf of an r.v. X be $f(x;\boldsymbol{\theta})$ with an unknown parameter vector $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^p$, where $\boldsymbol{\Theta}$ denotes the corresponding parameter space.
- Thus, we have a family of densities $\{f(x;\boldsymbol{\theta}):\ \ \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$.
- We need to select one member from the family as the pdf of X.
- This is equivalent to estimating the parameter vector $\boldsymbol{\theta}$.
- To this end, we take a random sample $X_1, \ldots, X_n$ from a population with the pdf $f(x;\boldsymbol{\theta})$.

1.1• Remarks

— In Chapters 1–2, we denote the pdf of an r.v. X by $f(x)$, while starting from Chapter 3 we denote it by $f(x;\boldsymbol{\theta})$ to emphasize its dependence on the parameter vector $\boldsymbol{\theta}$.

— For example, if $X \sim N(\mu, \sigma^2)$, we have

$$f(x;\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \quad \text{where} \quad \boldsymbol{\theta} = (\mu, \sigma^2)^\top.$$

— Let $2.18, 2.76, 1.80, 1.73, 1.13, 1.85, 2.02, 2.69, 1.66, 2.59$ be a random sample of size 10 from $N(\mu, \sigma^2)$, how to estimate μ and σ^2? This is the main topic of Chapter 3.

— An advanced reference book is: Lehmann, E.L. and Casella, G. (1998). *Theory of Point Estimation* (2-nd ed.). Springer, New York.

Definition 3.1 (A statistic). A function of one or more r.v.'s that does not depend on the unknown parameter vector is called a *statistic*. ‖

1.2• Comparison of Definition 3.1 with Definition 2.1 in §2.2

— In Definition 2.1, it is assumed that $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F(x)$; i.e., $\{X_i\}_{i=1}^n$ is a random sample from $F(x)$.

— In fact, the assumption of independence is not necessary.

— In other words, $\{X_i\}_{i=1}^n$ could be correlated or dependent.

1.3• Definitions of a point estimator and a point estimate

— If a statistic $Y = \varphi(\mathbf{x})$ is used to estimate the parameter θ, where $\mathbf{x} = (X_1, \ldots, X_n)^\top$, then the statistic is called a *point estimator* of θ, where Y is a *random variable*.

— If the observations of $X_1, \ldots, X_n$ are $x_1, \ldots, x_n$, then $y = \varphi(\boldsymbol{x})$ is called a *point estimate* of θ, where y is a *real number* and $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$.

1.4• Illustration examples

— For example, $\bar{X} = (1/n)\sum_{i=1}^n X_i$ is a point estimator of $\mu = E(X)$ and $\bar{x} = (1/n)\sum_{i=1}^n x_i$ is a point estimate of μ.

— Similarly, $S^2 = \sum_{i=1}^n (X_i - \bar{X})^2/(n-1)$ is a point estimator of $\sigma^2 = \text{Var}(X)$ and $s^2 = \sum_{i=1}^n (x_i - \bar{x})^2/(n-1)$ is a point estimate of σ^2.

1.5• How to understand a point estimator?

— A point estimator is a random variable.

— A point estimator is always related with the estimation of θ. For instance, $\bar{X}$ is a point estimator of $\mu = E(X)$ but $\sum_{i=1}^{n} X_i$ is not.

— Point estimator is not unique. For example, S^2 is an unbiased estimator of $\sigma^2 = \mathrm{Var}(X)$ while $\hat{\sigma}^2 = (1/n)\sum_{i=1}^{n}(x_i - \bar{x})^2$ is the moment estimator of σ^2 for any population. In particular, $\hat{\sigma}^2$ is the maximum likelihood estimator of σ^2 for the normal population.

3.1.2 Joint density and likelihood function

2• Difference between joint pdf and likelihood function

- Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the unknown parameter vector and $\boldsymbol{\Theta}$ is the parameter space.
- Let $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ be observations of $\mathbf{x} = (X_1, \ldots, X_n)^\top$, then the *joint density* of $\mathbf{x}$ is $f(\boldsymbol{x}; \boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i; \boldsymbol{\theta})$.
- Since $\boldsymbol{x}$ has been observed and its components are therefore fixed real numbers, we regard $f(\boldsymbol{x}; \boldsymbol{\theta})$ as a function of $\boldsymbol{\theta}$, and define

$$L(\boldsymbol{\theta}) = L(\boldsymbol{\theta}; \boldsymbol{x}) = f(\boldsymbol{x}; \boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i; \boldsymbol{\theta}), \quad \boldsymbol{\theta} \in \boldsymbol{\Theta},$$

  as the *likelihood function* of the random sample $\mathbf{x}$.
- Alternatively, $L(\boldsymbol{\theta})$ is also called the likelihood function of $\boldsymbol{\theta}$.

2.1• How to understand the likelihood function $L(\boldsymbol{\theta})$?

— The joint density $f(\boldsymbol{x}; \boldsymbol{\theta})$ is a term of Probability while the likelihood function $L(\boldsymbol{\theta})$ is a term of Statistics.

— $f(\boldsymbol{x}; \boldsymbol{\theta})$ emphasizes $\boldsymbol{x}$ while $L(\boldsymbol{\theta})$ emphasizes $\boldsymbol{\theta}$.

— In statistics, in general, $L(\boldsymbol{\theta})$ is concave. That is $\nabla^2 L(\boldsymbol{\theta}) \leqslant 0$. In particular, when $\boldsymbol{\theta}$ is one-dimensional, $L(\theta)$ is concave iff $L''(\theta) \leqslant 0$.

3• The log-likelihood function

- In practice, the natural logarithm of $L(\boldsymbol{\theta})$, called the *log-likelihood*, is mathematically much convenient to work with.

- We define $\ell(\boldsymbol{\theta}) \triangleq \log\{L(\boldsymbol{\theta})\} = \sum_{i=1}^{n} \log\{f(x_i;\boldsymbol{\theta})\}$ for $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.
- Note that there is no loss of information in using $\ell(\boldsymbol{\theta})$ instead of $L(\boldsymbol{\theta})$ because $\log(\cdot)$ is a monotonic increasing function.

3.1.3 Maximum likelihood estimator and maximum likelihood estimate

4• Definition

- Suppose that a statistic

$$\hat{\boldsymbol{\theta}} = \begin{pmatrix} \hat{\theta}_1 \\ \vdots \\ \hat{\theta}_p \end{pmatrix} = \begin{pmatrix} u_1(\mathbf{x}) \\ \vdots \\ u_p(\mathbf{x}) \end{pmatrix} \triangleq \boldsymbol{u}(\mathbf{x})$$

  satisfies

$$L(\hat{\boldsymbol{\theta}}) = \max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} L(\boldsymbol{\theta}).$$

- Statistically, we can equivalently write above equation as

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} L(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \ell(\boldsymbol{\theta}),$$

  where "arg" is the abbreviation of "argument".

- Then $\hat{\boldsymbol{\theta}} = \boldsymbol{u}(\mathbf{x})$ is called the *maximum likelihood estimator* (MLE) of $\boldsymbol{\theta}$ and $\boldsymbol{u}(\boldsymbol{x})$ is called a *maximum likelihood estimate* (mle) of $\boldsymbol{\theta}$.

4.1• Remarks

— Note that $L(\boldsymbol{\theta})$ and $\ell(\boldsymbol{\theta})$ share their maxima at the same value of $\boldsymbol{\theta}$, and it is usually easier to find the maximum of $\ell(\boldsymbol{\theta})$.

— In general, the MLE $\hat{\boldsymbol{\theta}}$ is the solution to the score equation

$$\nabla\ell(\boldsymbol{\theta}) \triangleq \frac{\partial\ell(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = \begin{pmatrix} \dfrac{\partial\ell(\boldsymbol{\theta})}{\partial\theta_1} \\ \vdots \\ \dfrac{\partial\ell(\boldsymbol{\theta})}{\partial\theta_p} \end{pmatrix} = \mathbf{0}_p. \tag{3.1}$$

— There is no guarantee that the MLE exists or if it does whether it is unique.

— Consider the special case of $p = 1$. Then (3.1) becomes

$$\ell'(\theta) = \frac{\mathrm{d}\ell(\theta)}{\mathrm{d}\theta} = 0.$$

If $L(\theta)$ is a monotonic function of θ, then the MLE $\hat{\theta}$ locates at the boundary of Θ or does not exist.

4.2• Stationary point, saddle point and critical point

— A point c satisfying $\varphi'(c) \doteq \varphi'(x)|_{x=c} = 0$ is called a *stationary point* of $\varphi(x)$.

— For instance, $L(\theta)$ and $\ell(\theta)$ have the same stationary points since

$$\ell'(\theta^*) = \ell'(\theta)|_{\theta=\theta^*} = \frac{L'(\theta)}{L(\theta)}\bigg|_{\theta=\theta^*} = 0;$$

i.e., $\ell'(\theta^*) = 0$ iff $L'(\theta^*) = 0$.

— It is possible for c to be a local rather than a global minimum or maximum or even to be a *saddle point*. For example, $\varphi(x) = x^3$ has a saddle point at 0.

— Let $\varphi(x)$ be defined on the closed interval $[a, b]$. Two endpoints a, b and any stationary points c are known as *critical points* of $\varphi(x)$.

5• Unrestricted MLE

Example 3.1 (Bernoulli distribution). Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta)$. Find the MLE of θ.

Solution. The parameter space $\Theta = \{\theta\colon\ 0 < \theta < 1\} = (0, 1)$. Note that the pmf of X_i is given by

X_i	0	1
$p(x_i;\theta) = \Pr(X_i = x_i)$	$1-\theta$	θ

.

Thus, we have $p(x_i;\theta) = \theta^{x_i}(1-\theta)^{1-x_i}$, $x_i = 0, 1$. The joint pmf is

$$p(\boldsymbol{x};\theta) = \prod_{i=1}^{n} p(x_i;\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i},$$

so that the likelihood function is given by

$$L(\theta) = \theta^{n\bar{x}}(1-\theta)^{n-n\bar{x}}, \quad 0 < \theta < 1,$$

where $\bar{x} = (1/n)\sum_{i=1}^{n} x_i$. Now

$$\ell(\theta) = n\bar{x}\log(\theta) + (n - n\bar{x})\log(1-\theta)$$

and

$$\ell'(\theta) = \frac{n\bar{x}}{\theta} - \frac{n - n\bar{x}}{1-\theta}. \tag{3.2}$$

Solving $\ell'(\theta) = 0$ for θ, we obtain the solution $\theta = \bar{x}$. To verify that it maximizes $\ell(\theta)$ or $L(\theta)$, we have two alternative methods.

Method I: To check that the second derivative of $\ell(\theta)$ evaluated at $\bar{x}$ is strictly negative; i.e., $\ell''(\bar{x}) < 0$. Now, for any $\theta \in (0, 1)$, uniformly we have

$$\frac{\mathrm{d}^2\ell(\theta)}{\mathrm{d}\theta^2} = -\left\{\frac{n\bar{x}}{\theta^2} + \frac{n - n\bar{x}}{(1-\theta)^2}\right\} < 0.$$

Therefore,

$$\hat{\theta} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is the MLE of θ and $\bar{x}$ is the mle of θ.

Method II: To check that $\ell'(\theta) > 0$ when $\theta < \bar{x}$ and $\ell'(\theta) < 0$ when $\theta > \bar{x}$. From (3.2), it is easy to check them.

In general, Method II is more convenient than Method I. However, in statistical practice, neither Method I nor Method II is necessary. ‖

Example 3.2 (Normal distribution). Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$. Find the MLEs of μ and σ^2.

Solution. Let $\boldsymbol{\theta} = (\mu, \sigma^2)^\top$. The parameter space is

$$\begin{aligned} \boldsymbol{\Theta} &= \{(\mu, \sigma^2)^\top : \ -\infty < \mu < \infty, \ \sigma^2 > 0\} \\ &= (-\infty, \infty) \times (0, \infty) = \mathbb{R} \times \mathbb{R}_+, \end{aligned}$$

and the likelihood function is

$$\begin{aligned} L(\mu, \sigma^2) &= \prod_{i=1}^{n} f(x_i; \mu, \sigma^2) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{-\frac{(x_i-\mu)^2}{2\sigma^2}\right\} \\ &= \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n \exp\left\{-\frac{\sum_{i=1}^n (x_i-\mu)^2}{2\sigma^2}\right\}. \end{aligned}$$

Then

$$\ell(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{\sum_{i=1}^n (x_i-\mu)^2}{2\sigma^2}.$$

By differentiating $\ell(\mu, \sigma^2)$ with respect to μ and σ^2 and letting them equal zeros, we have

$$\begin{aligned} \frac{\partial \ell(\mu, \sigma^2)}{\partial \mu} &= \frac{\sum_{i=1}^n (x_i-\mu)}{\sigma^2} = 0, \\ \frac{\partial \ell(\mu, \sigma^2)}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^n (x_i-\mu)^2}{2\sigma^4} = 0. \end{aligned}$$

The solutions are $\mu = \bar{x}$ and $\sigma^2 = \sum_{i=1}^n (x_i - \bar{x})^2/n$. Therefore,

$$\hat{\mu} = \bar{X} \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

are the MLEs of μ and σ^2, respectively. ‖

Example 3.3 (Uniform distribution with one unknown endpoint). Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} U(0, \theta]$, where $\theta > 0$. Find the MLE of θ.

Solution. The parameter space is $\Theta = (0, \infty) = \mathbb{R}_+$. The joint density of $\mathbf{x} = (X_1, \ldots, X_n)^\top$ is

$$f(\boldsymbol{x}; \theta) = \begin{cases} \dfrac{1}{\theta^n}, & \text{if } \ 0 < x_i \leqslant \theta, \ i = 1, \ldots, n, \\ 0, & \text{elsewhere.} \end{cases}$$

Then, the likelihood function is given by

$$L(\theta) = \begin{cases} \dfrac{1}{\theta^n}, & \text{if } \ \theta \geqslant x_{(n)} \doteq \max(x_1, \ldots, x_n), \\ 0, & \text{elsewhere.} \end{cases} \tag{3.3}$$

Note that $L(\theta)$ is a monotone and decreasing function of θ when $\theta \in [x_{(n)}, \infty)$ as shown in Figure 3.1, and arrives its maximum at $\theta = x_{(n)}$, thus $\hat{\theta} = X_{(n)}$ is the MLE of θ.

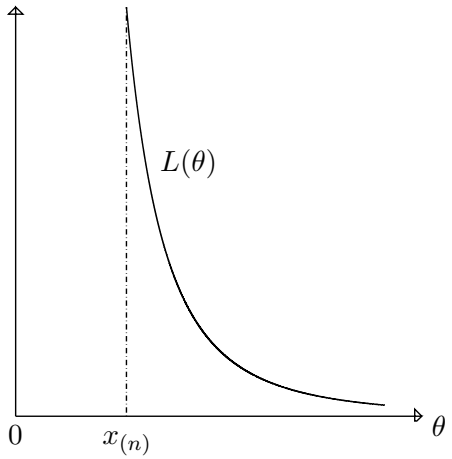

Figure 3.1 The likelihood function $L(\theta)$ defined by (3.3) is a monotone and decreasing function of θ when $\theta \in [x_{(n)}, \infty)$. ‖

5.1• Difference between maximum and supremum

— In Example 3.3, if we assume that $X_1, \ldots, X_n \overset{\text{iid}}{\sim} U(0, \theta)$, then the likelihood function (3.3) becomes

$$L(\theta) = \begin{cases} \dfrac{1}{\theta^n}, & \text{if } \theta > x_{(n)}, \\ 0, & \text{elsewhere.} \end{cases}$$

— Then, the MLE of θ does not exist.

— However, we can obtain

$$\sup L(\theta) = 1/x_{(n)}^n,$$

where "sup" is the abbreviation of "supremum".

— We should realize the difference between "max/min" and "sup/inf", where "inf" is the abbreviation of "infimum".

Example 3.4 (Uniform distribution with two unknown endpoints). Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} U[\theta - 0.5, \theta + 0.5]$, where $-\infty < \theta < \infty$. Find the MLE of θ.

Solution. The parameter space $\Theta = \mathbb{R}$. The joint density of $\mathbf{x}$ is

$$f(\boldsymbol{x}; \theta) = \prod_{i=1}^{n} f(x_i; \theta) = \prod_{i=1}^{n} I_{[\theta - 0.5,\, \theta + 0.5]}(x_i)$$

so that the likelihood is given by

$$\begin{aligned} L(\theta) &= I_{[x_{(n)} - 0.5,\, x_{(1)} + 0.5]}(\theta) \qquad (3.4) \\ &= \begin{cases} 1, & \text{if } x_{(n)} - 0.5 \leqslant \theta \leqslant x_{(1)} + 0.5, \\ 0, & \text{elsewhere.} \end{cases} \end{aligned}$$

In fact, (3.4) follows since $\prod_{i=1}^{n} I_{[\theta - 0.5,\, \theta + 0.5]}(x_i)$ is unity iff all $x_1, \ldots, x_n$ are in the interval $[\theta - 0.5,\ \theta + 0.5]$, which is true iff $\theta - 0.5 \leqslant x_{(1)}$ and $x_{(n)} \leqslant \theta + 0.5$ or $x_{(n)} - 0.5 \leqslant \theta \leqslant x_{(1)} + 0.5$. Therefore, any statistic $\hat{\theta}$ satisfying

$$X_{(n)} - 0.5 \leqslant \hat{\theta} \leqslant X_{(1)} + 0.5$$

is an MLE of θ. ‖

Example 3.5 (Laplace distribution). Let $X_1, \ldots, X_n$ be i.i.d. random variables with Laplace density (or double exponential density)

$$f(x; \theta) = \frac{1}{2} \mathrm{e}^{-|x - \theta|}, \quad -\infty < x < \infty, \quad -\infty < \theta < \infty.$$

Find the MLE of θ.

Solution. The parameter space $\Theta = \mathbb{R}$. The joint density of $\mathbf{x}$ is

$$f(\boldsymbol{x}; \theta) = \prod_{i=1}^{n} f(x_i; \theta) = \prod_{i=1}^{n} \frac{1}{2} \mathrm{e}^{-|x_i - \theta|}$$

so that the log-likelihood is given by $\ell(\theta) = -n \log(2) - \sum_{i=1}^{n} |x_i - \theta|$. The first derivative is

$$\ell'(\theta) = \sum_{i=1}^{n} \operatorname{sgn}(x_i - \theta), \qquad (3.5)$$

where $\text{sgn}(t) = 1$, 0, or -1 depending on whether $t > 0$, $t = 0$, or $t < 0$. Note that the absolute function

$$h(t) = |t| = \begin{cases} t, & \text{if } t > 0, \\ -t, & \text{if } t \leqslant 0. \end{cases}$$

When $t = 0$, $h(t)$ is not differentiable. When $t \neq 0$,

$$\begin{aligned} h'(t) &= \begin{cases} 1, & \text{if } t > 0, \\ -1, & \text{if } t < 0 \end{cases} \\ &= \text{sgn}(t). \end{aligned}$$

To get the solution to the score equation $\ell'(\theta) = 0$, we consider two cases.

- If n is even, then any point in the interval $(x_{(n/2)}, x_{(n/2+1)})$ is an mle of θ;
- If n is odd, then $\text{median}(x_1, \ldots, x_n)$ is the unique mle of θ because the median will make half the terms of the sum in expression (3.5) non-positive and half non-negative.

Therefore, the median($\mathbf{x}$) or any point in $(X_{(n/2)}, X_{(n/2+1)})$ is the MLE $\hat{\theta}$ of θ. ‖

5.2• Remarks on Example 3.5

— Let $n = 4$ and $x_1 = 0.1$, $x_2 = 0.2$, $x_3 = 0.3$, $x_4 = 0.8$. If let

$$\theta = \text{median}(x_1, \ldots, x_4) = \frac{0.2 + 0.3}{2} = 0.25,$$

then $\ell'(\theta) = \sum_{i=1}^{4} \text{sgn}(x_i - \theta) = -1 - 1 + 1 + 1 = 0$. If fact, *any point* in the open interval $(0.2, 0.3)$ is an mle of θ.

— Let $n = 3$ and $x_1 = -1$, $x_2 = 5$, $x_3 = 100$. If let $\theta = \text{median}(x_1, x_2, x_3) = 5$, then

$$\ell'(\theta) = \text{sgn}(-1 - 5) + \text{sgn}(5 - 5) + \text{sgn}(100 - 5) = -1 + 0 + 1 = 0.$$

Hence, $\text{median}(x_1, x_2, x_3)$ is the unique mle of θ.

6• Restricted MLE

- *Case 1: Equality constraints.* $\mathbf{A}\boldsymbol{\theta} = \boldsymbol{b}$, where $\mathbf{A}_{m\times p}$ and $\boldsymbol{b}_{m\times 1}$ are known, and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\top$ is an unknown parameter vector.
- *Case 2: Inequality constraints.* $\boldsymbol{a} \leqslant \mathbf{A}\boldsymbol{\theta} \leqslant \boldsymbol{b}$, where $\boldsymbol{a}_{m\times 1}$ is known.
- *Case 3: Convex constraint.* $\boldsymbol{\theta} \in \mathbb{S}$, where $\mathbb{S}$ is a convex set.

6.1• Definition of a convex set

— Let two points $C \in \mathbb{S}$ and $D \in \mathbb{S}$. If the segment of connecting the point C with the point D still belongs to $\mathbb{S}$, then $\mathbb{S}$ is called a convext set.

Example 3.6 (Multinomial distribution). Consider a multinomial experiment with n trials and p categories. The observed counts are $n_1, \ldots, n_p$ for the p categories. Let θ_j denote the cell probability of category j for $j = 1, \ldots, p$. We have $0 \leqslant \theta_j \leqslant 1$ and $\sum_{j=1}^p \theta_j = 1$. Find the MLE of θ_j subject to the equality constraint $\sum_{j=1}^p \theta_j = 1$.

Solution. Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\top$. The parameter vector space is

$$\mathbb{T}_p = \left\{\boldsymbol{\theta}\colon\ \theta_j \geqslant 0,\ j = 1, \ldots, p,\ \textstyle\sum_{j=1}^p \theta_j = 1\right\}, \tag{3.6}$$

which is the p-dimensional hyperplane. The joint pmf of $n_1, \ldots, n_p$ is

$$f(n_1, \ldots, n_p; \boldsymbol{\theta}) = \binom{n}{n_1, \ldots, n_p} \prod_{j=1}^p \theta_j^{n_j}, \quad n_j \geqslant 0, \quad \sum_{j=1}^p n_j = n.$$

The likelihood function of $\boldsymbol{\theta}$ is

$$L(\boldsymbol{\theta}) \propto \prod_{j=1}^p \theta_j^{n_j} = \left(\prod_{j=1}^{p-1} \theta_j^{n_j}\right) \left(1 - \sum_{j=1}^{p-1} \theta_j\right)^{n_p},$$

where

$$\theta_j \geqslant 0 \quad \text{and} \quad \sum_{j=1}^{p-1} \theta_j \leqslant 1. \tag{3.7}$$

Then

$$\ell(\boldsymbol{\theta}) = \sum_{j=1}^{p-1} n_j \log(\theta_j) + n_p \log\left(1 - \sum_{j=1}^{p-1} \theta_j\right).$$

By differentiating $\ell(\boldsymbol{\theta})$ with θ_j for $j = 1, \ldots, p-1$ and letting them equal zeros, we obtain

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_j} = \frac{n_j}{\theta_j} - \frac{n_p}{1 - \sum_{j=1}^{p-1} \theta_j} = \frac{n_j}{\theta_j} - \frac{n_p}{\theta_p} = 0, \quad j = 1, \ldots, p-1.$$

The solutions are given by

$$\hat{\theta}_j = \frac{n_j}{n}, \quad j = 1, \ldots, p-1,$$

which satisfy the constraints specified by (3.7). In addition, $\hat{\theta}_p = n_p/n$. ‖

6.2• Comments on Example 3.6

— Example 3.6 is a case of one equality constraint, in which we transfer the restricted case into an unrestricted case by substituting $\theta_p = 1 - \sum_{j=1}^{p-1} \theta_j$.

Example 3.7 (Normal mean with inequality constraints). Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, 1)$ subject to $a \leqslant \mu \leqslant b$, where a and b are two fixed constants. Find the MLE of μ.

Solution. The parameter space is $\Theta = [a, b]$. The likelihood function of μ is given by

$$L(\mu) = \prod_{i=1}^{n} \mathrm{e}^{-\frac{(x_i - \mu)^2}{2}}, \quad a \leqslant \mu \leqslant b$$

so that

$$\begin{aligned}
\ell(\mu) &= -\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2} \\
&= -\frac{1}{2} \sum_{i=1}^{n} (x_i^2 - 2x_i\mu + \mu^2) \\
&= -\frac{1}{2} \left(\sum_{i=1}^{n} x_i^2 - 2\mu \sum_{i=1}^{n} x_i + n\mu^2 \right) \\
&= -\frac{n}{2} \left\{ (\mu^2 - 2\mu\bar{x} + \bar{x}^2) - \bar{x}^2 + \frac{1}{n} \sum_{i=1}^{n} x_i^2 \right\} \\
&\propto -(\mu - \bar{x})^2, \quad a \leqslant \mu \leqslant b. \qquad (3.8)
\end{aligned}$$

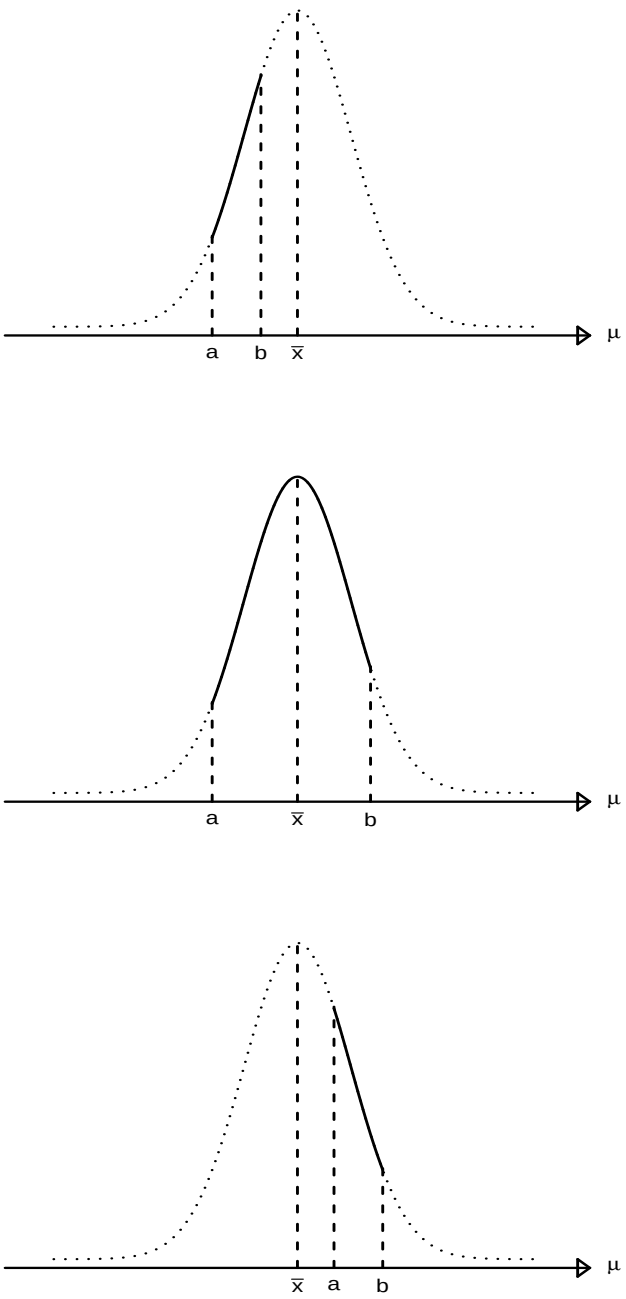

Figure 3.2 Plots of the log-likelihood function $\ell(\mu)$ defined by (B.12) for three cases. Top: $\bar{x} > b$; Middle: $a \leqslant \bar{x} \leqslant b$; Bottom: $\bar{x} < a$.

Figure 3.2 shows that $\ell(\mu)$ is a truncated quadratic function of μ. Hence

$$\begin{aligned} \mu &= \begin{cases} b, & \text{if } \bar{x} > b, \\ \bar{x}, & \text{if } a \leqslant \bar{x} \leqslant b, \\ a, & \text{if } \bar{x} < a \end{cases} \\ &= \text{median}(a, \bar{x}, b) \end{aligned}$$

is the restricted mle of μ and $\hat{\mu} = \text{median}(a, \bar{X}, b)$ is the restricted MLE of μ. As an exercise, to calculate $E(\hat{\mu})$ and $\text{Var}(\hat{\mu})$. ‖

不变性

3.1.4 The invariance property of MLE

7• Reparametrization via a one-to-one map

Theorem 3.1 (Invariance of MLE). Let $\hat{\boldsymbol{\theta}} = \boldsymbol{u}(X_1, \ldots, X_n)$ be the MLE of $\boldsymbol{\theta}_{p\times 1} \in \boldsymbol{\Theta}$. If $\boldsymbol{\eta}_{p\times 1} = \boldsymbol{h}(\boldsymbol{\theta}) = (h_1(\boldsymbol{\theta}), \ldots, h_p(\boldsymbol{\theta}))^\top$ is a one-to-one transformation between $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$, then $\hat{\boldsymbol{\eta}} = \boldsymbol{h}(\hat{\boldsymbol{\theta}})$ is the MLE of $\boldsymbol{\eta}$. ‖

Proof. Since $\boldsymbol{\eta} = \boldsymbol{h}(\boldsymbol{\theta})$ is a one-to-one map, we have $\boldsymbol{\theta} = \boldsymbol{h}^{-1}(\boldsymbol{\eta})$. The likelihood function is given by

$$L(\boldsymbol{\theta}) = L(\boldsymbol{h}^{-1}(\boldsymbol{\eta})) \triangleq L^*(\boldsymbol{\eta}).$$

We want to prove $L^*(\hat{\boldsymbol{\eta}}) \geqslant L^*(\boldsymbol{\eta})$ for all $\boldsymbol{\eta}$. In fact, we have

$$\begin{aligned} L^*(\hat{\boldsymbol{\eta}}) &= L^*(\boldsymbol{h}(\hat{\boldsymbol{\theta}})) = L\boldsymbol{h}^{-1}(\boldsymbol{h}(\hat{\boldsymbol{\theta}})) = L(\hat{\boldsymbol{\theta}}) \\ &\geqslant L(\boldsymbol{\theta}) = L^*(\boldsymbol{\eta}). \end{aligned}$$

Therefore, $\hat{\boldsymbol{\eta}} = \boldsymbol{h}(\hat{\boldsymbol{\theta}})$ is the MLE of $\boldsymbol{\eta}$. □

7.1• Understanding Theorem 3.1 through Figure 3.3

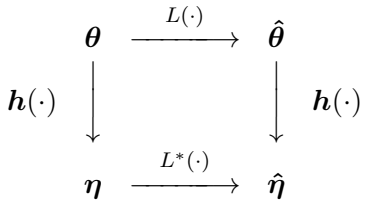

Figure 3.3 An illustration of Theorem 3.1.

7.2• Comments on Figure 3.3

— Figure 3.3 shows that Theorem 3.1 gives two ways to reach $\hat{\boldsymbol{\eta}}$.

— The first way is to first find the $\hat{\boldsymbol{\theta}}$ by maximizing the likelihood function $L(\boldsymbol{\theta})$, then to utilize the map $\boldsymbol{h}(\cdot)$ to obtain $\hat{\boldsymbol{\eta}} = \boldsymbol{h}(\hat{\boldsymbol{\theta}})$.

— The second way is to first utilize the map $\boldsymbol{h}(\cdot)$ to obtain a new parameter vector $\boldsymbol{\eta}$, then to find the $\hat{\boldsymbol{\eta}}$ by maximizing the likelihood function $L^*(\boldsymbol{\eta})$.

7.3• Two illustration examples

— Since $h(\sigma) = \sigma = \sqrt{\sigma^2}$ with $\sigma > 0$ is a one-to-one map between σ^2 and σ, it follows from Example 3.2 that $\hat{\sigma} = \{(1/n)\sum_{i=1}^n (X_i - \bar{X})^2\}^{1/2}$ different from $S = \{\sum_{i=1}^n (X_i - \bar{X})^2/(n-1)\}^{1/2}$, is an MLE of σ.

— Similarly, the MLE of, say $\log(\sigma^2)$, is $\log\left\{(1/n)\sum_{i=1}^n (X_i - \bar{X})^2\right\}$.

8• Can we extend Theorem 3.1?

- It is very natural to ask whether Theorem 3.1 still holds if the assumption that $\boldsymbol{\eta} = \boldsymbol{h}(\boldsymbol{\theta})$ is a one-to-one transformation is removed.

8.1• The MLE of variance in a Bernoulli distribution

— As a first example, assume an estimate of the variance; i.e., $\theta(1-\theta)$, of the Bernoulli(θ) distribution is desired.

— Example 3.1 gives the MLE of θ to be $\bar{X}$, but since $\theta(1-\theta)$ is not a one-to-one function of θ, Theorem 3.1 does not give the MLE of $\theta(1-\theta)$.

— Theorem 3.2 below will give such an estimator and it will be $\bar{X}(1-\bar{X})$.

8.2• The MLE of $\mu^2 + \sigma^2$ in normal distribution

— As a second example, consider the MLE of $\mu^2 + \sigma^2$ in Example 3.2.

— Since $\mu^2 + \sigma^2$ is not a one-to-one function of μ and σ^2, Theorem 3.1 does not give the MLE of $\mu^2 + \sigma^2$.

— Such an estimator will be obtainable from Theorem 3.2 below and it will be $\bar{X}^2 + (1/n)\sum_{i=1}^n (X_i - \bar{X})^2$.

Theorem 3.2 (Extension of Theorem 3.1). Let $\hat{\boldsymbol{\theta}}$ be the MLE of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\top \in \boldsymbol{\Theta}$. If $\boldsymbol{\eta}_{r\times 1} = \boldsymbol{h}(\boldsymbol{\theta}) = (h_1(\boldsymbol{\theta}), \ldots, h_r(\boldsymbol{\theta}))^\top$ for $1 \leqslant r \leqslant p$ is a many-to-few transformation between $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$, then $\hat{\boldsymbol{\eta}} = \boldsymbol{h}(\hat{\boldsymbol{\theta}}) = (h_1(\hat{\boldsymbol{\theta}}), \ldots, h_r(\hat{\boldsymbol{\theta}}))^\top$ is the MLE of $\boldsymbol{\eta}$. ‖

Proof. Let $\mathbb{H}$ denote the range space of the map $\boldsymbol{h}(\cdot) = (h_1(\cdot), \ldots, h_r(\cdot))^\top$. $\mathbb{H}$ is an r-dimensional space. Define

$$M(\boldsymbol{h}) = \max_{\{\boldsymbol{\theta}:\, \boldsymbol{h}(\boldsymbol{\theta})=\boldsymbol{h}\}} L(\boldsymbol{\theta}),$$

which is called the likelihood function induced by $\boldsymbol{h}(\cdot)$. It suffices to show

$$M(\boldsymbol{h}) \leqslant M(\boldsymbol{h}(\hat{\boldsymbol{\theta}})) \quad \text{for any } \boldsymbol{h} \in \mathbb{H},$$

which follows immediately from the inequality

$$\begin{aligned} M(\boldsymbol{h}) &= \max_{\{\boldsymbol{\theta}:\, \boldsymbol{h}(\boldsymbol{\theta})=\boldsymbol{h}\}} L(\boldsymbol{\theta}) \\ &\leqslant \max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} L(\boldsymbol{\theta}) = L(\hat{\boldsymbol{\theta}}) \\ &= \max_{\{\boldsymbol{\theta}:\, \boldsymbol{h}(\boldsymbol{\theta})=\boldsymbol{h}(\hat{\boldsymbol{\theta}})\}} L(\boldsymbol{\theta}) \\ &= M(\boldsymbol{h}(\hat{\boldsymbol{\theta}})), \end{aligned}$$

for any $\boldsymbol{h} \in \mathbb{H}$. □

8.3• Understanding Theorem 3.2

— This property of invariance of MLEs allows us in our discussion of maximum likelihood estimation to consider estimating $(\theta_1, \ldots, \theta_p)^\top$ rather than the more general $h_1(\theta_1, \ldots, \theta_p), \ldots, h_r(\theta_1, \ldots, \theta_p)$.

3.2 Moment Estimator

9• Three basic methods of estimation

- The first procedure for estimating parameters is the method of *maximum likelihood estimation.*
- The second procedure for estimating parameters is the *method of moments* proposed by the great British statistician Karl Pearson near the turn of the twentieth century.

- The third procedure is called *Bayesian estimation*.

10• Background for the maximum likelihood estimation

- Let $x_1 = 0.099$, $x_2 = -1.146$, $x_3 = -1.172$, $x_4 = -0.290$, $x_5 = 1.435$ and $x_6 = -0.657$ be corresponding observations of a random sample of size six from the population r.v. X.
- We guess that $X_1, \ldots, X_6 \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ or $X \sim N(\mu, \sigma^2)$ and we want to find the mles of μ and σ^2.
- We wonder if or not our guess is correct, which can be tested by statistical methods (e.g., the goodness-of-fit test, see §5.5); i.e.,

$$H_0\text{: The distribution of } X \text{ is normal}$$

against

$$H_1\text{: The distribution of } X \text{ is not normal.}$$

10.1• If H_0 is accepted, what can we do next step?

— Based on the observed data $\{x_i\}_{i=1}^6$, if H_0 is accepted, then by using the method of ML estimation as shown in Example 3.2, the mles of μ and σ^2 are given by

$$\bar{x} = -0.2885 \quad \text{and} \quad \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n} = 0.7954,$$

respectively.

— We could claim that $\{x_i\}_{i=1}^6$ are observation of a random sample of size six from the most possible population $N(-0.2885, 0.7954)$.

10.2• If H_0 is rejected, what can we do next step?

— One way is to guess another population distribution. If the new H_0 was accepted, we could repeat the above process.

— Alternatively, we can estimate the first and second moments of the unknown population distribution $F(\cdot)$ by using the *method of moments*.

— Of course, when the family of distribution is known but the parameters are unknown, the method of moments can also be applied.

11• Moment estimators

- By first equating the *sample moments*

$$\frac{1}{n}\sum_{i=1}^{n} X_i, \quad \frac{1}{n}\sum_{i=1}^{n} X_i^2, \quad \ldots, \quad \frac{1}{n}\sum_{i=1}^{n} X_i^r$$

to the corresponding *population moments*

$$E(X), \quad E(X^2), \quad \ldots, \quad E(X^r),$$

then solving the system of equations, we can obtain *moment estimators* of parameters.

- Specifically, if there are a total of r parameters, the moment estimators can be obtained from solving the system of equations:

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} X_i &= E(X), \\ \frac{1}{n}\sum_{i=1}^{n} X_i^2 &= E(X^2), \\ &\vdots \\ \frac{1}{n}\sum_{i=1}^{n} X_i^r &= E(X^r). \end{aligned}$$

p.59 大数定理，依概率收敛

Example 3.8 (Gamma distribution). Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Gamma}(\alpha, \beta)$. Find the moment estimators of α and β.

Solution. Let $X \sim \text{Gamma}(\alpha, \beta)$, from Appendix A.2.4, we have $E(X) = \alpha/\beta$ and $\text{Var}(X) = \alpha/\beta^2$. Thus

$$E(X^2) = \text{Var}(X) + \{E(X)\}^2 = \frac{\alpha(\alpha+1)}{\beta^2}.$$

The moment estimators of α and β must satisfy

$$\begin{aligned} \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i &= E(X) = \frac{\alpha}{\beta}, \quad \text{and} \\ \frac{1}{n}\sum_{i=1}^{n} X_i^2 &= E(X^2) = \frac{\alpha(\alpha+1)}{\beta^2}. \end{aligned}$$

Thus,

$$\hat{\beta}^{\mathrm{M}} = \frac{n\bar{X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \quad \text{and} \quad \hat{\alpha}^{\mathrm{M}} = \frac{n\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

are the corresponding moment estimators of α and β. ‖

Example 3.9 (Beta distribution). Let $x_1 = 0.42$, $x_2 = 0.10$, $x_3 = 0.65$ and $x_4 = 0.23$ be observations of random variables of size $n = 4$ from the pdf

$$f(x;\theta) = \theta x^{\theta-1}, \quad 0 \leqslant x \leqslant 1.$$

Find the moment estimate of θ.

Solution. Let $X \sim f(x;\theta)$, we have

$$E(X) = \int_0^1 x \cdot \theta x^{\theta-1}\,\mathrm{d}x = \frac{\theta}{\theta+1}.$$

Let $E(X)$ equal to the first sample moment

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{0.42 + 0.10 + 0.65 + 0.23}{4} = 0.35,$$

we obtain $\theta/(\theta+1) = \bar{x}$. Thus the moment estimate for θ is

$$\hat{\theta}^{\mathrm{M}} = \frac{\bar{x}}{1-\bar{x}} = \frac{0.35}{1-0.35} = 0.54.$$ ‖

Example 3.10 (Normal distribution). Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$. Find the moment estimators of μ and σ^2.

Solution. Let $X \sim N(\mu, \sigma^2)$, we have $E(X) = \mu$ and $E(X^2) = \sigma^2 + \mu^2$. The moment estimators of μ and σ^2 must satisfy

$$\bar{X} = \mu \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n} X_i^2 = \sigma^2 + \mu^2.$$

Hence,

$$\hat{\mu}^{\mathrm{M}} = \bar{X} \quad \text{and} \quad \hat{\sigma}^{2\mathrm{M}} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

are the corresponding moment estimators of μ and σ^2. ‖

11.1• The application range of the method of moments

— The method of moments can be applied to both *parametric* and *nonparametric* statistics.

11.2• What is the parametric statistics?

— Make inferences (i.e., estimation and testing hypothesis) on parameters in a known/specified family of distributions $\{f(x;\boldsymbol{\theta})\colon \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ based on an i.i.d. sample $\{x_i\}_{i=1}^n$ or more than one i.i.d. sample.

11.3• What is the nonparametric (or distribution-free) statistics?

— Make inferences (i.e., estimation and test) on an unknown distribution itself $F(\cdot)$ based on an i.i.d. sample $\{x_i\}_{i=1}^n$ or on two unknown distributions $F(\cdot)$ and $G(\cdot)$ based on two i.i.d. samples $\{x_i\}_{i=1}^n$ and $\{y_j\}_{j=1}^m$.

3.3 Bayesian Estimator

12• Three reference books for Bayesian statistics

- Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis* (2-nd ed.). Springer, New York, USA.
- Carlin, B.P. and Louis, T.A. (2008). *Bayesian Methods for Data Analysis* (3-rd ed.). Chapman & Hall/CRC (Texts in Statistical Science), Boca Raton, USA.
- Gelman, A., Carlin, J.P., Stern, H.S., Dunson, D.B., Vehtari, A. and Rubin, D.B. (2013). *Bayesian Data Analysis* (3-rd ed.). Chapman & Hall/CRC (Texts in Statistical Science), Boca Raton, USA.

13• Main features of Bayesian method

- In the ML estimation method and the method of moments, we have assumed that the parameters are *fixed* but *unknown* constants.
- In the Bayesian method, we assume that $\boldsymbol{\theta}$ is a random vector with a density $\pi(\boldsymbol{\theta})$, which is called the *prior density* of $\boldsymbol{\theta}$.

- Then the joint density or likelihood function (in the ML estimation method) of $\mathbf{x} = (X_1, \ldots, X_n)^\top$ becomes the conditional density (in the Bayesian method) of $\mathbf{x}$ given $\boldsymbol{\theta}$, denoted by $f(\boldsymbol{x}|\boldsymbol{\theta})$, where $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$.

13.1• The basic idea of Bayesian estimation

— The basic idea of Bayesian estimation is to utilize both the information from the prior density of $\boldsymbol{\theta}$ and the likelihood function of the observed data $\boldsymbol{x}$.

14• Three steps for determining Bayesian estimators

- Given a random sample $\mathbf{x} = (X_1, \ldots, X_n)^\top$, determine the joint density of $\mathbf{x}$ and $\boldsymbol{\theta}$:

$$\begin{aligned} f(\boldsymbol{x}, \boldsymbol{\theta}) &= \text{Likelihood} \times \text{Prior} \\ &= f(\boldsymbol{x}|\boldsymbol{\theta}) \times \pi(\boldsymbol{\theta}) \\ &= \left\{ \prod_{i=1}^{n} f(x_i|\boldsymbol{\theta}) \right\} \times \pi(\boldsymbol{\theta}), \end{aligned} \tag{3.9}$$

  where $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$.

- Determine the *posterior density* (i.e., the conditional density of $\boldsymbol{\theta}$ given $\mathbf{x} = \boldsymbol{x}$) of $\boldsymbol{\theta}$,

$$\begin{aligned} p(\boldsymbol{\theta}|\boldsymbol{x}) &= \frac{f(\boldsymbol{x}, \boldsymbol{\theta})}{f(\boldsymbol{x})} = c^{-1} f(\boldsymbol{x}, \boldsymbol{\theta}) \\ &\propto f(\boldsymbol{x}, \boldsymbol{\theta}) = \text{Likelihood} \times \text{Prior}, \end{aligned} \tag{3.10}$$

  where $f(\boldsymbol{x}) = \int_{\boldsymbol{\Theta}} f(\boldsymbol{x}, \boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} \;\hat{=}\; c$ is the normalizing constant of $p(\boldsymbol{\theta}|\boldsymbol{x})$ because $\mathbf{x} = \boldsymbol{x}$ is given.

- The Bayesian estimate of $\boldsymbol{\theta}$ (i.e., the conditional expectation of $\boldsymbol{\theta}$) is defined by

$$E(\boldsymbol{\theta}|\boldsymbol{x}) = \int_{\boldsymbol{\Theta}} \boldsymbol{\theta} \cdot p(\boldsymbol{\theta}|\boldsymbol{x}) \, \mathrm{d}\boldsymbol{\theta}. \tag{3.11}$$

Example 3.11 (Bernoulli–beta distribution). Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta)$ and the prior distribution of θ be $\text{Beta}(\alpha, \beta)$. Find the Bayesian estimate of θ.

Solution. Note that $f(x_i|\theta) = \theta^{x_i}(1-\theta)^{1-x_i}$, then the joint density of $\mathbf{x}$ and θ is

$$\begin{aligned} f(\boldsymbol{x}, \theta) &= \left\{ \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} \right\} \times \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha+x_+-1}(1-\theta)^{\beta+n-x_+-1}, \quad 0 < \theta < 1, \end{aligned}$$

where $x_+ \hat{=} \sum_{i=1}^{n} x_i$. The posterior density of θ is given by

$$p(\theta|\boldsymbol{x}) \propto \theta^{\alpha+x_+-1}(1-\theta)^{\beta+n-x_+-1}, \quad 0 < \theta < 1;$$

i.e., $\theta|\boldsymbol{x} \sim \text{Beta}(\alpha + x_+, \beta + n - x_+)$. Therefore,

$$E(\theta|\boldsymbol{x}) = \frac{\alpha + x_+}{\alpha + \beta + n}$$

is the Bayesian estimate of θ, and $(\alpha + n\bar{X})/(\alpha + \beta + n)$ is the Bayesian estimator of θ. ‖

Example 3.12 (Poisson–gamma distribution). Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Poisson}(\theta)$ and the prior distribution of θ be $\text{Gamma}(a, b)$. Find the Bayesian estimate of θ.

Solution. Note that $f(x_i|\theta) = \mathrm{e}^{-\theta}\theta^{x_i}/x_i!$, $x_i = 0, 1, 2, \ldots$, then the joint density of $\mathbf{x}$ and θ is

$$\begin{aligned} f(\boldsymbol{x}, \theta) &= \left\{ \prod_{i=1}^{n} \frac{\theta^{x_i}}{x_i!} \mathrm{e}^{-\theta} \right\} \times \frac{b^a}{\Gamma(a)} \theta^{a-1} \mathrm{e}^{-b\theta} \\ &= \frac{b^a}{\Gamma(a) \prod_{i=1}^{n} x_i!} \theta^{a+x_+-1} \mathrm{e}^{-(b+n)\theta}, \quad \theta > 0, \end{aligned}$$

where $x_+ \hat{=} \sum_{i=1}^{n} x_i$. The posterior density of θ is given by

$$p(\theta|\boldsymbol{x}) \propto \theta^{a+x_+-1} \mathrm{e}^{-(b+n)\theta}, \quad \theta > 0;$$

i.e., $\theta|\boldsymbol{x} \sim \text{Gamma}(a + x_+, b + n)$. Therefore,

$$E(\theta|\boldsymbol{x}) = \frac{a + x_+}{b + n}$$

is the Bayesian estimate of θ, and $(a+n\bar{X})/(b+n)$ is the Bayesian estimator of θ. ‖

15• Differences between MLE and Bayesian estimator

Table 3.1 A comparison of MLE with Bayesian estimator

	MLE	Bayesian estimator
1	$\boldsymbol{\theta}$: A fixed and unknown parameter vector	$\boldsymbol{\theta}$: A random vector with a prior density $\pi(\boldsymbol{\theta})$
2	$f(\boldsymbol{x};\boldsymbol{\theta})$: The joint density of $\mathbf{x} = (X_1, \ldots, X_n)^\top$	$f(\boldsymbol{x}\|\boldsymbol{\theta})$: The conditional density of $\mathbf{x}$ given $\boldsymbol{\theta}$
3	$L(\boldsymbol{\theta})$: Likelihood function	$p(\boldsymbol{\theta}\|\boldsymbol{x}) \propto L(\boldsymbol{\theta})\pi(\boldsymbol{\theta})$: Posterior density
4	$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} L(\boldsymbol{\theta})$: MLE	$\tilde{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} p(\boldsymbol{\theta}\|\boldsymbol{x})$: Posterior mode

15.1• The non-informative prior

— When the non-informative prior (i.e., $\pi(\boldsymbol{\theta}) \propto 1$) is taken as the prior of $\boldsymbol{\theta}$, or when $\pi(\boldsymbol{\theta})$ is flat, we have $\tilde{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}$.

16• Statistical interpretation of Bayesian estimator

- *Loss function*: $l(\boldsymbol{\theta}, \boldsymbol{a})$ computes the loss incurred when $\boldsymbol{\theta}$ is the true state of nature and the action $\boldsymbol{a} \in \mathcal{A}$ is taken.
- *Squared error loss*: $l(\boldsymbol{\theta}, \boldsymbol{a}) = \|\boldsymbol{\theta} - \boldsymbol{a}\|^2 = (\boldsymbol{\theta} - \boldsymbol{a})^\top(\boldsymbol{\theta} - \boldsymbol{a})$.
- *Posterior risk*: $\rho(p, \boldsymbol{a}) \triangleq \int_{\boldsymbol{\Theta}} \|\boldsymbol{\theta} - \boldsymbol{a}\|^2 p(\boldsymbol{\theta}|\boldsymbol{x})\, \mathrm{d}\boldsymbol{\theta}$.
- *Bayesian estimator*: $E(\boldsymbol{\theta}|\boldsymbol{x})$ is the action $\boldsymbol{a}^*$ such that the posterior risk reaches its minimum. That is,

$$E(\boldsymbol{\theta}|\boldsymbol{x}) = \boldsymbol{a}^* = \arg\min_{\boldsymbol{a}\in\mathcal{A}} \rho(p, \boldsymbol{a})$$

or

$$\rho(p, E(\boldsymbol{\theta}|\boldsymbol{x})) \leqslant \rho(p, \boldsymbol{a}), \quad \forall \boldsymbol{a} \in \mathcal{A}.$$

3.4 Properties of Estimators

3.4.1 Unbiasedness

17• Measures for comparing two point estimators

Definition 3.2 (Unbiased estimator and bias). An estimator $\varphi(\mathbf{x})$ is an *unbiased estimator* of the parameter θ if $E\{\varphi(\mathbf{x})\} = \theta$ for $\theta \in \Theta$. Otherwise, the estimator is biased and the bias is defined by

$$b(\theta) = E\{\varphi(\mathbf{x})\} - \theta, \tag{3.12}$$

where $\mathbf{x} = (X_1, \ldots, X_n)^\top$. ‖

Example 3.13 (Distribution with a finite second-order moment). Let X_1, $\ldots, X_n$ be a random sample from a population (which is not necessary to be a normal population) with mean μ and variance $\sigma^2 < \infty$. According to Eq.(2.9), we can see that

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{3.13}$$

are unbiased estimators of μ and σ^2, respectively. ‖

Example 3.14 (Uniform distribution). Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} U(0, \theta)$, then

1) the n-th order statistic $X_{(n)}$ is a biased estimator of θ;

2) $\frac{n+1}{n}X_{(n)}$ is an unbiased estimator of θ; and

3) $2\bar{X}$ is also an unbiased estimator of θ.

Solution. 1) From Example 2.16, we know that the pdf of $X_{(n)}$ is

$$f_n(x) = nx^{n-1}/\theta^n, \quad 0 < x < \theta.$$

Hence,

$$E\{X_{(n)}\} = \int_0^\theta x f_n(x)\,\mathrm{d}x = \frac{n}{n+1}\cdot\theta \neq \theta, \tag{3.14}$$

indicating that $X_{(n)}$ is a biased estimator of θ.

2) Clearly, $\frac{n+1}{n}X_{(n)}$ is an unbiased estimator of θ.

3) Since

$$E(X_1) = \int_0^\theta x_1 \cdot \frac{1}{\theta}\,\mathrm{d}x_1 = \frac{\theta}{2},$$

we have $E(2\bar{X}) = 2E(\bar{X}) = 2E(X_1) = \theta$. ‖

Definition 3.3 (MSE). Given an estimator $Y = \varphi(\mathbf{x})$ of θ, the *mean square error* (MSE) of the estimator is defined by

$$\text{MSE} = E\big\{\varphi(\mathbf{x}) - \theta\big\}^2. \qquad \|$$

17.1• Remarks on Definition 3.3

— It is easy to verify that

$$\begin{aligned} & \text{MSE} \\ = \; & E\big\{Y - E(Y) + E(Y) - \theta\big\}^2 \\ = \; & E\{Y - E(Y)\}^2 + \{E(Y) - \theta\}^2 + E\big[2\{Y - E(Y)\}\underbrace{\{E(Y) - \theta\}}_{\text{constant}}\big] \\ = \; & \text{Var}\{\varphi(\mathbf{x})\} + b^2(\theta). \end{aligned}$$

— Clearly, if an estimator $\varphi(\mathbf{x})$ is unbiased, then

$$\text{MSE} = \text{Var}\{\varphi(\mathbf{x})\}.$$

— Smaller MSE means greater precision.

3.4.2 Efficiency

18• Why need we the notion of efficiency?

- It is possible that there are several unbiased estimators for the same unknown parameter of interest.
- For instance, in Example 3.14, both $\frac{n+1}{n}X_{(n)}$ and $2\bar{X}$ are unbiased estimators of θ.
- Which one should we choose?
- Answer: The unbiased estimator with the *smaller* variance is the desired.
- Comparing two variances is equivalent to comparing two efficiencies.

18.1• Efficiency of an estimator

— Efficiency of an estimator $\hat{\theta}$ is proportional to the reciprocal of its variance:

$$\text{Eff}_{\hat{\theta}}(\theta) \propto \frac{1}{\text{Var}(\hat{\theta})}.$$

18.2• Relative efficiency of two estimators

Definition 3.4 (Relative efficiency). Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be two unbiased estimators for a parameter θ. If

$$\text{Var}(\hat{\theta}_1) < \text{Var}(\hat{\theta}_2), \tag{3.15}$$

we say that $\hat{\theta}_1$ is *more efficient* than $\hat{\theta}_2$. The *relative efficiency* of $\hat{\theta}_1$ to $\hat{\theta}_2$ is defined by the ratio

$$\frac{\text{Eff}_{\hat{\theta}_1}(\theta)}{\text{Eff}_{\hat{\theta}_2}(\theta)} = \frac{\text{Var}(\hat{\theta}_2)}{\text{Var}(\hat{\theta}_1)}. \qquad \|$$

Example 3.15 (Example 3.14 revisited). In Example 3.14, we have shown that $\hat{\theta}_1 = \frac{n+1}{n} X_{(n)}$ and $\hat{\theta}_2 = 2\bar{X}$ are two unbiased estimators of θ. Which estimator is more efficient?

Solution. From Appendix A.2.1, since $X_1 \sim U(0, \theta)$, we have $\text{Var}(X_1) = \theta^2/12$. Hence,

$$\text{Var}(\hat{\theta}_2) = \text{Var}(2\bar{X}) = \frac{4}{n^2} \cdot n\text{Var}(X_1) = \frac{\theta^2}{3n}.$$

On the other hand, similar to (3.14), we have

$$E(X_{(n)}^2) = \int_0^\theta x^2 \cdot \frac{nx^{n-1}}{\theta^n} \, \mathrm{d}x = \frac{n}{n+2}\theta^2. \tag{3.16}$$

Thus, based on (3.14) and (3.16), we obtain

$$\begin{aligned} \text{Var}(\hat{\theta}_1) &= \frac{(n+1)^2}{n^2}\text{Var}\{X_{(n)}\} \\ &= \frac{(n+1)^2}{n^2}\left[E(X_{(n)}^2) - \{E(X_{(n)})\}^2\right] \\ &= \frac{(n+1)^2}{n^2}\left\{\frac{n\theta^2}{n+2} - \frac{n^2}{(n+1)^2}\theta^2\right\} = \frac{\theta^2}{n(n+2)}. \end{aligned}$$

When $n > 1$, we have $\text{Var}(\hat{\theta}_1) < \text{Var}(\hat{\theta}_2)$, indicating that $\hat{\theta}_1$ has a smaller variance (and hence is more efficient) than $\hat{\theta}_2$. $\|$

19• Why need we the Cramér–Rao inequality

- Let $\mathcal{U} = \{\hat{\theta}\colon\ E(\hat{\theta}) = \theta\}$ denote the family of unbiased estimators of θ. The goal is to find the $\hat{\theta}^* \in \mathcal{U}$ with the smallest variance.
- Let $m = \#\mathcal{U}$ denote the number of elements in $\mathcal{U}$. If m is finite, we write $\mathcal{U} = \{\hat{\theta}_1, \ldots, \hat{\theta}_m\}$. Hence, we can choose the $\hat{\theta}_{k_0}$ such that

$$\mathrm{Var}(\hat{\theta}_{k_0}) \leqslant \mathrm{Var}(\hat{\theta}_j), \quad j \neq k_0, \quad j = 1, \ldots, m.$$

- If m is infinite, how to find the $\hat{\theta}^*$ with the smallest variance?

19.1• A motivation

— If we could find a constant c_0 satisfying

$$\mathrm{Var}(\hat{\theta}) \geqslant c_0, \quad \forall\, \hat{\theta} \in \mathcal{U},$$

then, this inequality can guide us to choose the $\hat{\theta}^*$ with variance being c_0.

— Thus, finding the $\hat{\theta}^*$ is equivalent to finding the lower bound c_0, which was found by Cramér and Rao.

— The c_0 is closely related to two new concepts: Score function and Fisher information.

19.2• Score function

— Let $X_1, \ldots, X_n$ be a random sample from the population r.v. X with density $f(x;\theta)$. Define $\mathbf{x} = (X_1, \ldots, X_n)^\top$ and $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ are their realizations. In the previous sections, we denote the likelihood function by

$$L(\theta) = L(\theta; x_1, \ldots, x_n) = L(\theta; \boldsymbol{x}) = \prod_{i=1}^{n} f(x_i;\theta).$$

— If we replace x_i in $L(\theta; x_1, \ldots, x_n)$ by X_i, then the resultant $L(\theta; \mathbf{x})$ is also a random variable and depends on the parameter θ.

— When the expectation and variance of a specific function of $L(\theta;\mathbf{x})$ are calculated, we also denote the likelihood function by

$$L(\theta) = L(\theta; X_1, \ldots, X_n) = L(\theta;\mathbf{x}) = \prod_{i=1}^{n} f(X_i;\theta)$$

to emphasize its dependence on $\mathbf{x}$.

— Let $\ell(\theta) = \log\{L(\theta)\}$ denote the log-likelihood function of θ, we call

$$S(\theta) = S(\theta;\mathbf{x}) \stackrel{\text{def}}{=} \frac{\mathrm{d}\ell(\theta)}{\mathrm{d}\theta} = \ell'(\theta) = \frac{L'(\theta)}{L(\theta)} \tag{3.17}$$

the *score function*.

19.3• Understanding the score function

— $S(\theta)$ is a function of θ.

— $S(\theta) = S(\theta;\mathbf{x})$ is also a function of $\mathbf{x}$ so that

$$E\{S(\theta)\} = E_{\mathbf{x}}\{S(\theta;\mathbf{x})\} = \int S(\theta;\boldsymbol{x}) \prod_{i=1}^{n} f(x_i;\theta)\,\mathrm{d}x_1 \cdots \mathrm{d}x_n.$$

— $S(\theta)$ is not a statistic because it depends on the unknown parameter θ.

19.4• Fisher information

— We call

$$I_n(\theta) = \mathrm{Var}\{S(\theta)\} = \mathrm{Var}_{\mathbf{x}}\{S(\theta;\mathbf{x})\} \tag{3.18}$$

the *Fisher information*, which is a way of measuring the amount of information that $\mathbf{x}$ carries about the unknown parameter θ.

— In many statistical problems, we have $E\{S(\theta)\} = 0$ so that (3.18) becomes

$$I_n(\theta) = E\{S^2(\theta;\mathbf{x})\} = E\left\{\left(\frac{\mathrm{d}\log L(\theta;\mathbf{x})}{\mathrm{d}\theta}\right)^2\right\}. \tag{3.19}$$

— However, it is possible in practice that $E\{S(\theta)\} \neq 0$ as shown in the following example.

Example 3.16 (Example 3.14 revisited). Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} U(0, \theta)$, where $\theta > 0$. Find the score function $S(\theta)$, $E\{S(\theta)\}$ and the Fisher information.

Solution. The population density is $f(x; \theta) = 1/\theta$, $x \in (0, \theta)$ depending on θ. We can rewrite $f(x; \theta) = (1/\theta) I_{(0,\theta)}(x)$ so that the likelihood function is

$$L(\theta) = \prod_{i=1}^{n} f(X_i; \theta) = \theta^{-n} \prod_{i=1}^{n} \xi_i,$$

where $\xi_i \triangleq I_{(0,\theta)}(X_i) \sim \text{Bernoulli}(p_i)$ with $p_i = \Pr(0 < X_i < \theta)$. From (3.17), we have

$$S(\theta; \mathbf{x}) = \frac{L'(\theta)}{L(\theta)} = \frac{-n\theta^{-n-1}}{\theta^{-n}} \prod_{i=1}^{n} \xi_i = -\frac{n}{\theta} \prod_{i=1}^{n} \xi_i.$$

Thus,

$$\begin{aligned} E\{S(\theta; \mathbf{x})\} &= -\frac{n}{\theta} \prod_{i=1}^{n} E(\xi_i) = -\frac{n}{\theta} \prod_{i=1}^{n} \Pr(0 < X_i < \theta) \\ &= -\frac{n}{\theta} \left(\int_0^{\theta} \frac{1}{\theta} \, \mathrm{d}x \right)^n = -\frac{n}{\theta} \neq 0. \end{aligned}$$

Similarly, we have $E\{S^2(\theta; \mathbf{x})\} = n^2/\theta^2$ and $I_n(\theta) = \text{Var}\{S(\theta; \mathbf{x})\} = 0$. ‖

19.5• A basic result on Bernoulli r.v. used in Example 3.16

— Let $\xi \sim \text{Bernoulli}(p)$, then $\xi \overset{\text{d}}{=} \xi^r$ for any positive integer r.

— Clearly, we have $E(\xi) = E(\xi^r)$.

20• The Cramér–Rao inequality

Theorem 3.3 (The general CR inequality). Let $\tau(\theta)$ be an arbitrary function of the unknown parameter θ. If (i) $\hat{\theta} = T(\mathbf{x})$ is an unbiased estimator of $\tau(\theta)$, and (ii) the support of the population density $f(x; \theta)$ does not depend on the parameter θ, then

$$\text{Var}(\hat{\theta}) \geqslant \frac{\{\tau'(\theta)\}^2}{I_n(\theta)}, \tag{3.20}$$

where $I_n(\theta)$ is the Fisher information. ‖

Proof. On the one hand,

$$1 = \int \prod_{i=1}^{n} f(x_i;\theta)\,\mathrm{d}x_1 \cdots \mathrm{d}x_n.$$

By differentiating both sides of this identity with respect to θ, we have

$$0 = \frac{\mathrm{d}}{\mathrm{d}\theta} \int \cdots \int L(\theta)\,\mathrm{d}x_1 \cdots \mathrm{d}x_n.$$

Since the supports of x_i's do not depend on the parameter θ, we can interchange differentiation and integration with respect to θ, yielding

$$\begin{aligned} 0 &= \int L'(\theta)\,\mathrm{d}x_1 \cdots \mathrm{d}x_n \\ &\overset{(3.17)}{=} \int S(\theta) \prod_{i=1}^{n} f(x_i;\theta)\,\mathrm{d}x_1 \cdots \mathrm{d}x_n \\ &= E\{S(\theta)\}. \end{aligned} \tag{3.21}$$

On the other hand, $\hat{\theta}$ is unbiased, then

$$\begin{aligned} \tau(\theta) = E(\hat{\theta}) &= \int T(\boldsymbol{x}) \prod_{i=1}^{n} f(x_i;\theta)\,\mathrm{d}x_1 \cdots \mathrm{d}x_n \\ &= \int T(\boldsymbol{x}) L(\theta)\,\mathrm{d}x_1 \cdots \mathrm{d}x_n. \end{aligned}$$

By differentiating both sides of the this equality with respect to θ, we obtain

$$\begin{aligned} \tau'(\theta) &= \int T(\boldsymbol{x}) L'(\theta)\,\mathrm{d}x_1 \cdots \mathrm{d}x_n \\ &= \int T(\boldsymbol{x}) \frac{L'(\theta)}{L(\theta)} \cdot L(\theta)\,\mathrm{d}x_1 \cdots \mathrm{d}x_n \\ &\overset{(3.17)}{=} \int T(\boldsymbol{x}) S(\theta) \cdot \prod_{i=1}^{n} f(x_i;\theta)\,\mathrm{d}x_1 \cdots \mathrm{d}x_n \\ &= E\{\hat{\theta} \times S(\theta)\} \\ &\overset{(3.21)}{=} \mathrm{Cov}\{\hat{\theta}, S(\theta)\}. \end{aligned}$$

By the Cauchy–Schwarz inequality,

$$\{\tau'(\theta)\}^2 = [\mathrm{Cov}\{\hat{\theta}, S(\theta)\}]^2 \leqslant \mathrm{Var}(\hat{\theta}) \times \mathrm{Var}\{S(\theta)\} = \mathrm{Var}(\hat{\theta}) \times I_n(\theta),$$

which indicates (3.20). □

20.1• Comments on Theorem 3.3

— The result in (3.20) is not valid if the support of $f(x;\theta)$ depends on θ, see Example 3.16.

— The Cauchy–Schwarz inequality states that $\{E(XY)\}^2 \leqslant E(X^2)E(Y^2)$ or equivalently $\{\mathrm{Cov}(X,Y)\}^2 \leqslant \mathrm{Var}(X)\cdot\mathrm{Var}(Y)$, see Theorem 1.5.

— The right hand side of (3.20) is called the *Cramér–Rao lower bound*.

— In particular, if $\tau(\theta)=\theta$, then (3.20) becomes

$$\mathrm{Var}(\hat{\theta}) \geqslant \frac{1}{I_n(\theta)}. \tag{3.22}$$

20.2• Two identities related to the Fisher information

— Theorem 3.4 below provides another way to calculate $I_n(\theta)$.

— That is, using (3.23) to calculate $I_n(\theta)$ is much easier than using (3.18).

Theorem 3.4 (Alternative expression). Let $I_n(\theta)$ denote the Fisher information. If $E\{S(\theta)\}=0$, then

$$I_n(\theta) = E\left\{-\frac{\mathrm{d}^2 \log L(\theta;\mathbf{x})}{\mathrm{d}\theta^2}\right\} = nI(\theta), \tag{3.23}$$

where

$$I(\theta) = E\left[\left\{\frac{\mathrm{d}\log f(X;\theta)}{\mathrm{d}\theta}\right\}^2\right] = E\left\{-\frac{\mathrm{d}^2\log f(X;\theta)}{\mathrm{d}\theta^2}\right\} \tag{3.24}$$

denotes the Fisher information for a single sample. ‖

Proof. From (3.21), we have

$$\begin{aligned}
0 &= \frac{\mathrm{d}}{\mathrm{d}\theta}\int S(\theta)L(\theta)\,\mathrm{d}x_1\cdots\mathrm{d}x_n \\
&= \int\left\{\frac{\mathrm{d}S(\theta)}{\mathrm{d}\theta}L(\theta) + S(\theta)L'(\theta)\right\}\mathrm{d}x_1\cdots\mathrm{d}x_n \\
&= E\left\{\frac{\mathrm{d}S(\theta)}{\mathrm{d}\theta}\right\} + \int S(\theta)S(\theta)L(\theta)\,\mathrm{d}x_1\cdots\mathrm{d}x_n \\
&= E\left\{\frac{\mathrm{d}^2\log L(\theta)}{\mathrm{d}\theta^2}\right\} + E\{S^2(\theta)\} \\
&= E\left\{\frac{\mathrm{d}^2\log L(\theta;\mathbf{x})}{\mathrm{d}\theta^2}\right\} + I_n(\theta).
\end{aligned}$$

Therefore, the first equation in (3.23) follows.

Since $L(\theta;\mathbf{x}) = \prod_{i=1}^{n} f(X_i;\theta)$, we have

$$\log L(\theta;\mathbf{x}) = \sum_{i=1}^{n} \log f(X_i;\theta),$$

and

$$\frac{\mathrm{d}^2 \log L(\theta;\mathbf{x})}{\mathrm{d}\theta^2} = \sum_{i=1}^{n} \frac{\mathrm{d}^2 \log f(X_i;\theta)}{\mathrm{d}\theta^2}.$$

Therefore,

$$\begin{aligned} I_n(\theta) &= E\left\{-\frac{\mathrm{d}^2 \log L(\theta;\mathbf{x})}{\mathrm{d}\theta^2}\right\} \\ &= \sum_{i=1}^{n} E\left\{-\frac{\mathrm{d}^2 \log f(X_i;\theta)}{\mathrm{d}\theta^2}\right\} \\ &= nE\left\{-\frac{\mathrm{d}^2 \log f(X;\theta)}{\mathrm{d}\theta^2}\right\} \\ &= nI(\theta). \end{aligned}$$

This means the second equation in (3.23). □

20.3• How to check the condition $E\{S(\theta)\} = 0$ in Theorem 3.4?

— If the support of the population density $f(x;\theta)$ does not depend on θ, $\Longrightarrow E\{S(\theta)\} = 0$.

— That is, that the support of $f(x;\theta)$ is free from θ is a sufficient condition for $E\{S(\theta)\} = 0$. We only need to check the support of $f(x;\theta)$.

20.4• How to understand (3.24)?

— In (3.17), we consider the case of $n = 1$ and we have

$$S(\theta) = \ell'(\theta) = \frac{\mathrm{d}\log f(X;\theta)}{\mathrm{d}\theta} = \frac{f'(X;\theta)}{f(X;\theta)},$$

where $f'(X;\theta)$ is the derivative of $f(X;\theta)$ with respect to θ.

— On the one hand, if $\ell'(\theta)$ is close to zero, then the r.v. X does not provide much information about θ.

— On the other hand, if $|\ell'(\theta)|$ or $\{\ell'(\theta)\}^2$ is large, then the r.v. X provides much information about θ.

— Thus, we can use $\{\ell'(\theta)\}^2$ to measure the amount of information provided by X.

— However, since X is a random variable, we should consider the average case. Thus, the Fisher information (for θ) contained in the r.v. X should be defined by

$$I(\theta) = E\{\ell'(\theta)\}^2,$$

which is (3.24).

21• UMVUE and efficient estimator

- The inequality (3.22) told us that for any $\hat{\theta} \in \mathcal{U} = \{\hat{\theta}\colon\ E(\hat{\theta}) = \theta\}$ with infinite elements, we have $\mathrm{Var}(\hat{\theta}) \geqslant 1/I_n(\theta)$, which can guide us to find a $\hat{\theta}^*$ such that

$$\hat{\theta}^* \in \mathcal{U} \quad \text{and} \quad \mathrm{Var}(\hat{\theta}^*) = \min_{\hat{\theta}\in\mathcal{U}} \mathrm{Var}(\hat{\theta}), \tag{3.25}$$

- This is a mathematical definition of a *uniformly minimum variance unbiased estimator* (UMVUE), which is to be shown in Definition 3.5.

- If $\hat{\theta}^*$ satisfies $\mathrm{Var}(\hat{\theta}^*) = 1/I_n(\theta)$, then $\hat{\theta}^*$ is called efficient estimator of θ. For the general case, see Definition 3.6.

Definition 3.5 (UMVUE). An estimator $\hat{\theta}^*$ is called a UMVUE of θ if it is unbiased and has the smallest variance among all unbiased estimators. ‖

Definition 3.6 (Efficient estimator). If an unbiased estimator $\hat{\theta} = T(\mathbf{x})$ for $\tau(\theta)$ has variance equal to the Cramér–Rao lower bound, then $\hat{\theta}$ is called an *efficient estimator* for $\tau(\theta)$. ‖

21.1• Efficient estimator versus UMVUE

— Obviously, an efficient estimator for $\tau(\theta)$ is a UMVUE for $\tau(\theta)$; i.e.,

$$\text{efficient estimator} \implies \text{UMVUE}.$$

— However, the converse is not always true; i.e.,

$$\text{efficient estimator} \not\Leftarrow \text{UMVUE}.$$

— In other words, it is possible that a UMVUE whose variance does not attain the CR lower bound. See Example 3.20 .

Example 3.17 (Bernoulli distribution). Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta)$. Then $\bar{X} = (1/n)\sum_{i=1}^n X_i$ is a UMVUE of θ.

Solution. Let $X \sim \text{Bernoulli}(\theta)$, then the pmf of X is $f(x;\theta) = \theta^x(1-\theta)^{1-x}$, $x = 0, 1$. Then, from (3.24), we have

$$I(\theta) = E\left\{\frac{\mathrm{d}\log f(X;\theta)}{\mathrm{d}\theta}\right\}^2 = E\left(\frac{X}{\theta} - \frac{1-X}{1-\theta}\right)^2$$

$$= E\left\{\frac{X-\theta}{\theta(1-\theta)}\right\}^2 = \frac{\text{Var}(X)}{\theta^2(1-\theta)^2} = \frac{1}{\theta(1-\theta)}$$

and

$$I_n(\theta) = nI(\theta) = \frac{n}{\theta(1-\theta)}.$$

Now, $\bar{X}$ is unbiased and

$$\text{Var}(\bar{X}) = \frac{\text{Var}(X)}{n} = \frac{\theta(1-\theta)}{n} = \frac{1}{I_n(\theta)};$$

i.e., the variance attains the CR lower bound. Then $\bar{X}$ is a UMVUE of θ. ‖

Example 3.18 (Normal distribution with known variance). Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma_0^2)$ with known σ_0^2 and unknown μ. Show that $\bar{X} = (1/n)\sum_{i=1}^n X_i$ is a UMVUE for μ.

Solution. Let $X \sim N(\mu, \sigma_0^2)$, then the pdf of X is

$$f(x;\mu) = \frac{1}{\sqrt{2\pi}\,\sigma_0}\exp\left\{-\frac{(x-\mu)^2}{2\sigma_0^2}\right\}$$

and

$$\log f(x;\mu) = -\log(\sqrt{2\pi}\,\sigma_0) - \frac{(x-\mu)^2}{2\sigma_0^2}.$$

From (3.24), we have

$$I(\mu) = E\left\{\frac{\mathrm{d}\log f(X;\mu)}{\mathrm{d}\mu}\right\}^2 = E\left(\frac{X-\mu}{\sigma_0^2}\right)^2 = \frac{1}{\sigma_0^2}.$$

and

$$I_n(\mu) = nI(\mu) = \frac{n}{\sigma_0^2}.$$

Now, $\bar{X}$ is unbiased and

$$\mathrm{Var}(\bar{X}) = \frac{\mathrm{Var}(X)}{n} = \frac{\sigma_0^2}{n} = \frac{1}{I_n(\mu)},$$

reaching the CR lower bound. Then $\bar{X}$ is a UMVUE of μ. ‖

21.2• Efficiency of an unbiased estimator

— In general, the *efficiency* of an unbiased estimator $\hat{\theta}$ for θ is defined by

$$\mathrm{eff}_{\hat{\theta}}(\theta) = \frac{\text{Cramér–Rao lower bound}}{\text{Actual variance}} = \frac{1/I_n(\theta)}{\mathrm{Var}(\hat{\theta})}. \tag{3.26}$$

Example 3.19 (Normal distribution with known mean). Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu_0, \theta)$ with known μ_0 and unknown θ. Calcualte the efficiency of $S^2 = \sum_{i=1}^n (X_i - \bar{X})^2/(n-1)$.

Solution. Let $X \sim N(\mu_0, \theta)$, then the pdf of X is

$$f(x;\theta) = \frac{1}{\sqrt{2\pi\theta}} \exp\left\{-\frac{(x-\mu_0)^2}{2\theta}\right\}$$

and

$$\log f(x;\theta) = -\frac{1}{2}\log(2\pi\theta) - \frac{(x-\mu_0)^2}{2\theta}.$$

From (3.24), we have

$$I(\theta) = E\left\{-\frac{\mathrm{d}^2\log f(X;\theta)}{\mathrm{d}\theta^2}\right\} = E\left\{-\frac{1}{2\theta^2} + \frac{(X-\mu_0)^2}{\theta^3}\right\} = \frac{1}{2\theta^2}.$$

and

$$I_n(\theta) = nI(\theta) = \frac{n}{2\theta^2}.$$

Since $(n-1)S^2/\theta \sim \chi^2(n-1)$, we have $E(S^2) = \theta$ and

$$\mathrm{Var}(S^2) = \frac{2\theta^2}{n-1} > \frac{2\theta^2}{n} = \frac{1}{I_n(\theta)}.$$

Therefore, S^2 is unbiased and its efficiency is

$$\mathrm{eff}_{S^2}(\theta) = \frac{1/I_n(\theta)}{\mathrm{Var}(S^2)} = \frac{n-1}{n} \to 1 \quad \text{as } n \to \infty;$$

i.e., S^2 is asymptotically efficient. ‖

Example 3.20 (Poisson distribution). Let $X \sim \mathrm{Poisson}(\theta)$ and $\tau(\theta) = \mathrm{e}^{-\theta}$. Define

$$\hat{\theta} = g(X) = \begin{cases} 1, & \text{if } X = 0, \\ 0, & \text{otherwise.} \end{cases}$$

Use Theorem 3.7 to show that $\hat{\theta}$ is the unique UMVUE of $\tau(\theta)$, but $\mathrm{Var}(\hat{\theta})$ is larger than the CR lower bound.

Solution. In Example 3.21 let $n = 1$, we know that $T(X) = X$ is sufficient for θ. Next, we need to prove that $T(X) = X$ is also complete. If

$$E\{h(X)\} = \sum_{x=0}^{\infty} h(x) \frac{\theta^x}{x!} \mathrm{e}^{-\theta} = 0,$$

for $\theta > 0$, we have

$$\sum_{x=0}^{\infty} h(x) \frac{\theta^x}{x!} = 0.$$

Since $\theta^x/x! > 0$ for any $\theta > 0$ and $x \geqslant 0$, we obtain $h(X) \equiv 0$. Then $T = X$ is also complete. Since $\hat{\theta} = g(X) = g(T)$ is unbiased for $\tau(\theta)$, it is the unique UMVUE for $\tau(\theta)$ according to Theorem 3.7.

Since $I(\theta) = 1/\theta$, and the CR lower bound is

$$\frac{\{\tau'(\theta)\}^2}{I(\theta)} = \theta \, \mathrm{e}^{-2\theta},$$

we have

$$\mathrm{Var}(\hat{\theta}) = \mathrm{e}^{-\theta}(1 - \mathrm{e}^{-\theta}) > \theta \, \mathrm{e}^{-2\theta},$$

which completes the proof. ‖

21.3• Is the UMVUE unique?

— If $\mathcal{U} = \{\hat{\theta}\colon\ E(\hat{\theta}) = \theta\}$ is not an empty set, then there exists *at most* one UMVUE of θ. In other words, the number of UMVUEs is zero or one.

21.4• How to find the unique UMVUE?

— In this subsection, we provide a sufficient condition; i.e.,

$$\text{if } \hat{\theta} \text{ is an efficient estimator } \Longrightarrow \hat{\theta} \text{ is the unique UMVUE.}$$

— §3.4.4 will provide a sufficient and necessary condition, which involves two important notions: Sufficiency (§3.4.3) and completeness (§3.4.3).

3.4.3 Sufficiency

22• Motivation from data reduction

- In many of the estimation problems, we need to summarize the information contained in the sample $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$.
- That is, we need to find some function of the sample that tells us just as much about θ as the sample itself.
- Such a function would be sufficient for estimation purposes and accordingly is called a *sufficient statistic*.

22.1• Raw data and reduced data

— Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x; \mu, \sigma^2)$, then $X_1, \ldots, X_n$ are called *raw data*.

— The quantities such as the sample mean $\bar{X}$, the sample variance S^2, the smallest order statistic $X_{(1)}$ and the largest order statistic $X_{(n)}$ are called *reduced data*.

— Given raw data, any reduce data can be determined uniquely; while the converse may not be true:

$$\text{raw data } \overrightarrow{\nleftarrow} \text{ reduced data.}$$

22.2• Intuitive interpretation on a sufficient statistic

— To estimate the population mean μ, we only need to use the reduced datum $\bar{X}$, which contains all information about μ.

— In other words, using $\bar{X}$ to estimate μ will not lose any information.

— However, using $\sum_{i=1}^{n-1} X_i/(n-1)$ to estimate μ will lose information from X_n.

— Hence, $\bar{X}$ is a sufficient estimator of μ while $\sum_{i=1}^{n-1} X_i/(n-1)$ is not a sufficient estimator of μ.

23• Single sufficient statistic

Definition 3.7 (Sufficient statistic). A statistic $T(\mathbf{x})$ is said to be a *sufficient statistic* of θ if the conditional distribution of $\mathbf{x}$, given $T(\mathbf{x}) = t$, does not depend on θ for any value of t. In discrete case, this means that

$$\Pr\{X_1 = x_1, \ldots, X_n = x_n; \theta | T(\mathbf{x}) = t\} = h(\boldsymbol{x})$$

does not depend on θ. ‖

23.1• Deeply understanding Definition 3.7

— The definition says that if you know the value of the sufficient statistic, then the sample values themselves are not needed and can tell you nothing more about θ.

— This is true since the distribution of the sample given the sufficient statistic does not depend on θ.

— The joint distribution of $\mathbf{x}$ and $T(\mathbf{x})$ is

$$\begin{aligned} & \Pr\{X_1 = x_1, \ldots, X_n = x_n, T(\mathbf{x}) = t; \theta\} \\ = \; & \Pr\{X_1 = x_1, \ldots, X_n = x_n; \theta | T(\mathbf{x}) = t\} \times \Pr\{T(\mathbf{x}) = t; \theta\} \\ = \; & h(\boldsymbol{x}) \times \Pr\{T(\mathbf{x}) = t; \theta\}, \end{aligned}$$

where the left-hand side is, in general, the joint distribution of $\mathbf{x}$ subject to the constraint $T(\mathbf{x}) = t$.

— Thus, the MLE $\hat{\theta}$ can be obtained by maximizing $\log[\Pr\{T(\mathbf{x}) = t; \theta\}]$.

— Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta)$ and $T(\mathbf{x}) = \sum_{i=1}^n X_i$. We have

$$\begin{aligned} & \Pr\{X_1 = x_1, \ldots, X_n = x_n, T(\mathbf{x}) = t; \theta\} \\ = \; & \Pr\{X_1 = x_1, \ldots, X_{n-1} = x_{n-1}, X_n = x_n = t - \textstyle\sum_{j=1}^{n-1} x_j; \theta\} \\ = \; & \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i} \\ = \; & \theta^t(1-\theta)^{n-t}. \end{aligned}$$

On the other hand, since $T(\mathbf{x}) = \sum_{i=1}^n X_i \sim \text{Binomial}(n, \theta)$, we obtain

$$\Pr\{T(\mathbf{x}) = t; \theta\} = \binom{n}{t}\theta^t(1-\theta)^{n-t}.$$

The MLE $\hat{\theta} = T(\mathbf{x})/n = \bar{X}$.

Example 3.21 (Poisson distribution). Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Poisson}(\theta)$, where $\theta > 0$. Show that $T(\mathbf{x}) = \sum_{i=1}^n X_i$ is a sufficient statistic of θ.

Solution. From Example 2.12, we have $T(\mathbf{x}) = \sum_{i=1}^n X_i \sim \text{Poisson}(n\theta)$. Since the conditional distribution

$$\begin{aligned} & \Pr\{X_1 = x_1, \ldots, X_n = x_n;\ \theta | T(\mathbf{x}) = t\} \\ = \; & \frac{\Pr(X_1 = x_1, \ldots, X_{n-1} = x_{n-1}, X_n = t - \sum_{i=1}^{n-1} x_i;\ \theta)}{\Pr(\sum_{i=1}^n X_i = t)} \\ = \; & \left(\prod_{i=1}^{n-1} \frac{\theta^{x_i}\,\mathrm{e}^{-\theta}}{x_i!}\right) \cdot \frac{\theta^{t-\sum_{i=1}^{n-1} x_i}\,\mathrm{e}^{-\theta}}{(t - \sum_{i=1}^{n-1} x_i)!} \Big/ \frac{(n\theta)^t\,\mathrm{e}^{-n\theta}}{t!} \\ = \; & \frac{t!}{x_1! \cdots x_{n-1}!(t - \sum_{i=1}^{n-1} x_i)!} \cdot \frac{1}{n^t} \end{aligned}$$

does not depend on θ for any value of t, $T(\mathbf{x})$ is a sufficient statistic of θ. ‖

Example 3.22 (Bernoulli distribution). Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta)$, where $\theta > 0$. Show that $T(\mathbf{x}) = \sum_{i=1}^n X_i$ is a sufficient statistic of θ.

Solution. Note that $T(\mathbf{x}) = \sum_{i=1}^n X_i \sim \text{Binomial}(n, \theta)$, we have

$$\begin{aligned}
& \Pr\{X_1 = x_1, \ldots, X_n = x_n;\ \theta | T(\mathbf{x}) = t\} \\
= & \frac{\Pr(X_1 = x_1, \ldots, X_n = x_n;\ \theta)}{\Pr\{T(\mathbf{x}) = t\}} \\
= & \frac{\prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} \\
= & \frac{\theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} \qquad [\because \textstyle\sum_{i=1}^n x_i = t] \\
= & \frac{\theta^t(1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} \\
= & \frac{1}{\binom{n}{t}},
\end{aligned}$$

which does not depend on θ for any value of t. Therefore, $T(\mathbf{x})$ is a sufficient statistic of θ. ‖

23.2• How to find a sufficient statistic?

Theorem 3.5 (Factorization theorem). A statistic $T(\mathbf{x})$ is a sufficient statistic of the unknown parameter θ iff the joint pdf (or pmf) can be written in the form

$$f(x_1, \ldots, x_n; \theta) = f(\boldsymbol{x}; \theta) = g(T(\boldsymbol{x}); \theta) \times h(\boldsymbol{x}), \tag{3.27}$$

where $h(\boldsymbol{x})$ does not depend on θ, $g(T; \theta)$ is a function of both T and θ, and it depends on $x_1, \ldots, x_n$ only through T. ‖

Proof. We give a proof for the discrete case.

" $\Longleftarrow$ " (Sufficiency). Assume that $\Pr(\mathbf{x} = \boldsymbol{x}; \theta) = g(T(\boldsymbol{x}); \theta) \times h(\boldsymbol{x})$. Note that

$$\begin{aligned}
\Pr\{T(\mathbf{x}) = t; \theta\} &= \sum_{T(\boldsymbol{x})=t} \Pr(\mathbf{x} = \boldsymbol{x}; \theta) \\
&= \sum_{T(\boldsymbol{x})=t} g(T(\boldsymbol{x}); \theta) \times h(\boldsymbol{x}) \\
&= g(t; \theta) \sum_{T(\boldsymbol{x})=t} h(\boldsymbol{x}),
\end{aligned}$$

we obtain

$$\begin{aligned}
\Pr\{\mathbf{x}=\boldsymbol{x};\theta|T(\mathbf{x})=t\} &= \begin{cases} 0, & \text{if } T(\boldsymbol{x})\neq t, \\ \dfrac{\Pr\{\mathbf{x}=\boldsymbol{x},T(\mathbf{x})=t;\theta\}}{\Pr\{T(\mathbf{x})=t;\theta\}}, & \text{if } T(\boldsymbol{x})=t, \end{cases} \\
&= \begin{cases} 0, & \text{if } T(\boldsymbol{x})\neq t, \\ \dfrac{\Pr(\mathbf{x}=\boldsymbol{x};\theta)}{\Pr\{T(\mathbf{x})=t;\theta\}}, & \text{if } T(\boldsymbol{x})=t, \end{cases} \\
&= \begin{cases} 0, & \text{if } T(\boldsymbol{x})\neq t, \\ \dfrac{h(\boldsymbol{x})}{\sum_{T(\boldsymbol{x})=t}h(\boldsymbol{x})}, & \text{if } T(\boldsymbol{x})=t. \end{cases}
\end{aligned}$$

It does not depend on θ, then $T(\mathbf{x})$ is sufficient for θ.

"$\Longrightarrow$" (Necessity). Assume that $T(\mathbf{x})$ is sufficient, then

$$\Pr(\mathbf{x}=\boldsymbol{x};\theta)=\Pr\{T(\mathbf{x})=t\}\times\Pr\{\mathbf{x}=\boldsymbol{x};\theta|T(\mathbf{x})=t\} \tag{3.28}$$

with $T(\boldsymbol{x})=t$. Let

$$\Pr\{T(\mathbf{x})=t\}=g(t;\theta) \quad \text{and} \quad \Pr\{\mathbf{x}=\boldsymbol{x};\theta|T(\mathbf{x})=t\}=h(\boldsymbol{x}),$$

then (3.28) becomes

$$\Pr(\mathbf{x}=\boldsymbol{x};\theta)=g(t;\theta)\times h(\boldsymbol{x})$$

and (3.27) follows. □

Example 3.23 (Normal distribution with known variance). Let $X_1,\ldots,X_n \overset{\text{iid}}{\sim} N(\theta,\sigma_0^2)$ with known σ_0^2. Then $\bar{X}$ is a sufficient statistic for θ.

Solution. The joint pdf of $X_1,\ldots,X_n$ is

$$\begin{aligned}
f(x_1,\ldots,x_n;\theta) &= \frac{1}{(\sqrt{2\pi}\,\sigma_0)^n}\exp\left\{-\frac{\sum_{i=1}^n(x_i-\theta)^2}{2\sigma_0^2}\right\} \\
&= \frac{1}{(\sqrt{2\pi}\,\sigma_0)^n}\exp\left\{-\frac{\sum_{i=1}^n(x_i-\bar{x}+\bar{x}-\theta)^2}{2\sigma_0^2}\right\} \\
&= \frac{1}{(\sqrt{2\pi}\,\sigma_0)^n}\exp\left\{-\frac{n(\bar{x}-\theta)^2}{2\sigma_0^2}\right\} \\
&\quad\times\exp\left\{-\frac{\sum_{i=1}^n(x_i-\bar{x})^2}{2\sigma_0^2}\right\}
\end{aligned}$$

Then $T=\bar{X}$ is sufficient for θ. ‖

Example 3.24 (Shift exponential distribution). Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x;\theta)$, where

$$f(x;\theta) = \begin{cases} \exp\{-(x-\theta)\}, & \text{if } x > \theta, \\ 0, & \text{elsewhere.} \end{cases}$$

Then $X_{(1)} = \min(X_1, \ldots, X_n)$ is a sufficient statistic for θ.

Solution. The joint pdf of $X_1, \ldots, X_n$ is

$$\begin{aligned} f(x_1, \ldots, x_n; \theta) &= \prod_{i=1}^{n} \exp\{-(x_i - \theta)\} \cdot I_{(\theta,\infty)}(x_i) \\ &= \mathrm{e}^{-\sum_{i=1}^{n} x_i + n\theta} \prod_{i=1}^{n} I_{(\theta,\infty)}(x_i) \\ &= \mathrm{e}^{n\theta} I_{(\theta,\infty)}(x_{(1)}) \times \mathrm{e}^{-\sum_{i=1}^{n} x_i}. \end{aligned}$$

Then $X_{(1)}$ is sufficient for θ. ‖

Example 3.25 (A special beta distribution). Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x;\theta)$, where

$$f(x;\theta) = \begin{cases} \theta x^{\theta-1}, & \text{if } 0 < x < 1, \\ 0, & \text{elsewhere,} \end{cases}$$

where $\theta > 0$. Then $\prod_{i=1}^{n} X_i$ is a sufficient statistic for θ.

Solution. The joint pdf of $X_1, \ldots, X_n$ is

$$f(x_1, \ldots, x_n; \theta) = \theta^n \left(\prod_{i=1}^{n} x_i \right)^{\theta-1} \times 1.$$

Then $\prod_{i=1}^{n} X_i$ is sufficient for θ. The function $h(\boldsymbol{x})$ in (3.27) may be a constant as shown in this example. ‖

23.3• Why do we need a sufficient statistic?

— Given a sufficient statistic T and an unbiased estimator Y of θ, we can immediately find another unbiased estimator Z with a smaller variance, see the beginning of §3.4.4.

— Usefulness in finding a unique UMVUE of θ: If a sufficient statistic $T(\mathbf{x})$ is also a complete statistic simultaneously, then we can immediately identify a unique UMVUE of θ, see Theorem 3.7.

23.4• Why does it take the name of sufficient statistic?

— For the normal population with known variance, we know that the sample mean $\bar{X}$ is an unbiased estimator of the population mean θ.

— From Example 3.23, we can see that $\bar{X}$ is also a sufficient statistic of θ.

— $\bar{X}$ contains *all/sufficient* information from the random sample $X_1, \ldots, X_n$ to estimate θ.

— $\sum_{i=1}^{n-1} X_i/(n-1)$ is also unbiased estimator of θ but it is not a sufficient statistic.

23.5• Is a sufficient statistic unique?

— First, we note that a sufficient statistic is not unique.

— If $Y_1 = T(\mathbf{x})$ is a sufficient statistic for θ and $Y_2 = g(Y_1)$, where $g(\cdot)$ is a *one-to-one* function, then Y_2 is also sufficient.

— For instance, in Example 3.23, $\sum_{i=1}^{n} X_i = n\bar{X}$ is another sufficient statistic of θ but $\bar{X}^2$ is not sufficient.

23.6• Sufficient statistic versus sufficient estimator

— An estimator is a meaningful statistic.

— In Example 3.23, $\bar{X}$ is a sufficient statistic of θ, and it is also a sufficient estimator of θ.

— Note that $\sum_{i=1}^{n} X_i$ is just a sufficient statistic of θ, not a sufficient estimator of θ.

23.7• Sufficient statistic versus unbiased estimator

— For the normal population with known variance, both $\bar{X}$ and $n\bar{X}$ are sufficient statistics for θ. The former is unbiased while the latter is biased.

— Both $\bar{X}$ and $\sum_{i=1}^{n-1} X_i/(n-1)$ are unbiased estimators for θ. The former is sufficient while the latter is not sufficient.

23.8• Statistic versus estimator

— An estimator $\Longrightarrow$ a statistic.

— Based on different criteria, estimators could be classified into:

$$\begin{cases} \text{biased estimator, unbiased estimator;} \\ \text{MLE, moment estimator, Bayesian estimator;} \\ \text{efficient estimator, UMVUE;} \\ \text{sufficient estimator, complete estimator.} \end{cases}$$

— For example,

$$\frac{1}{a}\sum_{i=1}^{n} X_i: \quad \text{statistic (for any non-zero constant } a),$$

$$\frac{1}{n}\sum_{i=1}^{n} X_i: \quad \begin{cases} \text{MLE,} \\ \text{moment estimator,} \\ \text{unbiased estimator,} \\ \text{UMVUE,} \\ \text{sufficient estimator.} \end{cases}$$

24• Joint sufficient statistics

- For some problems, no single sufficient statistic exists.
- However, there will always exist joint sufficient statistics.

Definition 3.8 (Joint sufficient statistics). Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x;\boldsymbol{\theta})$. The statistics $T_1(\mathbf{x}), \ldots T_r(\mathbf{x})$ are said to be *jointly sufficient* if the conditional distribution of $\mathbf{x}$, given $T_1 = t_1, \ldots, T_r = t_r$, does not depend on $\boldsymbol{\theta}$. ‖

Theorem 3.6 (Factorization theorem with joint sufficient statistics). A set of statistics $T_1(\mathbf{x}), \ldots, T_r(\mathbf{x})$ is jointly sufficient for the parameter vector $\boldsymbol{\theta}$ iff the joint pdf (or pmf) can be written in the form

$$f(x_1, \ldots, x_n; \boldsymbol{\theta}) = f(\boldsymbol{x}; \boldsymbol{\theta}) = g(T_1(\boldsymbol{x}), \ldots, T_r(\boldsymbol{x}); \boldsymbol{\theta}) \times h(\boldsymbol{x}), \tag{3.29}$$

where $h(\boldsymbol{x})$ does not depend on $\boldsymbol{\theta}$, $g(T_1, \ldots, T_r; \boldsymbol{\theta})$ depends on $x_1, \ldots, x_n$ only through $T_1, \ldots, T_r$. ‖

24.1• Comments on Theorem 3.6

— If $T_1(\mathbf{x}), \ldots, T_r(\mathbf{x})$ is a set of jointly sufficient statistics, then any set of one-to-one functions/transformations of $T_1(\mathbf{x}), \ldots, T_r(\mathbf{x})$ is also jointly sufficient.

— In addition, the function $h(\boldsymbol{x})$ in (3.29) may be a constant.

Example 3.26 (Normal distribution). Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$. Find jointly sufficient statistics for $\boldsymbol{\theta} = (\mu, \sigma^2)$.

Solution. The joint pdf of $X_1, \ldots, X_n$ is

$$\begin{aligned} f(\boldsymbol{x}; \boldsymbol{\theta}) &= \frac{1}{(\sqrt{2\pi\sigma^2})^n} \exp\left\{ -\frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2} \right\} \\ &= \frac{1}{(\sqrt{2\pi\sigma^2})^n} \exp\left(-\frac{\sum x_i^2 - 2\mu \sum x_i + n\mu^2}{2\sigma^2} \right) \\ &= (\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{\sum x_i^2 - 2\mu \sum x_i + n\mu^2}{2\sigma^2} \right) \times \frac{1}{(\sqrt{2\pi})^n}. \end{aligned}$$

Hence, $\sum X_i$ and $\sum X_i^2$ are jointly sufficient. It can be shown that $\bar{X}$ and $S^2 = \{1/(n-1)\} \sum (X_i - \bar{X})^2$ are one-to-one functions of $\sum X_i$ and $\sum X_i^2$; so $\bar{X}$ and S^2 are also jointly sufficient. ‖

3.4.4 Completeness

25• Why need we the notion of the complete statistic?

- Assume that $T(\mathbf{x})$ is *sufficient* for θ, and $Y(\mathbf{x})$ is an *unbiased estimator* of $\tau(\theta)$, a function of θ.
- Let $Z \hat{=} E(Y|T) = \varphi(T)$, then we have

$$\begin{aligned} E(Z) &\overset{(1.27)}{=} E(Y) = \tau(\theta), \quad \text{and} \qquad (3.30) \\ \text{Var}(Z) &= 0 + \text{Var}(Z) \\ &\leqslant E\Big\{ \underbrace{\text{Var}(Y|T)}_{\text{non-negative}} \Big\} + \text{Var}(Z) \\ &= E\{\text{Var}(Y|T)\} + \text{Var}\{E(Y|T)\} \overset{(1.28)}{=} \text{Var}(Y). \end{aligned}$$

- Thus, from a sufficient statistic T and an unbiased estimator Y, we can find a set of unbiased estimators $\{Z_j\}_{j=1}^m$ satisfying

$$\begin{aligned} Z_1 &= E(Y|T), \\ Z_2 &= E(Z_1|T), \\ Z_3 &= E(Z_2|T), \\ &\vdots \\ Z_m &= E(Z_{m-1}|T), \end{aligned}$$

  and $\mathrm{Var}(Z_m) \leqslant \mathrm{Var}(Z_{m-1}) \leqslant \cdots \leqslant \mathrm{Var}(Z_1) \leqslant \mathrm{Var}(Y)$.

- Let $\mathcal{U} = \{Z\colon\ E(Z) = \tau(\theta)\}$ and $\#\mathcal{U}$ is infinite.

- We wonder if we could find a $Z^* \in \mathcal{U}$ such that

$$\mathrm{Var}(Z^*) \leqslant \mathrm{Var}(Z), \quad \forall\, Z \in \mathcal{U}.$$

  In other words, Z^* is the UMVUE of $\tau(\theta)$.

- We wish that $Z^* = Z_m = \cdots = Z_1$. The notion of "complete statistic" facilitates this purpose.

26• Definition of a complete statistic

Definition 3.9 (Completeness). Let $X_1, \ldots, X_n$ denote a random sample from the pdf (or pmf) $f(x;\theta)$ with parameter space Θ and let $T(\mathbf{x})$ be a statistic, where $\mathbf{x} = (X_1, \ldots, X_n)^\top$. The statistic T is said to be *complete* if

$$E\{h(T)\} = 0 \quad \text{for all } \theta \in \Theta$$

implies that $h(T) = 0$ with probability 1; i.e.,

$$\Pr\{h(T) = 0\} = 1 \quad \text{for all } \theta \in \Theta,$$

where the function $h(T)$ is a statistic. ‖

26.1• Alternative statement

— Alternatively, we can say: T is complete iff the *only* unbiased estimator of 0 that is a function of T is the statistic that is identically 0 with probability 1.

27• How to understand the completeness?

- We need two "bridges" to reach the *uniqueness* of UMVUE.

27.1• Case I: $\mathcal{U} = \{\hat{\theta}\text{: } E(\hat{\theta}) = \theta\}$ with finite elements

— Let $\hat{\theta}_i \in \mathcal{U}$ for $i = 1, 2$.

— The first bridge is the *sufficiency*; i.e., suppose that we have found a sufficient statistic T for θ.

— From (3.30), we know that $Z_i = E(\hat{\theta}_i|T) = h_i(T)$, $i = 1, 2$, are two unbiased estimators of θ so that $E(Z_1 - Z_2) = \theta - \theta = 0$,

$$\text{Var}(Z_1) \leqslant \text{Var}(\hat{\theta}_1) \quad \text{and} \quad \text{Var}(Z_2) \leqslant \text{Var}(\hat{\theta}_2).$$

— Which one should we choose? Z_1 or Z_2?

— Of course, we choose Z_1 if $\text{Var}(Z_1) \leqslant \text{Var}(Z_2)$. Otherwise, we choose Z_2.

— In other words, we do not need the "second bridge" (i.e., the completeness) for Case I.

27.2• Case II: $\mathcal{U} = \{\hat{\theta}\text{: } E(\hat{\theta}) = \theta\}$ with infinite elements

— Let $\hat{\theta}_i \in \mathcal{U}$ for $i = 1, 2, \ldots$.

— Define $Z_i = E(\hat{\theta}_i|T) = h_i(T)$, $i = 1, 2, \ldots$, we have $E(Z_i - Z_j) = E\{h_i(T) - h_j(T)\} = \theta - \theta = 0$, and

$$\text{Var}(Z_i) \leqslant \text{Var}(\hat{\theta}_i), \quad i = 1, 2, \ldots.$$

— Which Z_i should we choose?

— Then, we wish to find a second "bridge" such that $Z_i = Z_j$ with probability 1. The second bridge is the completeness.

Example 3.27 (Uniform distribution). Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} U(0, \theta)$, where $\Theta = \{\theta\colon\ \theta > 0\}$. Show that $X_{(n)} = \max(X_1, \ldots, X_n)$ is complete.

Solution. We must show that if $E\{h(X_{(n)})\} = 0$ for all $\theta > 0$, then $\Pr\{h(X_{(n)}) = 0\} = 1$ for all $\theta > 0$. From Example 2.16, the density of $X_{(n)}$ is

$$f_n(x) = nx^{n-1}/\theta^n, \quad 0 < x < \theta.$$

Note that

$$E\{h(X_{(n)})\} = \int h(x) f_n(x)\,\mathrm{d}x = \int_0^\theta h(x)\theta^{-n} n x^{n-1}\,\mathrm{d}x,$$

and $E\{h(X_{(n)})\} = 0$ for all $\theta > 0$ when and only when

$$\int_0^\theta h(x) x^{n-1}\,\mathrm{d}x = 0 \quad \text{for all } \theta > 0.$$

Differentiating both sides of this identity with respect to θ produces

$$h(\theta)\theta^{n-1} = 0,$$

which in turn implies $h(\theta) = 0$ for all $\theta > 0$. ‖

28• How to find the unique UMVUE?

Theorem 3.7 (Lehmann–Scheffé theorem). Let $T(\mathbf{x})$ is a complete sufficient statistic for θ. If $g(T)$ is an unbiased estimator of $\tau(\theta)$, then $g(T)$ is the unique UMVUE for $\tau(\theta)$. ‖

Proof. Let Y be any unbiased estimator of $\tau(\theta)$ and let $\varphi(T) = E(Y|T)$, then

$$E\{\varphi(T)\} = \tau(\theta) \quad \text{and} \quad \mathrm{Var}\{\varphi(T)\} \leqslant \mathrm{Var}(Y).$$

Therefore,

$$E\{g(T) - \varphi(T)\} = \tau(\theta) - \tau(\theta) = 0 \quad \text{for all } \theta.$$

As T is complete, this implies that $g(T) = \varphi(T)$ with probability 1 and

$$\mathrm{Var}\{g(T)\} = \mathrm{Var}\{\varphi(T)\} \leqslant \mathrm{Var}(Y).$$

Consequently, $g(T)$ is the unique function of T which is unbiased and has a smaller variance than any other unbiased estimator has. Then $g(T)$ is the unique UMVUE of $\tau(\theta)$. □

Example 3.28 (Bernoulli distribution). Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta)$, where $\Theta = \{\theta\colon 0 < \theta < 1\}$. Show that the statistic $T = \sum_{i=1}^n X_i$ is a complete sufficient statistic for θ. Find the UMVUE for θ.

Solution. The joint pdf of $X_1, \ldots, X_n$ is

$$f(x_1, \ldots, x_n; \theta) = \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i} = \theta^t(1-\theta)^{n-t},$$

where $t = \sum_{i=1}^n x_i$. By Theorem 3.5, $T = \sum_{i=1}^n X_i$ is sufficient, and $T \sim \text{Binomial}(n, \theta)$. Now assume that a function $h(T)$ satisfies

$$E\{h(T)\} = \sum_{t=0}^n h(t)\Pr(T=t) = \sum_{t=0}^n h(t)\binom{n}{t}\theta^t(1-\theta)^{n-t} = 0, \qquad (3.31)$$

for $0 < \theta < 1$. Let $y = \theta/(1-\theta)$, then (3.31) becomes

$$\sum_{t=0}^n h(t)\binom{n}{t}y^t = 0, \quad y > 0.$$

A polynomial is identical to zero, then all coefficients are zero. Thus

$$h(t)\binom{n}{t} = 0 \quad \text{for } t = 0, 1, \ldots, n.$$

Hence $h(T) \equiv 0$. Then T is also complete. Since $\bar{X} = T/n$ is unbiased for θ, it is the unique UMVUE for θ according to Theorem 3.7. ‖

28.1• Remarks on Example 3.28

— Note that $T = \sum_{i=1}^n X_i$ is sufficient for θ and $Y = T/n$ is an unbiased estimator of θ.

— We have

$$Z_1 = E(Y|T) = E\left(\frac{T}{n}\middle|T\right) = \frac{T}{n} = Y \quad \text{and}$$

$$Z_2 = E(Z_1|T) = \frac{T}{n} = Y$$

so that $\mathcal{U} = \{\hat{\theta}\colon E(\hat{\theta}) = \theta\} = \{Y\}$ and $\#\mathcal{U} = 1$.

28.2• Remarks on Example 3.23

— From Example 3.23, we know that $T = \bar{X} = \sum_{i=1}^{n} X_i/n$ is sufficient for θ and $Y = T$ is an unbiased estimator of θ.

— We have $Z = E(Y|T) = E(T|T) = T$ so that $\mathcal{U} = \{\hat{\theta}\colon\ E(\hat{\theta}) = \theta\} = \{\bar{X}\}$ and $\#\mathcal{U} = 1$.

3.5 Limiting Properties of MLE

29• MLE weakly converges in probability to its true value

- In §3.1.4, we have stated the invariance property of MLE. In this section, we introduce limiting properties of MLE.
- We rewrite Definition 2.3 as follows: A sequence of r.v.'s $\{X_n\}_{n=1}^{\infty}$ is said to weakly converge in probability to an r.v. X, denoted by $X_n \xrightarrow{\text{P}} X$, if for any $\varepsilon > 0$, $\lim_{n\to\infty} \Pr(|X_n - X| \geqslant \varepsilon) = 0$.
- Let $\{X_n\}_{n=1}^{\infty}$ be i.i.d. from a population with pdf $f(x;\theta)$. Let $\hat{\theta}_n$ be the MLE of θ based on $X_1, \ldots, X_n$. Then under certain regularity conditions, we have

$$\hat{\theta}_n \xrightarrow{\text{P}} \theta \quad \text{as} \quad n \to \infty. \tag{3.32}$$

29.1• MLE also converges in distribution to its true value

— The conclusion in (3.32) states that when $n \to \infty$, the MLE $\hat{\theta}_n$ weakly converges in probability to the true value of the parameter.

— From Property 2.1 in §2.5.3, we obtain $\hat{\theta}_n \xrightarrow{\text{L}} \theta$; i.e., the MLE $\hat{\theta}_n$ converges in distribution to the true value of the parameter.

30• MLE is asymptotically normally distributed

- Let $\{X_n\}_{n=1}^{\infty} \overset{\text{iid}}{\sim} f(x;\theta)$ and $\hat{\theta}_n$ be the MLE of θ based on $X_1, \ldots, X_n$.
- Let $S(\theta;\mathbf{x})$ with $\mathbf{x} = (X_1, \ldots, X_n)^{\top}$ and $I_n(\theta) = nI(\theta)$ denote the score function and the Fisher information, respectively.

- If $E\{S(\theta;\mathbf{x})\} = 0$ and $\mathrm{Var}\{S(\theta;\mathbf{x})\} = nI(\theta)$, then

$$\frac{S(\theta;\mathbf{x})}{\sqrt{nI(\theta)}} \xrightarrow{\mathrm{L}} N(0,1) \quad \text{as} \quad n \to \infty \tag{3.33}$$

and

$$\{nI(\theta)\}^{1/2}(\hat{\theta}_n - \theta) \xrightarrow{\mathrm{L}} N(0,1) \quad \text{as} \quad n \to \infty. \tag{3.34}$$

- The corresponding proofs of (3.33) and (3.34) are given in §3.6.

30.1• Remarks on (3.34)

— The MLE $\hat{\theta}_n$ is an asymptotically unbiased estimator of θ.

— The MLE $\hat{\theta}_n$ is an asymptotically UMVUE because it reaches the CR lower bound in the sense that

$$\lim_{n\to\infty} \mathrm{eff}_{\hat{\theta}_n}(\theta) = \lim_{n\to\infty} \frac{1/I_n(\theta)}{\mathrm{Var}(\hat{\theta}_n)} = 1.$$

— The MLE $\hat{\theta}_n$ is asymptotically normally distributed.

Example 3.29 (A special beta distribution). Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x;\theta)$, where

$$f(x;\theta) = \begin{cases} \theta x^{\theta-1}, & \text{if } 0 < x < 1, \\ 0, & \text{elsewhere,} \end{cases}$$

and $\Theta = \{\theta\colon\ \theta > 0\}$. Find the MLE of θ and study its limiting properties.

Solution. The likelihood function is

$$L(\theta) = \prod_{i=1}^{n} \theta x_i^{\theta-1} = \theta^n \left(\prod_{i=1}^{n} x_i\right)^{\theta-1}.$$

The MLE of θ is $\hat{\theta}_n = -n/\sum_{i=1}^{n} \log X_i$. Since

$$I(\theta) = E\left\{-\frac{\mathrm{d}^2 \log f(X;\theta)}{\mathrm{d}\theta^2}\right\} = \frac{1}{\theta^2},$$

we have

$$(n/\theta^2)^{1/2}(\hat{\theta}_n - \theta) \xrightarrow{\mathrm{L}} N(0,1).$$

For large n, we have approximately

$$(n/\hat{\theta}_n^2)^{1/2}(\hat{\theta}_n - \theta) \dot{\sim} N(0,1). \qquad \|$$

3.6 Some Challenging Questions

Example 3.30 (Grouped Dirichlet distribution). Let $(x_1, \ldots, x_4, x_{12}, x_{34})$ be observed values of random variables $(X_1, \ldots, X_4, X_{12}, X_{34})$, respectively. Assume that the likelihood function of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_4)^\top$ is

$$L(\boldsymbol{\theta}) = \left(\prod_{i=1}^{4} \theta_i^{x_i} \right) \cdot (\theta_1 + \theta_2)^{x_{12}} (\theta_3 + \theta_4)^{x_{34}}, \quad \boldsymbol{\theta} \in \mathbb{T}_4.$$

Find the MLE of $\boldsymbol{\theta}$ subject to the constraints $\theta_i \geqslant 0$ and $\sum_{i=1}^{4} \theta_i = 1$.

Solution. The log-likelihood function is given by

$$\begin{aligned} \ell(\boldsymbol{\theta}) &= \sum_{i=1}^{4} x_i \log \theta_i + x_{12} \log(\theta_1 + \theta_2) + x_{34} \log(\theta_3 + \theta_4) \\ &= \sum_{i=1}^{3} x_i \log \theta_i + x_4 \log(1 - \theta_1 - \theta_2 - \theta_3) \\ &\quad + x_{12} \log(\theta_1 + \theta_2) + x_{34} \log(1 - \theta_1 - \theta_2). \end{aligned}$$

Solving the following system of equations

$$\begin{aligned} \frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_1} &= \frac{x_1}{\theta_1} - \frac{x_4}{1 - \theta_1 - \theta_2 - \theta_3} \\ &\quad + \frac{x_{12}}{\theta_1 + \theta_2} - \frac{x_{34}}{1 - \theta_1 - \theta_2} = 0, \qquad (3.35) \\ \frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_2} &= \frac{x_2}{\theta_2} - \frac{x_4}{1 - \theta_1 - \theta_2 - \theta_3} \\ &\quad + \frac{x_{12}}{\theta_1 + \theta_2} - \frac{x_{34}}{1 - \theta_1 - \theta_2} = 0, \\ \frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_3} &= \frac{x_3}{\theta_3} - \frac{x_4}{1 - \theta_1 - \theta_2 - \theta_3} = 0, \end{aligned}$$

we obtain

$$\begin{aligned} \frac{x_1}{\theta_1} &= \frac{x_2}{\theta_2} = \frac{x_1 + x_2}{\theta_1 + \theta_2}, \\ \frac{x_3}{\theta_3} &= \frac{x_4}{\theta_4} = \frac{x_3 + x_4}{\theta_3 + \theta_4}. \end{aligned}$$

Hence, from (3.35), we have

$$\frac{x_1+x_2}{\theta_1+\theta_2}-\frac{x_3+x_4}{\theta_3+\theta_4}+\frac{x_{12}}{\theta_1+\theta_2}-\frac{x_{34}}{\theta_3+\theta_4}=0,$$

or

$$\frac{x_1+x_2+x_{12}}{\theta_1+\theta_2}=\frac{x_3+x_4+x_{34}}{\theta_3+\theta_4}=\frac{N}{1},$$

where $N \,\hat{=}\, \sum_{i=1}^4 x_i + x_{12} + x_{34}$, resulting in

$$\theta_1+\theta_2=\frac{x_1+x_2+x_{12}}{N}.$$

Therefore, the MLE of θ_i is

$$\hat{\theta}_i=\frac{X_i}{N}\left\{\frac{X_1+X_2+X_{12}}{X_1+X_2}\cdot I_{(1\leqslant i\leqslant 2)}+\frac{X_3+X_4+X_{34}}{X_3+X_4}\cdot I_{(3\leqslant i\leqslant 4)}\right\}. \qquad \|$$

31• Proof of (3.33) and (3.34)

31.1• Recall the central limit theorem

— Let $\{Y_n\}_{n=1}^{\infty}$ be i.i.d. random variables with the common mean μ and common variance $\sigma^2>0$.

— The central limit theorem presented in Theorem 2.9 states that

$$\frac{\frac{1}{\sqrt{n}}\sum_{i=1}^n Y_i-\sqrt{n}\mu}{\sigma}=\frac{\sqrt{n}(\sum_{i=1}^n Y_i/n-\mu)}{\sigma}\xrightarrow{\mathrm{L}} N(0,1) \qquad (3.36)$$

as $n\to\infty$.

31.2• Proof of (3.33)

— The likelihood function of θ is $L(\theta;\mathbf{x})=\prod_{i=1}^n f(X_i;\theta)$ so that the log-likelihood function is $\ell(\theta;\mathbf{x})=\sum_{i=1}^n \log f(X_i;\theta)$.

— The score function is

$$S(\theta;\mathbf{x})=\frac{\mathrm{d}\ell(\theta;\mathbf{x})}{\mathrm{d}\theta}=\sum_{i=1}^n\frac{\mathrm{d}\log f(X_i;\theta)}{\mathrm{d}\theta}\,\hat{=}\,\sum_{i=1}^n Y_i, \qquad (3.37)$$

so that

$$E\{S(\theta;\mathbf{x})\}=nE(Y_1)\quad\text{and}\quad \mathrm{Var}\{S(\theta;\mathbf{x})\}=n\mathrm{Var}(Y_1), \qquad (3.38)$$

where $\{Y_i\}_{i=1}^{\infty}$ be i.i.d. random variables with the common mean

$$\mu = E(Y_1) \overset{(3.38)}{=} \frac{E\{S(\theta;\mathbf{x})\}}{n} = 0 \tag{3.39}$$

and the common variance

$$\sigma^2 = \mathrm{Var}(Y_1) \overset{(3.38)}{=} \frac{\mathrm{Var}\{S(\theta;\mathbf{x})\}}{n} = \frac{I_n(\theta)}{n} = I(\theta). \tag{3.40}$$

— Thus

$$\begin{aligned}
\frac{S(\theta;\mathbf{x}) - E\{S(\theta;\mathbf{x})\}}{\sqrt{\mathrm{Var}\{S(\theta;\mathbf{x})\}}} &\overset{(3.39)}{=} \frac{S(\theta;\mathbf{x})}{\sqrt{nI(\theta)}} \\
&\overset{(3.37)}{=} \frac{\sum_{i=1}^n Y_i}{\sqrt{nI(\theta)}} \overset{(3.40)}{=} \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^n Y_i}{\sigma} \\
&\overset{(3.36)}{=} \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^n Y_i - \sqrt{n}\times 0}{\sigma} \\
&\overset{\mathrm{L}}{\to} N(0,1) \quad \text{as} \quad n\to\infty,
\end{aligned}$$

which completes the proof of (3.33). □

31.3• Proof of (3.34)

— By applying the first-order Taylor expansion to the score function $S(\theta;\mathbf{x})$ around the MLE $\hat{\theta}_n$ and noting $S(\hat{\theta}_n;\mathbf{x}) = 0$, we have

$$S(\theta;\mathbf{x}) = S(\hat{\theta}_n;\mathbf{x}) + (\theta - \hat{\theta}_n)\frac{\mathrm{d}S(\theta;\mathbf{x})}{\mathrm{d}\theta}\bigg|_{\theta=\theta^*} \hat{=}\ 0 + (\theta - \hat{\theta}_n)H(\theta^*;\mathbf{x}),$$

where θ^* is a point between θ and $\hat{\theta}_n$. Thus

$$\frac{S(\theta;\mathbf{x})}{\sqrt{nI(\theta)}} = \sqrt{nI(\theta)}\,(\hat{\theta}_n - \theta) \times \frac{-H(\theta^*;\mathbf{x})/n}{I(\theta)}.$$

— We only need to prove that

$$-\frac{H(\theta^*;\mathbf{x})}{n} \overset{\mathrm{P}}{\to} I(\theta) \quad \text{as} \quad n\to\infty. \tag{3.41}$$

— According to the weak law of large number (see, Theorem 2.7), we have

$$-\frac{H(\theta;\mathbf{x})}{n} = -\frac{1}{n}\cdot\frac{\mathrm{d}S(\theta;\mathbf{x})}{\mathrm{d}\theta} \overset{(3.37)}{=} \frac{1}{n}\sum_{i=1}^{n} -\frac{\mathrm{d}^2\log f(X_i;\theta)}{\mathrm{d}\theta^2}$$

$$\hat{=} \frac{1}{n}\sum_{i=1}^{n} Z_i \xrightarrow{\mathrm{P}} E(Z_1) \quad \text{as} \quad n\to\infty. \tag{3.42}$$

From (3.24), we obtain

$$E(Z_1) = E\left\{-\frac{\mathrm{d}^2\log f(X_1;\theta)}{\mathrm{d}\theta^2}\right\} = I(\theta).$$

From (3.32), since $\hat{\theta}_n \xrightarrow{\mathrm{P}} \theta$ as $n\to\infty$, we have

$$-\frac{H(\theta^*;\mathbf{x})}{n} = -\frac{H(\theta;\mathbf{x})}{n}\times\frac{H(\theta^*;\mathbf{x})}{H(\theta;\mathbf{x})} \qquad \text{[by using (3.42)]}$$

$$\xrightarrow{\mathrm{P}} I(\theta)\times 1 = I(\theta) \quad \text{as} \quad n\to\infty,$$

implying (3.41). □

Exercise 3

3.1 Let $X_1,\ldots,X_n \overset{\text{iid}}{\sim} U[\theta_1,\theta_2]$. Find the MLEs of θ_1 and θ_2.

3.2 A sample of size n_1 is drawn from $N(\mu_1,\sigma_1^2)$. A second sample of size n_2 is drawn from $N(\mu_2,\sigma_2^2)$. Assume that the two samples are independent.

(a) What is the MLE of $\theta=\mu_1-\mu_2$?

(b) If we assume that the total sample size $n=n_1+n_2$ is fixed, how should the n observations be approximately divided between the two populations in order to minimize the variance of the $\hat{\theta}$?

3.3 The joint pmf of N_1, N_2, N_3 and N_4 is assumed to be

$$p(n_1,\ldots,n_4;\boldsymbol{\theta}) = \binom{n}{n_1,\ldots,n_4}\prod_{i=1}^{4}\theta_i^{n_i},$$

where $n_i\geqslant 0$, $\sum_{i=1}^{4} n_i = n$, and $\boldsymbol{\theta}=(\theta_1,\ldots,\theta_4)^\top\in\mathbb{T}_4$. Let $\theta_1=\alpha\beta$, $\theta_2=\alpha(1-\beta)$, $\theta_3=(1-\alpha)\beta$, and $\theta_4=(1-\alpha)(1-\beta)$, where $0<\alpha<1$ and $0<\beta<1$. Find the MLEs of α and β.

3.4 Let $X_{i1}, \ldots, X_{in} \overset{\text{iid}}{\sim} N(\mu_i, \sigma^2)$ for $i = 1, \ldots, 4$, where $\mu_1 = a + b + c$, $\mu_2 = a + b - c$, $\mu_3 = a - b + c$, and $\mu_4 = a - b - c$. The four samples are independent. What are the MLEs of a, b, c and σ^2?

3.5 Let $X_1, \ldots, X_n \sim U[\mu - \sqrt{3}\,\sigma,\ \mu + \sqrt{3}\,\sigma]$, where $\mu \in \mathbb{R}$ and $\sigma > 0$.

(a) Find the MLEs of μ and σ.

(b) Find the moment estimators of μ and σ.

3.6 Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x;\theta)$ with $f(x;\theta) = \mathrm{e}^{-(x-\theta)}$ for $x \geqslant \theta$ and $\theta \in \mathbb{R}$.

(a) Find the MLE of θ.

(b) Find the moment estimator of θ.

(c) Using the prior density $\pi(\theta) = \mathrm{e}^{-\theta} I_{(0,\infty)}(\theta)$, find the Bayesian estimator of θ.

3.7 Let $X \sim \text{Bernoulli}(\theta)$. Let $t_1(X) = X$ and $t_2(X) = 1/2$.

(a) Are both $t_1(X)$ and $t_2(X)$ unbiased?

(b) Compare the MSE of $t_1(X)$ with that of $t_2(X)$.

3.8 Let $\{Y = 1\}$ denote the class of people who possess a sensitive characteristic (e.g., drug-taking, shoplifting, driving under influence and so on) and $\{Y = 0\}$ denote the complementary class. Let W be a non-sensitive dichotomous variate and be independent of Y. The interviewer should select a suitable W so that the proportion $p = \Pr(W = 1)$ can be estimated easily. Without loss of generality, p is assumed to be known. For example, we may define $W = 1$ if the respondent was born between August and December and $W = 0$ otherwise. Hence, it is reasonable to assume that $p \approx 5/12 = 0.41667$. Our aim is to estimate the proportion $\pi = \Pr(Y = 1)$.

To collect sensitive information, the interviewer may adopt the format at the left-hand side of Table 3.2. The interviewee is then asked to put a tick in either the open circle or in the triangle formed by the three solid dots in Table 3.2 according to his/her truthful answer. In this case, $\{Y = 0,\ W = 0\}$ means that the interviewee was neither a drug user nor born between August and December. That is, $\{Y = 0,\ W = 0\}$ represents a non-sensitive subclass. On the other hand, a tick in

the triangle may possibly indicates the interviewee was born between August and December (i.e., $\{W = 1\}$). Therefore, respondents who are drug users are well covered their true identities by those who are between–August–December born non-drug users, and are willing to circle the triangle formed by the three dots. Such a design encourages the respondents to not only participate in the survey but also provide their truthful responses.

Table 3.2 The triangular model and its cell probabilities

Category	$W=0$	$W=1$			$W=0$	$W=1$	Total
$Y=0$	◯	●		$Y=0$	$(1-\pi)(1-p)$	$(1-\pi)p$	$1-\pi$
$Y=1$	●	●		$Y=1$	$\pi(1-p)$	πp	π
				Total	$1-p$	p	1

Note: Please truthfully put a tick in the circle (i.e., ◯) or circle the triangle formed by the three dots (i.e., ●).

Let $Y_{\text{obs}} = \{y_i\colon\ i = 1, \ldots, n\}$ denote the observed data for n respondents with $y_i = 1$ if the i-th respondent puts a tick in the triangle; $y_i = 0$ otherwise.

(a) Find the MLE $\hat{\pi}$ of π.

(b) Find the expectation of $\hat{\pi}$.

3.9 A discrete random variable Y is said to follow a *zero-truncated binomial* (ZTB) distribution if its pmf is

$$\Pr(Y = y) = \binom{m}{y}\pi^y(1-\pi)^{m-y}/[1-(1-\pi)^m], \quad 1 \leqslant y \leqslant m,$$

where $\pi \in (0, 1)$ is an unknown parameter, and m is a known positive integer. We will write $Y \sim \text{ZTB}(m, \pi)$.

Let $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \text{ZTB}(m, \pi)$. Find the MLE of π by using the Fisher scoring algorithm.

3.10 Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu_0, \theta)$, where μ_0 is known and $\theta > 0$.

(a) Find the MLE $\hat{\theta}$ of θ?

(b) What is the asymptotic distribution of $\sqrt{n}(\hat{\theta} - \theta)$?

3.11 Let $X_1, \ldots, X_n$ be a random sample from a distribution with density

$$f(x;\theta) = \frac{g(x)}{h(\theta)}, \quad a(\theta) \leqslant x \leqslant b(\theta),$$

where $g(x)$ is a function of x only and $h(\theta) = \int_{a(\theta)}^{b(\theta)} g(x)dx$ a function of θ only. Let $a^{-1}(\theta)$ and $b^{-1}(\theta)$ be the inverse functions of $a(\theta)$ and $b(\theta)$, respectively. Prove that

(a) If $a(\theta)$ and $b(\theta)$ are monotone-increasing and monotone-decreasing functions of θ, respectively, then the sufficient statistic for θ is $\hat{\theta} = \min\{a^{-1}(X_{(1)}),\ b^{-1}(X_{(n)})\}$, where $X_{(1)}$ and $X_{(n)}$ are the smallest and largest order statistics, respectively.

(b) If $a(\theta)$ and $b(\theta)$ are monotone-decreasing and monotone-increasing functions of θ, respectively, then the sufficient statistic for θ is $\hat{\theta} = \max\{a^{-1}(X_{(1)}),\ b^{-1}(X_{(n)})\}$.

(c) The $\hat{\theta}$ is also the MLE of θ.

3.12 Let $Y = 1$ if a respondent is a drug user and $Y = 0$ otherwise. Let U denote the number of travel out of Hong Kong per year for the same respondent in a population in Hong Kong. Obviously, Y is a sensitive binary r.v. (thus it is not observable if the question is asked directly) and U is a non-sensitive random variable. Define $X = Y + U$. Let $Y \sim \text{Bernoulli}(\theta)$, $U \sim \text{Poisson}(\lambda)$, and $Y \perp\!\!\!\perp U$. The interviewer could ask the i-th respondent to report the sum $X_i = U_i + Y_i$ according to his/her truthful answer, $i = 1, \ldots, n$. Let the observed data be $X_1, \ldots, X_n$.

(a) Find the moment estimators of θ and λ.

(b) Find the MLEs of θ and λ.

3.13 Let $X_1, \ldots, X_n$ be a random sample from $f(x;\theta) = \mathrm{e}^{-(x-\theta)} I_{(\theta,\infty)}(x)$ for $-\infty < \theta < \infty$ and $Y_1 = \min(X_1, \ldots, X_n)$.

(a) Show that Y_1 is a complete sufficient statistic for θ.

(b) Find the function of Y_1 which is the unique UMVUE of θ.

3.14 Let a random sample of size n be taken from a discrete distribution with pmf $f(x;\theta) = 1/\theta$, $x = 1, 2, \ldots, \theta$, where θ is an unknown positive integer.

(a) Show that the largest observation $X_{(n)} \,\hat{=}\, Y$ is a complete sufficient statistic for θ.

(b) Prove that

$$\frac{Y^{n+1} - (Y-1)^{n+1}}{Y^n - (Y-1)^n}$$

is the unique UMVUE of θ.

3.15 Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta)$. Define $\tau(\theta) = \text{Var}(X) = \theta(1-\theta)$.

(a) Find the Cramér–Rao lower bound for the unbiased estimator of $\tau(\theta)$.

(b) Find the unique UMVUE of $\tau(\theta)$ if such exists.

3.16 Let $X_i \sim \text{Poisson}(\lambda_i)$ for $i = 0, 1, 2$, and X_0, X_1, X_2 are independent. Define $Y_1 = X_0 + X_1$ and $Y_2 = X_0 + X_2$. Then $(Y_1, Y_2)^\top$ is said to follow the two-dimensional Poisson distribution with parameters $(\lambda_0, \lambda_1, \lambda_2)$, denoted by $(Y_1, Y_2)^\top \sim \text{MP}_2(\lambda_0, \lambda_1, \lambda_2)$.

(a) Find the joint probability mass function of $(Y_1, Y_2)^\top$.

(b) Let $\mathbf{y}_1, \ldots, \mathbf{y}_n \overset{\text{iid}}{\sim} \text{MP}_2(\lambda_0, \lambda_1, \lambda_2)$, where $\mathbf{y}_j = (Y_{1j}, Y_{2j})^\top$ and $\boldsymbol{y}_j = (y_{1j}, y_{2j})^\top$ denotes the realization of $\mathbf{y}_j$, $j = 1, \ldots, n$. Furthermore, let $\min(\boldsymbol{y}_j) = \min(y_{1j}, y_{2j}) = 0$ for all $j = 1, \ldots, n$. Find the MLEs of $(\lambda_0, \lambda_1, \lambda_2)$.

Chapter 4

Confidence Interval Estimation

4.1 Introduction

1• Why isn't a point estimator enough?

- Point estimators, no matter how they are determined (e.g., MLEs, moment estimators, or Bayesian estimators), share the fundamental weakness: They cannot provide the precision of the estimators.
- For instance, let $X_1 \ldots, X_n \overset{\text{iid}}{\sim} \text{Poisson}(\lambda)$. Then, $\hat{\lambda} = \bar{X}$ is not only the MLE and but also the moment estimator of the parameter λ.
- Suppose that a sample of size six is taken from Poisson(λ) and $\hat{\lambda} = 6.8$.
- Does it follow that the true λ is likely to be close to $\hat{\lambda}$ — say, in the interval [6.7, 6.9] or is the estimation process so imprecise that λ might actually be as small as 1.0 or as large as 12.0?
- Unfortunately, point estimators do not allow us to make those kinds of extrapolations.

1.1• Random interval

— An interval is called a *random interval* if at least one of its end points is a random variable.

1.2• Confidence interval

— The usual way to quantify the amount of uncertainty in an estimator is to construct a *confidence interval* (CI).

— In principle, CIs are ranges of numbers that have a high probability of "containing" the unknown parameter as an interior point.

— By looking at the width of a CI, we can get a good sense of the estimator's precision.

2• Two-sided confidence interval

- Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x;\theta)$ and $\mathbf{x} = (X_1, \ldots, X_n)^\top$.
- Let $T_1(\mathbf{x})$ and $T_2(\mathbf{x})$ be two statistics such that $T_1 \leqslant T_2$ and

$$\Pr(T_1 \leqslant \theta \leqslant T_2) = 1 - \alpha.$$

- Then the random interval $[T_1, T_2]$ is called a $100(1-\alpha)\%$ CI for θ, $1-\alpha$ is called the *confidence coefficient/level*, and T_1 and T_2 are called the lower and upper confidence limits/bounds, respectively.
- A value $[t_1, t_2]$ of the random interval $[T_1, T_2]$ is also called a $100(1-\alpha)$ percent CI for θ.

2.1• One-sided CI

— Assume that $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x;\theta)$.

— Let $T_1(\mathbf{x})$ be a statistic for which $\Pr(T_1 \leqslant \theta) = 1 - \alpha$, then T_1 is called a one-sided lower confidence limit for θ.

— Similarly, let $T_2(\mathbf{x})$ be a statistic for which $\Pr(\theta \leqslant T_2) = 1 - \alpha$, then T_2 is called a one-sided upper confidence limit for θ.

3• Pivotal quantity

- To find a CI for a parameter, we need to introduce the concept of *pivot*.
- The use of pivots for the construction of CIs or confidence sets can be traced as far back as Fisher (1930), who used the term of *inverse probability*.

- Based on pivotal quantities, Barnard (1949, 1980) proposed a so-called *pivotal inference*, which is closely related with the theory of *structural inference* of Fraser (1968, 1979).
- Berger and Wolpert (1984) discussed the strengths and weaknesses of these methods.

枢轴变量

Definition 4.1 (Pivotal quantity). Assume that $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x;\theta)$ and $T = T(X_1, \ldots, X_n)$ is a sufficient statistic of θ. Let $P = P(T, \theta)$ be a function of T and θ. If the distribution of P does not depend on θ, then P is called a *pivotal quantity* or *pivot*. ‖

Example 4.1 (Exponential distribution). Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Exponential}(\theta)$. Find a $100(1-\alpha)\%$ CI for θ.

Solution. The pdf of $X \sim \text{Exponential}(\theta)$ is $\theta\, \mathrm{e}^{-\theta x}$, $x > 0$, $\theta > 0$. Using the property stated in Appendix A.2.3, we have

$$n\bar{X} = \sum_{i=1}^{n} X_i \sim \text{Gamma}(n, \theta).$$

By Property 1 in Appendix A.2.4, we obtain

$$2\theta n\bar{X} \sim \text{Gamma}(2n/2, 1/2) = \chi^2(2n).$$

Then $2\theta n\bar{X}$ is a pivotal quantity. Let $\chi^2(\alpha, \nu)$ denote the upper α-th quantile satisfying

$$\Pr\{\chi^2(\nu) > \chi^2(\alpha, \nu)\} = \alpha. \tag{4.1}$$

Thus, using the equal-probability (or equal-tail) method, we have

$$\begin{aligned} 1-\alpha &= \Pr\left\{\chi^2(1-\alpha/2, 2n) \leqslant 2\theta n\bar{X} \leqslant \chi^2(\alpha/2, 2n)\right\} \\ &= \Pr\left\{\frac{\chi^2(1-\alpha/2, 2n)}{2n\bar{X}} \leqslant \theta \leqslant \frac{\chi^2(\alpha/2, 2n)}{2n\bar{X}}\right\}; \end{aligned}$$

that is,

$$[L_p, U_p] = \left[\frac{\chi^2(1-\alpha/2, 2n)}{2n\bar{X}},\ \frac{\chi^2(\alpha/2, 2n)}{2n\bar{X}}\right]$$

is a $100(1-\alpha)\%$ CI for θ. ‖

3.1• The shortest CI

— Many textbooks and monographes adopt the equal-probability method.

— However, in general, the equal-probability CI $[L_p, U_p]$ is not optimal because its width $U_p - L_p$ is not the shortest. For more details, see §4.7.

4• A general method of constructing a pivotal quantity

- Let $X \sim f(x;\theta)$ and the cdf be denoted by $F(x)$ or $F(x;\theta)$.
- We assume that $F(x)$ is continuous. ???
- We first prove that $W \,\hat{=}\, F(X) \sim U(0,1)$. In fact,

$$\begin{aligned}
\Pr(W \leqslant w) &= \Pr\{F(X) \leqslant w\} \quad [\because F^{-1} \text{ is also increasing}] \\
&= \Pr\{F^{-1}(F(X)) \leqslant F^{-1}(w)\} \\
&= \Pr\{X \leqslant F^{-1}(w)\} \qquad [\because F \text{ is the cdf of } X] \\
&= F[F^{-1}(w)] \;=\; w,
\end{aligned}$$

  implying $W \sim U(0,1)$.

- Next, it is easy to see that

$$V \,\hat{=}\, -\log(W) = -\log\{F(X)\} = -\log\{F(X;\theta)\}$$

  follows the standard exponential distribution with pdf

$$g_V(v) = h_W(w) \times \left|\frac{\mathrm{d}w}{\mathrm{d}v}\right| = 1 \times \left|\frac{\mathrm{d}\,\mathrm{e}^{-v}}{\mathrm{d}v}\right| = \mathrm{e}^{-v}, \qquad v > 0.$$

- Namely, $V \sim \text{Exponential}(1) = \text{Gamma}(1,1)$.
- According to Property 1 in Appendix A.2.4, we have

$$2V \sim \text{Gamma}\left(\frac{2}{2}, \frac{1}{2}\right) = \chi^2(2). \tag{4.2}$$

4.1• Sample size is n

— When sample size is n; i.e., $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x;\theta)$, from (4.2), we have

$$-2\sum_{i=1}^{n} \log F(X_i;\theta) \sim \chi^2(2n), \tag{4.3}$$

which is a pivotal quantity.

4.2 The Confidence Interval of Normal Mean

4.2.1 The variance is known

5• The case of known variance

- Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma_0^2)$, where σ_0^2 is known.
- It was shown that $\bar{X}$ is a sufficient statistic of μ in Example 3.23.
- Thus, we need to find a pivotal quantity P being a function of both $\bar{X}$ and μ, such that the distribution of P does not depend on the parameter μ.
- By using (2.13),
$$P = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma_0} \sim N(0, 1).$$
That is P is a pivotal quantity.
- Thus, we have
$$\Pr\left\{-z_{\alpha/2} \leqslant \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma_0} \leqslant z_{\alpha/2}\right\} = 1 - \alpha$$
or
$$\Pr\left(\bar{X} - z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}} \leqslant \mu \leqslant \bar{X} + z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}\right) = 1 - \alpha,$$
where z_α denotes the upper α-th quantile of $N(0, 1)$ such that $\Pr(Z > z_\alpha) = \alpha$.
- Therefore,
$$\left[\bar{X} - z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}},\ \bar{X} + z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}\right] \tag{4.4}$$
is a $100(1 - \alpha)\%$ CI for μ.

5.1• CI must be functions of the sufficient statistic

— Define $Y \hat{=} \sum_{i=1}^{n} X_i/(n-1)$, we have
$$\frac{Y - E(Y)}{\sqrt{\text{Var}(Y)}} = \frac{Y - \mu}{\sigma_0/\sqrt{n-1}} \sim N(0, 1),$$

and it seems that

$$\left[Y - z_{\alpha/2}\frac{\sigma_0}{\sqrt{n-1}},\ Y + z_{\alpha/2}\frac{\sigma_0}{\sqrt{n-1}}\right] \tag{4.5}$$

is also a $100(1-\alpha)\%$ CI for μ.

— How to compare (4.5) with (4.4)?

— The Y is not a sufficient statistic of μ but $\bar{X}$ is.

5.2• The key for constructing a CI is to find a pivot

— For example, from $X_i \sim N(\mu, \sigma_0^2)$, we have $(X_i - \mu)/\sigma_0 \sim N(0,1)$. We wonder whether or not $(X_i - \mu)/\sigma_0$ is a pivotal quantity?

— If yes, we would obtain

$$\Pr\left(-z_{\alpha/2} \leqslant \frac{X_i - \mu}{\sigma_0} \leqslant z_{\alpha/2}\right) = 1 - \alpha,$$

and hence we could claim that $[X_i - z_{\alpha/2}\sigma_0,\ X_i + z_{\alpha/2}\sigma_0]$ $(i = 1, \ldots, n)$ are also $100(1-\alpha)\%$ CIs for μ.

— Of course, the answer is no because a single X_i is not sufficient for μ.

Example 4.2 (Numerical illustration with known variance). If a random sample of size $n = 20$ from $N(\mu, \sigma_0^2)$ with $\sigma_0^2 = 225$ has the mean $\bar{x} = 64.3$, construct a 95% CI for the population mean μ.

Solution. Substituting $n = 20$, $\bar{x} = 64.3$, $\sigma_0 = 15$, and $z_{0.025} = 1.96$ into (4.4), we get

$$64.3 - 1.96 \cdot \frac{15}{\sqrt{20}} \leqslant \mu \leqslant 64.3 + 1.96 \cdot \frac{15}{\sqrt{20}},$$

which reduces to $57.7 \leqslant \mu \leqslant 70.9$. ‖

4.2.2 The variance is unknown

6• The case of unknown variance

- Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, where σ^2 is unknown.

- It has been shown that $\mathbf{t}(\mathbf{x}) = (T_1(\mathbf{x}), T_2(\mathbf{x}))^\top \hat{=} (\bar{X}, S^2)^\top$ is jointly sufficient for $\boldsymbol{\theta} = (\mu, \sigma^2)^\top$ in Example 3.26.
- Define $P \hat{=} \sqrt{n}(\bar{X} - \mu)/S$. It is easy to rewrite

$$P = \frac{\sqrt{n}(1, 0)\{\mathbf{t}(\mathbf{x}) - \boldsymbol{\theta}\}}{\sqrt{(0, 1)\mathbf{t}(\mathbf{x})}},$$

  which is a function of both $\mathbf{t}(\mathbf{x})$ and $\boldsymbol{\theta}$.
- By using (2.14),

$$P = \frac{\sqrt{n}(\bar{X} - \mu)}{S} \sim t(n - 1).$$

  That is, P is a pivotal quantity.
- Thus, we have

$$\Pr\left\{-t(\alpha/2, n - 1) \leqslant \frac{\sqrt{n}(\bar{X} - \mu)}{S} \leqslant t(\alpha/2, n - 1)\right\} = 1 - \alpha$$

  or

$$\Pr\left\{\bar{X} - t(\alpha/2, n - 1)\frac{S}{\sqrt{n}} \leqslant \mu \leqslant \bar{X} + t(\alpha/2, n - 1)\frac{S}{\sqrt{n}}\right\} = 1 - \alpha,$$

  where $t(\alpha, n-1)$ denotes the upper α-th quantile of $t(n-1)$ such that $\Pr\{t(n - 1) > t(\alpha, n - 1)\} = \alpha$.
- Therefore,

$$\left[\bar{X} - t(\alpha/2, n - 1)\frac{S}{\sqrt{n}},\ \bar{X} + t(\alpha/2, n - 1)\frac{S}{\sqrt{n}}\right] \tag{4.6}$$

  is a $100(1 - \alpha)\%$ CI for μ.

Example 4.3 (Numerical illustration with unknown variance). If a random sample of size $n = 12$ from $N(\mu, \sigma^2)$ has the mean $\bar{x} = 66.3$ and $s = 8.4$, construct a 95% CI for the population mean μ.

Solution. Substituting $n = 12$, $\bar{x} = 66.3$, $s = 8.4$, and $t(0.025, 11) = 2.201$ into (4.6), the 95% CI for μ becomes

$$66.3 - 2.201 \cdot \frac{8.4}{\sqrt{12}} \leqslant \mu \leqslant 66.3 + 2.201 \cdot \frac{8.4}{\sqrt{12}}$$

or simply $61.0 \leqslant \mu \leqslant 71.6$. ‖

4.3 The Confidence Interval of the Difference of Two Normal Means

7• The issues

- Let $X_{11}, \ldots, X_{1n_1}$ and $X_{21}, \ldots, X_{2n_2}$ be two independent samples from $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, respectively.
- The objective of this section is to construct CI of $\mu_1 - \mu_2$.
- We consider three cases.

7.1• Two variances are known

加工误差和两种药物的效果

— When σ_1^2 and σ_2^2 are known, we know that $\bar{X}_i$ is sufficient for μ_i for $i = 1, 2$. Note that $\bar{X}_1 \perp\!\!\!\perp \bar{X}_2$, then

$$\mathbf{t}(\mathbf{x}) = \mathbf{t}(\mathbf{x}_1, \mathbf{x}_2) = (T_1(\mathbf{x}_1), T_2(\mathbf{x}_2))^\top \hat{=} (\bar{X}_1, \bar{X}_2)^\top$$

is jointly sufficient for $\boldsymbol{\mu} = (\mu_1, \mu_2)^\top$, where $\mathbf{x}_i = (X_{i1}, \ldots, X_{in_i})^\top$ for $i = 1, 2$.

— Define $Z \hat{=} \{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)\}/\sigma_0$, where $\sigma_0 \hat{=} (\sigma_1^2/n_1 + \sigma_2^2/n_2)^{1/2}$.

— It is easy to rewrite

$$Z = \frac{(1, -1)\{\mathbf{t}(\mathbf{x}) - \boldsymbol{\mu}\}}{\sigma_0},$$

which is a function of both $\mathbf{t}(\mathbf{x})$ and $\boldsymbol{\mu}$.

— Since $\bar{X}_i \sim N(\mu_i, \sigma_i^2/n_i)$ for $i = 1, 2$ and $\bar{X}_1 \perp\!\!\!\perp \bar{X}_2$, we obtain

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma_0} \sim N(0, 1). \tag{4.7}$$

That is, Z is a pivotal quantity.

— Similar to (4.4), a $100(1 - \alpha)\%$ CI for $\mu_1 - \mu_2$ is given by

$$\left[\bar{X}_1 - \bar{X}_2 - z_{\alpha/2} \cdot \sigma_0,\ \bar{X}_1 - \bar{X}_2 + z_{\alpha/2} \cdot \sigma_0\right]. \tag{4.8}$$

7.2• Two variances are unknown but equal

— We consider the case that $\sigma_1^2 = \sigma_2^2 = \sigma^2$ are unknown.

— By using (2.10)

$$\frac{(n_i - 1)S_i^2}{\sigma^2} \sim \chi^2(n_i - 1), \quad i = 1, 2, \quad \text{and} \quad S_1^2 \perp\!\!\!\perp S_2^2, \tag{4.9}$$

where $S_i^2 = \sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2/(n_i - 1)$, $i = 1, 2$.

— Let

$$S_p^2 = \frac{\sum_{i=1}^{2}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2}{n_1 + n_2 - 2} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n' - 1}$$

denote the pooled sample variance, where $n' = n_1 + n_2 - 1$, then S_p^2 is an unbiased estimator of the common variance σ^2 and

$$Y \triangleq \frac{(n' - 1)S_p^2}{\sigma^2} = \sum_{i=1}^{2} \frac{(n_i - 1)S_i^2}{\sigma^2} \overset{(4.9)}{\sim} \chi^2(n' - 1). \tag{4.10}$$

— From (4.7) and (4.10), we know $Z \perp\!\!\!\perp Y$ and

$$\frac{Z}{\sqrt{Y/(n' - 1)}} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p(1/n_1 + 1/n_2)^{1/2}} \sim t(n' - 1). \tag{4.11}$$

— Therefore, a $100(1 - \alpha)\%$ CI for $\mu_1 - \mu_2$ is given by

$$\bar{X}_1 - \bar{X}_2 \mp t(\alpha/2, n' - 1) \cdot S_p(1/n_1 + 1/n_2)^{1/2}. \tag{4.12}$$

7.3• Two variances are unknown

— When σ_1^2 and σ_2^2 are unknown, we can show that (see Exercise 4.1)

$$T_{\text{Welch}} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \;\dot{\sim}\; t(\nu), \tag{4.13}$$

where

$$\nu = \left\{ \frac{c^2}{n_1 - 1} + \frac{(1 - c)^2}{n_2 - 1} \right\}^{-1} \quad \text{and} \quad c = \frac{S_1^2/n_1}{S_1^2/n_1 + S_2^2/n_2}.$$

— As an adjusted version of (4.11), the T_{Welch} is sometimes called Welch's approximate t.

4.4 The Confidence Interval of Normal Variance

4.4.1 The mean is known

8• The case of known mean

- Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu_0, \sigma^2)$, where μ_0 is known.
- Since the joint density of $\mathbf{x} = (X_1, \ldots, X_n)^\top$ can be decomposed into

$$\begin{aligned} f(\boldsymbol{x}; \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{-\frac{(x_i - \mu_0)^2}{2\sigma^2}\right\} \\ &= (\sigma^2)^{-n/2} \exp\left\{-\frac{\sum_{i=1}^{n}(x_i - \mu_0)^2}{2\sigma^2}\right\} \times (2\pi)^{-n/2}, \\ &= g(T(\boldsymbol{x}); \sigma^2) \times h(\boldsymbol{x}), \end{aligned}$$

  by the factorization theorem, we know that $T(\mathbf{x}) \mathrel{\hat{=}} \sum_{i=1}^{n}(X_i - \mu_0)^2$ is a sufficient statistic of σ^2.

- Define

$$P \mathrel{\hat{=}} \sum_{i=1}^{n} \left(\frac{X_i - \mu_0}{\sigma}\right)^2 = \frac{T(\mathbf{x})}{\sigma^2},$$

  which is a function of both $T(\mathbf{x})$ and σ^2.

- Note that $(X_i - \mu_0)/\sigma \overset{\text{iid}}{\sim} N(0, 1)$, then

$$P = \sum_{i=1}^{n} \left(\frac{X_i - \mu_0}{\sigma}\right)^2 \sim \chi^2(n).$$

  That is, P is a pivotal quantity.

- Thus

$$\begin{aligned} 1 - \alpha &= \Pr\left\{\chi^2(1 - \alpha/2, n) \leqslant P \leqslant \chi^2(\alpha/2, n)\right\} \\ &= \Pr\left\{\frac{\sum_{i=1}^{n}(X_i - \mu_0)^2}{\chi^2(\alpha/2, n)} \leqslant \sigma^2 \leqslant \frac{\sum_{i=1}^{n}(X_i - \mu_0)^2}{\chi^2(1 - \alpha/2, n)}\right\}. \end{aligned}$$

- A $100(1 - \alpha)\%$ CI for σ^2 is given by

$$\left[\frac{\sum_{i=1}^{n}(X_i - \mu_0)^2}{\chi^2(\alpha/2, n)}, \frac{\sum_{i=1}^{n}(X_i - \mu_0)^2}{\chi^2(1 - \alpha/2, n)}\right]. \tag{4.14}$$

4.4.2 The mean is unknown

9• The case of unknown mean

- Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, where μ is unknown.
- It has been shown that $\mathbf{t}(\mathbf{x}) = (T_1(\mathbf{x}), T_2(\mathbf{x}))^\top \triangleq (\bar{X}, S^2)^\top$ is jointly sufficient for $\boldsymbol{\theta} = (\mu, \sigma^2)^\top$ in Example 3.26.
- Define $P \triangleq (n-1)S^2/\sigma^2$. It is easy to rewrite

$$P = \frac{(0, n-1)\mathbf{t}(\mathbf{x})}{(0,1)\boldsymbol{\theta}},$$

  which is a function of both $\mathbf{t}(\mathbf{x})$ and $\boldsymbol{\theta}$.
- By using (2.10),

$$P = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1).$$

  That is, P is a pivotal quantity.
- Thus

$$\begin{aligned} 1-\alpha &= \Pr\left\{\chi^2(1-\alpha/2, n-1) \leqslant \frac{(n-1)S^2}{\sigma^2} \leqslant \chi^2(\alpha/2, n-1)\right\} \\ &= \Pr\left\{\frac{(n-1)S^2}{\chi^2(\alpha/2, n-1)} \leqslant \sigma^2 \leqslant \frac{(n-1)S^2}{\chi^2(1-\alpha/2, n-1)}\right\}. \end{aligned}$$

- A $100(1-\alpha)\%$ CI for σ^2 is given by

$$\left[\frac{(n-1)S^2}{\chi^2(\alpha/2, n-1)}, \frac{(n-1)S^2}{\chi^2(1-\alpha/2, n-1)}\right]. \tag{4.15}$$

4.5 The Confidence Interval of the Ratio of Two Normal Variances

10• The issue

- Let $X_{11}, \ldots, X_{1n_1}$ and $X_{21}, \ldots, X_{2n_2}$ be two independent samples from $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, respectively.
- The objective is to construct a CI of σ_1^2/σ_2^2.

- Define $\nu_i = n_i - 1$ for $i = 1, 2$. Theorem 2.4 states that

$$\frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2} \sim F(\nu_1, \nu_2).$$

- Thus, we can write

$$1 - \alpha = \Pr\left\{ f(1 - \alpha/2, \nu_1, \nu_2) \leqslant \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2} \leqslant f(\alpha/2, \nu_1, \nu_2) \right\},$$

  where $f(\alpha, \nu_1, \nu_2)$ denotes the upper α-th quantile of $F(\nu_1, \nu_2)$ such that $\Pr\{F(\nu_1, \nu_2) > f(\alpha, \nu_1, \nu_2)\} = \alpha$.

- Since

$$f(1 - \alpha/2, \nu_1, \nu_2) = \frac{1}{f(\alpha/2, \nu_2, \nu_1)}, \tag{4.16}$$

  it follows that

$$\left[\frac{S_1^2}{S_2^2} \cdot f^{-1}(\alpha/2, \nu_1, \nu_2),\ \frac{S_1^2}{S_2^2} \cdot f(\alpha/2, \nu_2, \nu_1) \right] \tag{4.17}$$

  is a $100(1 - \alpha)\%$ CI for σ_1^2/σ_2^2.

10.1• Proof of (4.16)

— Note that

$$\frac{1}{F(\nu_2, \nu_1)} \overset{\mathrm{d}}{=} \frac{1}{\dfrac{\chi^2(\nu_2)/\nu_2}{\chi^2(\nu_1)/\nu_1}} = \frac{\chi^2(\nu_1)/\nu_1}{\chi^2(\nu_2)/\nu_2} \overset{\mathrm{d}}{=} F(\nu_1, \nu_2). \tag{4.18}$$

— On the one hand, the definition of $f(\alpha/2, \nu_2, \nu_1)$ indicates that

$$\frac{\alpha}{2} = \Pr\{F(\nu_2, \nu_1) > f(\alpha/2, \nu_2, \nu_1)\},$$

we obtain

$$\begin{aligned} 1 - \frac{\alpha}{2} &= \Pr\{F(\nu_2, \nu_1) \leqslant f(\alpha/2, \nu_2, \nu_1)\} \\ &= \Pr\left\{ \frac{1}{F(\nu_2, \nu_1)} \geqslant f^{-1}(\alpha/2, \nu_2, \nu_1) \right\} \\ &\overset{(4.18)}{=} \Pr\left\{ F(\nu_1, \nu_2) \geqslant f^{-1}(\alpha/2, \nu_2, \nu_1) \right\}. \end{aligned} \tag{4.19}$$

— On the other hand, the definition of $f(1-\alpha/2, \nu_1, \nu_2)$ means that

$$1 - \frac{\alpha}{2} = \Pr\{F(\nu_1, \nu_2) > f(1-\alpha/2, \nu_1, \nu_2)\}. \tag{4.20}$$

— By comparing (4.20) with (4.19), we immediately obtain (4.16). □

4.6 Large-Sample Confidence Intervals 找近似的置信区间

11• Three methods for constructing approximate CIs

- In the previous sections, we discussed the construction of *exact* CIs of parameters of interest in independent normal populations.
- The aim of this section is to introduce three methods to construct approximate CIs of parameters in other populations for large sample sizes.

12• Method I: Based on the central limit theorem

- Let $\{X_n\}_{n=1}^{\infty}$ be i.i.d. from a population with mean μ and variance $\sigma^2 > 0$.
- When σ^2 is a function of μ, we could denote σ^2 by $\sigma^2(\mu)$.
- Let $\bar{X}_n = \sum_{i=1}^{n} X_i/n$, according the central limit theorem (Theorem 2.9), we have

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma(\mu)} \dot{\sim} N(0,1). \tag{4.21}$$

12.1• The first approximate CI for μ

— Based on (4.21), the first approximate $100(1-\alpha)\%$ CI for μ can be constructed as

$$\begin{aligned} 1-\alpha &\doteq \Pr\left\{-z_{\alpha/2} \leqslant \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma(\mu)} \leqslant z_{\alpha/2}\right\} \\ &= \Pr(L_1 \leqslant \mu \leqslant U_1), \end{aligned} \tag{4.22}$$

where both L_1 and U_1 are functions of $\bar{X}_n$, $z_{\alpha/2}$ and n.

12.2• The second approximate CI for μ

— If it is very difficult to find the the lower bound L_1 and the upper bound U_1 in (4.22) from the two inequalities

$$-z_{\alpha/2} \leqslant \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma(\mu)} \leqslant z_{\alpha/2},$$

we could replace $\sigma(\mu)$ by its MLE $\hat{\sigma} = \sigma(\hat{\mu})$.

— Thus, the second approximate $100(1-\alpha)\%$ CI for μ is given by

$$[L_2, U_2] = \left[\bar{X}_n - z_{\alpha/2} \cdot \frac{\hat{\sigma}}{\sqrt{n}},\ \bar{X}_n + z_{\alpha/2} \cdot \frac{\hat{\sigma}}{\sqrt{n}}\right]. \tag{4.23}$$

Example 4.4 (Bernoulli distribution). Let $\{X_n\}_{n=1}^{\infty} \overset{\text{iid}}{\sim} \text{Bernoulli}(\mu)$ with parameter $\mu = \Pr(X_1 = 1)$.

1) Using (4.22) to find the approximate $100(1-\alpha)\%$ CI $[L_1, U_1]$ for μ.

2) Using (4.23) to find the approximate $100(1-\alpha)\%$ CI $[L_2, U_2]$ for μ.

Solution. We have $\mu = E(X_1)$ and $\sigma^2 = \sigma^2(\mu) = \text{Var}(X_1) = \mu(1-\mu)$. Based on $X_1, \ldots, X_n$, the MLE of μ is $\hat{\mu} = \bar{X}_n$.

1) From (4.21), we obtain

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sqrt{\mu(1-\mu)}} \dot{\sim} N(0,1).$$

Thus,

$$\begin{aligned} 1-\alpha &\doteq \Pr\left\{\left|\frac{\bar{X}_n - \mu}{\sqrt{\mu(1-\mu)/n}}\right| \leqslant z_{\alpha/2}\right\} \\ &= \Pr\{(\bar{X}_n - \mu)^2 \leqslant z_{\alpha/2}^2 \mu(1-\mu)/n\} \\ &= \Pr\{(1+z_*)\mu^2 - (2\bar{X}_n + z_*)\mu + \bar{X}_n^2 \leqslant 0\}, \end{aligned} \tag{4.24}$$

where $z_* = z_{\alpha/2}^2/n$. Solving the quadratic inequality inside the probability in (4.24), we obtain the first approximate $100(1-\alpha)\%$ CI of μ as follows:

$$\begin{aligned} [L_1, U_1] &= \frac{2\bar{X}_n + z_* \pm \sqrt{(2\bar{X}_n + z_*)^2 - 4(1+z_*)\bar{X}_n^2}}{2(1+z_*)} \\ &= \frac{2\bar{X}_n + z_* \pm \sqrt{4z_*\bar{X}_n(1-\bar{X}_n) + z_*^2}}{2(1+z_*)}, \end{aligned}$$

which is within $[0, 1]$.

2) Note that $\hat{\sigma} = \sqrt{\hat{\mu}(1-\hat{\mu})}$. From (4.23),

$$[L_2, U_2] = \left[\bar{X}_n - z_{\alpha/2} \cdot \sqrt{\frac{\hat{\mu}(1-\hat{\mu})}{n}} \,,\; \bar{X}_n + z_{\alpha/2} \cdot \sqrt{\frac{\hat{\mu}(1-\hat{\mu})}{n}} \right].$$

is the second approximate $100(1-\alpha)\%$ CI for μ. Sometimes, the CI $[L_2, U_2]$ is also called the $100(1-\alpha)\%$ Wald-type CI of μ, which may be beyond $[0, 1]$. ‖

13• Method II: Based on the asymptotic normality of $S(\theta; \mathbf{x})$

- Let $\{X_n\}_{n=1}^{\infty} \overset{\text{iid}}{\sim} f(x; \theta)$ and $\hat{\theta}_n$ be the MLE of θ based on $\{X_i\}_{i=1}^{n}$.
- If $E\{S(\theta; \mathbf{x})\} = 0$ and $\text{Var}\{S(\theta; \mathbf{x})\} = nI(\theta)$, from (3.33), we have

$$\frac{S(\theta; \mathbf{x})}{\sqrt{nI(\theta)}} \dot{\sim} N(0, 1), \tag{4.25}$$

where $S(\theta; \mathbf{x})$ denotes the score function with $\mathbf{x} = (X_1, \ldots, X_n)^\top$, $nI(\theta)$ is the Fisher information, and

$$I(\theta) = E\left\{ -\frac{\mathrm{d}^2 \log f(X; \theta)}{\mathrm{d}\theta^2} \right\}.$$

13.1• An approximate CI for θ

— Based on (4.25), an approximate $100(1-\alpha)\%$ CI for θ is given by

$$\begin{aligned} 1 - \alpha &\doteq \Pr\left\{ -z_{\alpha/2} \leqslant \frac{S(\theta; \mathbf{x})}{\sqrt{nI(\theta)}} \leqslant z_{\alpha/2} \right\} \\ &= \Pr(L_3 \leqslant \theta \leqslant U_3), \end{aligned} \tag{4.26}$$

where both L_3 and U_3 are functions of the θ's sufficient statistic $T(\mathbf{x})$, $z_{\alpha/2}$ and n.

Example 4.5 (Exponential distribution). Let $\{X_n\}_{n=1}^{\infty} \overset{\text{iid}}{\sim} f(x; \theta) = \theta\, \mathrm{e}^{-\theta x}$ for $x > 0$ and $\theta > 0$.

1) Find the MLE of θ based on $X_1, \ldots, X_n$.

2) Based on (4.26), to construct an approximate $100(1-\alpha)\%$ CI for θ.

Solution. 1) Based on $\mathbf{x} = (X_1, \ldots, X_n)^\top$, the log-likelihood function of θ is

$$\ell(\theta) = n \log \theta - \theta \sum_{i=1}^{n} x_i.$$

Let $\ell'(\theta) = n/\theta - \sum_{i=1}^{n} x_i = 0$, the resulting MLE of θ is $\hat{\theta}_n = 1/\bar{X}_n$.

2) The score function is given by

$$S(\theta; \mathbf{x}) = \ell'(\theta) = \frac{n}{\theta} - \sum_{i=1}^{n} x_i.$$

Since $I(\theta) = \theta^{-2}$, from (4.26), we have

$$1 - \alpha \doteq \Pr\left\{ -z_{\alpha/2} \leqslant \frac{n/\theta - n\bar{X}_n}{\sqrt{n\theta^{-2}}} \leqslant z_{\alpha/2} \right\}.$$

Hence

$$[L_3, U_3] = \left[\frac{1 - z_{\alpha/2}/\sqrt{n}}{\bar{X}_n}, \ \frac{1 + z_{\alpha/2}/\sqrt{n}}{\bar{X}_n} \right]$$

is a large-sample $100(1-\alpha)\%$ CI for θ. ‖

14• Method Ⅲ: Based on the asymptotic normality of $\hat{\theta}_n$

- Let $\{X_n\}_{n=1}^{\infty} \overset{\text{iid}}{\sim} f(x; \theta)$ and $\hat{\theta}_n$ be the MLE of θ based on $\{X_i\}_{i=1}^{n}$.
- From (3.34), we have

$$\{nI(\theta)\}^{1/2}(\hat{\theta}_n - \theta) \dot{\sim} N(0, 1). \tag{4.27}$$

14.1• The first approximate CI for θ

— Based on (4.27), the first approximate $100(1-\alpha)\%$ CI for θ can be constructed as

$$\begin{aligned} 1 - \alpha &\doteq \Pr[-z_{\alpha/2} \leqslant \{nI(\theta)\}^{1/2}(\hat{\theta}_n - \theta) \leqslant z_{\alpha/2}] \\ &= \Pr(L_4 \leqslant \theta \leqslant U_4), \end{aligned} \tag{4.28}$$

where both L_4 and U_4 are functions of $\hat{\theta}_n$, $z_{\alpha/2}$ and n.

14.2• The second approximate CI for θ

— If it is very difficult to find the the lower bound L_4 and the upper bound U_4 in (4.28) from the two inequalities

$$-z_{\alpha/2} \leqslant \{nI(\theta)\}^{1/2}(\hat{\theta}_n - \theta) \leqslant z_{\alpha/2},$$

we could replace $I(\theta)$ by its MLE $I(\hat{\theta}_n)$.

— Thus, the second approximate $100(1-\alpha)\%$ CI for θ is given by

$$[L_5, U_5] = \left[\hat{\theta}_n - z_{\alpha/2}\Big/\sqrt{nI(\hat{\theta}_n)}\,,\ \hat{\theta}_n + z_{\alpha/2}\Big/\sqrt{nI(\hat{\theta}_n)}\right]. \qquad (4.29)$$

Example 4.6 (Example 4.5 revisited). Based on (4.28), to construct an approximate $100(1-\alpha)\%$ CI for θ.

Solution. From Example 4.5, we have $\hat{\theta}_n = 1/\bar{X}_n$ and $I(\theta) = \theta^{-2}$. From (4.28), we obtain

$$1 - \alpha \doteq \Pr\left\{-z_{\alpha/2} \leqslant \frac{\sqrt{n}}{\theta}\left(\frac{1}{\bar{X}_n} - \theta\right) \leqslant z_{\alpha/2}\right\}.$$

Hence,

$$[L_4, U_4] = \left[\frac{1/\bar{X}_n}{1 + z_{\alpha/2}/\sqrt{n}},\ \frac{1/\bar{X}_n}{1 - z_{\alpha/2}/\sqrt{n}}\right]$$

is a large-sample $100(1-\alpha)\%$ CI for θ.

The exact equal-tail CI $[L_p, U_p]$ presented in Example 4.1, $[L_3, U_3]$ given in Example 4.5 and $[L_4, U_4]$ in this example are totally different. ‖

4.7 The Shortest Confidence Interval

Example 4.7 (The shortest CI of μ when σ^2 is known). Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma_0^2)$, where σ_0^2 is known. Prove that

1) The equal-tail CI of μ is the shortest.

2) The equal-tail CI and the equal-height CI of μ are identical.

Proof. Let $Y_i = X_i/\sigma_0$, then $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} N(\mu/\sigma_0, 1)$. Without loss of generality, we set $\sigma_0 = 1$. Let $\Phi(\cdot)$ denote the cdf of $N(0,1)$ and

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

be the pdf of $N(0,1)$.

1) The pivot is $Z = \sqrt{n}(\bar{X} - \mu) \sim N(0,1)$, and we have $-Z \sim N(0,1)$. Let $\bar{X} + a$ and $\bar{X} + b$ denote the lower bound and upper bound of the $100(1-\alpha)\%$ CI of μ, we have

$$\begin{aligned} 1 - \alpha &= \Pr(\bar{X} + a \leqslant \mu \leqslant \bar{X} + b) \\ &= \Pr(a \leqslant \mu - \bar{X} \leqslant b) \\ &= \Pr\{\sqrt{n}a \leqslant \sqrt{n}(\mu - \bar{X}) \leqslant \sqrt{n}b\} \\ &= \Pr(\sqrt{n}a \leqslant -Z \leqslant \sqrt{n}b), \end{aligned}$$

which is equivalent to

$$\alpha_1 = \Pr(-Z < \sqrt{n}a) = \Phi(\sqrt{n}a) \quad \text{and} \tag{4.30}$$

$$\begin{aligned} \alpha - \alpha_1 &= \Pr(-Z > \sqrt{n}b) \\ &= \Pr(Z < -\sqrt{n}b) = \Phi(-\sqrt{n}b) \end{aligned} \tag{4.31}$$

for any $\alpha_1 \in [0, \alpha]$. From (4.30) and (4.31), we have

$$a = \Phi^{-1}(\alpha_1)/\sqrt{n} \quad \text{and} \quad b = -\Phi^{-1}(\alpha - \alpha_1)/\sqrt{n}.$$

The width of the CI $[\bar{X} + a, \bar{X} + b]$ is a function of the unknown α_1; i.e.,

$$L(\alpha_1) = b - a = -\frac{1}{\sqrt{n}} \left\{\Phi^{-1}(\alpha_1) + \Phi^{-1}(\alpha - \alpha_1)\right\}. \tag{4.32}$$

We would like to find the $\hat{\alpha}_1$ such that $L(\alpha_1)$ is minimized:

$$\hat{\alpha}_1 = \arg \min_{0 \leqslant \alpha_1 \leqslant \alpha} L(\alpha_1).$$

To find the $\hat{\alpha}_1$, we first prove the following formula

$$\frac{\mathrm{d}\Phi^{-1}(y)}{\mathrm{d}y} = \frac{1}{\phi(\Phi^{-1}(y))}. \tag{4.33}$$

In fact, by defining $x = \Phi^{-1}(y)$, we have $y = \Phi(x)$ so that $\mathrm{d}y = \phi(x)\mathrm{d}x$. Thus

$$\frac{\mathrm{d}\Phi^{-1}(y)}{\mathrm{d}y} = \frac{\mathrm{d}x}{\mathrm{d}y} = \frac{1}{\mathrm{d}y/\mathrm{d}x} = \frac{1}{\phi(x)} = \frac{1}{\phi(\Phi^{-1}(y))},$$

Now, we find the $\hat{\alpha}_1$. Let

$$\begin{aligned} 0 = \frac{\mathrm{d}L(\alpha_1)}{\mathrm{d}\alpha_1} &\overset{(4.32)}{=} -\frac{1}{\sqrt{n}}\left\{\frac{\mathrm{d}\Phi^{-1}(\alpha_1)}{\mathrm{d}\alpha_1} + \frac{\mathrm{d}\Phi^{-1}(\alpha-\alpha_1)}{\mathrm{d}\alpha_1}\right\} \\ &\overset{(4.33)}{=} -\frac{1}{\sqrt{n}}\left\{\frac{1}{\phi(\Phi^{-1}(\alpha_1))} - \frac{1}{\phi(\Phi^{-1}(\alpha-\alpha_1))}\right\}, \end{aligned}$$

we have

$$\phi(\Phi^{-1}(\alpha_1)) = \phi(\Phi^{-1}(\alpha-\alpha_1)),$$

or

$$\Phi^{-1}(\alpha_1) = \Phi^{-1}(\alpha-\alpha_1).$$

Thus, $\alpha_1 = \alpha/2$, implying that the shortest CI

$$[\bar{X}+a, \bar{X}+b] = [\bar{X} + \Phi^{-1}(\alpha/2)/\sqrt{n},\ \bar{X} - \Phi^{-1}(\alpha/2)/\sqrt{n}\,] \tag{4.34}$$

is arrived at $\alpha_1 = \alpha/2$. This shortest CI is also equal-tail CI of μ.

2) On the other hand, the CI in (4.34) is symmetric about $\bar{X}$. In addition $\phi(\cdot)$ is symmetric about y-axis, so the CI is also an equal-height CI. □

Exercise 4

4.1 Show that the distribution of T_{Welch} defined in (4.13) can be approximated by a t-distribution with ν degrees of freedom, where

$$\nu = \left\{\frac{c^2}{n_1-1} + \frac{(1-c)^2}{n_2-1}\right\}^{-1} \quad \text{and} \quad c = \frac{S_1^2/n_1}{S_1^2/n_1 + S_2^2/n_2}.$$

[HINT: Let $W = S_1^2/n_1 + S_2^2/n_2$ and approximate W/g by $\chi^2(f)$, where g and f are two unknown quantities, which can be determined by moment methods]

4.2 Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Poisson}(\lambda)$. Let $n = 100$, $\lambda = 6.25$, and $\alpha = 0.05$.

(a) Find an approximate equal-tail $100(1-\alpha)\%$ CI for λ based on (4.22).

(b) By replacing $z_{\alpha/2}$ and $-z_{\alpha/2}$ with z_{α_2} and $z_{1-\alpha+\alpha_2}$ in (4.22), respectively, find the shortest $100(1-\alpha)\%$ CI for λ, where $0 \leqslant \alpha_2 \leqslant \alpha$.

4.3 Let $3.3, -0.3, -0.6, -0.9$ be a random sample from $N(\mu, \sigma^2)$.

(a) If $\sigma = 3$, find a 90% CI of μ.

(b) What would be the CI of μ if σ were unknown.

4.4 Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, where both μ and σ^2 are unknown. Find the sample size n such that $0.95 = \Pr(L \leqslant \sigma/5)$, where L denotes the length of a 90% CI of μ.

4.5 To test two promising new lines of hybrid corn under normal farming conditions, a seed company selected eight farms at random in Iowa and planted both lines in experimental plots on each farm. The yields (converted to bushels per acre) for the eight locations were

$$\text{Line } A: \quad 86 \quad 87 \quad 56 \quad 93 \quad 84 \quad 93 \quad 75 \quad 79$$

$$\text{Line } B: \quad 80 \quad 79 \quad 58 \quad 91 \quad 77 \quad 82 \quad 74 \quad 66$$

Assuming that the two yields are jointly normally distributed, i.e.,

$$\begin{pmatrix} A \\ B \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix}, \begin{pmatrix} \sigma_A^2 & \rho\sigma_A\sigma_B \\ \rho\sigma_A\sigma_B & \sigma_B^2 \end{pmatrix} \right),$$

estimate the difference between the mean yields by a 95% CI.

4.6 Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x;\theta) = \theta x^{\theta-1}$, where $0 < x < 1$ and $\theta > 0$ is the unknown parameter.

(a) Find a pivot, and use it to find a $100(1-\alpha)\%$ equal-tail CI of θ.

(b) Construct the $100(1-\alpha)\%$ shortest CI for θ.

[HINT: Use the result in (4.3)]

Chapter 5

Hypothesis Testing

5.1 Introduction

1• Two important areas of statistical inference

- The first one is the estimation of parameters (Chapters 3 and 4).
- The second is the testing of hypotheses (Chapter 5).

2• What is hypothesis testing?

- Based on observations from a random sample, statisticians follow a formal process to determine whether or not to reject a null hypothesis.
- This process is called hypothesis testing.

2.1• Four steps of a hypothesis testing

— *State the hypotheses.* This involves stating the null and alternative hypotheses. The hypotheses are stated in such a way that they are mutually exclusive. That is, if one is true, the other must be false.

— *Formulate an analysis plan.* The analysis plan describes how to use sample data to evaluate the null hypothesis. The evaluation often focuses around a single test statistic.

— *Analyze sample data.* Find the value of the test statistic (mean score, proportion, t-statistic, z-score, etc.) described in the analysis plan.

— *Interpret results.* Apply the decision rule described in the analysis plan. If the value of the test statistic is unlikely, based on the null hypothesis, reject the null hypothesis.

5.1.1 Several basic notions

3• Simple hypothesis versus composite hypothesis

- A *statistical hypothesis* is an assumption about a population parameter. This assumption may or may not be true.
- A researcher might conduct a statistical experiment to test the validity of this hypothesis.
- If the statistical hypothesis completely specifies the population distribution, it is called a *simple hypothesis*.
- Otherwise, it is called a *composite hypothesis*.

3.1• An illustration example

— Let $X_1, \ldots, X_n$ be a random sample from a population with pdf/pmf $f(x;\theta)$, where $\theta \in \Theta = [2, \infty)$.

— Consider the following hypotheses:

$$H_0\colon\ \theta = 2 \quad \text{against} \quad H_1\colon\ \theta > 2. \tag{5.1}$$

— Let $\Theta_0 = \{2\}$ and $\Theta_1 = (2, \infty)$, then hypotheses (5.1) can be rewritten as

$$H_0\colon\ \theta \in \Theta_0 \quad \text{against} \quad H_1\colon\ \theta \in \Theta_1. \tag{5.2}$$

— In (5.1) or (5.2), H_0 is a simple hypothesis since it completely specifies the population distribution, while H_1 is a composite hypothesis.

4• Null hypothesis versus alternative hypothesis

- The first, the hypothesis being tested, is called the *null hypothesis*, denoted by H_0. 原假设
- The second is called the *alternative hypothesis*, denoted by H_1 or H_a. 备择假设

不交即可，无需张满完全空间

4.1• Several illustration examples

— In (5.1), H_0 is a simple null hypothesis, while H_1 is a composite alternative hypothesis.

— In the second example, we are testing the simple null hypothesis H_0: $\theta = 0.90$ against the simple alternative hypothesis H_1: $\theta = 0.60$, where θ is the parameter of a binomial population.

— In the third example, we are testing the composite null hypothesis H_0: $\theta \geqslant 4200$ against the composite alternative hypothesis H_1: $\theta < 4000$, where θ is the parameter of an exponential population.

4.2• The decision rule

sequential

— Two opposite hypotheses H_0 and H_1 divide the parameter space Θ into two subsets as shown in (5.2).

不接受不同于拒绝

— The thinking is that if the null hypothesis is false, then the alternative hypothesis is true, and vice versa.

— We often say that H_0 is tested against/versus H_1.

— If the null hypothesis H_0 is not rejected, we say that H_0 is accepted.

5• Rejection region versus acceptance region

- Let $\mathbb{S}$ be the set of all possible values of $\mathbf{x} = (X_1, \ldots, X_n)^\top$.
- A test partitions $\mathbb{S}$ into two subsets: $\mathbb{C}$ and its complement $\mathbb{C}'$. That is, $\mathbb{S} = \mathbb{C} \cup \mathbb{C}'$.

5.1• Rejection/critical region $\mathbb{C}$

拒绝域和接受域

— We reject H_0 or accept H_1 if $\boldsymbol{x} = (x_1, \ldots, x_n)^\top \in \mathbb{C}$.

5.2• Acceptance region $\mathbb{C}'$

— We accept H_0 or reject H_1 if $\boldsymbol{x} = (x_1, \ldots, x_n)^\top \in \mathbb{C}'$.

5.1.2 Type I error and Type Ⅱ error

6• Cost for making a decision

Table 5.1 Correct decision and wrong decision

Being adjudged to be	The man is crimeless	The man commits a crime
Guilty (yes)	Type I error	Correct decision
Guiltless (no)	Correct decision	Type Ⅱ error

6.1• The rule of thumb

— Any decision/action will have two outcomes: correct decision or wrong decision.

— We should enhance the probability of making correct decision and reduce the probability of making a wrong decision.

7• Why need we distinguish two kinds of decision error?

- Depending on the seriousness in practice, decision errors can be classified into Type I error (α) and Type Ⅱ error (β).
- For example,

  $\alpha = \Pr(\text{being adjudged to be guilty} \mid \text{the man is crimeless})$ and

  $\beta = \Pr(\text{being adjudged to be guiltless} \mid \text{the man commits a crime})$.
- In this case, the Type I error is more serious than the Type Ⅱ error.
- Therefore, we should control α within an acceptable level (say, $\leqslant 5\%$), and minimize β.

7.1• Type I error, Type I error function and Type I error rate

— Rejection of the null hypothesis H_0 when it is true is called *Type I error*.

— The probability of making a Type I error is denoted by

$$\begin{aligned} \alpha(\theta) = \Pr(\text{Type I error}) &= \Pr(\text{rejecting } H_0 \mid H_0 \text{ is true}) \\ &= \Pr(\mathbf{x} \in \mathbb{C} \mid \theta \in \Theta_0), \end{aligned} \tag{5.3}$$

which is a function of θ defined in Θ_0.

— Usually, $\alpha(\theta)$ is called the *Type I error function.*

— When $\Theta_0 = \{\theta_0\}$, $\alpha(\theta) = \alpha(\theta_0) \hat{=} \alpha$ is called the *Type I error rate.*

7.2• Type II error, Type II error function and Type II error rate

— Acceptance of the H_0 when it is false is called *Type II error.*

— The probability of making a Type II error is denoted by

$$\begin{aligned} \beta(\theta) = \Pr(\text{Type II error}) &= \Pr(\text{accepting } H_0 \mid H_0 \text{ is false}) \\ &= \Pr(\mathbf{x} \in \mathbb{C}' \mid \theta \in \Theta_1), \end{aligned} \tag{5.4}$$

which is a function of θ defined in Θ_1.

— $\beta(\theta)$ is called the *Type II error function.*

— When $\Theta_1 = \{\theta_1\}$, $\beta(\theta) = \beta(\theta_1) \hat{=} \beta$ is called the *Type IIerror rate.*

7.3• Summary of these notions

Table 5.2 Type I error function and Type II error function

	H_0 is true ($\theta \in \Theta_0$)	H_0 is false ($\theta \in \Theta_1$)
Reject H_0 ($\boldsymbol{x} \in \mathbb{C}$)	$\alpha(\theta)$	Correct decision
Accept H_0 ($\boldsymbol{x} \in \mathbb{C}'$)	Correct decision	$\beta(\theta)$

Example 5.1 (Discrete distribution). Let θ be the recovery probability of a new medication and X denote the observed number of recoveries in $n = 20$ trials. Suppose that a manufacturer of the medication wants to test H_0: $\theta = 0.90$ against H_1: $\theta = 0.60$. His test statistic is X and he will accept H_0 if $x > 14$; otherwise, he will reject it. Find α and β.

Solution. Note that $X \sim \text{Binomial}(20, \theta)$, then $x = 0, 1, \ldots, 20$. The acceptance region for the null hypothesis is $\mathbb{C}' = \{x\colon\ x = 15, 16, 17, 18, 19, 20\}$, and the rejection region is $\mathbb{C} = \{x\colon\ x = 0, 1, 2, \ldots, 14\}$. We obtain

$$\alpha = \alpha(0.9) = \Pr(X \leqslant 14 | \theta = 0.90) = \sum_{x=0}^{14} \binom{20}{x} 0.9^x 0.1^{20-x} = 0.0113$$

and

$$\beta = \beta(0.6) = \Pr(X > 14|\theta = 0.60) = \sum_{x=15}^{20} \binom{20}{x} 0.6^x 0.4^{20-x} = 0.1255. \quad \|$$

Example 5.2 (Continuous distribution). Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma_0^2)$, where σ_0^2 is known. Suppose that we want to test H_0: $\mu = \mu_0$ against H_1: $\mu = \mu_1$.

1) Find the value of K such that $\mathbb{C} = \{\boldsymbol{x} = (x_1, \ldots, x_n)^\top : \bar{x} > K\}$ provides a critical region of the Type I error rate $\alpha = 0.05$ for sample size n.

2) Determine the minimum sample size n such that the Type II error rate $\beta = 0.06$.

Solution. 1) Note that $\bar{X} \sim N(\mu, \sigma_0^2/n)$, we have

$$\frac{\bar{X} - \mu}{\sigma_0/\sqrt{n}} \overset{\text{d}}{=} Z \sim N(0, 1).$$

Therefore,

$$\begin{aligned} \alpha &\overset{(5.3)}{=} \Pr(\bar{X} > K|\mu = \mu_0) = \Pr\left(\frac{\bar{X} - \mu}{\sigma_0/\sqrt{n}} > \frac{K - \mu}{\sigma_0/\sqrt{n}}\middle|\mu = \mu_0\right) \\ &= \Pr\left(Z > \frac{K - \mu_0}{\sigma_0/\sqrt{n}}\right) = \Pr(Z > z_\alpha), \end{aligned}$$

and we obtain $\frac{K-\mu_0}{\sigma_0/\sqrt{n}} = z_\alpha$ or

$$K = \mu_0 + z_\alpha \cdot \frac{\sigma_0}{\sqrt{n}} = \mu_0 + z_{0.05} \cdot \frac{\sigma_0}{\sqrt{n}} = \mu_0 + 1.645 \cdot \frac{\sigma_0}{\sqrt{n}}. \tag{5.5}$$

2) Based on (5.4), we obtain

$$\begin{aligned} \beta &= \Pr(\bar{X} \leqslant K|\mu = \mu_1) = \Pr\left(\frac{\bar{X} - \mu}{\sigma_0/\sqrt{n}} \leqslant \frac{K - \mu}{\sigma_0/\sqrt{n}}\middle|\mu = \mu_1\right) \\ &= \Pr\left(Z \leqslant \frac{K - \mu_1}{\sigma_0/\sqrt{n}}\right) = 1 - \Pr\left(Z > \frac{K - \mu_1}{\sigma_0/\sqrt{n}}\right); \end{aligned}$$

i.e.,

$$1 - \beta = \Pr\left(Z > \frac{K - \mu_1}{\sigma_0/\sqrt{n}}\right) = \Pr(Z > z_{1-\beta}).$$

Hence, $\frac{K-\mu_1}{\sigma_0/\sqrt{n}} = z_{1-\beta}$. By replacing K with (5.5), we obtain

$$n = \left\{ \frac{\sigma_0(z_\alpha - z_{1-\beta})}{\mu_1 - \mu_0} \right\}^2.$$

When $\sigma_0 = 1$, $\mu_1 = 11$, $\mu_0 = 10$, since $z_\alpha = z_{0.05} = 1.645$ and $z_{1-\beta} = z_{0.94} = -1.55477$, we obtain $n = 10.24$, or 11 rounded up to the nearest integer. ‖

5.1.3 Power function 势函数

8• Definition of the power function

- Let $\mathbb{C}$ be the critical region of a test for testing H_0 against H_1, then the function

$$p(\theta) = \Pr(\text{rejecting } H_0 \mid \theta) = \Pr(\mathbf{x} \in \mathbb{C} \mid \theta) \tag{5.6}$$

  is the *power function* of the test.

- Thus, the values of the power function are the probabilities of rejecting the null hypothesis H_0 for various values of the parameter θ.

8.1• Recall the role of the mean square error in estimation

— The MSE of the estimator $\hat{\theta}$ is defined by $\text{MSE}(\hat{\theta}) = \text{Var}(\hat{\theta}) + \{E(\hat{\theta}) - \theta\}^2$, where θ is the parameter of interest.

— MSE is a measure in evaluating the goodness of an estimator $\hat{\theta}$ or in comparing two competing estimators $\hat{\theta}_1$ and $\hat{\theta}_2$.

— For example, if $\text{MSE}(\hat{\theta}_1) \leqslant \text{MSE}(\hat{\theta}_2)$, then $\hat{\theta}_1$ is better than $\hat{\theta}_2$.

8.2• The role of the power function in hypothesis testing

— The power function plays the same role in hypothesis testing as that (MSE) played in estimation.

— The power function is a golden standard in assessing the goodness of a test T or in comparing two competing tests T_1 and T_2.

8.3• Relationship between power and Type I/II error functions

— When $\theta \in \Theta_0$, we have

$$p(\theta) = \Pr(\mathbf{x} \in \mathbb{C} \mid \theta \in \Theta_0) \overset{(5.3)}{=} \alpha(\theta). \tag{5.7}$$

— When $\theta \in \Theta_1$, we obtain

$$\begin{aligned} p(\theta) &= \Pr(\mathbf{x} \in \mathbb{C} \mid \theta \in \Theta_1) \\ &= 1 - \Pr(\mathbf{x} \in \mathbb{C}' \mid \theta \in \Theta_1) \overset{(5.4)}{=} 1 - \beta(\theta). \end{aligned}$$

— When $\theta \notin \Theta_0 \cup \Theta_1$, the power function $p(\theta)$ is defined by (5.6).

8.4• How to choose a good test?

— Ideally, we should minimize both $\alpha(\theta)$ and $\beta(\theta)$. However, in general, this is impossible.

— In practice, the Type I error is usually more serious than the Type II error.

— We may fix the probability of Type I error at a preassigned (small) level α^* $(0 < \alpha^* < 1)$, then minimize the probability of Type II error.

— That is, consider the tests with

$$\sup_{\theta \in \Theta_0} p(\theta) \overset{(5.7)}{=} \alpha(\theta) \leqslant \alpha^*$$

and choose the one with the probability of Type II error $\beta(\theta)$ being minimized.

8.5• Comparison of two tests T_1 and T_2

— If $\alpha_{T_1}(\theta)$, $\alpha_{T_2}(\theta) \leqslant \alpha^*$ and $\beta_{T_1}(\theta) \leqslant \beta_{T_2}(\theta)$, then T_1 is better than T_2.

Example 5.3 (Example 5.1 revisited). Suppose that we had wanted to test the null hypothesis H_0: $\theta \geqslant 0.9$ against the alternative hypothesis H_1: $\theta < 0.9$. Investigate the power function corresponding to the same test criterion as in Example 5.1, where we accept H_0 if $x > 14$ and reject it if $x \leqslant 14$. As before, x is the observed number of successes (recoveries) in $n = 20$ trials.

Solution. Note that $\Theta_0 = [0.9, 1)$ and $\Theta_1 = (0, 0.9)$, the power function is

$$p(\theta) = \Pr(X \leqslant 14|\theta) = \sum_{x=0}^{14} \binom{20}{x} \theta^x (1-\theta)^{20-x}.$$

Table 5.3 displays the values of $p(\theta)$ for various θ.

Table 5.3 The power function $p(\theta)$

θ	0.00	0.05	0.10	0.15	0.20	0.25	0.30
$p(\theta)$	1.0000	1.0000	1.0000	0.9999	0.9999	0.9999	0.9999
θ	0.35	0.40	0.45	0.50	0.55	0.60	0.65
$p(\theta)$	0.9996	0.9983	0.9935	0.9793	0.9446	0.8744	0.7546
θ	0.70	0.75	0.80	0.85	0.90	0.95	1.00
$p(\theta)$	0.5836	0.3828	0.1957	0.0673	0.0112	0.0003	0.0000

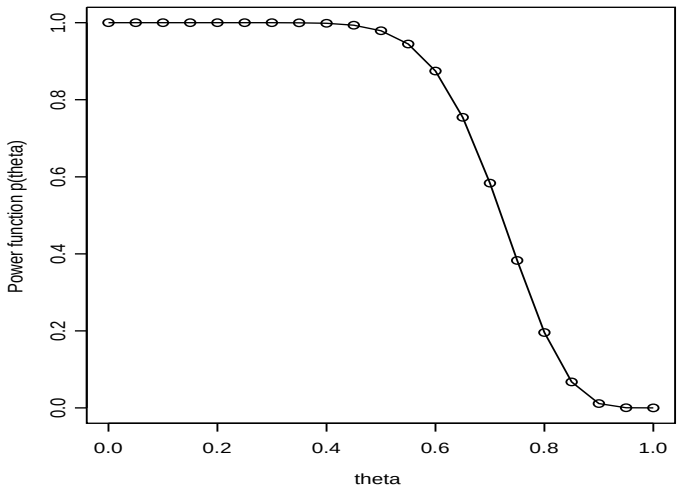

Figure 5.1 The power function $p(\theta)$ in Example 5.3. ‖

5.2 The Neyman–Pearson Lemma

9• Objective of this section

- If there are a set of tests $\{T_j\}_{j=1}^{\infty}$ to test H_0 against H_1, we would like to identify the *most powerful test* (MPT) for the case where both H_0 and H_1 are simple, and

一致最优检验

- to identify *the uniformly most powerful test* (UMPT) for the case where both H_0 and H_1 are composite.

5.2.1 Simple null hypothesis versus simple alternative

10• Definition of size of a test

Definition 5.1 (Size of a test). Consider to test H_0: $\theta \in \Theta_0$ against H_1: $\theta \in \Theta_1 = \Theta - \Theta_0$. A test φ with critical region $\mathbb{C}$ is said to have size α if

$$\sup_{\theta\in\Theta_0} p_\varphi(\theta) = \sup_{\theta\in\Theta_0} \Pr(\mathbf{x} \in \mathbb{C} \mid \theta) = \sup_{\theta\in\Theta_0} \alpha_\varphi(\theta) = \alpha, \tag{5.8}$$

where $\mathbf{x} = (X_1, \ldots, X_n)^\top$. ‖

10.1• A special case

— When H_0 is a simple null hypothesis; i.e., when $\Theta_0 = \{\theta_0\}$, from (5.8), we have

$$\sup_{\theta\in\Theta_0} p_\varphi(\theta) = p_\varphi(\theta_0) = \alpha_\varphi(\theta_0) = \alpha.$$

— In this case, the size of a test is identical to the Type I error rate.

11• What is the most powerful test?

Definition 5.2 (Most powerful test). A test φ with critical region $\mathbb{C}$ is said to be the *most powerful test* with size α for testing H_0: $\theta = \theta_0$ against H_1: $\theta = \theta_1$, if

$$\text{(i)} \qquad p_\varphi(\theta_0) = \alpha, \tag{5.9}$$

and

$$\text{(ii)} \qquad p_\varphi(\theta_1) \geqslant p_\psi(\theta_1) \tag{5.10}$$

for any other test ψ with

$$p_\psi(\theta_0) \leqslant \alpha, \tag{5.11}$$

where θ_0 and θ_1 are two given values. ‖

11.1• Some remarks on Definition 5.2

— A test φ is the most powerful of size α if it has size α and if among all other tests with size α or less it has the largest power.

— Note that (5.10) implies $\beta_\varphi(\theta_1) = 1 - p_\varphi(\theta_1) \leqslant 1 - p_\psi(\theta_1) = \beta_\psi(\theta_1)$.

— Then the test φ has the minimum of the probability of Type II error.

— Moreover, $\mathbb{C}$ is called the best critical region of size α.

12• How to construct the most powerful test?

- A so-called Neyman–Pearson Lemma can be used to find the most powerful test with size α.

12.1• Karl Pearson (27 March 1857 – 27 April 1936)

— He was an English mathematician and biostatistician.

— He has been credited with establishing the discipline of mathematical statistics.

— He was the founder of the world's first university statistics department at University College London in 1911.

— He is known for Pearson distribution, Pearson's r, Pearson's χ^2 test, phi coefficient, and principal components analysis.

12.2• Egon S. Pearson (11 August 1895 – 12 June 1980)

— He was the only son of Karl Pearson and, like his father, a leading British statistician.

— He went to Winchester School and Trinity College, Cambridge, and succeeded his father as professor of statistics at University College London and as editor of the journal Biometrika.

— He is best known for development of the Neyman–Pearson Lemma of statistical hypothesis testing.

12.3• Jerzy Neyman (16 April 1894 – 05 August 1981)

— He was a Polish American mathematician and statistician.

— He spent a couple of years in London and Paris on a fellowship to study statistics with Karl Pearson and Émile Borel, and spent most of his career at the University of California, Berkeley.

— He first introduced the modern concept of a CI into statistics.

Lemma 5.1 (Neyman–Pearson Lemma). Assume that $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x;\theta)$. Let the likelihood function be $L(\theta) = L(\theta; \boldsymbol{x})$. Then a test φ with critical region

$$\mathbb{C} = \left\{ \boldsymbol{x} = (x_1, \ldots, x_n)^\top \colon \frac{L(\theta_0)}{L(\theta_1)} \leqslant k \right\} \tag{5.12}$$

and size α is the most powerful test of size α for testing $H_0\colon \theta = \theta_0$ against $H_1\colon \theta = \theta_1$, where k is a value determined by the size α. ‖

Proof. The test φ is of size α, hence (5.9) holds. Now we consider any test ψ with critical region $\mathbb{A}$ satisfying (5.11), we would prove that (5.10) holds. From (5.12), we have

$$\mathbb{C}' = \left\{ \boldsymbol{x} = (x_1, \ldots, x_n)^\top \colon \frac{L(\theta_0)}{L(\theta_1)} > k \right\}. \tag{5.13}$$

Note that

$$\mathbb{C} = \mathbb{C} \cap \mathbb{A}' + \mathbb{C} \cap \mathbb{A} \quad \text{and} \quad \mathbb{A} = \mathbb{A} \cap \mathbb{C}' + \mathbb{A} \cap \mathbb{C}, \tag{5.14}$$

we obtain

$$\begin{aligned}
& p_\varphi(\theta_1) - p_\psi(\theta_1) \\
= \;& \Pr(X \in \mathbb{C}|\theta_1) - \Pr(X \in \mathbb{A}|\theta_1) \;=\; \int_{\mathbb{C}} L(\theta_1)\,\mathrm{d}x - \int_{\mathbb{A}} L(\theta_1)\,\mathrm{d}x \\
\overset{(5.14)}{=} \;& \int_{\mathbb{C}\cap\mathbb{A}'} L(\theta_1)\,\mathrm{d}x - \int_{\mathbb{A}\cap\mathbb{C}'} L(\theta_1)\,\mathrm{d}x \qquad \text{[from (5.12) and (5.13)]} \\
\geqslant \;& \frac{1}{k}\int_{\mathbb{C}\cap\mathbb{A}'} L(\theta_0)\,\mathrm{d}x - \frac{1}{k}\int_{\mathbb{A}\cap\mathbb{C}'} L(\theta_0)\,\mathrm{d}x \\
= \;& \frac{1}{k}\int_{\mathbb{C}} L(\theta_0)\,\mathrm{d}x - \frac{1}{k}\int_{\mathbb{A}} L(\theta_0)\,\mathrm{d}x \\
= \;& \frac{1}{k}\Big\{ \Pr(X \in \mathbb{C}|\theta_0) - \Pr(X \in \mathbb{A}|\theta_0) \Big\} \qquad \text{[from (5.9) and (5.11)]} \\
\geqslant \;& \frac{1}{k}(\alpha - \alpha) = 0.
\end{aligned}$$

□

Example 5.4 (Continuous case). Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\theta, 1)$.

1) Find the most powerful test of size α for testing H_0: $\theta = \theta_0$ against H_1: $\theta = \theta_1\ (> \theta_0)$.

2) Find the power function.

Solution. 1) Let φ be a test with critical region satisfying (5.12). The likelihood function is given by

$$L(\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(x_i - \theta)^2\right\} = (2\pi)^{-n/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^2\right\}.$$

Then

$$\begin{aligned} \frac{L(\theta_0)}{L(\theta_1)} &= \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\{(x_i - \theta_0)^2 - (x_i - \theta_1)^2\}\right] \\ &= \exp\left\{(\theta_0 - \theta_1)\sum_{i=1}^{n} x_i - n(\theta_0^2 - \theta_1^2)/2\right\} \leqslant k \end{aligned}$$

is equivalent to

$$\bar{x} \geqslant \frac{\log(k)}{n(\theta_0 - \theta_1)} + \frac{\theta_0 + \theta_1}{2} \;\hat{=}\; c.$$

To determine c, we consider the size

$$\begin{aligned} \alpha &= \Pr(\bar{X} \geqslant c \mid \theta = \theta_0) \\ &= \Pr\{\sqrt{n}(\bar{X} - \theta_0) \geqslant \sqrt{n}(c - \theta_0)\} \\ &= \Pr\{Z \geqslant \sqrt{n}(c - \theta_0)\} \\ &= \Pr(Z \geqslant z_\alpha). \end{aligned}$$

Then, $\sqrt{n}(c - \theta_0) = z_\alpha$ or $c = \theta_0 + z_\alpha/\sqrt{n}$. Thus, the test with critical region $\mathbb{C} = \{\boldsymbol{x}\colon\ \bar{x} \geqslant \theta_0 + z_\alpha/\sqrt{n}\,\}$ is the most powerful test of size α.

2) The power function

$$\begin{aligned} p_\varphi(\theta) &= \Pr(\mathbf{x} \in \mathbb{C} \mid \theta) \\ &= \Pr(\bar{X} \geqslant \theta_0 + z_\alpha/\sqrt{n} \mid \theta) \\ &= \Pr\{\sqrt{n}(\bar{X} - \theta) \geqslant z_\alpha + \sqrt{n}(\theta_0 - \theta) \mid \theta\} \\ &= \Pr\{Z \geqslant z_\alpha + \sqrt{n}(\theta_0 - \theta)\} \\ &= 1 - \Phi\{z_\alpha + \sqrt{n}(\theta_0 - \theta)\} \end{aligned} \tag{5.15}$$

is monotonic increasing in θ, which is shown in Figure 5.2.

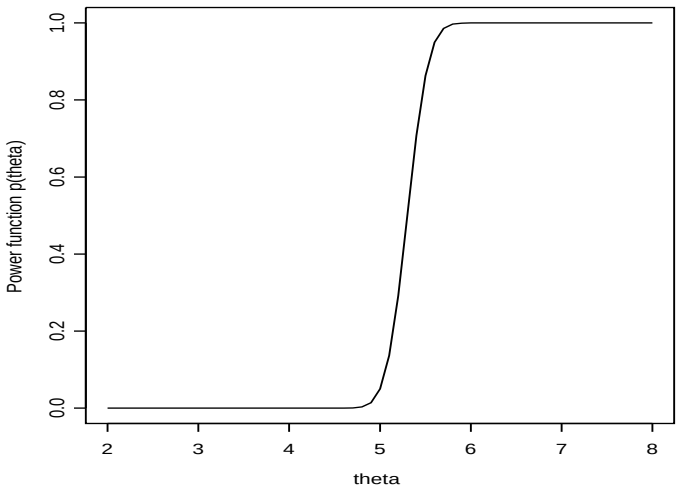

Figure 5.2 The power function $p_\varphi(\theta)$ defined in (5.15). ‖

12.4• A Summary on Example 5.4

— For a simple H_0 and a simple H_1, finding the most powerful test with size α is equivalent to finding its critical region $\mathbb{C}$ defined by (5.12); i.e., finding k or its function.

— Using the definition of size α (or Type I error rate), we can find the k.

Example 5.5 (Discrete case). Let X be a discrete random variable, x be its observation, and the pmf of X be given by

x	1	2	3	4	5	6	7	8	9	10
$f(x;0)$	0	0.58	0.02	0.05	0.03	0.11	0.01	0.07	0.04	0.09
$f(x;1)$	0.6	0	0.06	0.08	0.03	0.01	0.04	0.12	0.02	0.04

1) Find the most powerful test of size 0.1 for testing H_0: $\theta = 0$ against H_1: $\theta = 1$.

2) Calculate the Type Ⅱ error rate.

Solution. 1) First, we calculate the ratio $f(x;0)/f(x;1)$ for all x; i.e.,

x	1	2	3	4	5	6	7	8	9	10
$\frac{f(x;0)}{f(x;1)}$	0	∞	$\frac{1}{3}$	$\frac{5}{8}$	1	11	$\frac{1}{4}$	$\frac{7}{12}$	2	$\frac{9}{4}$

Second, sort the likelihood ratios in increasing order

x	1	7	3	8	4	5	9	10	6	2
$\frac{f(x;0)}{f(x;1)}$	0	$\frac{1}{4}$	$\frac{1}{3}$	$\frac{7}{12}$	$\frac{5}{8}$	1	2	$\frac{9}{4}$	11	∞

By the Neyman–Pearson lemma, a test of size α with critical region

$$\mathbb{C} = \{x\colon\ f(x;0)/f(x;1) \leqslant k\}$$

is the most powerful test of α. The $\mathbb{C}$ should be one of the following sets:

$$\emptyset, \{1\}, \{1,7\}, \{1,7,3\}, \{1,7,3,8\}, \ldots, \{1,7,3,8,\ldots,6,2\}.$$

We can use the size α to determine the $\mathbb{C}$. Now $\alpha = 0.1$, we have

$$\begin{aligned}
\alpha(0) &= \Pr(X \in \mathbb{C}|0) \\
&= \Pr(X=1|0) + \Pr(X=7|0) + \Pr(X=3|0) + \Pr(X=8|0) \\
&= 0 + 0.01 + 0.02 + 0.07 = 0.1.
\end{aligned}$$

We can choose $k \in [7/12, 5/8)$, the critical region is

$$\mathbb{C} = \{x\colon\ f(x;0)/f(x;1) \leqslant k\} = \{1,7,3,8\}.$$

2) The acceptance region is $\mathbb{C}' = \{4,5,9,10,6,2\}$ so that the Type II error rate is given by

$$\begin{aligned}
\beta(1) &= \Pr(X \in \mathbb{C}'|\theta = 1) \\
&= \Pr(X=4|1) + \Pr(X=5|1) + \Pr(X=9|1) \\
&\quad + \Pr(X=10|1) + \Pr(X=6|1) + \Pr(X=2|1) \\
&= 0.08 + 0.03 + 0.02 + 0.04 + 0.01 + 0 = 0.18.
\end{aligned}$$

‖

Example 5.6 (Binomial distribution). Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Binomial}(1, \theta)$. Find the most powerful test of size α for testing H_0: $\theta = \theta_0$ against H_1: $\theta = \theta_1 \, (> \theta_0)$.

Solution. The likelihood function is $L(\theta) = \theta^{n\bar{x}}(1-\theta)^{n-n\bar{x}}$ so that

$$\begin{aligned} \frac{L(\theta_0)}{L(\theta_1)} &= \frac{\theta_0^{n\bar{x}}(1-\theta_0)^{n-n\bar{x}}}{\theta_1^{n\bar{x}}(1-\theta_1)^{n-n\bar{x}}} \\ &= \left\{\frac{\theta_0(1-\theta_1)}{\theta_1(1-\theta_0)}\right\}^{n\bar{x}} \left(\frac{1-\theta_0}{1-\theta_1}\right)^n \leqslant k, \end{aligned} \tag{5.16}$$

where $\bar{x} = (1/n)\sum_{i=1}^n x_i$. Since $\theta_1 > \theta_0$, we have $1-\theta_1 < 1-\theta_0$, $\theta_0(1-\theta_1) < \theta_1(1-\theta_0)$ and

$$\log\left\{\frac{\theta_0(1-\theta_1)}{\theta_1(1-\theta_0)}\right\} < 0.$$

Thus, (5.16) is equivalent to

$$n\bar{x} \geqslant \frac{\log(k) - n\log\left(\dfrac{1-\theta_0}{1-\theta_1}\right)}{\log\left\{\dfrac{\theta_0(1-\theta_1)}{\theta_1(1-\theta_0)}\right\}} \mathrel{\hat{=}} c.$$

Note that $n\bar{X} = \sum_{i=1}^n X_i \sim \text{Binomial}(n, \theta)$. When H_0 is true, we have $n\bar{X} \overset{\text{d}}{=} B_0 \sim \text{Binomial}(n, \theta_0)$. Thus, the c can be determined by the size

$$\alpha = \Pr(n\bar{X} \geqslant c \mid \theta = \theta_0) = \Pr(B_0 \geqslant c).$$

Therefore, a test with critical region $\mathbb{C} = \{\boldsymbol{x}\colon\ n\bar{x} \geqslant c\}$ is the most powerful test of size α.

Let $\theta_0 = 1/4$ and $n = 10$. We must find the c satisfying

$$\alpha = \Pr(B_0 \geqslant c) = \sum_{x=c}^{10} \binom{10}{x} 0.25^x 0.75^{10-x}.$$

If $\alpha = 0.0197$, then $c = 6$, and if $\alpha = 0.0781$, then $c = 5$. For $\alpha = 0.05$, there is no critical region $\mathbb{C}$ and a constant k of the form given in the Neyman–Pearson Lemma.

In this example our random variables are discrete, and for discrete random variables it is not possible to find a k and $\mathbb{C}$ satisfying (5.12) for an arbitrary fixed $\alpha \in (0, 1)$. ‖

5.2.2 Composite hypotheses

13• Definition of the uniformly most powerful test

Definition 5.3 (Uniformly most powerful test). A test φ with critical region $\mathbb{C}$ is said to be the *uniformly most powerful test* (UMPT) of size α for testing H_0: $\theta \in \Theta_0$ against H_1: $\theta \in \Theta_1 = \Theta - \Theta_0$, if

$$\text{(i)} \qquad \sup_{\theta \in \Theta_0} p_\varphi(\theta) = \sup_{\theta \in \Theta_0} \alpha_\varphi(\theta) = \alpha, \tag{5.17}$$

and

$$\text{(ii)} \qquad p_\varphi(\theta) \geqslant p_\psi(\theta) \tag{5.18}$$

for all $\theta \in \Theta_1$ and for any other test ψ with $\sup_{\theta \in \Theta_0} p_\psi(\theta) \leqslant \alpha$. ‖

13.1• Four comments on UMPT

— A test φ is the UMPT of size α if it has size α and if among all other tests with size α or less it has the largest power for all alternative values of θ.

— The adverb 'uniformly' refers to 'all' alternative θ values.

— The critical region $\mathbb{C}$ of the UMPT *does not* depend on any θ in Θ_1.

— The UMPT may not exist for some testing problems. However, when it does exist, we can see that it is quite a nice test since among all tests with size α or less it has the greatest chance of rejecting H_0 when H_0 is false.

13.2• What is the difference between the MPT and the UMPT?

— For the MPT, both H_0: $\theta = \theta_0$ and H_1: $\theta = \theta_1$ are simple hypotheses;

— while for the UMPT, both H_0: $\theta \in \Theta_0$ and H_1: $\theta \in \Theta_1$ are composite hypotheses.

14• How to find the UMPT?

Example 5.7 (Example 5.4 revisited). Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\theta, 1)$. Find the UMPT of size α for testing H_0: $\theta \leqslant \theta_0$ against H_1: $\theta > \theta_0$.

Solution. Step 1: From Example 5.4, the test φ with critical region

$$\mathbb{C} = \{\boldsymbol{x}\text{: } \bar{x} \geqslant \theta_0 + z_\alpha/\sqrt{n}\,\}$$

is the MPT of size α for testing H_{0s}: $\theta = \theta_0$ against H_{1s}: $\theta = \theta_1 > \theta_0$. Therefore, we obtain

$$p_\varphi(\theta_0) \overset{(5.9)}{=} \alpha \qquad \text{and} \tag{5.19}$$

$$p_\varphi(\theta) \overset{(5.15)}{=} 1 - \Phi\{z_\alpha + \sqrt{n}(\theta_0 - \theta)\}. \tag{5.20}$$

Step 2: Since the critical region $\mathbb{C}$ depends only on n, θ_0, α and the fact $\theta_1 > \theta_0$, but not on the value of θ_1, the test φ is also the UMPT of size α for testing

$$H_{0s}\text{: } \theta = \theta_0 \quad \text{against} \quad H_1\text{: } \theta > \theta_0. \tag{5.21}$$

Thus, we only need to prove both (5.17) and (5.18). The latter can be obtained immediately since φ is the UMPT of size α for testing (5.21). Thus, we only need to prove (5.17).

Step 3: It follows from (5.20) that

$$\begin{aligned}
\sup_{\theta \in \Theta_0} p_\varphi(\theta) &= \sup_{\theta \leqslant \theta_0} [1 - \Phi\{z_\alpha + \sqrt{n}(\theta_0 - \theta)\}] \\
&= \max_{\theta \leqslant \theta_0} [1 - \Phi\{z_\alpha + \sqrt{n}(\theta_0 - \theta)\}] \\
&= 1 - \min_{\theta \leqslant \theta_0} \Phi\{z_\alpha + \sqrt{n}(\theta_0 - \theta)\} \\
&= 1 - \Phi(z_\alpha) \qquad [\because \Phi(-x) \text{ is a decreasing function of } x] \\
&= 1 - (1 - \alpha) \;=\; \alpha \overset{(5.19)}{=} p_\varphi(\theta_0),
\end{aligned}$$

implying (5.17). Then, the test φ is also the UMPT of size α for testing

$$H_0\text{: } \theta \leqslant \theta_0 \quad \text{against} \quad H_1\text{: } \theta > \theta_0. \qquad \|$$

14.1• Two other cases

— Similarly, we can find the UMPT of size α for testing

$$H_0\text{: } \theta \geqslant \theta_0 \quad \text{against} \quad H_1\text{: } \theta < \theta_0.$$

— However, there is no UMPT of size α for testing

$$H_0\colon\ \theta = \theta_0 \quad \text{against} \quad H_1\colon\ \theta \neq \theta_0.$$

Example 5.8 (Normal distribution with known mean). Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu_0, \sigma^2)$. Find the UMPT of size α for testing $H_0\colon\ \sigma^2 \leqslant \sigma_0^2$ against $H_1\colon\ \sigma^2 > \sigma_0^2$.

Solution. Step 1: We consider to test

$$H_{0s}\colon\ \sigma^2 = \sigma_0^2 \quad \text{against} \quad H_{1s}\colon\ \sigma^2 = \sigma_1^2 > \sigma_0^2. \tag{5.22}$$

The likelihood function is

$$L(\sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{\sum_{i=1}^n (x_i - \mu_0)^2}{2\sigma^2}\right\},$$

so that $\sum_{i=1}^n (X_i - \mu_0)^2$ is a sufficient statistic of σ^2. Then

$$\frac{L(\sigma_0^2)}{L(\sigma_1^2)} = \left(\frac{\sigma_1^2}{\sigma_0^2}\right)^{n/2} \exp\left\{\frac{1}{2}\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2}\right)\sum_{i=1}^n (x_i - \mu_0)^2\right\} \leqslant k$$

is equivalent to

$$\sum_{i=1}^n (x_i - \mu_0)^2 \geqslant \frac{2\log(k) - n\log(\sigma_1^2/\sigma_0^2)}{\sigma_1^{-2} - \sigma_0^{-2}} \ \hat{=}\ c,$$

where $\sigma_1^{-2} - \sigma_0^{-2} < 0$ since $\sigma_1^2 > \sigma_0^2$.

Note that $\sum_{i=1}^n (X_i - \mu_0)^2/\sigma^2 \sim \chi^2(n)$. When H_{0s} in (5.22) is true, we have

$$\frac{\sum_{i=1}^n (X_i - \mu_0)^2}{\sigma_0^2} \sim \chi^2(n).$$

Thus, the c can be determined by the size

$$\begin{aligned}
\alpha &= \Pr\left\{\sum_{i=1}^n (X_i - \mu_0)^2 \geqslant c \middle| \sigma^2 = \sigma_0^2\right\} \\
&= \Pr\left\{\frac{\sum_{i=1}^n (X_i - \mu_0)^2}{\sigma^2} \geqslant \frac{c}{\sigma^2} \middle| \sigma^2 = \sigma_0^2\right\} \\
&= \Pr\left\{\frac{\sum_{i=1}^n (X_i - \mu_0)^2}{\sigma_0^2} \geqslant \frac{c}{\sigma_0^2}\right\} \\
&= \Pr\{\chi^2(n) \geqslant c/\sigma_0^2\} \\
&= \Pr\{\chi^2(n) \geqslant \chi^2(\alpha, n)\},
\end{aligned}$$

i.e., $c = \sigma_0^2\chi^2(\alpha, n)$. Therefore, by the Neyman–Pearson lemma, a test φ with critical region

$$\mathbb{C} = \{\boldsymbol{x}\colon\ \textstyle\sum_{i=1}^n (x_i - \mu_0)^2 \geqslant \sigma_0^2\chi^2(\alpha, n)\}$$

is the MPT of size α for testing (5.22).

Step 2: Since the critical region $\mathbb{C}$ depends only on n, σ_0^2, α and the fact $\sigma_1^2 > \sigma_0^2$, but not on the value of σ_1^2, the test φ is also the UMPT of size α for testing

$$H_{0s}\colon\ \sigma^2 = \sigma_0^2 \quad \text{against} \quad H_1\colon\ \sigma^2 > \sigma_0^2.$$

Step 3: The supremum of the power function is given by

$$\begin{aligned}
\sup_{\sigma^2 \leqslant \sigma_0^2} p_\varphi(\sigma^2) &= \sup_{\sigma^2 \leqslant \sigma_0^2} \Pr\{\textstyle\sum_{i=1}^n (X_i - \mu_0)^2 \geqslant c | \sigma^2\} \\
&= \sup_{\sigma^2 \leqslant \sigma_0^2} \Pr\left\{ \frac{\sum_{i=1}^n (X_i - \mu_0)^2}{\sigma^2} \geqslant \frac{c}{\sigma^2} \right\} \\
&= \sup_{\sigma^2 \leqslant \sigma_0^2} \Pr\{\chi^2(n) \geqslant c/\sigma^2\} \\
&= \max_{\sigma^2 \leqslant \sigma_0^2} [1 - \Pr\{\chi^2(n) < c/\sigma^2\}] \\
&= 1 - \min_{\sigma^2 \leqslant \sigma_0^2} \Pr\{\chi^2(n) < c/\sigma^2\} \\
&= 1 - \Pr\{\chi^2(n) < c/\sigma_0^2\} \\
&= 1 - \Pr\{\chi^2(n) < \chi^2(\alpha, n)\} \\
&= \Pr\{\chi^2(n) \geqslant \chi^2(\alpha, n)\} \\
&= \alpha = p_\varphi(\sigma_0^2),
\end{aligned}$$

where the fact that $\Pr\{\chi^2(n) < c/\sigma^2\}$ is a decreasing function of σ^2 is utilized. Then, the test φ is also the UMPT of size α for testing

$$H_0\colon\ \sigma^2 \leqslant \sigma_0^2 \quad \text{against} \quad H_1\colon\ \sigma^2 > \sigma_0^2. \qquad \|$$

14.2• A summary of three steps for finding the UMPT

— Step 1: Given two composite hypotheses $H_0\colon\ \theta \in \Theta_0$ and $H_1\colon\ \theta \in \Theta_1$, we first consider two simple hypotheses $H_{0s}\colon\ \theta = \theta_0 \in \Theta_0$ and $H_{1s}\colon\ \theta = \theta_1 \in \Theta_1$. By using the Neyman–Pearson Lemma, we can find the MPT φ of size α with critical region $\mathbb{C}$.

— Step 2: If the $\mathbb{C}$ is free from θ_1, then φ is the UMPT of size α for testing H_{0s}: $\theta = \theta_0 \in \Theta_0$ against H_1: $\theta \in \Theta_1$.

— Step 3: If $\sup_{\theta\in\Theta_0} p_\varphi(\theta) = \alpha = p_\varphi(\theta_0)$, then φ is the UMPT of size α for testing H_0: $\theta \in \Theta_0$ against H_1: $\theta \in \Theta_1$.

5.3 Likelihood Ratio Test

15• Why need we the likelihood ratio test?

- The Neyman–Pearson Lemma provides a means of constructing the most powerful critical region for testing a simple H_0 against a simple H_1, but it does not always apply to composite hypotheses.
- For example, there is no UMPT of size α for testing the two-sided hypotheses: H_0: $\theta = \theta_0$ against H_1: $\theta \neq \theta_0$ for the normal population.
- If a UMPT of size α does not exist, we may employ the *likelihood ratio test* (LRT).

15.1• Difference between LRT and MPT/UMPT

— The LRT is a general method/tool for finding a test statistic or constructing the critical region of a test, it can be applied to any H_0 and H_1, but the resulting test may not be optimal;

— while the MPT or UMPT emphasizes that the derived test of size α has the *highest power* among a class of tests with size less than or equal to α, but UMPT may not exist.

5.3.1 Likelihood ratio statistic

16• Distinguishing the whole parameter space from $\Theta_0 \cup \Theta_1$

- Suppose that we wish to test H_0: $\theta \in \Theta_0$ against H_1: $\theta \in \Theta_1$, where Θ_0 and Θ_1 are disjoint; i.e., $\Theta_0 \cap \Theta_1 = \emptyset$.
- Let Θ^* be the parameter space. Define $\Theta \hat{=} \Theta_0 \cup \Theta_1$, then $\Theta \subseteq \Theta^*$.
- For instance, in Example 5.9, $\Theta = \Theta^*$; while in Example 5.10, $\Theta \subset \Theta^*$.

16.1• Definition of LR statistic

— Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x;\theta)$ and $x_1, \ldots, x_n$ be their realizations.

— Define $\mathbf{x} = (X_1, \ldots, X_n)^\top$, $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$, and

$$L(\theta) = L(\theta; \boldsymbol{x}) = \prod_{i=1}^{n} f(x_i;\theta)$$

is the likelihood function of θ.

— The ratio

$$\lambda(\boldsymbol{x}) = \frac{\sup_{\theta\in\Theta_0} L(\theta)}{\sup_{\theta\in\Theta} L(\theta)}$$

is referred to as a value of the LR statistic

$$\lambda(\mathbf{x}) = \frac{L(\hat{\theta}^{\mathrm{R}})}{L(\hat{\theta})}, \tag{5.23}$$

where $\hat{\theta}^{\mathrm{R}}$ denotes the restricted MLE of θ in Θ_0.

— If $\Theta = \Theta^*$, then $\hat{\theta}$ in (5.23) is the unrestricted MLE of θ; if $\Theta \subset \Theta^*$, then $\hat{\theta}$ is the restricted MLE of θ in Θ.

— Obviously, $0 < \lambda(\boldsymbol{x}) \leqslant 1$.

16.2• Difference between supremum and maximum

— The short answer is that if the maximum (or minimum) exists, there is no difference. But it is possible that the supremum (or infimum) exists while the maximum (or minimum) does not.

— For example, let $\Phi(x)$ be the cdf of the standard normal distribution, then we have

$$\sup_{x\in\mathbb{R}} \Phi(x) = 1 \quad \text{and} \quad \inf_{x\in\mathbb{R}} \Phi(x) = 0;$$

while neither $\max_{x\in\mathbb{R}} \Phi(x)$ nor $\min_{x\in\mathbb{R}} \Phi(x)$ exists.

— Let $\phi(x) = (1/\sqrt{2\pi})\exp(-x^2/2)$ be the pdf of the $N(0,1)$, then we have

$$\max_{x\in\mathbb{R}} \phi(x) = \sup_{x\in\mathbb{R}} \phi(x) = \frac{1}{\sqrt{2\pi}} \quad \text{and} \quad \inf_{x\in\mathbb{R}} \phi(x) = 0;$$

while $\min_x \phi(x)$ does not exist.

16.3• LR statistic must be a function of a sufficient statistic

— Both $\hat{\theta}^{\mathrm{R}}$ and $\hat{\theta}$ should be function of a sufficient statistic $T = T(\mathbf{x})$ of θ, the LR statistic $\lambda(\mathbf{x})$ or its logarithm depends on $\mathbf{x}$ only through T.

— Therefore, we can write $\log\{\lambda(\mathbf{x})\} = h(T)$ or $\log\{\lambda(\boldsymbol{x})\} = h(t)$ for some function $h(\cdot)$.

— In general, we consider three cases: $h(t)$ is monotone, concave or convex.

5.3.2 Likelihood ratio test

17• When H_0 is true, what would we expect?

- When H_0: $\theta \in \Theta_0$ is true, since θ is not uniformly distributed in Θ_0, we wonder where is the most possible place that θ locates inside Θ_0?
- Intuitively, θ should be near to its MLE since we use its MLE to estimate the true value of θ.
- Note that $\hat{\theta}^{\mathrm{R}}$ is the restricted MLE of $\theta \in \Theta_0$.
- In addition, $\hat{\theta}$ is the restricted MLE of $\theta \in \Theta = \Theta_0 \cup \Theta_1$.
- Therefore, when H_0 is true, $\hat{\theta}^{\mathrm{R}}$ should be the global maximum; i.e., $\hat{\theta}^{\mathrm{R}} = \hat{\theta}$.
- When H_0 is false, $\hat{\theta}^{\mathrm{R}}$ is the local maximum; while $\hat{\theta}$ is the global maximum.

18• Determination of the critical region

- If H_0 is true, we would expect $L(\hat{\theta}^{\mathrm{R}})$ to be close to $L(\hat{\theta})$, so that $\lambda(\boldsymbol{x})$ would be close to 1.
- If $\lambda(\boldsymbol{x})$ is very small, we may suspect the null hypothesis (i.e., would reject H_0).
- Therefore, the critical region that H_0 is rejected is

$$\mathbb{C} = \{\boldsymbol{x}\colon\ \lambda(\boldsymbol{x}) \leqslant \lambda_\alpha\}, \quad 0 < \lambda_\alpha < 1, \tag{5.24}$$

where $\lambda(\boldsymbol{x}) \neq 1$.

- The LRT of size α is a test with critical region $\mathbb{C}$ given by (5.24), and λ_α is determined by

$$\sup_{\theta\in\Theta_0} \Pr\{\lambda(\mathbf{x}) \leqslant \lambda_\alpha|\theta\} = \alpha.$$

18.1• A special case of the LRT

— When testing a simple null hypothesis against a simple alternative hypothesis, the LRT will lead to the same test as that given by the Neyman–Pearson Lemma.

Example 5.9 (Normal population with unit variance). Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\theta, 1)$. Find the LRT of size α for testing H_0: $\theta = \theta_0$ against H_1: $\theta \neq \theta_0$.

Solution. Note that $\Theta_0 = \{\theta_0\}$, $\Theta_1 = (-\infty, \theta_0)\cup(\theta_0, \infty)$, then $\Theta = \Theta_0\cup\Theta_1 = (-\infty, \infty) = \Theta^*$.

Step 1: Calculate $\lambda(\boldsymbol{x})$. Since the likelihood function is

$$L(\theta) = (2\pi)^{-n/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^2\right\},$$

we have

$$\sup_{\theta=\theta_0} L(\theta) = L(\theta_0) = (2\pi)^{-n/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(x_i - \theta_0)^2\right\} \quad \text{and}$$

$$\sup_{\theta\in\Theta} L(\theta) = L(\bar{x}) = (2\pi)^{-n/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2\right\}.$$

Note that

$$\lambda(\boldsymbol{x}) = \frac{L(\theta_0)}{L(\bar{x})} = \exp\left\{-\frac{n(\bar{x} - \theta_0)^2}{2}\right\} \overset{(5.24)}{\leqslant} \lambda_\alpha$$

is equivalent to

$$\sqrt{n}|\bar{x} - \theta_0| \geqslant \sqrt{-2\log(\lambda_\alpha)} \doteq c.$$

Step 2: Find the critical region $\mathbb{C}$. The LRT of size α has the critical region

$$\mathbb{C} = \{\boldsymbol{x}\colon \sqrt{n}|\bar{x} - \theta_0| \geqslant c\}.$$

To determine the c, let $Z \sim N(0, 1)$. From

$$\alpha = \Pr(\mathbf{x} \in \mathbb{C}|\theta = \theta_0) = \Pr(\sqrt{n}|\bar{X} - \theta_0| \geqslant c|\theta_0) = \Pr(|Z| \geqslant c),$$

we obtain $c = z_{\alpha/2}$. Thus, the H_0 must be rejected if $\sqrt{n}|\bar{x} - \theta_0| \geqslant z_{\alpha/2}$. ‖

Example 5.10 (Normal population with mean zero). Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(0, \theta)$ with $\theta > 0$. Find the LRT of size α for testing H_0: $\theta = \theta_0$ against H_1: $\theta < \theta_0$, where $\theta_0 > 0$.

Solution. Note that $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = (0, \theta_0)$, then $\Theta = \Theta_0 \cup \Theta_1 = (0, \theta_0] \subset (0, \infty) = \Theta^*$.

Step 1: Calculate $\lambda(\boldsymbol{x})$. The likelihood function is

$$L(\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\theta}} \exp\left(-\frac{x_i^2}{2\theta}\right) = (2\pi\theta)^{-n/2} \exp\left(-\frac{nt}{2\theta}\right),$$

where $t \triangleq \sum_{i=1}^{n} x_i^2/n$. It is clear that $T = \sum_{i=1}^{n} X_i^2/n$ is a sufficient statistic of θ and

$$\frac{nT}{\theta} = \sum_{i=1}^{n} \frac{X_i^2}{\theta} \sim \chi^2(n). \tag{5.25}$$

It is easy to see that

$$\sup_{\theta=\theta_0} L(\theta) = L(\theta_0) = (2\pi\theta_0)^{-n/2} \exp\left(-\frac{nt}{2\theta_0}\right).$$

From the log-likelihood

$$\ell(\theta) = -\frac{n}{2}\log(2\pi\theta) - \frac{nt}{2\theta},$$

we obtain

$$\begin{aligned} \frac{\mathrm{d}\ell(\theta)}{\mathrm{d}\theta} &= -\frac{n}{2\theta} + \frac{nt}{2\theta^2}, \quad \text{and} \\ \frac{\mathrm{d}^2\ell(\theta)}{\mathrm{d}\theta^2} &= \frac{n}{\theta^2}\left(\frac{1}{2} - \frac{t}{\theta}\right) < 0 \quad \text{if} \quad \theta < 2t. \end{aligned}$$

Hence, $\ell(\theta)$ is strictly concave when $0 < \theta < 2t$ and has the maximum at $\theta = t$. On the other hand, the restricted MLE $\hat{\theta}$ of θ in $\Theta = (0, \theta_0]$ is

$$\hat{\theta} = \min(t, \theta_0) = \begin{cases} t, & \text{if } t \leqslant \theta_0, \\ \theta_0, & \text{if } t > \theta_0, \end{cases}$$

so that

$$L(\hat{\theta}) = \sup_{\theta \leqslant \theta_0} L(\theta) = \begin{cases} L(t), & \text{if } t \leqslant \theta_0, \\ L(\theta_0), & \text{if } t > \theta_0. \end{cases}$$

We have

$$\lambda(\boldsymbol{x}) = \frac{L(\theta_0)}{L(\hat{\theta})} = \begin{cases} \dfrac{L(\theta_0)}{L(t)} = \left(\dfrac{t}{\theta_0}\right)^{n/2} \mathrm{e}^{-\frac{n}{2}(\frac{t}{\theta_0}-1)}, & \text{if } t \leqslant \theta_0, \\ 1, & \text{if } t > \theta_0. \end{cases}$$

Step 2: Find the critical region $\mathbb{C}$. From (5.24), the LRT of size α has the critical region

$$\begin{aligned} \mathbb{C} &= \left\{\boldsymbol{x}\colon\, t \leqslant \theta_0 \quad \text{and} \quad \left(\frac{t}{\theta_0}\right)^{n/2} \mathrm{e}^{-\frac{n}{2}(\frac{t}{\theta_0}-1)} \leqslant \lambda_\alpha\right\} \\ &= \left\{\boldsymbol{x}\colon\, \frac{t}{\theta_0} \leqslant 1 \quad \text{and} \quad \left(\frac{t}{\theta_0}\right)^{n/2} \mathrm{e}^{-\frac{n}{2}(\frac{t}{\theta_0}-1)} \leqslant \lambda_\alpha\right\} \\ &= \{\boldsymbol{x}\colon\, 0 < y \leqslant 1 \quad \text{and} \quad y^{n/2}\,\mathrm{e}^{-n(y-1)/2} \leqslant \lambda_\alpha\} \\ &= \{\boldsymbol{x}\colon\, 0 < y \leqslant 1 \quad \text{and} \quad h(y) \leqslant \lambda_\alpha\}, \end{aligned} \tag{5.26}$$

where $y \,\hat{=}\, t/\theta_0 > 0$ and $h(y) \,\hat{=}\, y^{n/2}\,\mathrm{e}^{-n(y-1)/2}$.

Step 2(a): Check if or not $h(y)$ is log-concave. Define

$$H(y) \,\hat{=}\, \log\{h(y)\} = \frac{n\log(y)}{2} - \frac{n(y-1)}{2}.$$

Letting $H'(y) = n/(2y) - n/2 = 0$, we obtain $y = 1$. In addition $H''(y) = -(n/2)y^{-2} < 0$. Therefore, $h(y)$ is strictly log-concave and has a maximum at $y = 1$.

Step 2(b): Find an equivalent $\mathbb{C}$ *involving* T *and* k. Hence $0 < y \leqslant 1$ and $h(y) \leqslant \lambda_\alpha$ if and only if $y \leqslant k$, where k is a constant satisfying $0 < k \leqslant 1$. Thus, (5.26) becomes

$$\mathbb{C} = \{\boldsymbol{x}\colon\, t/\theta_0 \leqslant k\}, \tag{5.27}$$

where $0 < k \leqslant 1$.

Step 2(c): Find the constant k. When H_0 is true, from (5.25), we have $nT/\theta_0 \sim \chi^2(n)$. The k can be determined by

$$\begin{aligned} \alpha &= \Pr(\mathbf{x} \in \mathbb{C}|H_0 \text{ is true}) = \Pr(T/\theta_0 \leqslant k|H_0 \text{ is true}) \\ &= \Pr(nT/\theta_0 \leqslant nk|H_0 \text{ is true}) = \Pr\{\chi^2(n) \leqslant nk\}, \end{aligned}$$

or equivalently $1-\alpha = \Pr\{\chi^2(n) > nk\} = \Pr\{\chi^2(n) > \chi^2(1-\alpha, n)\}$, we obtain $nk = \chi^2(1-\alpha, n)$ or $k = \chi^2(1-\alpha, n)/n$. The critical region (5.27) becomes

$$\mathbb{C} = \left\{\boldsymbol{x}\colon\ \textstyle\sum_{i=1}^n x_i^2 \leqslant \theta_0 \chi^2(1-\alpha, n)\right\}.$$

The null hypothesis H_0 must be rejected if $\sum_{i=1}^n x_i^2 \leqslant \theta_0 \chi^2(1-\alpha, n)$. ‖

Example 5.11 (Exponential population). Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Exponential}(\theta)$ with pdf $f(x;\theta) = \theta\, \mathrm{e}^{-\theta x}$ for $x > 0$ and $\theta > 0$. Find the LRT of size α for testing $H_0\colon\ \theta \leqslant \theta_0$ against $H_1\colon\ \theta > \theta_0$.

Solution. Note that $\Theta_0 = (0, \theta_0]$ and $\Theta_1 = (\theta_0, \infty)$, then $\Theta = (0, \infty) = \Theta^*$.

Step 1: Calculate $\lambda(\boldsymbol{x})$. The likelihood function is given by

$$L(\theta) = \prod_{i=1}^n \theta\, \mathrm{e}^{-\theta x_i} = \theta^n\, \mathrm{e}^{-\theta n \bar{x}}$$

so that $\bar{X} = (1/n)\sum_{i=1}^n X_i$ is a sufficient statistic of θ. The log-likelihood function is $\ell(\theta) = n \log\theta - \theta n \bar{x}$. We have

$$\ell'(\theta) = \frac{n}{\theta} - n\bar{x} \quad \text{and} \quad \ell''(\theta) = -n\theta^{-2} < 0 \quad \forall \theta \in \Theta.$$

Hence, $\ell(\theta)$ is strictly concave and has the maximum at $\hat{\theta} = 1/\bar{x}$. On the other hand, the restricted MLE $\hat{\theta}^{\mathrm{R}}$ of θ in Θ_0 is given by

$$\hat{\theta}^{\mathrm{R}} = \begin{cases} 1/\bar{x}, & \text{if } \theta_0 \geqslant 1/\bar{x}, \\ \theta_0, & \text{if } \theta_0 < 1/\bar{x}. \end{cases}$$

Therefore,

$$\max_{\theta\in\Theta} L(\theta) = L(\hat{\theta}) = (1/\bar{x})^n\, \mathrm{e}^{-n},$$

and

$$\max_{\theta\leqslant\theta_0} L(\theta) = L(\hat{\theta}^{\mathrm{R}}) = \begin{cases} L(\hat{\theta}) = (1/\bar{x})^n\, \mathrm{e}^{-n}, & \text{if } \theta_0 \geqslant 1/\bar{x}, \\ L(\theta_0) = \theta_0^n\, \mathrm{e}^{-\theta_0 n\bar{x}}, & \text{if } \theta_0 < 1/\bar{x}, \end{cases}$$

so that

$$\lambda(\boldsymbol{x}) = \begin{cases} 1, & \text{if } \theta_0 \geqslant 1/\bar{x}, \\ \dfrac{\theta_0^n\, \mathrm{e}^{-\theta_0 n\bar{x}}}{(1/\bar{x})^n\, \mathrm{e}^{-n}}, & \text{if } \theta_0 < 1/\bar{x}. \end{cases}$$

Step 2: Find the critical region $\mathbb{C}$. From (5.24), the LRT of size α has the critical region

$$\begin{aligned}\mathbb{C} &= \{\boldsymbol{x}\colon\ \theta_0 < 1/\bar{x} \quad \text{and} \quad (\theta_0\bar{x})^n\,\mathrm{e}^{-\theta_0 n\bar{x}+n} \leqslant \lambda_\alpha\} \\ &= \{\boldsymbol{x}\colon\ \theta_0\bar{x} < 1 \quad \text{and} \quad (\theta_0\bar{x})^n\,\mathrm{e}^{-n(\theta_0\bar{x}-1)} \leqslant \lambda_\alpha\} \\ &= \{\boldsymbol{x}\colon\ 0 < y < 1 \quad \text{and} \quad y^n\,\mathrm{e}^{-n(y-1)} \leqslant \lambda_\alpha\} \\ &= \{\boldsymbol{x}\colon\ 0 < y < 1 \quad \text{and} \quad h(y) \leqslant \lambda_\alpha\}, \end{aligned} \tag{5.28}$$

where $y \,\hat{=}\, \theta_0\bar{x} > 0$ and $h(y) \,\hat{=}\, y^n\,\mathrm{e}^{-n(y-1)}$.

Step 2(a): Check if or not $h(y)$ *is log-concave*. Define

$$H(y) \,\hat{=}\, \log\{h(y)\} = n\log(y) - n(y-1).$$

Letting $H'(y) = n/y - n = 0$, we obtain $y = 1$. In addition

$$H''(y) = -ny^{-2} < 0.$$

Therefore, $h(y)$ is strictly log-concave and has a maximum at $y = 1$.

Step 2(b): Find an equivalent $\mathbb{C}$ *involving* $\bar{X}$ *and* k. Hence $0 < y < 1$ and $h(y) \leqslant \lambda_\alpha$ if and only if $y \leqslant k$, where k is a constant satisfying $0 < k < 1$. Thus, (5.28) becomes

$$\mathbb{C} = \{\boldsymbol{x}\colon\ \theta_0\bar{x} \leqslant k\},$$

where $0 < k < 1$.

Step 2(c): Find the constant k. We recall from Example 4.1 that

$$2\theta n\bar{X} \sim \chi^2(2n).$$

From (5.8), the k can be determined by the size

$$\begin{aligned}\alpha &= \sup_{\theta\in\Theta_0} \Pr(\mathbf{x} \in \mathbb{C}|\theta) \\ &= \sup_{\theta\leqslant\theta_0} \Pr(\theta_0\bar{X} \leqslant k|\theta) \\ &= \sup_{\theta\leqslant\theta_0} \Pr(2\theta n\bar{X} \leqslant 2\theta nk/\theta_0|\theta) \\ &= \max_{\theta\leqslant\theta_0} \Pr\{\chi^2(2n) \leqslant 2\theta nk/\theta_0|\theta\} \\ &= \Pr\{\chi^2(2n) \leqslant 2nk\},\end{aligned}$$

or equivalently

$$1-\alpha = \Pr\{\chi^2(2n) > 2nk\} = \Pr\{\chi^2(2n) > \chi^2(1-\alpha, 2n)\},$$

we obtain $2nk = \chi^2(1-\alpha, 2n)$ or $k = \chi^2(1-\alpha, 2n)/2n$. The null hypothesis H_0 is rejected when $\bar{X} \leqslant \chi^2(1-\alpha, 2n)/(2n\theta_0)$. ‖

18.2• A summary of two steps for finding the LRT

— Step 1: Calculate the LR statistic $\lambda(\boldsymbol{x})$, where $\lambda(\mathbf{x}) = h(T)$ with $T = T(\mathbf{x})$ being a sufficient statistic of θ.

— Step 2: Find the critical region $\mathbb{C}$

Step 2(a): Check if or not $h(t)$ is monotone or log-concave;

Step 2(b): Find an equivalent $\mathbb{C}$ involving the sufficient statistic T and a constant k;

Step 2(c): Find the constant k via the definition of size α by noting that a pivotal quantity $P = P(T, \theta)$ follows a certain standard distribution.

5.4 Tests on Normal Means

5.4.1 One–sample normal test when variance is known

19• The issue

- Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma_0^2)$, where σ_0^2 is known.
- Suppose that we want to test the null hypothesis H_0: $\mu = \mu_0$ against one of the alternatives $\mu \neq \mu_0$, $\mu > \mu_0$, or $\mu < \mu_0$.

20• The critical region (CR) approach

20.1• Step 1: To find a test statistic

— Since $\bar{X}$ is a sufficient statistic of μ and the distribution of

$$Z = \frac{\bar{X} - \mu}{\sigma_0/\sqrt{n}}$$

is $N(0,1)$ that does not depend on the unknown parameter μ, we know that Z is a pivotal quantity.

— The test statistic is

$$Z_0 \doteq \frac{\bar{X}-\mu_0}{\sigma_0/\sqrt{n}} = \frac{(\bar{X}-\mu)+(\mu-\mu_0)}{\sigma_0/\sqrt{n}} = Z + \frac{\mu-\mu_0}{\sigma_0/\sqrt{n}}.$$

— When H_0 is true, i.e., $\mu = \mu_0$, we obtain

$$Z_0 = Z \sim N(0,1). \tag{5.29}$$

20.2• Step 2: To determine a critical region of size α

— Since

$$\alpha = \Pr(|Z| \geqslant z_{\alpha/2}), \quad \alpha = \Pr(Z \geqslant z_\alpha), \quad \alpha = \Pr(Z \leqslant -z_\alpha),$$

the critical regions of size α (shown in Figures 5.3, 5.4 and 5.5) for the corresponding alternatives $\mu \neq \mu_0$, $\mu > \mu_0$, or $\mu < \mu_0$ are given by

$$\mathbb{C}_1 = \{\boldsymbol{x}\colon\ |z_0| \geqslant z_{\alpha/2}\}, \quad \mathbb{C}_2 = \{\boldsymbol{x}\colon\ z_0 \geqslant z_\alpha\}, \quad \mathbb{C}_3 = \{\boldsymbol{x}\colon\ z_0 \leqslant -z_\alpha\},$$

respectively, where

$$z_0 = \frac{\bar{x}-\mu_0}{\sigma_0/\sqrt{n}}. \tag{5.30}$$

20.3• The relationship between a pivot and a test statistic

— Let $T = T(\mathbf{x})$ be a sufficient statistic of θ. As a function of both T and θ, a pivotal quantity $P = P(T,\theta)$ is incomputable.

— A test statistic does not depend on the θ, so it is computable.

— If $P = P(T,\theta)$ is a pivot and $H_0\colon\ \theta = \theta_0$, then $P_0 = P(T,\theta_0)$ is the corresponding test statistic.

— When H_0 is true, we have $P_0 = P$.

Example 5.12 (Numerical illustration). Let $X_1,\ldots,X_n \overset{\text{iid}}{\sim} N(\mu,\sigma_0^2)$, and we observed $n = 25$, $\bar{x} = 8.091$, and $\sigma_0 = 0.16$. Test $H_0\colon\ \mu = \mu_0 = 8$ against $H_1\colon\ \mu \neq \mu_0 = 8$ at the 0.01 level of significance.

Solution. Since $\alpha = 0.01$, we have $z_{\alpha/2} = z_{0.005} = 2.575$. We will reject the H_0 if $z_0 \leqslant -2.575$ or $z_0 \geqslant 2.575$, where z_0 is defined by (5.30). Note that $z_0 = 2.84 > 2.575$, the null hypothesis must be rejected. ‖

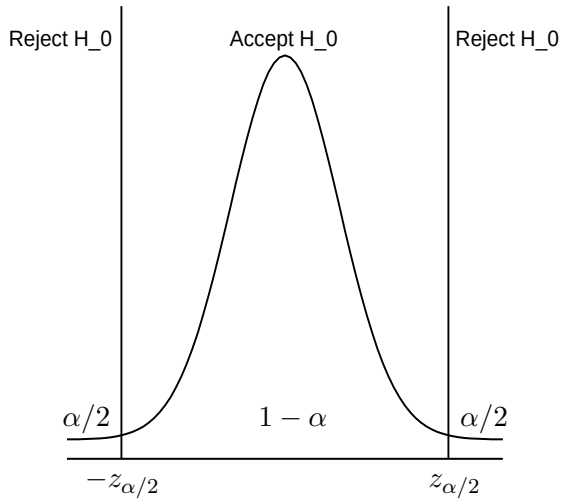

Figure 5.3 The critical region $\mathbb{C}_1$ for a two-tailed test.

21• The p-value approach

- The corresponding p-values can be calculated by

$$\begin{aligned} p\text{-value} &= 2\Pr(Z \geqslant |z_0|) && \text{if } H_1 : \mu \neq \mu_0, \\ &= \Pr(Z^2 \geqslant z_0^2) \\ &= \Pr\{\chi^2(1) \geqslant z_0^2\}, \\ p\text{-value} &= \Pr(Z \geqslant z_0), && \text{if } H_1 : \mu > \mu_0, \\ p\text{-value} &= \Pr(Z \leqslant z_0), && \text{if } H_1 : \mu < \mu_0, \end{aligned}$$

  where Z is specified by (5.29) and z_0 given by (5.30) denotes the observed value of Z_0.

- When p-value $\geqslant \alpha$ (level of significance), we cannot reject the H_0.
- When p-value $< \alpha$, we reject H_0.

Example 5.13 (Example 5.12 revisited). Calculate the p-value for testing H_0: $\mu = \mu_0 = 8$ against H_1: $\mu \neq \mu_0 = 8$ at the 0.01 level of significance.

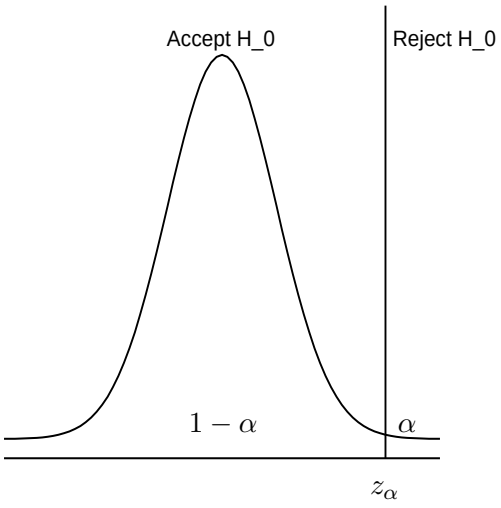

Figure 5.4 The critical region $\mathbb{C}_2$ for a one-tailed test (H_1: $\mu > \mu_0$).

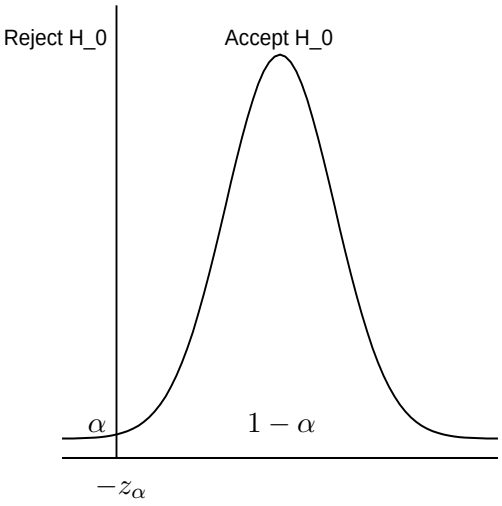

Figure 5.5 The critical region $\mathbb{C}_3$ for a one-tailed test (H_1: $\mu < \mu_0$).

Solution. Since $z_0 = 2.84$ and $\alpha = 0.01$, we have

$$\begin{aligned} p\text{-value} &= 2\Pr(Z \geqslant |z_0|) = 2\Pr(Z \geqslant 2.84) = 2\{1 - \Pr(Z < 2.84)\} \\ &= 2\{1 - \Phi(2.84)\} = 2(1 - 0.9977443) = 0.0045 < 0.01. \end{aligned}$$

Alternatively,

$$p\text{-value} = \Pr\{\chi^2(1) \geqslant 2.84^2\} = 1 - 0.9954886 = 0.0045 < 0.01,$$

so that the H_0 must be rejected. ‖

22• Difference of the CR approach from the p-value approach

- Two approaches should be equivalent.
- The p-value reflects the extent of rejecting/accepting the H_0. For example, we strongly reject H_0 if p-value < 0.001.
- The p-value approach conveys much information and is thus preferred.

5.4.2 One–sample t test

23• The issue

- Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ with unknown variance σ^2.
- Suppose that we want to test the null hypothesis H_0: $\mu = \mu_0$ against one of the alternatives $\mu \neq \mu_0$, $\mu > \mu_0$, or $\mu < \mu_0$.

24• The critical region approach

24.1• Step 1: To find a test statistic

— Since $(\bar{X}, S^2)$ are joint sufficient statistics for (μ, σ^2) and the distribution of

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

is $t(n-1)$ that does not depend on the unknown parameters (μ, σ^2), we know that T is a pivotal quantity.

— The test statistic is

$$T_1 \triangleq \frac{\bar{X} - \mu_0}{S/\sqrt{n}} = \frac{(\bar{X} - \mu) + (\mu - \mu_0)}{S/\sqrt{n}} = T + \frac{\mu - \mu_0}{S/\sqrt{n}}.$$

— When H_0 is true, i.e., $\mu = \mu_0$, we obtain

$$T_1 = T \sim t(n-1). \tag{5.31}$$

24.2• Step 2: To determine a critical region of size α

— Since

$$\begin{aligned} \alpha &= \Pr\{|T| \geqslant t(\alpha/2, n-1)\}, \\ \alpha &= \Pr\{T \geqslant t(\alpha, n-1)\}, \\ \alpha &= \Pr\{T \leqslant -t(\alpha, n-1)\}, \end{aligned}$$

the critical regions of size α for the corresponding alternatives $\mu \neq \mu_0$, $\mu > \mu_0$, or $\mu < \mu_0$ are given by

$$\begin{aligned} \mathbb{C}_1 &= \{\boldsymbol{x}\colon\ |t_1| \geqslant t(\alpha/2, n-1)\}, \\ \mathbb{C}_2 &= \{\boldsymbol{x}\colon\ t_1 \geqslant t(\alpha, n-1)\}, \\ \mathbb{C}_3 &= \{\boldsymbol{x}\colon\ t_1 \leqslant -t(\alpha, n-1)\}, \end{aligned}$$

respectively, where

$$t_1 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}. \tag{5.32}$$

Example 5.14 (Numerical illustration). Let $X_1, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$, and we observed $n = 5$, $\bar{x} = 183.1$, and $s = 8.2$. Test H_0: $\mu = \mu_0 = 185$ against H_1: $\mu < \mu_0 = 185$ at the 0.05 level of significance.

Solution. Since $\alpha = 0.05$, we have $t(\alpha, n-1) = t(0.05, 4) = 2.132$. We will reject the null hypothesis if $t_1 \leqslant -2.132$, where t_1 is defined by (5.32). Since $t_1 = -0.518 > -2.132$, the null hypothesis cannot be rejected. ‖

25• The p-value approach

- The corresponding p-values are given by

$$\begin{aligned} p\text{-value} &= 2\Pr(T \geqslant |t_1|) && \text{if } H_1 : \mu \neq \mu_0, \\ p\text{-value} &= \Pr(T \geqslant t_1), && \text{if } H_1 : \mu > \mu_0, \\ p\text{-value} &= \Pr(T \leqslant t_1), && \text{if } H_1 : \mu < \mu_0, \end{aligned}$$

  where T is specified by (5.31) and t_1 given by (5.32) denotes the observed value of T_1.

Example 5.15 (Example 5.14 revisited). Calculate the p-value for testing H_0: $\mu = \mu_0 = 185$ against H_1: $\mu < \mu_0 = 185$ at the 0.05 level of significance.

Solution. Since $n = 5$, $t_1 = -0.518$ and $\alpha = 0.05$, we have

$$\begin{aligned} p\text{-value} &= \Pr(T \leqslant t_1) = \Pr\{t(n-1) \leqslant -0.518\} \\ &= \Pr\{t(4) \leqslant -0.518\} = 0.3159 > 0.05 \end{aligned}$$

so that the H_0 cannot be rejected. ‖

5.4.3 Two–sample t test

26• The issue

- Let $X_{i1}, \ldots, X_{in_i} \overset{\text{iid}}{\sim} N(\mu_i, \sigma^2)$ for $i = 1, 2$, and the two random samples be independent.
- Suppose that we want to test the null hypothesis H_0: $\mu_1 - \mu_2 = \delta$ against one of the alternative $\mu_1 - \mu_2 \neq \delta$, $\mu_1 - \mu_2 > \delta$, or $\mu_1 - \mu_2 < \delta$.

27• The critical region approach

27.1• Step 1: To find a test statistic

— Set $n' = n_1 + n_2 - 1$. From (4.11), the distribution of

$$T^* = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{1/n_1 + 1/n_2}}$$

is $t(n' - 1)$ that does not depend on the means μ_1, μ_2 and the common unknown variance σ^2. In other words, T^* is a pivotal quantity.

— The test statistic is

$$\begin{aligned} T_2 &\hat{=} \frac{(\bar{X}_1 - \bar{X}_2) - \delta}{S_p\sqrt{1/n_1 + 1/n_2}} \\ &= \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2) + (\mu_1 - \mu_2) - \delta}{S_p\sqrt{1/n_1 + 1/n_2}} \\ &= T^* + \frac{(\mu_1 - \mu_2) - \delta}{S_p\sqrt{1/n_1 + 1/n_2}}, \end{aligned}$$

where

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n'-1}$$

denotes the pooled sample variance.

— When H_0 is true, i.e., $\mu_1 - \mu_2 = \delta$, we obtain

$$T_2 = T^* \sim t(n'-1). \tag{5.33}$$

27.2• Step 2: To determine a critical region of size α

— Since

$$\begin{aligned} \alpha &= \Pr\{|T^*| \geqslant t(\alpha/2, n'-1)\}, \\ \alpha &= \Pr\{T^* \geqslant t(\alpha, n'-1)\}, \\ \alpha &= \Pr\{T^* \leqslant -t(\alpha, n'-1)\}, \end{aligned}$$

the critical regions of size α for the corresponding alternatives $\mu_1 - \mu_2 \neq \delta$, $\mu_1 - \mu_2 > \delta$, or $\mu_1 - \mu_2 < \delta$ are given by

$$\begin{aligned} \mathbb{C}_1 &= \{(\boldsymbol{x}_1, \boldsymbol{x}_2)\colon\ |t_2| \geqslant t(\alpha/2, n'-1)\}, \\ \mathbb{C}_2 &= \{(\boldsymbol{x}_1, \boldsymbol{x}_2)\colon\ t_2 \geqslant t(\alpha, n'-1)\}, \\ \mathbb{C}_3 &= \{(\boldsymbol{x}_1, \boldsymbol{x}_2)\colon\ t_2 \leqslant -t(\alpha, n'-1)\}, \end{aligned}$$

respectively, where $\boldsymbol{x}_i = (X_{i1}, \ldots, X_{in_i})^\top$ for $i = 1, 2$, and

$$t_2 = \frac{(\bar{x}_1 - \bar{x}_2) - \delta}{s_p\sqrt{1/n_1 + 1/n_2}}. \tag{5.34}$$

Example 5.16 (Numerical illustration). Let $X_{i1}, \ldots, X_{in_i}$ be two independent random samples from $N(\mu_i, \sigma^2)$, $i = 1, 2$, and we observed $n_1 = n_2 = 4$, $\bar{x}_1 = 546$, $\bar{x}_2 = 492$, $s_1 = 31$, and $s_2 = 26$. Test H_0: $\mu_1 - \mu_2 = 0$ against H_1: $\mu_1 - \mu_2 > 0$ at the 0.05 level of significance.

Solution. Since $\alpha = 0.05$, we have $t(\alpha, n_1 + n_2 - 2) = t(0.05, 6) = 1.943$. We will reject the H_0 if $t_2 \geqslant 1.943$, where t_2 is defined by (5.34). We calculate

$$s_p = \sqrt{\frac{3(31)^2 + 3(26)^2}{4+4-2}} = 28.609 \quad \text{and} \quad t_2 = \frac{(546-492)-0}{28.609\sqrt{1/4+1/4}} = 2.67.$$

Since $t_2 = 2.67 > 1.943$, the null hypothesis must be rejected. ‖

28• The p-value approach

- The corresponding p-values can be calculated by

$$\begin{aligned} p\text{-value} &= 2\Pr(T^* \geqslant |t_2|) && \text{if } H_1 : \mu_1 - \mu_2 \neq \delta, \\ p\text{-value} &= \Pr(T^* \geqslant t_2), && \text{if } H_1 : \mu_1 - \mu_2 > \delta, \\ p\text{-value} &= \Pr(T^* \leqslant t_2), && \text{if } H_1 : \mu_1 - \mu_2 < \delta, \end{aligned}$$

  where T^* is specified by (5.33) and t_2 given by (5.34) denotes the observed value of T_2.

Example 5.17 (Example 5.16 revisited). Calculate the p-value for testing H_0: $\mu_1 - \mu_2 = 0$ against H_1: $\mu_1 - \mu_2 > 0$ at the 0.05 level of significance.

Solution. Since $n_1 = n_2 = 4$, $t_2 = 2.67$ and $\alpha = 0.05$, we have

$$\begin{aligned} p\text{-value} &= \Pr(T^* \geqslant t_2) = \Pr\{t(n_1 + n_2 - 2) \geqslant 2.67\} \\ &= \Pr\{t(6) \geqslant 2.67\} = 1 - \Pr\{t(6) < 2.67\} \\ &= 1 - 0.9815 = 0.0185 < 0.05, \end{aligned}$$

so that the H_0 must be rejected. ‖

5.5 Goodness of Fit Test

5.5.1 Introduction

29• Background

- Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F(x; \boldsymbol{\theta})$ and $x_1, \ldots, x_n$ denote their realizations, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)^\top$ is the parameter vector.
- In practice, suppose that we have observed a random sample $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$.
- In general, the true cdf $F(x; \boldsymbol{\theta})$ of the population random variable X is always *unknown* to us.

29.1• Assumption

— In the previous sections or chapters, we assumed that $F(x;\boldsymbol{\theta})$ is equal to $F_0(x;\boldsymbol{\theta})$, which is the cdf of a specific distribution (e.g., the normal distribution $N(\mu,\sigma^2)$ with known or unknown parameters).

29.2• Verification

— We wonder if this assumption is valid.

— Therefore, before performing a data analysis, we need to verify this assumption based on the observed data $\boldsymbol{x}$.

29.3• Statistical issue

— Given $\boldsymbol{x}=(x_1,\ldots,x_n)^\top$, we wish to test the null hypothesis

$$H_0\colon\ F(x;\boldsymbol{\theta})=F_0(x;\boldsymbol{\theta}) \tag{5.35}$$

against the alternative hypothesis

$$H_1\colon\ F(x;\boldsymbol{\theta})\neq F_0(x;\boldsymbol{\theta}). \tag{5.36}$$

— To this end, Karl Pearson suggested a chi-square goodness-of-fit test.

30• A review on multinomial distribution

- Let n be a positive integer. If a random vector $(Y_1,\ldots,Y_m)^\top$ has the following joint density

$$f(y_1,\ldots,y_m;p_1,\ldots,p_m)=\binom{n}{y_1,\ldots,y_m}\prod_{j=1}^m p_j^{y_j},$$

where $y_j\geqslant 0$, $\sum_{j=1}^m y_j=n$, $p_j\geqslant 0$ and $\sum_{j=1}^m p_j=1$. We will write

$$(Y_1,\ldots,Y_m)^\top\sim\text{Multinomial}(n;p_1,\ldots,p_m).$$

- The binomial distribution is a special case of the multinomial distribution with $m=2$.

Theorem 5.1 (Large sample chi-square distribution). Let $(N_1, \ldots, N_m)^\top \sim \text{Multinomial}(n; p_1, \ldots, p_m)$, where $n = \sum_{j=1}^m N_j$ and $\sum_{j=1}^m p_j = 1$. Define

$$Q_n = \sum_{j=1}^{m} \frac{(N_j - np_j)^2}{np_j},$$

then Q_n has a limiting distribution, as n approaches infinity, the chi-square distribution with $m-1$ degrees of freedom, i.e.,

$$Q_n \xrightarrow{\text{L}} \chi^2(m-1) \quad \text{as} \quad n \to \infty. \qquad \|$$

Proof. A formal/full proof of this theorem is beyond the scope of this textbook. We give a proof for the case of $m = 2$. Since $N_1 \sim \text{Binomial}(n, p_1)$, then from the central limit theorem, we have

$$Y_n \doteq \frac{N_1 - np_1}{\sqrt{np_1(1-p_1)}} \xrightarrow{\text{L}} N(0,1) \quad \text{as} \quad n \to \infty.$$

Example 2.7 implies $Y_n^2 \xrightarrow{\text{L}} \chi^2(1)$ as $n \to \infty$. Now $N_2 = n - N_1$ and $p_2 = 1 - p_1$, we obtain

$$\begin{aligned} Q_n &= \frac{(N_1 - np_1)^2}{np_1} + \frac{(N_2 - np_2)^2}{np_2} \\ &= \frac{(N_1 - np_1)^2}{np_1} + \frac{(n - N_1 - n + np_1)^2}{n(1-p_1)} \\ &= \frac{(N_1 - np_1)^2}{np_1(1-p_1)} = Y_n^2 \xrightarrow{\text{L}} \chi^2(1) \quad \text{as} \quad n \to \infty. \end{aligned}$$ □

30.1• Understanding Theorem 5.1

— When $n \to \infty$, note that $n = \sum_{j=1}^m N_j$, then the observed value of N_j is very large.

— Since $N_j \sim \text{Binomial}(n, p_j)$, Example 2.18 shows $N_j \dot{\sim} \text{Poisson}(np_j)$, provided that np_j is finite as $n \to \infty$.

— In addition,

$$\frac{N_j - np_j}{\sqrt{np_j}} \dot{\sim} N(0,1) \quad \text{and} \quad \left(\frac{N_j - np_j}{\sqrt{np_j}}\right)^2 \dot{\sim} \chi^2(1).$$

— Q_n asymptotically follows the χ^2 distribution with $m-1$ (rather than m) degrees of freedom, since there is one constraint $n = \sum_{j=1}^m N_j$.

5.5.2 The chi-square test for totally known distribution

31• The issue

- In this subsection, we consider testing (5.35) against (5.36), where both the distribution family $\{F_0(x;\boldsymbol{\theta})\colon \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ and the parameter vector $\boldsymbol{\theta}$ are *totally known*.
- For such situations, we could denote $F_0(x;\boldsymbol{\theta})$ by $F_0(x)$.

32• Pearson's chi-square test

32.1• Partition of the sample space

— To apply Theorem 5.1, we first partition the sample space $\mathbb{S}$ into $\mathbb{A}_1$, …, $\mathbb{A}_m$ such that $\mathbb{A}_1$, …, $\mathbb{A}_m$ are mutually exclusive and $\mathbb{S} = \cup_{j=1}^m \mathbb{A}_j$.

— Next, we take a random sample $X_1, \ldots, X_n$ from the population r.v. X with the true cdf $F(x;\boldsymbol{\theta})$.

— Let N_j denote the number of $X_1, \ldots, X_n$ that fall in the set $\mathbb{A}_j$, then

$$(N_1, \ldots, N_m)^\top \sim \text{Multinomial}(n; p_1, \ldots, p_m),$$

where

$$p_j = \Pr(X \in \mathbb{A}_j) = \int_{\mathbb{A}_j} \mathrm{d}F(x;\boldsymbol{\theta}).$$

— Since $F(x;\boldsymbol{\theta})$ is unknown, we can estimate p_j by $\hat{p}_j = N_j/n$, where $n = \sum_{j=1}^m N_j$.

32.2• Equivalent hypotheses

— Define

$$p_{j0} = \int_{\mathbb{A}_j} \mathrm{d}F_0(x;\boldsymbol{\theta}). \tag{5.37}$$

— Then, testing (5.35) against (5.36) reduces to testing

$$H_0\colon\ p_j = p_{j0}, \qquad \text{for all } j = 1, \ldots, m-1 \quad \text{against} \tag{5.38}$$

$$H_1\colon\ p_j \neq p_{j0}, \qquad \text{for at least one of } j = 1, \ldots, m-1. \tag{5.39}$$

— When H_0 is true, we have

$$Q_n = \sum_{j=1}^{m} \frac{(N_j - np_{j0})^2}{np_{j0}} \dot{\sim} \chi^2(m-1) \quad \text{as} \quad n \to \infty. \tag{5.40}$$

32.3• Remarks on Q_n

— On the one hand, we can rewrite (5.40) as

$$Q_n = \sum_{j=1}^{m} \frac{(O_j - E_j)^2}{E_j},$$

where

$$\begin{aligned} O_j &\hat{=} N_j = \text{observed number of } X_1, \ldots, X_n \text{ that fall in } \mathbb{A}_j \quad \text{and} \\ E_j &\hat{=} np_{j0} = \text{expectation number of } X_1, \ldots, X_n \text{ that fall in } \mathbb{A}_j. \end{aligned}$$

— On the other hand, we can also rewrite (5.40) as

$$Q_n = \sum_{j=1}^{m} \frac{(\hat{p}_j - p_{j0})^2}{p_{j0}/n},$$

which can be viewed as some kind of "standardized distance" or deviance or discrepancy, since $\sum_{j=1}^{m}(\hat{p}_j - p_{j0})^2$ denotes the Euclidean distance between two vectors $(\hat{p}_1, \ldots, \hat{p}_m)^\top = (N_1/n, \ldots, N_m/n)^\top$ and $(p_{10}, \ldots, p_{m0})^\top$.

— Q_n measures how well the distribution $F_0(x; \boldsymbol{\theta})$ fit the observations $\boldsymbol{x}$.

32.4• The critical region

— When H_0 is true, Q_n should be very small, or tends to 0.

— Then, the critical region of the chi-square test is

$$\begin{aligned} \mathbb{C} &= \{(n_1, \ldots, n_m)^\top \colon Q_n \geqslant c\} \\ &= \{(n_1, \ldots, n_m)^\top \colon Q_n \geqslant \chi^2(\alpha, m-1)\}, \end{aligned} \tag{5.41}$$

where $n_1, \ldots, n_m$ are observed values of $N_1, \ldots, N_m$, and c is determined by the size

$$\alpha = \Pr(Q_n \geqslant c | H_0) = \Pr\{\chi^2(m-1) \geqslant \chi^2(\alpha, m-1)\}.$$

— The test based on Q_n is called Pearson's chi-square goodness-of-fit test.

32.5• Other discrepancies

— We may consider the following discrepancies

$$Q_n^* = \sum_{j=1}^{m} \left| \frac{N_j - np_{j0}}{\sqrt{np_{j0}}} \right| \quad \text{and}$$

$$Q_n^{**} = \sum_{j=1}^{m} \hat{p}_j \log\left(\frac{\hat{p}_j}{p_{j0}}\right) = \frac{1}{n}\sum_{j=1}^{m} N_j \log\left(\frac{N_j}{np_{j0}}\right).$$

— However, the first question is how to derive the asymptotic distributions of Q_n^* and Q_n^{**}.

— If we could find their asymptotic distributions, the second question is how to compare the three tests associated with the three test statistics.

Example 5.18 (Discrete uniform distribution). In an experiment of casting a die, random experiments are repeated independently under the same conditions for 60 times. Suppose that experimental frequencies of outcomes 1, 2, 3, 4, 5, 6 are 13, 19, 11, 8, 5, 4 as shown in the following table.

Outcome (j)	1	2	3	4	5	6	Total
Frequency (N_j)	13	19	11	8	5	4	60

Would

$$H_0\text{: } \Pr(X = j) = 1/6, \quad j = 1, \ldots, 6$$

be accepted at the 0.05 significance level?

Solution. Now $n = 60$, $m = 6$ and $np_{j0} = 60 \times \frac{1}{6} = 10$, $j = 1, \ldots, 6$. According to (5.40) and (5.41), we have

$$\begin{aligned} Q_{60} &= \sum_{j=1}^{6} \frac{(N_j - np_{j0})^2}{np_{j0}} \\ &= \frac{(13-10)^2}{10} + \frac{(19-10)^2}{10} + \frac{(11-10)^2}{10} \\ &\quad + \frac{(8-10)^2}{10} + \frac{(5-10)^2}{10} + \frac{(4-10)^2}{10} \\ &= 15.6 > \chi^2(0.05, 5) = 11.07. \end{aligned}$$

Thus, H_0 is rejected at the 0.05 significance level. ‖

Example 5.19 (Beta distribution). A point is to be selected from the unit interval $(0, 1)$ randomly. Let

$$\begin{aligned} \mathbb{A}_1 &= (0, 1/4], & \mathbb{A}_2 &= (1/4, 1/2], \\ \mathbb{A}_3 &= (1/2, 3/4] \quad \text{and} & \mathbb{A}_4 &= (3/4, 1). \end{aligned}$$

Random experiments are repeated independently for 80 times under the same conditions. Observed frequencies that these points fall into $\mathbb{A}_1, \mathbb{A}_2, \mathbb{A}_3$ and $\mathbb{A}_4$ are 6, 18, 20 and 36, respectively. Test

$$\begin{aligned} H_0&: \quad \text{The cdf is Beta}(2, 1) \quad \text{against} \\ H_1&: \quad \text{The cdf is not Beta}(2, 1) \end{aligned}$$

at the 0.025 significance level.

Solution. $F_0(x; \boldsymbol{\theta})$ is the beta distribution Beta$(2, 1)$ with density $2x$. From (5.37), we calculate

$$\begin{aligned} p_{10} &= \int_0^{1/4} 2x \, \mathrm{d}x = \frac{1}{16}, \quad p_{20} = \int_{1/4}^{1/2} 2x \, \mathrm{d}x = \frac{3}{16}, \\ p_{30} &= \int_{1/2}^{3/4} 2x \, \mathrm{d}x = \frac{5}{16}, \quad p_{40} = \int_{3/4}^{1} 2x \, \mathrm{d}x = \frac{7}{16}. \end{aligned}$$

Now $n = 80$ and $m = 4$. According to (5.40) and (5.41), we have

$$\begin{aligned} Q_{80} &= \sum_{j=1}^{4} \frac{(N_j - np_{j0})^2}{np_{j0}} \\ &= \frac{(6-5)^2}{5} + \frac{(18-15)^2}{15} + \frac{(20-25)^2}{25} + \frac{(36-35)^2}{35} \\ &= 1.83 \\ &< \chi^2(0.025, 3) \\ &= 9.3484. \end{aligned}$$

Thus, we cannot reject H_0 at the 0.025 significance level. ‖

5.5.3 The chi-square test for known distribution family with unknown parameters

33• The issue

- In this subsection, we consider testing (5.35) against (5.36), where only the distribution family $\{F_0(x;\boldsymbol{\theta})\colon \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ is known, while the parameter vector $\boldsymbol{\theta}$ is *unknown*.
- In such situations, p_{j0} defined by (5.37) depends on $\boldsymbol{\theta}$, so we can write

$$p_{j0} = \int_{\mathbb{A}_j} \mathrm{d}F_0(x;\boldsymbol{\theta}) = p_{j0}(\boldsymbol{\theta}) = p_{j0}(\theta_1,\ldots,\theta_q). \tag{5.42}$$

34• Pearson's chi-square test

34.1• MLEs of $\boldsymbol{\theta}$ and p_{j0}

— Based on the *raw data* $\boldsymbol{x} = (x_1,\ldots,x_n)^\top$, using the ML estimation method, we can obtain the MLE $\hat{\boldsymbol{\theta}}$.

— For example, if $F_0(x;\boldsymbol{\theta})$ is the cdf of $N(\mu,\sigma^2)$, we have

$$\hat{\mu} = \bar{x} \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2.$$

— Let the MLE of θ_k be $\hat{\theta}_k$, $k = 1,\ldots,q$.

— By the invariance property of MLE (see §3.1.4), the MLE of p_{j0} in (5.42) is given by

$$\hat{p}_{j0} = p_{j0}(\hat{\theta}_1,\ldots,\hat{\theta}_q), \quad j = 1,\ldots,m-1.$$

34.2• Equivalent hypotheses

— From (5.42), we know that (5.38) and (5.39) become

$$H_0\colon\ p_j = \hat{p}_{j0}, \qquad \text{for all } j = 1,\ldots,m-1 \quad \text{against}$$

$$H_1\colon\ p_j \neq \hat{p}_{j0}, \qquad \text{for at least one of } j = 1,\ldots,m-1.$$

— Similar to (5.40), under H_0, we have

$$\hat{Q}_n = \sum_{j=1}^{m}\frac{(N_j - n\hat{p}_{j0})^2}{n\hat{p}_{j0}} \dot{\sim} \chi^2(m-q-1) \quad \text{as} \quad n\to\infty. \tag{5.43}$$

Example 5.20 (Poisson distribution). Suppose that in 200 passages of one author's works, the word "also" showed the following frequency distribution

Number of "also" (j)	0	1	2	3	4	5	6	7	Total
Frequency of passages (N_j)	22	53	58	39	20	5	2	1	200

Is it reasonable to assume that the distribution of the number of times the author uses "also" in a passage is Poisson (the approximate significance level is taken to be 0.05)?

Solution. $F_0(x;\boldsymbol{\theta})$ is the cdf of Poisson(λ). We want to test

$$\begin{aligned} H_0&: \text{ The distribution is Poisson}(\lambda) \qquad \text{against} \\ H_1&: \text{ The distribution is not Poisson}(\lambda). \end{aligned}$$

Now $n = 200$, the MLE of λ is

$$\begin{aligned} \hat{\lambda} &= \bar{x} = 0 \times \frac{22}{200} + 1 \times \frac{53}{200} + 2 \times \frac{58}{200} + 3 \times \frac{39}{200} \\ &\quad + 4 \times \frac{20}{200} + 5 \times \frac{5}{200} + 6 \times \frac{2}{200} + 7 \times \frac{1}{200} \\ &= 2.05. \end{aligned}$$

We calculate $\{p_{j0}\}_{j=0}^{7}$ according to (5.42) and their MLEs:

$$\begin{aligned} p_{j0} &= p_{j0}(\lambda) = \frac{\lambda^j}{j!}\,\mathrm{e}^{-\lambda}, \quad j = 0,1,\ldots,6, \quad p_{7,0} = 1 - \sum_{j=0}^{6} p_{j0}, \\ \hat{p}_{j0} &= p_{j0}(\hat{\lambda}) = \frac{\hat{\lambda}^j}{j!}\,\mathrm{e}^{-\hat{\lambda}}, \quad j = 0,1,\ldots,6, \quad \hat{p}_{7,0} = 1 - \sum_{j=0}^{6} \hat{p}_{j0}, \end{aligned}$$

and obtain

j	0	1	2	3	4	5	6	7 ($\geqslant 7$)
N_j	22	53	58	39	20	5	2	1
$\hat{p}_{j0}$	0.1287	0.2639	0.2705	0.1848	0.0947	0.0388	0.0132	0.0054
$n\hat{p}_{j0}$	25.747	52.781	54.101	36.969	18.946	7.7681	2.6541	1.0800
							11.5022	

If the expected frequencies in some classes are less than 5, they should be combined with an adjacent class. Therefore, we combine the last 3 classes. According to (5.43), $m = 6$, we have

$$\hat{Q}_{200} = \sum_{j=0}^{5} \frac{(N_j - n\hat{p}_{j0})^2}{n\hat{p}_{j0}} = 2.052 < \chi^2(0.05, 6-1-1) = 9.488.$$

Thus, we cannot reject H_0 at the 0.05 approximate significance level. ‖

Example 5.21 (Normal distribution). Assume that $x_1, \ldots, x_n$ ($n = 40$) are given by

```
2.2070 2.3364 1.8048 1.7401 2.0898 2.1319 1.5457 1.9230
2.2654 2.0438 1.9670 2.0039 1.9003 1.9510 2.1060 1.2724
1.9019 1.8046 1.7856 1.9629 1.8754 1.9128 2.4564 2.0458
2.0421 1.6903 1.9279 1.6062 2.5279 2.5894 1.4994 1.9955
2.0977 1.4951 2.2308 2.2313 1.7722 2.4016 2.2167 2.0653
```

Please perform a normality test by means of the χ^2 goodness-of-fit test.

Solution. $F_0(x; \mu, \sigma^2)$ is the cdf of $N(\mu, \sigma^2)$. We want to test

H_0: The distribution is $N(\mu, \sigma^2)$ against
H_1: The distribution is not $N(\mu, \sigma^2)$.

Now $n = 40$, the MLEs of μ and σ are given by

$$\hat{\mu} = \bar{x} = 1.9856 \quad \text{and} \quad \hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} = 0.28718.$$

To calculate p_{j0} and $\hat{p}_{j0}$, we let the n sample values $x_1, \ldots, x_n$ be grouped into k categories. For example, the j-th category could be taken as all those observations falling in the interval $(a_{j-1}, a_j]$, $j = 1, \ldots, m$, where

$$-\infty = a_0 < a_1 < a_2 < \cdots < a_{m-1} < a_m = +\infty.$$

Then

$$\begin{aligned} p_{j0} &= p_{j0}(\mu, \sigma^2) = \int_{a_{j-1}}^{a_j} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \mathrm{d}x \\ &= \Phi\left(\frac{a_j - \mu}{\sigma}\right) - \Phi\left(\frac{a_{j-1} - \mu}{\sigma}\right), \quad \text{and} \\ \hat{p}_{j0} &= p_{j0}(\hat{\mu}, \hat{\sigma}^2) = \Phi\left(\frac{a_j - \hat{\mu}}{\hat{\sigma}}\right) - \Phi\left(\frac{a_{j-1} - \hat{\mu}}{\hat{\sigma}}\right). \end{aligned}$$

Furthermore, let $a_0 = -\infty$, $a_6 = +\infty$, $(a_1, \ldots, a_5) = (1.5, 1.8, 2.0, 2.2, 2.5)$, so that $(N_1, \ldots, N_6) = (4, 5, 12, 9, 8, 2)$. Then

$$\begin{aligned}
\hat{p}_{1,0} &= \Phi\left(\frac{a_1 - \hat{\mu}}{\hat{\sigma}}\right) - \Phi\left(\frac{a_0 - \hat{\mu}}{\hat{\sigma}}\right) \\
&= \Phi\left(\frac{1.5 - 1.9856}{0.28718}\right) - \Phi\left(\frac{-\infty - 1.9856}{0.28718}\right) \\
&= 0.045427, \\
\hat{p}_{2,0} &= 0.213623, \\
\hat{p}_{3,0} &= 0.252341, \\
\hat{p}_{4,0} &= 0.260946, \\
\hat{p}_{5,0} &= 0.191033, \\
\hat{p}_{6,0} &= 0.036631.
\end{aligned}$$

Thus

$$\begin{aligned}
\hat{Q}_{40} &= \frac{(4 - 40 * 0.045427)^2}{40 * 0.045427} + \frac{(5 - 40 * 0.213623)^2}{40 * 0.213623} \\
&\quad + \frac{(12 - 40 * 0.252341)^2}{40 * 0.252341} + \frac{(9 - 40 * 0.260946)^2}{40 * 0.260946} \\
&\quad + \frac{(8 - 40 * 0.191033)^2}{40 * 0.191033} + \frac{(2 - 40 * 0.036631)^2}{40 * 0.036631} \\
&= 2.622397 + 1.470622 + 0.233800 \\
&\quad + 0.118493 + 0.016838 + 0.195163 \\
&= 4.6573 < 7.8147 \\
&= \chi^2(0.05, 3) = \chi^2(\alpha, m - 1 - 2),
\end{aligned}$$

we cannot reject H_0. ‖

Exercise 5

5.1 In Example 5.5, if $\mathbb{C} = \{1, 7, 3, 8, 4\}$, how to find the Type I error rate $\alpha(0)$ and the Type II error rate $\beta(1)$?

5.2 Let $Y \sim \text{Binomial}(n, \theta)$. We reject H_0: $\theta = 0.5$ and accept H_1: $\theta > 0.5$ if $Y \geqslant c$. Consider the normal approximation to the binomial

distribution, please find n and c to give a power function $p(\theta)$ with $p(0.5) = 0.1$ and $p(2/3) = 0.95$.

5.3 Let $X_1, \ldots, X_n$ be a random sample from Gamma$(2, \theta)$ with pdf

$$f(x;\theta) = \begin{cases} \dfrac{\theta^2}{\Gamma(2)} x\, \mathrm{e}^{-\theta x}, & \text{if } x > 0, \\ 0, & \text{otherwise,} \end{cases}$$

where $\theta > 0$.

(a) Find the pdf of $Y = \sum_{i=1}^{n} X_i$.

(b) Find the MPT of size α for testing H_0: $\theta = \theta_0 (= 1)$ against H_1: $\theta = \theta_1 (> 1)$.

(c) Express the power function as an integral.

5.4 Let $X_1, \ldots, X_n$ be a random sample from

$$f(x;\theta) = \begin{cases} \theta(1-x)^{\theta-1}, & \text{if } 0 < x < 1, \\ 0, & \text{otherwise,} \end{cases}$$

where $\theta > 0$.

(a) Find the MPT of size α for testing H_0: $\theta = \theta_0 (= 1)$ against H_1: $\theta = \theta_1 (> 1)$.

(b) Find the LRT for testing H_0: $\theta = 1$ against H_1: $\theta \neq 1$.

5.5 Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\theta, 1)$. Find the UMPT of size α for testing H_0: $\theta \geqslant \theta_0$ against H_1: $\theta < \theta_0$.

5.6 Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ with unknown mean μ. Find the LRT with size α for testing H_0: $\sigma^2 = \sigma_0^2$ against one of the alternative $\sigma^2 \neq \sigma_0^2$, $\sigma^2 > \sigma_0^2$, or $\sigma^2 < \sigma_0^2$.

5.7 Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ be independent random samples from $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$, respectively, where σ^2 is the common but unknown variance. Find the LRT for testing H_0: $\mu_1 = \mu_2 = 0$ against H_1: $\mu_1 \neq \mu_2$ or $\mu_1 = \mu_2 \neq 0$.

[HINT: Express the LR ratio as $1/(1 + \frac{F}{n-1})$, where $F \sim F(2, 2n-2)$ under H_0]

5.8 Given two random samples of size n_1 and n_2 from two normal distributions $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$. Find the LRT for testing H_0: $\sigma_1^2 = \sigma_2^2$ against one of the alternative $\sigma_1^2 \neq \sigma_2^2$, $\sigma_1^2 > \sigma_2^2$, or $\sigma_1^2 < \sigma_2^2$.

5.9 The number of successes in n trials is to be used to test the null hypothesis that the parameter θ of a binomial population equals $1/2$ against the alternative that it does not equal to $1/2$.

(a) Find an expression for the LR statistic.

(b) Show that the critical region of the LRT can be written as $x\log(x) + (n-x)\log(n-x) \geqslant c$, where x is the observed number of successes.

(c) Study the graph of $f(x) = x\log(x) + (n-x)\log(n-x)$, in particular its minimum and its symmetry, to show that the critical region of this LRT can also be written as $|x - n/2| \geqslant c$, where c is a constant that depends on the size of the critical region.

5.10 Mendelian theory indicates that the shape and color of a certain variety of pea ought to be grouped into four groups, "round and yellow," "round and green," "angular and yellow," and "angular and green," according to the ratios $9:3:3:1$. For $n = 556$ peas, the following were observed (the last column gives the expected number):

Round and yellow	315	312.75
Round and green	108	104.25
Angular and yellow	101	104.25
Angular and green	32	34.75

Are the data consistent at the size of 0.05 with the null hypothesis H_0: $p_1 = 9/16, p_2 = 3/16, p_3 = 3/16$ and $p_4 = 1/16$?

5.11 A die was cast 300 times with the following results:

Occurrence:	1	2	3	4	5	6
Frequency:	43	49	56	45	66	41

Are the data consistent at the size of 0.05 with the null hypothesis that the die is true?

Chapter 6

Critical Regions and p-values for Skew Null Distributions

6.1 Tests on Normal Variances

6.1.1 One–sample chi-square test

1• The issue

- Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, where μ and σ^2 are unknown.
- Suppose that we want to test the null hypothesis H_0: $\sigma^2 = \sigma_0^2$ against one of the alternatives H_1: $\sigma^2 \neq \sigma_0^2$, H_2: $\sigma^2 > \sigma_0^2$, or H_3: $\sigma^2 < \sigma_0^2$.

2• The critical region approach

2.1• Step 1: To find a test statistic

— Since the distribution of

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$$

is $\chi^2(n-1)$ that does not depend on the unknown parameters (μ, σ^2), we know that χ^2 is a pivotal quantity.

— The test statistic is

$$\chi_1^2 \,\hat{=}\, \frac{(n-1)S^2}{\sigma_0^2} = \frac{(n-1)S^2}{\sigma^2} \cdot \frac{\sigma^2}{\sigma_0^2} = \chi^2 \cdot \frac{\sigma^2}{\sigma_0^2}.$$

— When H_0 is true, i.e., $\sigma^2 = \sigma_0^2$, we obtain

$$\chi_1^2 = \chi^2 \sim \chi^2(n-1). \tag{6.1}$$

2.2• Step 2: To determine a critical region of size α

— Since

$$\begin{aligned}
\alpha &= \alpha_1 + \alpha_2 = \Pr(\chi_1^2 \leqslant k_1) + \Pr(\chi_1^2 \geqslant k_2), && \text{for } H_1,\\
\alpha &= \Pr\{\chi_1^2 \geqslant \chi^2(\alpha, n-1)\}, && \text{for } H_2,\\
\alpha &= \Pr\{\chi_1^2 \leqslant \chi^2(1-\alpha, n-1)\}, && \text{for } H_3,
\end{aligned}$$

where

$$k_1 = \chi^2(1-\alpha_1, n-1) \quad \text{and} \quad k_2 = \chi^2(\alpha_2, n-1), \tag{6.2}$$

the critical regions of size α for the three alternatives H_1, H_2 or H_3 are given by

$$\begin{aligned}
\mathbb{C}_1 &= \{\boldsymbol{x}\colon\ \chi_{1,\,\text{obs}}^2 \leqslant k_1 \quad \text{or} \quad \chi_{1,\,\text{obs}}^2 \geqslant k_2\}, \qquad (6.3)\\
\mathbb{C}_2 &= \{\boldsymbol{x}\colon\ \chi_{1,\,\text{obs}}^2 \geqslant \chi^2(\alpha, n-1)\},\\
\mathbb{C}_3 &= \{\boldsymbol{x}\colon\ \chi_{1,\,\text{obs}}^2 \leqslant \chi^2(1-\alpha, n-1)\},
\end{aligned}$$

respectively, where $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$,

$$\chi_{1,\,\text{obs}}^2 = \frac{(n-1)s^2}{\sigma_0^2}, \tag{6.4}$$

denotes the observed value of the test statistic χ_1^2.

2.3• Step 3: To determine (k_1, k_2) via the equal-height approach

— Almost all textbooks simply use the equal-tail approach by letting $\alpha_1 = \alpha_2 = \alpha/2$ so that

$$k_1 = \chi^2(1-\alpha/2, n-1) \quad \text{and} \quad k_2 = \chi^2(\alpha/2, n-1). \tag{6.5}$$

— However, for skew densities like this one, the equal-tail approach leading to the results in (6.5) is incorrect likelihood ratio test.

— In the follows, we introduce the equal-height approach for determining (k_1, k_2), see Figure 6.1.

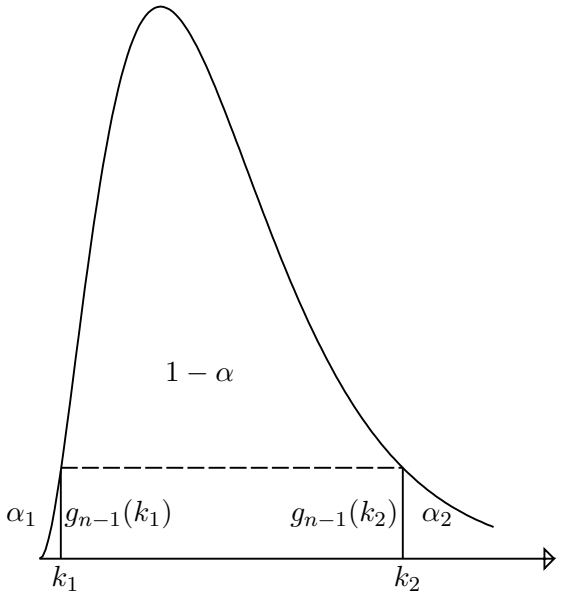

Figure 6.1 The critical region $\mathbb{C}_1$ defined by (6.3) for a two-tailed χ^2 test, where $g_{n-1}(\cdot)$ denotes the density function of the $\chi^2(n-1)$, and $\alpha_1 + \alpha_2 = \alpha$.

— We denote the density function of the $\chi^2(n-1)$ by

$$g_{n-1}(x) = \frac{2^{-(n-1)/2}}{\Gamma((n-1)/2)} x^{(n-1)/2-1}\, \mathrm{e}^{-x/2}, \quad x > 0,$$

whose mode is $n-3$.

— The equal-height approach requires to determine k_1 and k_2 satisfying

$$g_{n-1}(k_1) = g_{n-1}(k_2),$$

subject to $k_1 < n-3 < k_2$.

— According to (6.2), finding (k_1, k_2) is equivalent to finding (α_1, α_2).

— Since $\alpha_2 = \alpha - \alpha_1$, we only need to find α_1, which satisfies

$$g_{n-1}(\chi^2(1-\alpha_1, n-1)) = g_{n-1}(\chi^2(\alpha-\alpha_1, n-1)). \tag{6.6}$$

— Especially, when $g_{n-1}(\cdot)$ is skew toward the left, we always have $0 < \alpha_1 < \alpha/2$.

— Thus, we can use the grid method or the bisection method to find α_1.

2.4• R codes

```
function(n, alone)
{  # Function name: Compute.alpha.one.chisq.test(n, alone)
   # n is the sample size
   # alone is a vector consisting of the grid points
   # over the interval [0.001, 0.025], e.g.,
   # alone = seq(0.025, 0.001, -0.0001)
   v <- n - 1; al <- 0.05
   L <- length(alone)
   k1 <- qchisq(alone, v)
   k2 <- qchisq(1 - al + alone, v)
   gk1 <- dchisq(k1, v)
   gk2 <- dchisq(k2, v)
   error <- abs(gk1 - gk2)
   result <- matrix(c(alone, k1, k2, gk1, gk2, error), L, 6)
   return(result)   }
```

Example 6.1 (Numerical illustration). Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, and we observed $n = 25$ and $s^2 = 1600$. Test H_0: $\sigma^2 = \sigma_0^2 = 900$ against H_1: $\sigma^2 \neq \sigma_0^2$ at the 0.05 level of significance.

Solution. Since $\alpha = 0.05$, from Table 6.1, we have $\alpha_1 = 0.0138$, $k_1 = 11.361$ and $k_2 = 37.817$. In this example, the mode of the density of $\chi^2(n-1)$ is $n - 3 = 22$. We will reject the H_0 if $\chi^2_{1,\text{obs}} \leqslant 11.361$ or $\chi^2_{1,\text{obs}} \geqslant 37.817$, where $\chi^2_{1,\text{obs}}$ is defined by (6.4). Since $\chi^2_{1,\text{obs}} = 42.667 > 37.817$, the null hypothesis is rejected.

Table 6.1 Calculation of α_1 from (6.6)

α_1	k_1	k_2	$g_{n-1}(k_1)$	$g_{n-1}(k_2)$	Error
0.0250	12.4012	39.364	0.01323	0.00609	0.0071416
0.0230	12.2454	39.047	0.01244	0.00653	0.0059176
0.0210	12.0793	38.751	0.01164	0.00696	0.0046744
0.0190	11.9010	38.472	0.01080	0.00739	0.0034095
0.0170	11.7082	38.209	0.00994	0.00782	0.0021199
0.0150	11.4974	37.960	0.00904	0.00824	0.0008017
0.0140	11.3840	37.840	0.00858	0.00845	0.0001305
0.0138	11.3610	37.817	0.00849	0.00849	4.7961×10^{-6}

NOTE: Error $= |g_{n-1}(k_1) - g_{n-1}(k_2)|$. ‖

3• The p-value approach

- The corresponding p-values can be calculated by

$$\begin{aligned} p\text{-value} &= p_1 + p_2 = \Pr(\chi_1^2 \leqslant b_1) + \Pr(\chi_1^2 \geqslant b_2), \text{ for } H_1, \\ p\text{-value} &= \Pr(\chi_1^2 \geqslant \chi_{1,\,\text{obs}}^2), \qquad \text{for } H_2, \\ p\text{-value} &= \Pr(\chi_1^2 \leqslant \chi_{1,\,\text{obs}}^2), \qquad \text{for } H_3, \end{aligned} \tag{6.7}$$

  where p_1, p_2, b_1 and b_2 are shown in Figure 6.2, χ_1^2 is specified by (6.1) and $\chi_{1,\,\text{obs}}^2$ is given by (6.4).

- To calculate the two-tailed p-value defined by (6.7), we consider two cases.

3.1• Case I: $\chi_{1,\,\text{obs}}^2 < n-3$

— In this case, we define $b_1 = \chi_{1,\,\text{obs}}^2$ as shown in Figure 6.2.

— The value of b_2 can be obtained by solving

$$g_{n-1}(b_2) = g_{n-1}(b_1) \quad \text{subject to} \quad b_2 > n-3.$$

3.2• Case II: $\chi_{1,\,\text{obs}}^2 > n-3$

— In this case, we define $b_2 = \chi_{1,\,\text{obs}}^2$ as shown in Figure 6.3.

— The value of b_1 can be determined by

$$g_{n-1}(b_1) = g_{n-1}(b_2) \quad \text{subject to} \quad 0 < b_1 < n-3. \tag{6.8}$$

Example 6.2 (Example 6.1 revisited). Calculate the p-value for testing H_0: $\sigma^2 = \sigma_0^2 = 900$ against H_1: $\sigma^2 \neq \sigma_0^2$ at the 0.05 level of significance.

Solution. Note that $n = 25$ and $\chi_{1,\,\text{obs}}^2 = 42.667 > 22 = n-3$, then $b_2 = \chi_{1,\,\text{obs}}^2 = 42.667$, $g_{n-1}(b_2) = g_{24}(42.667) = 0.0028345$. Now (6.8) becomes $g_{24}(b_1) = 0.0028345$. From Table 6.2, we obtain $b_1 = 9.4095 < 22$. From (6.7), we have

$$\begin{aligned} p\text{-value} &= p_1 + p_2 \\ &= \Pr\{\chi^2(n-1) \leqslant b_1\} + \Pr\{\chi^2(n-1) \geqslant b_2\} \\ &= \Pr\{\chi^2(24) \leqslant 9.4095\} + \Pr\{\chi^2(24) \geqslant 42.667\} \\ &= 0.0034162 + 0.010854 = 0.01427 < 0.05, \end{aligned}$$

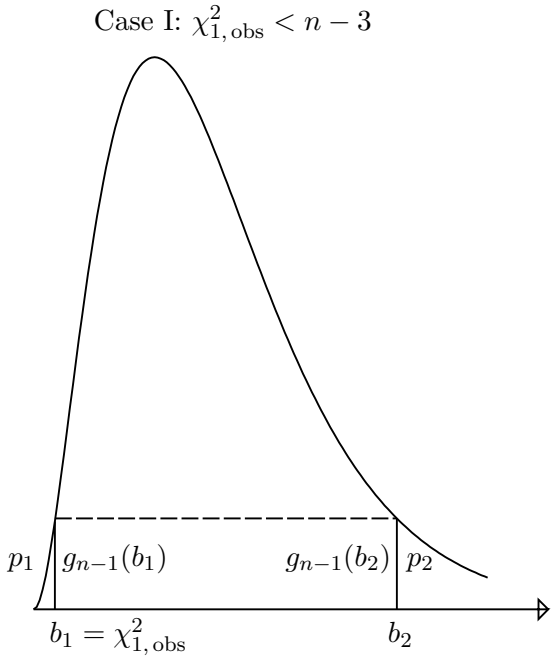

Figure 6.2 The p-value defined by (6.7) for a two-tailed χ^2 test for Case I: $\chi^2_{1,\,\mathrm{obs}}$ given by (6.4) is at the left tail, where $g_{n-1}(\cdot)$ denotes the density of the $\chi^2(n-1)$ with mode $n-3$.

so that the H_0 must be rejected.

Table 6.2 Calculation of b_1 from $g_{24}(b_1) = 0.0028345$

b_1	Error	b_1	Error
11	4.2971×10^{-3}	9.6	3.7808×10^{-4}
10	1.2866×10^{-3}	9.5	1.7535×10^{-4}
9.9	1.0444×10^{-3}	9.4	1.8020×10^{-5}
9.8	8.1244×10^{-4}	9.41	9.0208×10^{-7}
9.7	5.9038×10^{-4}	9.4095	4.6199×10^{-8}

NOTE: Error $= |g_{24}(b_1) - 0.0028345|$. ‖

6.1.2 Two-sample F test

4• The issue

- Let $X_{i1}, \ldots, X_{in_i} \overset{\text{iid}}{\sim} N(\mu_i, \sigma_i^2)$ for $i = 1, 2$, and the two random samples be independent, where μ_i and σ_i^2 are unknown.

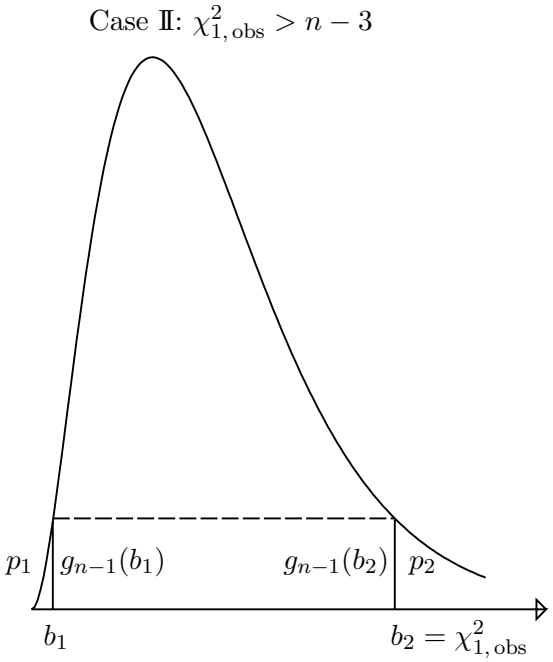

Figure 6.3 The p-value defined by (6.7) for a two-tailed χ^2 test for Case II: $\chi^2_{1,\,\mathrm{obs}}$ given by (6.4) is at the right tail, where $g_{n-1}(\cdot)$ denotes the density of the $\chi^2(n-1)$ with mode $n-3$.

- Suppose that we want to test the null hypothesis H_0: $\sigma_1^2 = \sigma_2^2$ against one of the alternatives H_1: $\sigma_1^2 \neq \sigma_2^2$, H_2: $\sigma_1^2 > \sigma_2^2$, or H_3: $\sigma_1^2 < \sigma_2^2$.

5• The critical region approach

5.1• Step 1: To find a test statistic

— Define $\nu_i = n_i - 1$ for $i = 1, 2$. Since the distribution of

$$F = \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2}$$

is $F(\nu_1, \nu_2)$ that does not depend on the unknown parameters (μ_i, σ_i^2), we know that F is a pivotal quantity.

— The test statistic is

$$F_0 \,\hat{=}\, \frac{S_1^2}{S_2^2} = \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2} \cdot \frac{\sigma_1^2}{\sigma_2^2} = F \cdot \frac{\sigma_1^2}{\sigma_2^2}.$$

— When H_0 is true, i.e., $\sigma_1^2 = \sigma_2^2$, we have

$$F_0 = F \sim F(\nu_1, \nu_2). \tag{6.9}$$

5.2• Step 2: To determine a critical region of size α

— Since

$$\begin{aligned} \alpha &= \alpha_1 + \alpha_2 = \Pr(F_0 \leqslant k_1) + \Pr(F_0 \geqslant k_2), && \text{for } H_1, \\ \alpha &= \Pr\{F_0 \geqslant f(\alpha, \nu_1, \nu_2)\}, && \text{for } H_2, \\ \alpha &= \Pr\{F_0 \leqslant f(1-\alpha, \nu_1, \nu_2)\}, && \text{for } H_3, \end{aligned}$$

where

$$k_1 = f(1-\alpha_1, \nu_1, \nu_2) \quad \text{and} \quad k_2 = f(\alpha_2, \nu_1, \nu_2), \tag{6.10}$$

the critical regions of size α for the three alternatives H_1, H_2 or H_3 are given by

$$\begin{aligned} \mathbb{C}_1 &= \{(\boldsymbol{x}_1, \boldsymbol{x}_2)\colon\ f_0 \leqslant k_1 \text{ or } f_0 \geqslant k_2\}, \\ \mathbb{C}_2 &= \{(\boldsymbol{x}_1, \boldsymbol{x}_2)\colon\ f_0 \geqslant f(\alpha, \nu_1, \nu_2)\}, \\ \mathbb{C}_3 &= \{(\boldsymbol{x}_1, \boldsymbol{x}_2)\colon\ f_0 \leqslant f(1-\alpha, \nu_1, \nu_2)\}, \end{aligned} \tag{6.11}$$

respectively, where $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{in_i})^\top$ for $i = 1, 2$,

$$f_0 = s_1^2/s_2^2, \tag{6.12}$$

denotes the observed value of the test statistic F_0.

5.3• Step 3: To determine (k_1, k_2) via the equal-height approach

— Almost all textbooks simply use the equal-tail approach by letting $\alpha_1 = \alpha_2 = \alpha/2$ so that

$$k_1 = f(1-\alpha/2, \nu_1, \nu_2) \quad \text{and} \quad k_2 = f(\alpha/2, \nu_1, \nu_2). \tag{6.13}$$

— However, for skew densities like this one, the equal-tail approach leading to the results in (6.13) is incorrect likelihood ratio test.

— In the follows, we introduce the equal-height approach for determining (k_1, k_2), see Figure 6.4.

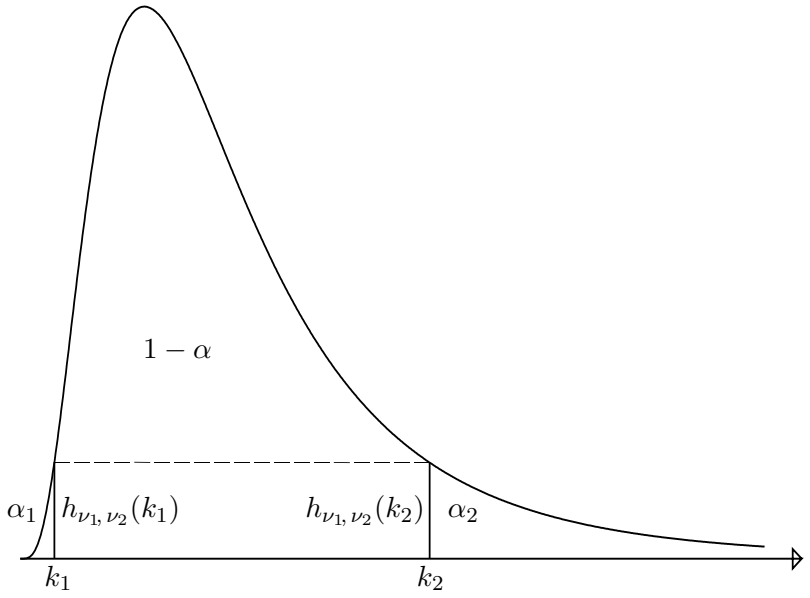

Figure 6.4 The critical region $\mathbb{C}_1$ defined by (6.11) for a two-tailed F test, where $h_{\nu_1,\nu_2}(\cdot)$ denotes the density function of the $F(\nu_1,\nu_2)$, and $\alpha_1+\alpha_2=\alpha$.

— We denote the density function of the $F(\nu_1,\nu_2)$ by

$$h_{\nu_1,\nu_2}(x)=\frac{\Gamma(\frac{\nu_1+\nu_2}{2})}{\Gamma(\frac{\nu_1}{2})\Gamma(\frac{\nu_2}{2})}\left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}}x^{\frac{\nu_1}{2}-1}\left(1+\frac{\nu_1 x}{\nu_2}\right)^{-\frac{\nu_1+\nu_2}{2}},\quad x>0,$$

whose mode is $m_0 \hat{=} (\nu_1-2)\nu_2/\{\nu_1(\nu_2+2)\}$.

— The equal-height approach requires to determine k_1 and k_2 satisfying

$$h_{\nu_1,\nu_2}(k_1)=h_{\nu_1,\nu_2}(k_2)$$

subject to $k_1<m_0<k_2$.

— According to (6.10), finding (k_1,k_2) is equivalent to finding (α_1,α_2).

— Since $\alpha_2=\alpha-\alpha_1$, we only need to find α_1, which satisfies

$$h_{\nu_1,\nu_2}(f(1-\alpha_1,\nu_1,\nu_2))=h_{\nu_1,\nu_2}(f(\alpha-\alpha_1,\nu_1,\nu_2)). \tag{6.14}$$

— Especially, when $h_{\nu_1,\nu_2}(\cdot)$ is skew toward the left, we always have $0<\alpha_1<\alpha/2$.

— Thus, we can use the grid method or the bisection method to find α_1.

5.4• R codes

```
function(n1, n2, alone)
{  # Function name: Compute.alpha.one.F.test(n1, n2, alone)
   # n1 and n2 are the sample sizes of the two samples
   # alone is a vector consisting of the grid points
   # over the interval [0.0001, 0.025], e.g.,
   # seq(0.025, 0.001, -0.0001)
   v1 <- n1 - 1; v2 <- n2 - 1; al <- 0.05
   L <- length(alone)
   k1 <- qf(alone, v1, v2)
   k2 <- qf(1 - al + alone, v1, v2)
   hk1 <- df(k1, v1, v2)
   hk2 <- df(k2, v1, v2)
   error <- abs(hk1 - hk2)
   result <- matrix(c(alone, k1, k2, hk1, hk2, error), L, 6)
   return(result)  }
```

Example 6.3 (Numerical illustration). Let $X_{i1}, \ldots, X_{in_i}$ be two independent random samples from $N(\mu_i, \sigma_i^2)$, $i = 1, 2$, and we observed $n_1 = 13$, $s_1^2 = 37853.17$, $n_2 = 7$, $s_2^2 = 15037.00$. Test H_0: $\sigma_1^2 = \sigma_2^2$ against H_1: $\sigma_1^2 \neq \sigma_2^2$ at $\alpha = 0.05$.

Solution. Since $\alpha = 0.05$, from Table 6.3, we have $\alpha_1 = 0.000657$, $k_1 = 0.10889$ and $k_2 = 4.0233$. The mode of the density of $F(12, 6)$ is $m_0 = 0.625$. We will reject the H_0 if $f_0 \leqslant 0.10889$ or $f_0 \geqslant 4.0233$, where f_0 is defined by (6.12). Since $f_0 = 2.5173 \in (0.10889, 4.0233)$, we accept the H_0.

Table 6.3 Calculation of α_1 from (6.14)

α_1	k_1	k_2	$h_{\nu_1,\nu_2}(k_1)$	$h_{\nu_1,\nu_2}(k_2)$	Error
0.025	0.26822	5.3662	0.31283	0.011359	0.30147
0.015	0.23158	4.6648	0.23300	0.017738	0.21526
0.005	0.17370	4.1885	0.11615	0.024730	0.09142
0.0005	0.10269	4.0176	0.02286	0.028046	0.00518
0.0006	0.10678	4.0212	0.02614	0.027971	0.00182
0.00065	0.10864	4.0230	0.02772	0.027934	0.00021
0.000657	0.10889	4.0233	0.02794	0.027928	1.4163×10^{-5}

NOTE: Error $= |h_{\nu_1,\nu_2}(k_1) - h_{\nu_1,\nu_2}(k_2)|$. ‖

6• The p-value region approach

- The corresponding p-values can be calculated by

$$\begin{aligned} p\text{-value} &= p_1 + p_2 = \Pr(F_0 \leqslant b_1) + \Pr(F_0 \geqslant b_2), && \text{for } H_1, \quad (6.15)\\ p\text{-value} &= \Pr(F_0 \geqslant f_0), && \text{for } H_2,\\ p\text{-value} &= \Pr(F_0 \leqslant f_0), && \text{for } H_3, \end{aligned}$$

  where p_1, p_2, b_1 and b_2 are shown in Figure 6.5, F_0 is specified by (6.9) and f_0 is given by (6.12).

- To calculate the two-tailed p-value defined by (6.15), we consider two cases.

6.1• Case I: f_0 is at the left tail

— The value of b_2 at the right tail can be obtained by (see Figure 6.5(i))

$$h_{\nu_1,\nu_2}(b_2) = h_{\nu_1,\nu_2}(f_0) \quad \text{subject to} \quad b_2 > m_0.$$

6.2• Case II: f_0 is at the right tail

— The value of b_1 at the left tail can be determined by (see Figure 6.5(ii))

$$h_{\nu_1,\nu_2}(b_1) = h_{\nu_1,\nu_2}(f_0) \quad \text{subject to} \quad 0 < b_2 < m_0. \qquad (6.16)$$

Example 6.4 (Example 6.3 revisited). Calculating the p-value for testing H_0: $\sigma_1^2 = \sigma_2^2$ against H_1: $\sigma_1^2 \neq \sigma_2^2$ at $\alpha = 0.05$.

Solution. Note that $n_1 = 13$, $n_2 = 7$ and $f_0 = 2.5173 > 0.625 = \text{mode}$, then $h_{\nu_1,\nu_2}(f_0) = h_{12,6}(2.5173) = 0.10241$. Now (6.16) becomes $h_{\nu_1,\nu_2}(b_1) = 0.10241$. From Table 6.4, we have $b_1 = 0.1658 < 0.625$. From (6.15), we obtain

$$\begin{aligned} p\text{-value} &= p_1 + p_2\\ &= \Pr\{F_0 \leqslant b_1\} + \Pr\{F_0 \geqslant b_2\}\\ &= \Pr\{F_0 \leqslant b_1\} + \Pr\{F_0 \geqslant f_0\}\\ &= \Pr\{F(12,6) \leqslant 0.1658\} + \Pr\{F(12,6) \geqslant 2.5173\}\\ &= 0.0041373 + 1 - 0.86694 = 0.1372 > 0.05, \end{aligned}$$

so that the H_0 cannot be rejected.

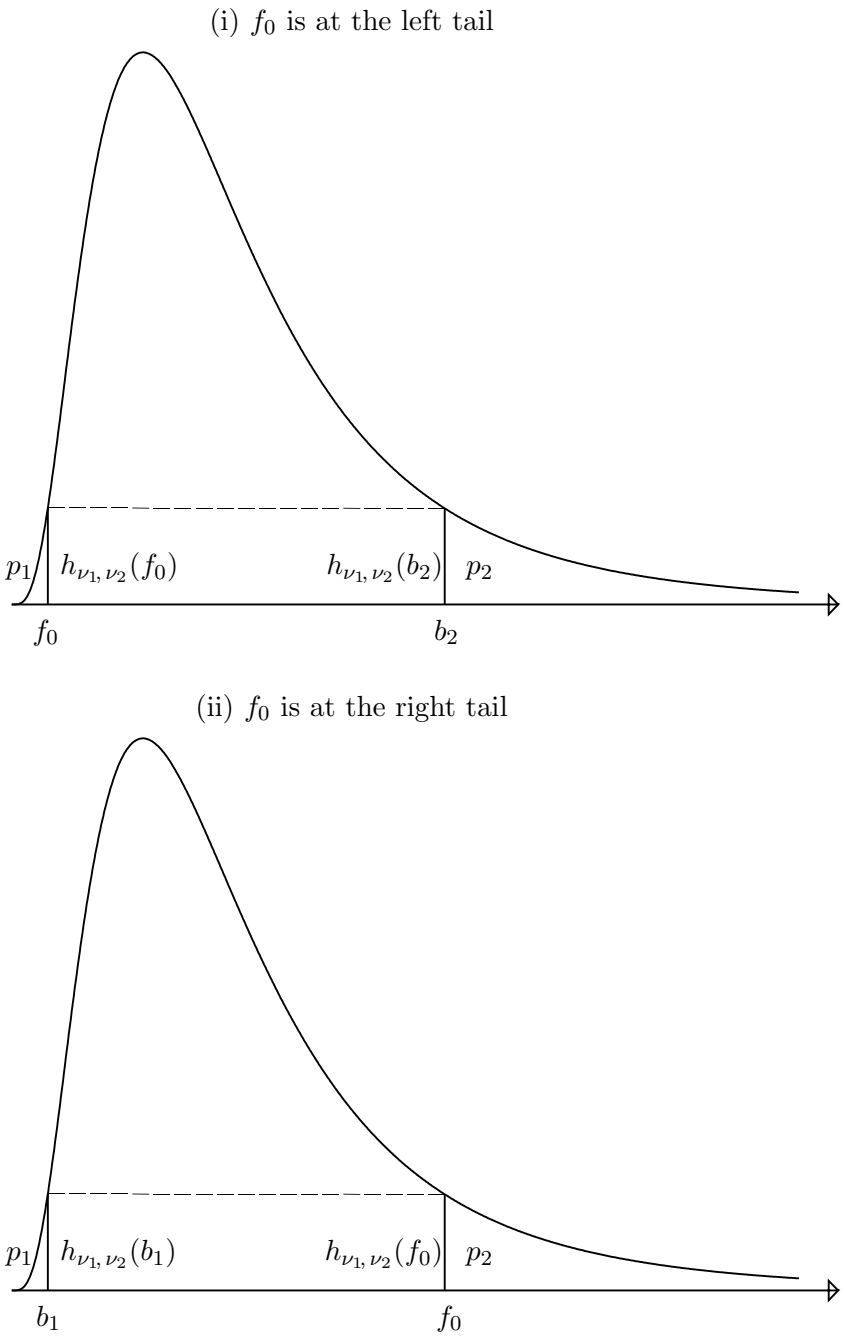

Figure 6.5 The p-value defined by (6.15) for a two-tailed F test, where $h_{\nu_1,\nu_2}(\cdot)$ denotes the density of the $F(\nu_1,\nu_2)$ with mode $m_0 = (\nu_1-2)\nu_2/\{\nu_1(\nu_2+2)\}$. (i) f_0 given by (6.12) is at the left tail; (ii) f_0 is at the right tail.

Table 6.4 Calculation of b_1 from $h_{\nu_1,\nu_2}(b_1) = 0.10241$

b_1	Error	b_1	Error
0.625	0.59141	0.170	0.007185
0.325	0.32767	0.168	0.003705
0.200	0.06411	0.166	0.000271
0.190	0.04426	0.1659	1.0091×10^{-4}
0.180	0.02523	0.1658	6.9435×10^{-5}

NOTE: Error $= |h_{\nu_1,\nu_2}(b_1) - 0.10241|$. ‖

APPENDIX A

Basic Statistical Distributions

A.1 Discrete distributions

A.1.1 Finite discrete distribution

Notation: $X \sim \text{FDiscrete}_n(\boldsymbol{x}, \boldsymbol{p})$, $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$, $\boldsymbol{p} = (p_1, \ldots, p_n)^\top \in \mathbb{T}_n \hat{=} \{(p_1, \ldots, p_n)\colon\ p_i \geqslant 0, \sum_{i=1}^n p_i = 1\}$.

Density: $\Pr(X = x_i) = p_i, \quad i = 1, \ldots, n.$

Moments: $E(X) = \sum_{i=1}^n x_i p_i$, $\text{Var}(X) = \sum_{i=1}^n x_i^2 p_i - (\sum_{i=1}^n x_i p_i)^2$.

Note: The *uniform discrete* distribution is a special case of the finite discrete distribution with $p_i = 1/n$ for all i.

Sampling: `sample(x, size, replace = FALSE, prob = NULL)` takes a sample of the specified size from the elements of `x` using either with or without replacement.

Examples:
```
> sample(c(0,1), 100, replace= T, prob=c(0.8, 0.2))
> sample(1:20, 4)    # the default: replace= F
```

A.1.2 Hypergeometric distribution

Notation: $X \sim \text{Hgeometric}(m, n, k)$, m, n, k are positive integers.

Density: $\text{Hgeometric}(x|m, n, k) = \binom{m}{x}\binom{n}{k-x}/\binom{m+n}{k}$,
where $x = \max(0, k-n), \ldots, \min(m, k)$.

Moments: $E(X) = km/N'$, $\text{Var}(X) = kmn(N'-k)/[N'^2(N'-1)]$,
where $N' \hat{=} m + n$.

Computing:
`> prod(5:1)` $= 5!$
`> prod(20:16)` $= 20 \times 19 \times 18 \times 17 \times 16$
`> choose(40,5)` $= \binom{40}{5}$

Functions:
```
dhyper(x, m, n, k)
phyper(q, m, n, k)
qhyper(p, m, n, k)
rhyper(nn, m, n, k)
```

A.1.3 Poisson distribution

Notation: $X \sim \text{Poisson}(\lambda)$, $\lambda > 0$

Density: $\text{Poisson}(x|\lambda) = \lambda^x \, \mathrm{e}^{-\lambda}/x!$, $x = 0, 1, 2, \ldots$

Moments: $E(X) = \lambda$, $\text{Var}(X) = \lambda$.

Properties:
- If $\{X_i\}_{i=1}^n \overset{\text{ind}}{\sim} \text{Poisson}(\lambda_i)$, then

$$\begin{aligned} \textstyle\sum_{i=1}^n X_i &\sim \text{Poisson}(\textstyle\sum_{i=1}^n \lambda_i), \quad \text{and} \\ (X_1, \ldots, X_n) | (\textstyle\sum_{i=1}^n X_i = m) &\sim \text{Multinomial}_n(m, \boldsymbol{p}), \end{aligned}$$

  where $\boldsymbol{p} = (\lambda_1, \ldots, \lambda_n)^\top / \sum_{i=1}^n \lambda_i$;

- The Poisson and gamma distribution have relationship:

$$\textstyle\sum_{x=k}^{\infty} \text{Poisson}(x|\lambda) = \int_0^\lambda \text{Gamma}(y|k, 1) \, \mathrm{d}y.$$

Functions:
```
dpois(x, lambda)
ppois(q, lambda)
qpois(p, lambda)
rpois(n, lambda)
```

```
================================================================
> x <- 0:20
> plot(x, dpois(x, 4), type="h")              # histogram-like
                                                 # Figure A.1
****************************************************************
```

A.1.4 Binomial distribution

Notation: $X \sim \text{Binomial}(n, p)$, n is a positive integer, $p \in (0, 1)$.

Density: $\text{Binomial}(x|n, p) = \binom{n}{x} p^x (1-p)^{n-x}$, $x = 0, 1, \ldots, n$.

Moments: $E(X) = np$, $\text{Var}(X) = np(1-p)$.

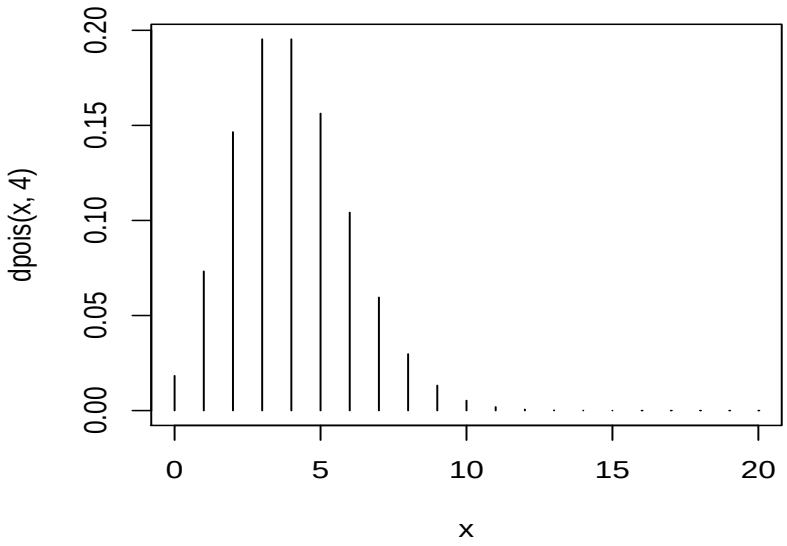

Figure A.1 Point probabilities of Poisson(4).

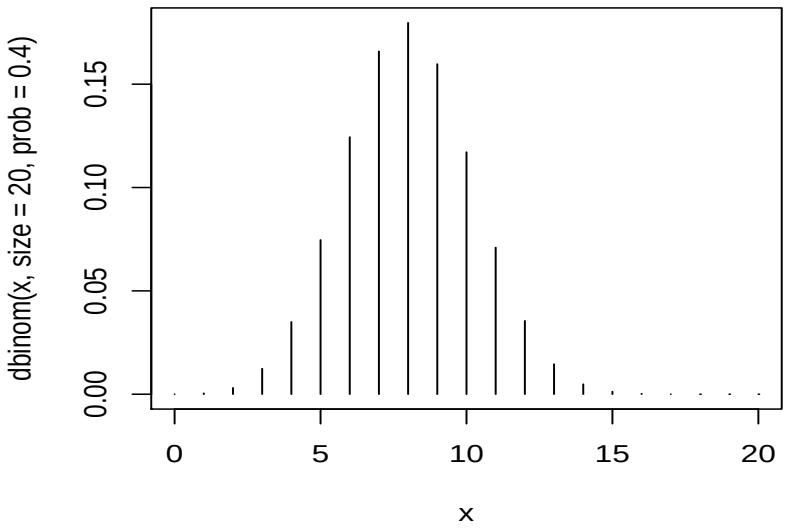

Figure A.2 Point probabilities of Binomial$(20, 0.4)$.

Properties: • If $\{X_i\}_{i=1}^d \overset{\text{ind}}{\sim} \text{Binomial}(n_i, p)$, then

$$\sum_{i=1}^d X_i \sim \text{Binomial}(\sum_{i=1}^d n_i, p);$$

• The binomial and beta distribution have relationship:

$$\sum_{x=0}^k \text{Binomial}(x|n,p) = \int_0^{1-p} \text{Beta}(x|n-k, k+1)\,\mathrm{d}x,$$

where $0 \leqslant k \leqslant n$.

Note: When $n = 1$, binomial distribution is called *Bernoulli* distribution.

Functions:
```
dbinom(x, size, prob)     # size= n,  prob= p
pbinom(q, size, prob)
qbinom(p, size, prob)
rbinom(nn, size, prob)
```

```
============================================================
> x <- 0:20
> plot(x, dbinom(x, size=20, prob=0.4), type="h")
                                                  # Figure A.2
************************************************************
```

A.1.5 Multinomial distribution

Notation: $\mathbf{x} = (X_1, \ldots, X_d)^\top \sim \text{Multinomial}(n; p_1, \ldots, p_d)$ or
$\mathbf{x} = (X_1, \ldots, X_d)^\top \sim \text{Multinomial}_d(n, \boldsymbol{p})$,
n is a positive integer, $\boldsymbol{p} = (p_1, \ldots, p_d)^\top \in \mathbb{T}_d$,

Density: $\text{Multinomial}_d(\boldsymbol{x}|n, \boldsymbol{p}) = \binom{n}{x_1, \ldots, x_d} \prod_{i=1}^d p_i^{x_i}$,
$\boldsymbol{x} = (x_1, \ldots, x_d)^\top$, $x_i \geqslant 0$, $\sum_{i=1}^d x_i = n$.

Moments: $E(X_i) = np_i$, $\text{Var}(X_i) = np_i(1-p_i)$, $\text{Cov}(X_i, X_j) = -np_ip_j$.

Note: The binomial distribution is a special case of the multinomial with $d = 2$.

Functions:
```
dmultinom(x, size = NULL, prob) # size= n, prob=  p
rmultinom(nn, size, prob)
```

A.2 Continuous distributions

A.2.1 Uniform distribution

Notation: $X \sim U(a, b)$, $a < b$

Density: $U(x|a, b) = 1/(b - a)$, $x \in (a, b)$.

Moments: $E(X) = (a + b)/2$, $\text{Var}(X) = (b - a)^2/12$.

Properties: If $Y \sim U(0, 1)$, then $X = a + (b - a)Y \sim U(a, b)$.

Functions:
```
dunif(x, min= 0, max= 1)     # min= a,  max= b
punif(q, min= 0, max= 1)
qunif(p, min= 0, max= 1)
runif(n, min= 0, max= 1)
```

A.2.2 Beta distribution

Notation: $X \sim \text{Beta}(a, b)$, $a > 0, b > 0$.

Density: $\text{Beta}(x|a, b) = x^{a-1}(1 - x)^{b-1}/B(a, b)$, $0 < x < 1$.

Moments: $E(X) = a/(a + b)$, $E(X^2) = a(a + 1)/[(a + b)(a + b + 1)]$, $\text{Var}(X) = ab/[(a + b)^2(a + b + 1)]$.

Properties: If $Y_1 \sim \text{Gamma}(a, 1)$, $Y_2 \sim \text{Gamma}(b, 1)$, and $Y_1 \perp\!\!\!\perp Y_2$, then $Y_1/(Y_1 + Y_2) \sim \text{Beta}(a, b)$.

Note: When $a = b = 1$, $\text{Beta}(1, 1) = U(0, 1)$.

Functions:
```
dbeta(x, shape1, shape2)   # shape1= a,  shape2= b
pbeta(q, shape1, shape2)
qbeta(p, shape1, shape2)
rbeta(n, shape1, shape2)
```

A.2.3 Exponential distribution

Notation: $X \sim \text{Exponential}(\beta)$, rate parameter $\beta > 0$.

Density: $\text{Exponential}(x|\beta) = \beta\, \mathrm{e}^{-\beta x}$, $x > 0$.

Moments: $E(X) = 1/\beta$, $\text{Var}(X) = 1/\beta^2$.

Properties: • If $U \sim U(0,1)$, then $-\frac{\log U}{\beta} \sim$ Exponential(β);

• If $\{X_i\}_{i=1}^n \overset{\text{iid}}{\sim}$ Exponential(β), then $\sum_{i=1}^n X_i \sim$ Gamma(n, β).

Functions:
```
dexp(x, rate= 1)          # rate= β
pexp(q, rate= 1)
qexp(p, rate= 1)
rexp(n, rate= 1)
```

A.2.4 Gamma distribution

Notation: $X \sim$ Gamma(α, β), shape parameter $\alpha > 0$, rate parameter $\beta > 0$.

Density: $\text{Gamma}(x|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \, \mathrm{e}^{-\beta x}$, $x > 0$.

Moments: $E(X) = \alpha/\beta$, $\text{Var}(X) = \alpha/\beta^2$.

Properties: • If $X \sim$ Gamma(α, β) and $c > 0$, then $cX \sim$ Gamma($\alpha, \beta/c$);

• If $\{X_i\}_{i=1}^n \overset{\text{ind}}{\sim}$ Gamma(α_i, β), then $\sum X_i \sim$ Gamma($\sum \alpha_i, \beta$);

• $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$, $\Gamma(1) = 1$ and $\Gamma(1/2) = \sqrt{\pi}$.

Note: Gamma($1, \beta$) = Exponential(β). Gamma($\nu/2, 1/2$) = $\chi^2(\nu)$.

Functions:
```
dgamma(x, shape, rate= 1)     # shape= α,   rate= β
pgamma(q, shape, rate= 1)
qgamma(p, shape, rate= 1)
rgamma(n, shape, rate= 1)
```

A.2.5 Chi-square distribution

Notation: $X \sim \chi^2(n) \equiv$ Gamma($\frac{n}{2}, \frac{1}{2}$), degree of freedom $n > 0$.

Density: $\chi^2(x|n) = \frac{2^{-n/2}}{\Gamma(n/2)} x^{n/2-1} \, \mathrm{e}^{-x/2}$, $x > 0$.

Moments: $E(X) = n$, $\text{Var}(X) = 2n$.

Properties: • If $Y \sim N(0,1)$, then $X = Y^2 \sim \chi^2(1)$;

• If $\{X_j\}_{j=1}^m \overset{\text{ind}}{\sim} \chi^2(n_j)$, then $\sum_{j=1}^m X_j \sim \chi^2(\sum_{j=1}^m n_j)$.

Functions:
```
dchisq(x, df)          # df = n
pchisq(q, df)
qchisq(p, df)
rchisq(nn, df)
```

```
================================================================
> x <- seq(0.01, 25, 0.1)
> par(mfrow=c(2, 2))                              # Figure A.3
> curve(dchisq(x, df= 1), from=0.1, to = 25)
> curve(dchisq(x, df= 2), from=0.1, to = 25)
> curve(dchisq(x, df= 3), from=0.1, to = 25)
> curve(dchisq(x, df= 4), from=0.1, to = 25)
****************************************************************
```

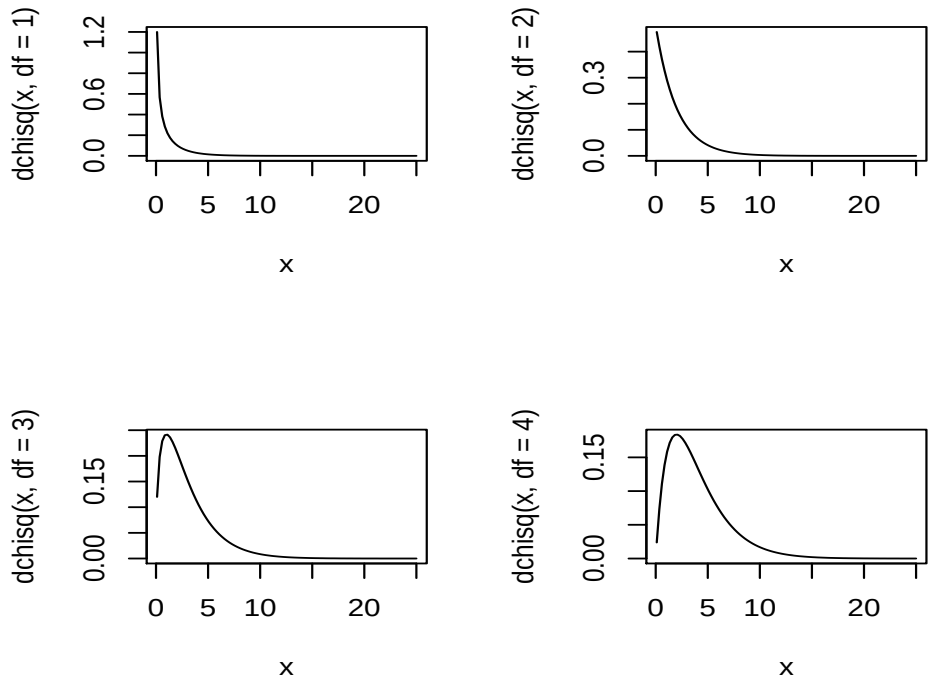

Figure A.3 Density functions of $\chi^2(n)$ for $n = 1, 2, 3, 4$.

A.2.6 t- or Student's t-distribution

Notation: $X \sim t(n)$, n is a positive integer.

Density: $t(x|n) = \dfrac{\Gamma(\frac{n+1}{2})}{\sqrt{\pi n}\,\Gamma(\frac{n}{2})}\left(1+\dfrac{x^2}{n}\right)^{-\frac{n+1}{2}}, \quad -\infty < x \leqslant \infty.$

Moments: $E(X) = 0$ (if $n > 1$), $\mathrm{Var}(X) = \frac{n}{n-2}$ (if $n > 2$).

Properties: Let $Z \sim N(0,1)$, $Y \sim \chi^2(n)$, and $Z \perp\!\!\!\perp Y$, then

$$\frac{Z}{\sqrt{Y/n}} \sim t(n).$$

Note: When $n = 1$, $t(n) = t(1)$ is called *standard Cauchy distribution*, whose mean and variance do not exist.

Functions:
```
dt(x, df)          # df = n
pt(q, df)
qt(p, df)
rt(nn, df)
```

A.2.7 F or Fisher's F-distribution

Notation: $X \sim F(n_1, n_2)$, n_1, n_2 are positive integers.

Density: $F(x|n_1,n_2) = \frac{(n_1/n_2)^{n_1/2}}{B(\frac{n_1}{2},\frac{n_2}{2})} x^{\frac{n_1}{2}-1}(1+\frac{n_1 x}{n_2})^{-\frac{n_1+n_2}{2}}$, $x > 0$.

Moments: $E(X) = \frac{n_2}{n_2-2}$ (if $n_2 > 2$), $\mathrm{Var}(X) = \frac{2n_2^2(n_1+n_2-2)}{n_1(n_2-4)(n_2-2)^2}$ (if $n_2 > 4$).

Properties: Let $Y_i \sim \chi^2(n_i)$, $i = 1, 2$, and $Y_1 \perp\!\!\!\perp Y_2$, then

$$\frac{Y_1/n_1}{Y_2/n_2} \sim F(n_1, n_2).$$

Functions:
```
df(x, df1, df2)          # df1= n1,   df2= n2
pf(q, df1, df2)
qf(p, df1, df2)
rf(n, df1, df2)
```

A.2.8 Normal or Gaussian distribution

Notation: $X \sim N(\mu, \sigma^2)$, $-\infty < \mu < \infty$, $\sigma^2 > 0$.

Density: $N(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp[-\frac{(x-\mu)^2}{2\sigma^2}]$, $-\infty < x < \infty$.

Moments: $E(X) = \mu$, $\text{Var}(X) = \sigma^2$.

Properties:
- If $\{X_i\} \overset{\text{ind}}{\sim} N(\mu_i, \sigma_i^2)$, then $\sum a_i X_i \sim N(\sum a_i \mu_i, \sum a_i^2 \sigma_i^2)$;
- If $X_1|X_2 \sim N(X_2, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$, then

$$X_1 \sim N(\mu_2, \sigma_1^2 + \sigma_2^2).$$

Functions:
```
dnorm(x, mean=0, sd= 1)   # mean= μ, sd= σ
pnorm(q, mean=0, sd= 1)
qnorm(p, mean=0, sd= 1)
rnorm(n, mean=0, sd= 1)
```

```
================================================================
> x <- seq(-4, 4, 0.1)
> plot(x, dnorm(x), type="l",
                          ylab="Density function of N(0,1)")
# Note that this is the letter "l", not the digit "1"
                                                  # Figure A.4
----------------------------------------------------------------
# An alternative way of creating the plot is

> curve(dnorm(x), from=-4, to = 4,
                          ylab="Density function of N(0,1)")
****************************************************************
```

A.2.9 Multivariate normal or Gaussian distribution

Notation: $\mathbf{x} = (X_1, \ldots, X_d)^\top \sim N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ or $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\mu} \in \mathbb{R}^d$, $\boldsymbol{\Sigma} > 0$.

Density: $N_d(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(\sqrt{2\pi})^d |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\}$, $\boldsymbol{x} \in \mathbb{R}^d$.

Moments: $E(\mathbf{x}) = \boldsymbol{\mu}$, $\text{Var}(\mathbf{x}) = \boldsymbol{\Sigma}$.

Functions: Producing one or more samples from the specified multivariate normal distribution

```
mvrnorm(n= 1, mu, Sigma, tol= 1e-6, empirical= F)
rmvn(n, mu, V)
```

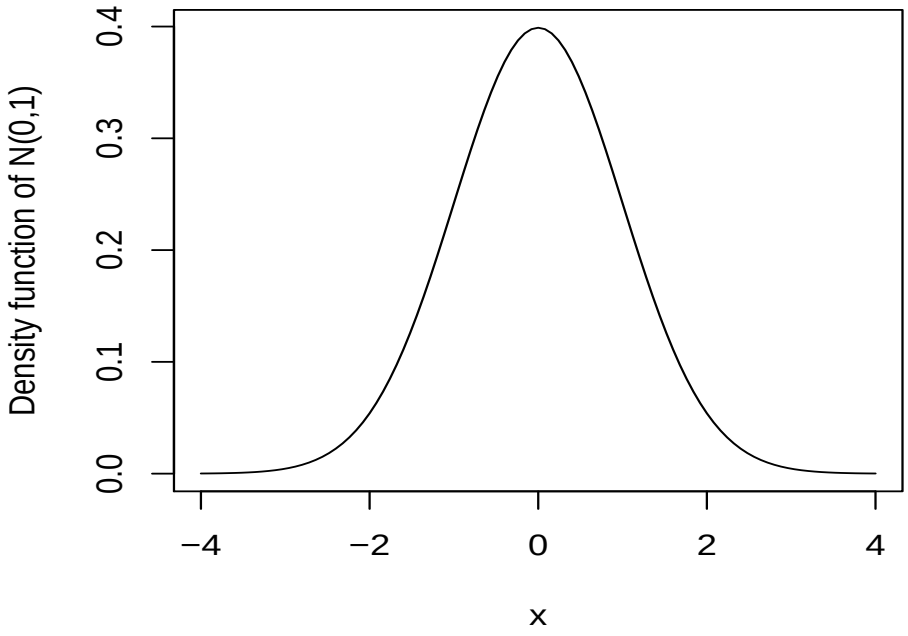

Figure A.4 Density functions of $N(0, 1)$.

APPENDIX B

The Newton–Raphson and Fisher Scoring Algorithms

B.1 Newton's Method for Root Finding

1• The issue

- Let g: $[a, b] \to \mathbb{R}$ be a differentiable function defined on the interval $[a, b]$ with values in the real line $\mathbb{R}$.
- Assume that we want to find the root $x^{(\infty)}$ of the equation

$$g(x) = 0. \tag{B.1}$$

2• Formulation of Newton's method

- Let $x^{(t)}$ denote the t-th approximation of the root $x^{(\infty)}$.
- The first-order Taylor expansion of $g(x)$ in the neighborhood of the $x^{(t)}$ gives

$$g(x) = \underbrace{g(x^{(t)}) + (x - x^{(t)})g'(x^{(t)})}_{h(x)} + R_1,$$

  where R_1 is the remainder and

$$h(x) \doteq g(x^{(t)}) + (x - x^{(t)})g'(x^{(t)})$$

  is the tangent line through the point $(x^{(t)}, g(x^{(t)}))$ with slope $g'(x^{(t)})$ as shown in Figure B.1.
- In the neighborhood of the $x^{(t)}$, we use the straight line $h(x)$ to approximate the curve $g(x)$.
- Let $h(x) = 0$, we obtain the $(t + 1)$-th approximation of $x^{(\infty)}$ as

$$x^{(t+1)} = x^{(t)} - \frac{g(x^{(t)})}{g'(x^{(t)})}, \qquad t = 0, 1, 2, \ldots, \tag{B.2}$$

  which defines Newton's method as illustrated by Figure B.1.

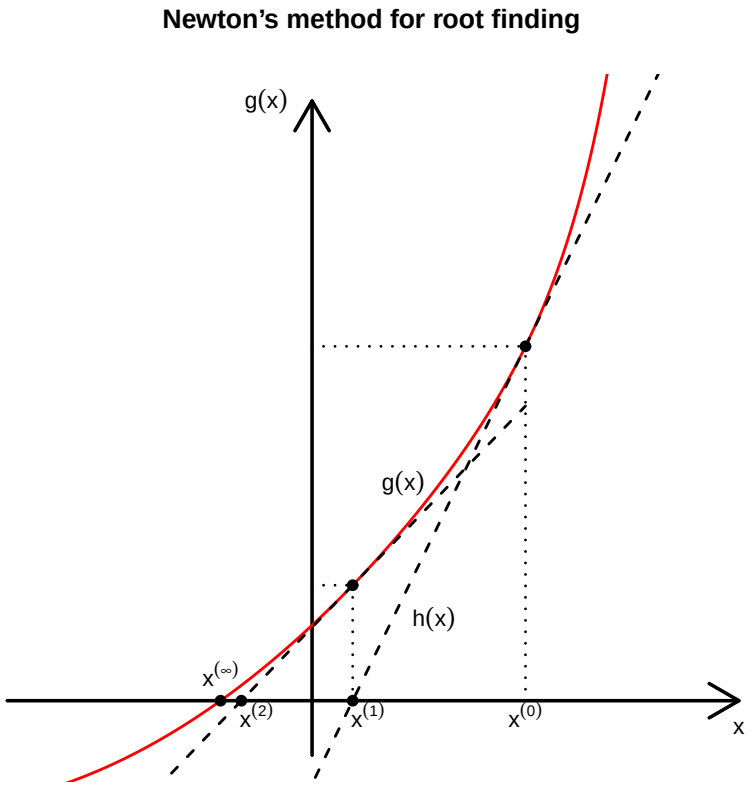

Figure B.1 Starting from the initial value $x^{(0)}$, we draw a segment parallel to the y-axis by connecting two points $(x^{(0)}, 0)$ and $(x^{(0)}, g(x^{(0)}))$. The point $(x^{(1)}, 0)$ is the intersection of the x-axis and the tangent line drawn through the point $(x^{(0)}, g(x^{(0)}))$, thus, we obtain the first approximation $x^{(1)}$ of the root $x^{(\infty)}$ of the equation $g(x) = 0$. Continue this process, we can obtain the second approximation $x^{(2)}$, the third approximation $x^{(3)}$ till $x^{(\infty)}$. Newton's method fails to converge if $x^{(0)}$ is chosen too far from $x^{(\infty)}$.

2.1• Basic idea of Newton's method

— Using the tangent line to approximate the curve locally, and then finding the root of the tangent line iteratively.

2.2• Comments on the name

— In numerical analysis, Newton's method (also known as the Newton–Raphson algorithm), named after Isaac Newton and Joseph Raphson, is a method for finding successively better approximations to the roots (or zeroes) of a real-valued function $g(x)$.

2.3• Comments on the history

— The name "Newton's method" is derived from Isaac Newton's description of a special case of the method in "De analysi per aequationes numero terminorum infinitas" (written in 1669, published in 1711 by William Jones) and in "De metodis fluxionum et serierum infinitarum" (written in 1671, translated and published as "Method of Fluxions" in 1736 by John Colson).

— However, his method differs substantially from the modern method given above: Newton applies the method only to polynomials. He does not compute the successive approximations $x^{(t)}$, but computes a sequence of polynomials, and only at the end arrives at an approximation for the root $x^{(\infty)}$. Newton views the method as purely algebraic and makes no mention of the connection with calculus. Newton may have derived his method from a similar but less precise method by Vieta. Newton's method was used by 17-th century Japanese mathematician Seki Kōwa to solve single-variable equations, though the connection with calculus was missing.

— Newton's method was first published in 1685 in "A Treatise of Algebra both Historical and Practical" by John Wallis. In 1690, Joseph Raphson published a simplified description in "Analysis aequationum universalis". Raphson again viewed Newton's method purely as an algebraic method and restricted its use to polynomials, but he describes the method in terms of the successive approximations $x^{(t)}$ instead of the more complicated sequence of polynomials used by Newton. Finally, in 1740, Thomas Simpson described Newton's method as an iterative method for solving general nonlinear equations using calculus, essentially giving the description above. In the same publication, Simpson also gives the generalization to systems of two equations and notes that Newton's method can be used for solving optimization problems by setting the gradient to zero.

— Arthur Cayley in 1879 in "The Newton-Fourier imaginary problem" was the first to notice the difficulties in generalizing Newton's method to complex roots of polynomials with degree greater than 2 and complex initial values. This opened the way to the study of the theory of iterations of rational functions.

3• An illustrative example

Example B.1 (Numerical illustration). Using Newton's method to find the unique root of the equation $g(x) = 0$ on the interval $(0, \infty)$, where

$$g(x) = 1.95 - \mathrm{e}^{-2/x} - 2\,\mathrm{e}^{-x^4}.$$

Solution. It is easy to verify that

$$g'(x) = -2x^{-2}\,\mathrm{e}^{-2/x} + 8x^3\,\mathrm{e}^{-x^4}.$$

Let $x^{(0)} = 1$. By using a calculator, from (B.2), we obtain

$$\begin{aligned}
x^{(1)} &= x^{(0)} - \frac{g(x^{(0)})}{g'(x^{(0)})} = 1 - \frac{1.0789}{2.6724} = 0.596273,\\
x^{(2)} &= x^{(1)} - \frac{g(x^{(1)})}{g'(x^{(1)})} = x^{(1)} - \frac{0.1526}{1.2981} = 0.478749,\\
x^{(3)} &= x^{(2)} - \frac{g(x^{(2)})}{g'(x^{(2)})} = x^{(2)} - \frac{0.0370}{0.6991} = 0.425798,\\
x^{(4)} &= x^{(3)} - \frac{g(x^{(3)})}{g'(x^{(3)})} = x^{(3)} - \frac{0.0056}{0.4969} = 0.414628,\\
x^{(5)} &= x^{(4)} - \frac{g(x^{(4)})}{g'(x^{(4)})} = x^{(4)} - \frac{2.0768 \times 10^{-4}}{0.460135} = 0.414177,\\
x^{(6)} &= x^{(5)} - \frac{g(x^{(5)})}{g'(x^{(5)})} = x^{(5)} - \frac{3.2684 \times 10^{-7}}{0.458687} = 0.414176, \quad \text{and}\\
x^{(7)} &= x^{(6)} - \frac{g(x^{(6)})}{g'(x^{(6)})} = x^{(6)} - \frac{8.1334 \times 10^{-13}}{0.458685} = 0.414176.
\end{aligned}$$

Thus, $x^{(\infty)} = x^{(7)} = 0.414176$ is the unique root of $g(x) = 0$. ‖

3.1• R codes

```
Newton.method <- function(x0, NumNM)
{  # =============== Aim ====================================
   # Find the root of the equation:
   # g(x)  = 1.95 - exp(-2/x) - 2*exp(-x^4) = 0
   # =============== Input ==================================
   # x0    = initial value of x, x0 = 1
   # NumNM = the number of iterations in the Newton method
```

```
    # ============== Output ================================
    # X     = storing the approximation roots of the equation
    #         g(x) = 0
    # ======================================================
    options(width=68, digits=6)
    x <- x0
    X <- matrix(0, NumNM, 1)
    for (tt in 1:NumNM) {
        gx <- 1.95 - exp(-2/x) - 2*exp(-x^4)
        gpx <- -2*exp(-2/x)/(x^2) + 8*x^3*exp(-x^4)
        x <- x - gx/gpx
        X[tt, 1] <- x
    }
    return(X)
}
```

3.2• Output with initial value $x^{(0)} = 1$

— Let $x^{(0)} = 1$ and NumNM $= 10$, by implementing the above R program, we obtain

—
```
> Newton.method(1, 10)}
           [,1]
 [1,] 0.596273
 [2,] 0.478749
 [3,] 0.425798
 [4,] 0.414628
 [5,] 0.414177
 [6,] 0.414176
 [7,] 0.414176
 [8,] 0.414176
 [9,] 0.414176
[10,] 0.414176
```

— Obviously, if the error is set to be 10^{-6}, the Newton method converged to 0.414176 in 7 iterations.

3.3• Output with initial value $x^{(0)} = 1.5$

— Let $x^{(0)} = 1.5$ and NumNM $= 10$, by implementing the above R program, we obtain

```
— > Newton.method(1.5, 10)
                [,1]
 [1,]  2.78971e+01
 [2,]  4.53963e+02
 [3,]  9.92302e+04
 [4,]  4.67745e+09
 [5,]  1.03923e+19
 [6,]  5.12998e+37
 [7,]  1.25004e+75
 [8,] 7.42238e+149
 [9,]          NaN
[10,]          NaN
```

— Obviously, the Newton method fails to converge.

B.2 Newton's Method for Calculating MLE

4• The issue

- Let $\ell(\theta)$ be the log-likelihood function, which is twice continuously differentiable and concave.
- Suppose that we want to find the MLE of the parameter θ:

$$\hat{\theta} = \arg \max_{\theta \in \Theta} \ell(\theta), \tag{B.3}$$

where Θ is the parameter space.

5• Newton's method

- The MLE $\hat{\theta}$ in (B.3) is the solution of

$$\ell'(\theta) = 0. \tag{B.4}$$

- By comparing (B.4) with (B.1); i.e., by replacing $g(\cdot)$ in (B.2) with $\ell'(\cdot)$, we obtain

$$\theta^{(t+1)} = \theta^{(t)} - \frac{\ell'(\theta^{(t)})}{\ell''(\theta^{(t)})}, \qquad t = 0, 1, 2, \ldots, \tag{B.5}$$

which defines Newton's method for finding MLE.

6• Application to zero-truncated Poisson model

- A discrete r.v. X is said to follow a *zero-truncated Poisson* (ZTP) distribution or positive Poisson distribution, denoted by $X \sim \text{ZTP}(\lambda)$ for $\lambda > 0$, if its pmf is

$$\Pr(X = x) = c \cdot \frac{\lambda^x \, \mathrm{e}^{-\lambda}}{x!}, \quad c \doteq \frac{1}{1 - \mathrm{e}^{-\lambda}}, \quad x = 1, 2, \ldots. \tag{B.6}$$

- This pmf can be viewed as the same as the non-truncated Poisson with an adjustment factor c such that the sum of all terms adds up to 1.

- If λ is large the adjustment factor has little or no effect, but if λ is small (< 4) the adjustment is much greater.

6.1• Background

— The standard Poisson distribution is defined for non-negative integer values of x including $x = 0$.

— However, there are many situations in which the class $x = 0$ is not included in the sample data.

— The ZTP distribution/regression is used to model count data for which the value zero cannot occur.

6.2• Three real examples

— A study of length of hospital stay, in days, as a function of age, kind of health insurance and whether or not the patient died while in the hospital. Length of hospital stay is recorded as a minimum of at least one day.

— A study of the number of journal articles published by tenured faculty as a function of discipline (fine arts, science, social science, humanities, medical, etc). To get tenure faculty must publish. Thus, there are no tenured faculty with zero publication.

— A study by the county traffic court on the number of tickets received by teenagers as predicted by school performance, amount of driver training and gender. Only individuals who have received at least one citation are in the traffic court files.

6.3• The Newton–Raphson algorithm

— Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{ZTP}(\lambda)$ and $Y_{\text{obs}} = \{x_1, \ldots, x_n\}$ denote the observed data, where x_i is the realization of X_i.

— The observed-data likelihood function for λ is given by

$$L(\lambda|Y_{\text{obs}}) = \prod_{i=1}^{n} \frac{\lambda^{x_i}\, \mathrm{e}^{-\lambda}}{(1 - \mathrm{e}^{-\lambda})x_i!}$$

so that the log-likelihood function is

$$\ell(\lambda|Y_{\text{obs}}) \propto n\Big\{\bar{x}\log(\lambda) - \lambda - \log(1 - \mathrm{e}^{-\lambda})\Big\}, \tag{B.7}$$

where $\bar{x} = (1/n)\sum_{i=1}^{n} x_i$.

— The score equation $\ell'(\lambda|Y_{\text{obs}}) = 0$ yields

$$g(\lambda) \;\hat{=}\; \lambda - \bar{x}(1 - \mathrm{e}^{-\lambda}) = 0. \tag{B.8}$$

— Since $g'(\lambda) = 1 - \bar{x}\,\mathrm{e}^{-\lambda}$, the Newton–Raphson algorithm for solving (B.8) is given by

$$\begin{aligned} \lambda^{(t+1)} &= \lambda^{(t)} - \frac{g(\lambda^{(t)})}{g'(\lambda^{(t)})} \\ &= \frac{\bar{x}\{1 - \exp(-\lambda^{(t)}) - \lambda^{(t)}\exp(-\lambda^{(t)})\}}{1 - \bar{x}\exp(-\lambda^{(t)})}. \end{aligned} \tag{B.9}$$

6.4• Cholera data in an Indian village

— Consider the cholera data in an Indian village given in the first two rows of Table B.1.

Table B.1 Observed counts for the cholera data fitted with a ZTP model by excluding the class $x = 0$

x	0	1	2	3	4	$\geqslant 5$	Total
n_x	168	32	16	6	1	0	223

SOURCE: Table 1 of Meng (1997).

— We exclude the class $x = 0$.

6.5• Limitation with the Newton–Raphson algorithm

— Now $\bar{x} = 1.563636$. With $\lambda^{(0)} = 2$, the iterative scheme (B.9) converged to $\hat{\lambda} = 0.972178$ in five iterations, while with $\lambda^{(0)} = 0.4$, the algorithm converged to a wrong solution $\hat{\lambda} = 0$ in nine iterations.

— In fact, the Newton–Raphson algorithm will fail for any initial value in the interval (0, 0.4470139].

— This is not surprising because choosing $\lambda^{(0)} \leqslant \log(\bar{x}) = 0.4470139$ will make the denominator of (B.9) negative or even zero (Meng, 1997, p. 14).

— For more complicated problems, especially high-dimensional cases, the Newton–Raphson algorithm can be very sensitive to the initial value and sometimes fails to converge.

6.6• R codes

```
ZTP.model.NR <- function(lambda0, NumNR)
{  # ================ Aim ====================================
   # Find the MLE of \lambda in (B.7)
   # ================ Input ==================================
   # lambda0 = initial value of \lambda
   # NumNR   = the number of iterations in NR algorithm
   # ================ Output =================================
   # TH      = storing the approximation MLEs of \lambda
   # ==========================================================
   options(width=68, digits=6)
   x <- 1:5
   nx <- c(32, 16, 6, 1, 0)
   n <- sum(nx)
   xbar <- sum(x*nx)/n
   lambda <- lambda0
   TH <- matrix(0, NumNR, 1)
   for (tt in 1:NumNR) {
        a <- exp(-lambda)
        lambda <- xbar*(1-a -lambda*a)/(1-xbar*a)
        TH[tt, 1] <- lambda
   }
   return(TH)
}
```

6.7• Output with initial value $\lambda^{(0)} = 2$

— Let $\lambda^{(0)} = 2$ and NumNR $= 8$, by implementing the above R program, we obtain

—
```
> ZTP.model.NR(2, 8)
          [,1]
[1,] 1.178093
[2,] 0.993280
[3,] 0.972486
[4,] 0.972178
[5,] 0.972178
[6,] 0.972178
[7,] 0.972178
[8,] 0.972178
```

— Obviously, if the error is set to be 10^{-6}, the Newton–Raphson algorithm converged to 0.972178 in 5 iterations.

6.8• Output with initial value $\lambda^{(0)} = 0.4$

— Let $\lambda^{(0)} = 0.4$ and NumNR $= 10$, by implementing the above R program, we obtain

—
```
> ZTP.model.NR(0.4, 10)
               [,1]
 [1,] -1.99940e+00
 [2,] -1.24242e+00
 [3,] -6.51359e-01
 [4,] -2.59075e-01
 [5,] -6.08971e-02
 [6,] -4.56287e-03
 [7,] -2.86041e-05
 [8,] -1.13485e-09
 [9,] -6.60835e-17
[10,] -1.83328e-16
```

— Obviously, the Newton–Raphson algorithm converged to the wrong value of zero.

B.3 The Newton–Raphson Algorithm for High-dimensional Cases

7• The issue

- Let Y_{obs} be the observed data, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)^\top \in \boldsymbol{\Theta}$ be the parameter vector of interest, and $\boldsymbol{\Theta}$ is the parameter space.
- Suppose that we want to find the MLEs

$$\hat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \ell(\boldsymbol{\theta}|Y_{\text{obs}}),$$

where $\ell(\boldsymbol{\theta}|Y_{\text{obs}})$ is the observed-data log-likelihood function.

8• The observed and expected information matrices

- Let $\ell(\boldsymbol{\theta}|Y_{\text{obs}})$ be a twice continuously differentiable and concave function and ∇ denote the derivative operator.
- Mathematically,

$$\nabla \ell(\boldsymbol{\theta}|Y_{\text{obs}}) = \frac{\partial \ell(\boldsymbol{\theta}|Y_{\text{obs}})}{\partial \boldsymbol{\theta}} = \begin{pmatrix} \dfrac{\partial \ell(\boldsymbol{\theta}|Y_{\text{obs}})}{\partial \theta_1} \\ \vdots \\ \dfrac{\partial \ell(\boldsymbol{\theta}|Y_{\text{obs}})}{\partial \theta_m} \end{pmatrix}$$

is called the *gradient vector* and

$$\nabla^2 \ell(\boldsymbol{\theta}|Y_{\text{obs}}) = \frac{\partial^2 \ell(\boldsymbol{\theta}|Y_{\text{obs}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} = \begin{pmatrix} \dfrac{\partial \ell(\boldsymbol{\theta}|Y_{\text{obs}})}{\partial \theta_1^2} & \cdots & \dfrac{\partial \ell(\boldsymbol{\theta}|Y_{\text{obs}})}{\partial \theta_1 \partial \theta_m} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial \ell(\boldsymbol{\theta}|Y_{\text{obs}})}{\partial \theta_m \partial \theta_1} & \cdots & \dfrac{\partial \ell(\boldsymbol{\theta}|Y_{\text{obs}})}{\partial \theta_m^2} \end{pmatrix}$$

is called the *Hessian matrix*.

- Statistically, $\nabla \ell(\boldsymbol{\theta}|Y_{\text{obs}})$ is called the *score vector*, the negative Hessian matrix, denoted by $\mathbf{I}(\boldsymbol{\theta}|Y_{\text{obs}}) = -\nabla^2 \ell(\boldsymbol{\theta}|Y_{\text{obs}})$, which is also called the *observed information matrix* and

$$\mathbf{J}(\boldsymbol{\theta}) = E\{\mathbf{I}(\boldsymbol{\theta}|Y_{\text{obs}})\} = -\int \nabla^2 \ell(\boldsymbol{\theta}|Y_{\text{obs}}) f(Y_{\text{obs}}|\boldsymbol{\theta}) \, \mathrm{d}Y_{\text{obs}}$$

is called the *Fisher/expected information matrix*.

9• The Newton–Raphson algorithm

- Consider the second-order Taylor expansion of $\ell(\boldsymbol{\theta}|Y_{\text{obs}})$ around $\boldsymbol{\theta}^{(t)}$:

$$\begin{aligned}\ell(\boldsymbol{\theta}|Y_{\text{obs}}) &= \ell(\boldsymbol{\theta}^{(t)}|Y_{\text{obs}}) + (\boldsymbol{\theta}-\boldsymbol{\theta}^{(t)})^\top \nabla\ell(\boldsymbol{\theta}^{(t)}|Y_{\text{obs}}) \\ &\quad + \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}^{(t)})^\top \nabla^2\ell(\boldsymbol{\theta}^{(t)}|Y_{\text{obs}})(\boldsymbol{\theta}-\boldsymbol{\theta}^{(t)}) + R_2 \\ &\hat{=} h(\boldsymbol{\theta}) + R_2,\end{aligned}$$

  where R_2 is the remainder.

- In the neighborhood of the $\boldsymbol{\theta}^{(t)}$, we use the quadratic function $h(\boldsymbol{\theta})$ to approximate the log-likelihood function $\ell(\boldsymbol{\theta}|Y_{\text{obs}})$.

- Solving the system of equations

$$\mathbf{0}_m = \frac{\partial h(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \nabla\ell(\boldsymbol{\theta}^{(t)}|Y_{\text{obs}}) + \nabla^2\ell(\boldsymbol{\theta}^{(t)}|Y_{\text{obs}})(\boldsymbol{\theta}-\boldsymbol{\theta}^{(t)}), \tag{B.10}$$

  we obtain

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \mathbf{I}^{-1}(\boldsymbol{\theta}^{(t)}|Y_{\text{obs}})\nabla\ell(\boldsymbol{\theta}^{(t)}|Y_{\text{obs}}), \tag{B.11}$$

  which defines the NR algorithm, where $\boldsymbol{\theta}^{(0)}$ is the initial values of $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^{(t)}$ is the t-th approximation of $\hat{\boldsymbol{\theta}}$.

9.1• Definition of derivatives of a vector

— To derive (B.10), we define

$$\frac{\partial \boldsymbol{b}_{m\times 1}}{\partial x} = \begin{pmatrix} \frac{\partial b_1}{\partial x} \\ \vdots \\ \frac{\partial b_m}{\partial x} \end{pmatrix}, \qquad \frac{\partial b}{\partial \boldsymbol{x}_{n\times 1}} = \begin{pmatrix} \frac{\partial b}{\partial x_1} \\ \vdots \\ \frac{\partial b}{\partial x_n} \end{pmatrix},$$

$$\frac{\partial \boldsymbol{b}^\top}{\partial x} = \left(\frac{\partial b_1}{\partial x}, \ldots, \frac{\partial b_m}{\partial x}\right), \quad \frac{\partial b}{\partial \boldsymbol{x}^\top} = \left(\frac{\partial b}{\partial x_1}, \ldots, \frac{\partial b}{\partial x_n}\right),$$

$$\underbrace{\frac{\partial \boldsymbol{b}^\top}{\partial \boldsymbol{x}}}_{n\times m} = \left(\frac{\partial b_1}{\partial \boldsymbol{x}}, \ldots, \frac{\partial b_m}{\partial \boldsymbol{x}}\right) = \begin{pmatrix} \frac{\partial b_1}{\partial x_1} & \cdots & \frac{\partial b_m}{\partial x_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial b_1}{\partial x_n} & \cdots & \frac{\partial b_m}{\partial x_n} \end{pmatrix}. \tag{B.12}$$

9.2• Some important results on vector derivatives

— We have

$$\frac{\partial(\boldsymbol{a}^\top \boldsymbol{x})}{\partial \boldsymbol{x}} = \frac{\partial(\boldsymbol{x}^\top \boldsymbol{a})}{\partial \boldsymbol{x}} = \boldsymbol{a}_{n\times 1}, \tag{B.13}$$

$$\frac{\partial(\mathbf{A}\boldsymbol{x})}{\partial \boldsymbol{x}^\top} = \mathbf{A}_{m\times n}, \qquad \frac{\partial(\mathbf{A}\boldsymbol{x})^\top}{\partial \boldsymbol{x}} = \mathbf{A}^\top, \tag{B.14}$$

$$\frac{\partial(\boldsymbol{x}^\top \mathbf{C}\boldsymbol{x})}{\partial \boldsymbol{x}} = (\mathbf{C}_{n\times n} + \mathbf{C}^\top)\boldsymbol{x}. \tag{B.15}$$

9.3• Proof of (B.13)

— The l-th component of the left–hand side of (B.13) is

$$\frac{\partial(\boldsymbol{a}^\top \boldsymbol{x})}{\partial x_l} = \frac{\partial(\boldsymbol{x}^\top \boldsymbol{a})}{\partial x_l} = \frac{\partial(\sum_{i=1}^n a_i x_i)}{\partial x_l} = a_l, \quad l = 1, \ldots, n,$$

which implies (B.13). □

9.4• Proof of (B.14)

— We only prove the second formula in (B.14). Let

$$\mathbf{A} = \begin{pmatrix} \boldsymbol{a}_{(1)}^\top \\ \vdots \\ \boldsymbol{a}_{(m)}^\top \end{pmatrix}.$$

Then $\mathbf{A}^\top = (\boldsymbol{a}_{(1)}, \ldots, \boldsymbol{a}_{(m)})$.

— Define $\boldsymbol{b} = \mathbf{A}\boldsymbol{x}$, we have $b_j = \boldsymbol{a}_{(j)}^\top \boldsymbol{x}$ for $j = 1, \ldots, m$. Hence,

$$\begin{aligned} \frac{\partial(\mathbf{A}\boldsymbol{x})^\top}{\partial \boldsymbol{x}} &= \frac{\partial \boldsymbol{b}^\top}{\partial \boldsymbol{x}} \\ &\overset{\text{(B.12)}}{=} \left(\frac{\partial b_1}{\partial \boldsymbol{x}}, \ldots, \frac{\partial b_m}{\partial \boldsymbol{x}}\right) \\ &= \left(\frac{\partial(\boldsymbol{a}_{(1)}^\top \boldsymbol{x})}{\partial \boldsymbol{x}}, \ldots, \frac{\partial(\boldsymbol{a}_{(m)}^\top \boldsymbol{x})}{\partial \boldsymbol{x}}\right) \\ &\overset{\text{(B.13)}}{=} (\boldsymbol{a}_{(1)}, \ldots, \boldsymbol{a}_{(m)}) = \mathbf{A}^\top, \end{aligned}$$

which implies (B.14). □

9.5• Proof of (B.15)

— Let

$$\mathbf{C}_{n\times n} = (c_{ij}) = \begin{pmatrix} \boldsymbol{c}_{(1)}^\top \\ \vdots \\ \boldsymbol{c}_{(n)}^\top \end{pmatrix} = (\boldsymbol{c}_1, \ldots, \boldsymbol{c}_n),$$

we only need to prove that the l-th component of the left-hand side of (B.15) is

$$\frac{\partial(\boldsymbol{x}^\top \mathbf{C}\boldsymbol{x})}{\partial x_l} = (\boldsymbol{c}_{(l)}^\top + \boldsymbol{c}_l^\top)\boldsymbol{x} \qquad \text{for } l = 1, \ldots, n. \tag{B.16}$$

— Since

$$\begin{aligned} \boldsymbol{x}^\top \mathbf{C}\boldsymbol{x} &= \sum_{i=1}^n \sum_{j=1}^n x_i c_{ij} x_j = \sum_{i=1}^n x_i^2 c_{ii} + \sum_{i,j=1}^n \sum_{i\neq j} x_i c_{ij} x_j \\ &= x_l^2 c_{ll} + \sum_{j=1,\, j\neq l}^n x_l c_{lj} x_j + \sum_{i=1,\, i\neq l}^n x_i c_{il} x_l + c^*, \end{aligned}$$

where c^* is a constant, not depending on x_l, we have

$$\begin{aligned} \frac{\partial(\boldsymbol{x}^\top \mathbf{C}\boldsymbol{x})}{\partial x_l} &= 2x_l c_{ll} + \sum_{j=1,\, j\neq l}^n c_{lj} x_j + \sum_{i=1,\, i\neq l}^n x_i c_{il} \\ &= \sum_{j=1}^n c_{lj} x_j + \sum_{i=1}^n x_i c_{il} \\ &= (c_{l1}, \ldots, c_{ln})\boldsymbol{x} + (c_{1l}, \ldots, c_{nl})\boldsymbol{x} = (\boldsymbol{c}_{(l)}^\top + \boldsymbol{c}_l^\top)\boldsymbol{x}, \end{aligned}$$

which implies (B.16). □

9.6• Proof of (B.10)

— Let $\mathbf{B} = \nabla^2 \ell(\boldsymbol{\theta}^{(t)}|Y_{\text{obs}})$. Thus,

$$\begin{aligned} \mathbf{0}_m &= \frac{\partial h(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \\ &\overset{\text{(B.13) \& (B.15)}}{=} \mathbf{0} + \nabla \ell(\boldsymbol{\theta}^{(t)}|Y_{\text{obs}}) + \frac{1}{2}(\mathbf{B} + \mathbf{B}^\top)(\boldsymbol{\theta} - \boldsymbol{\theta}^{(t)}) \\ &= \nabla \ell(\boldsymbol{\theta}^{(t)}|Y_{\text{obs}}) + \mathbf{B}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(t)}), \end{aligned}$$

implying (B.10). □

10• Application to univariate t distribution

- The density of $X \sim t(\mu, \sigma^2, \nu)$ with $\mu \in \mathbb{R}$ and $\sigma^2 \in \mathbb{R}_+$ is

$$t(x|\mu, \sigma^2, \nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\nu\pi}\sigma}\left\{1 + \frac{(x-\mu)^2}{\nu\sigma^2}\right\}^{-\frac{\nu+1}{2}}, \quad x \in \mathbb{R}.$$

  When $\mu = 0$ and $\sigma^2 = 1$, it is called the standard t distribution, denoted by $t(\nu)$.

- Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} t(\mu, \sigma^2, \nu)$, where (μ, σ^2) are two unknown parameters and $\nu > 0$ is known.

- The aim is to find the MLEs of μ and σ^2.

10.1• The Newton–Raphson algorithm

— Let $Y_{\text{obs}} = \{x_1, \ldots, x_n\}$ and $\boldsymbol{\theta} = (\mu, \sigma^2)^\top$.

— The observed-data likelihood function is

$$\begin{aligned} L(\boldsymbol{\theta}|Y_{\text{obs}}) &= \prod_{i=1}^{n} \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\nu\pi}\sigma}\left\{1 + \frac{(x_i-\mu)^2}{\nu\sigma^2}\right\}^{-\frac{\nu+1}{2}} \\ &\propto (\sigma^2)^{-n/2} \prod_{i=1}^{n}\left\{1 + \frac{(x_i-\mu)^2}{\nu\sigma^2}\right\}^{-\frac{\nu+1}{2}} \end{aligned}$$

  so that the log-likelihood function is given by

$$\ell(\boldsymbol{\theta}|Y_{\text{obs}}) = -\frac{n}{2}\log(\sigma^2) - \frac{\nu+1}{2}\sum_{i=1}^{n}\log\left\{1 + \frac{(x_i-\mu)^2}{\nu\sigma^2}\right\}.$$

— The score vector is

$$\begin{aligned} \nabla\ell(\boldsymbol{\theta}|Y_{\text{obs}}) &= \begin{pmatrix} \dfrac{\partial\ell(\boldsymbol{\theta}|Y_{\text{obs}})}{\partial\mu} \\ \dfrac{\partial\ell(\boldsymbol{\theta}|Y_{\text{obs}})}{\partial\sigma^2} \end{pmatrix} \\ &= \begin{pmatrix} (\nu+1)\displaystyle\sum_{i=1}^{n}\frac{x_i-\mu}{\nu\sigma^2 + (x_i-\mu)^2} \\ -\dfrac{n}{2\sigma^2} + \dfrac{\nu+1}{2\sigma^2}\displaystyle\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\nu\sigma^2 + (x_i-\mu)^2} \end{pmatrix}. \end{aligned}$$

— Clearly, we cannot obtain a closed-form solution to the score equation $\nabla\ell(\boldsymbol{\theta}|Y_{\text{obs}}) = \mathbf{0}$.

— To apply the Newton–Raphson algorithm (B.11), we need to calculate the inverse of the observed information matrix

$$\mathbf{I}(\boldsymbol{\theta}|Y_{\text{obs}}) = -\begin{pmatrix} \dfrac{\partial^2\ell(\boldsymbol{\theta}|Y_{\text{obs}})}{\partial\mu^2} & \dfrac{\partial^2\ell(\boldsymbol{\theta}|Y_{\text{obs}})}{\partial\mu\partial\sigma^2} \\ \dfrac{\partial^2\ell(\boldsymbol{\theta}|Y_{\text{obs}})}{\partial\mu\partial\sigma^2} & \dfrac{\partial^2\ell(\boldsymbol{\theta}|Y_{\text{obs}})}{\partial(\sigma^2)^2} \end{pmatrix},$$

where

$$\begin{cases} \dfrac{\partial^2\ell(\boldsymbol{\theta}|Y_{\text{obs}})}{\partial\mu^2} &= -(\nu+1)\displaystyle\sum_{i=1}^{n}\frac{\nu\sigma^2-(x_i-\mu)^2}{\{\nu\sigma^2+(x_i-\mu)^2\}^2}, \\ \dfrac{\partial^2\ell(\boldsymbol{\theta}|Y_{\text{obs}})}{\partial\mu\partial\sigma^2} &= -\nu(\nu+1)\displaystyle\sum_{i=1}^{n}\frac{x_i-\mu}{\{\nu\sigma^2+(x_i-\mu)^2\}^2}, \\ \dfrac{\partial^2\ell(\boldsymbol{\theta}|Y_{\text{obs}})}{\partial(\sigma^2)^2} &= \dfrac{n}{2\sigma^4}-\dfrac{\nu+1}{2\sigma^4}\displaystyle\sum_{i=1}^{n}\frac{(x_i-\mu)^2\{2\nu\sigma^2+(x_i-\mu)^2\}}{\{\nu\sigma^2+(x_i-\mu)^2\}^2}. \end{cases}$$

B.4 The Fisher Scoring Algorithm

11• The Fisher scoring algorithm

- A variant of the NR algorithm is the (Fisher) scoring algorithm, where the observed information matrix in (B.11) is replaced by the Fisher information matrix:

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \mathbf{J}^{-1}(\boldsymbol{\theta}^{(t)})\nabla\ell(\boldsymbol{\theta}^{(t)}|Y_{\text{obs}}).$$

12• Application to zero-truncated Poisson model

- From (B.7), the first and second derivatives of the log-likelihood function are given by

$$\frac{\mathrm{d}\ell(\lambda|Y_{\text{obs}})}{\mathrm{d}\lambda} = n\left(\frac{\bar{x}}{\lambda}-\frac{1}{1-\mathrm{e}^{-\lambda}}\right)$$

and

$$\frac{\mathrm{d}^2\ell(\lambda|Y_{\text{obs}})}{\mathrm{d}\lambda^2} = n\left\{-\frac{\bar{x}}{\lambda^2}+\frac{\mathrm{e}^{-\lambda}}{(1-\mathrm{e}^{-\lambda})^2}\right\},$$

respectively.

- From (B.6), we have $E(X) = \lambda/(1 - \mathrm{e}^{-\lambda})$. Thus, the Fisher information is

$$J(\lambda) = E\left\{-\frac{\mathrm{d}^2\ell(\lambda|Y_{\text{obs}})}{\mathrm{d}\lambda^2}\right\} = \frac{n(1 - \mathrm{e}^{-\lambda} - \lambda\,\mathrm{e}^{-\lambda})}{\lambda(1 - \mathrm{e}^{-\lambda})^2}.$$

- Let $\lambda^{(0)}$ be the initial value of the MLE $\hat{\lambda}$. If $\lambda^{(t)}$ denotes the t-th approximation of $\hat{\lambda}$, then, its $(t+1)$-th approximation can be obtained by the following Fisher scoring algorithm:

$$\begin{aligned}\lambda^{(t+1)} &= \lambda^{(t)} + J^{-1}(\lambda^{(t)})\frac{\mathrm{d}\ell(\lambda^{(t)}|Y_{\text{obs}})}{\mathrm{d}\lambda} \\ &= \lambda^{(t)} + \frac{\{1 - \mathrm{e}^{-\lambda^{(t)}}\}\{\bar{x} - \bar{x}\,\mathrm{e}^{-\lambda^{(t)}} - \lambda^{(t)}\}}{1 - \mathrm{e}^{-\lambda^{(t)}} - \lambda^{(t)}\,\mathrm{e}^{-\lambda^{(t)}}}.\end{aligned}$$

12.1• R codes

```
ZTP.model.FS <- function(lambda0, NumFS)
{  # ================ Aim ====================================
   # Find the MLE of \lambda in (B.7)
   # ================ Input ==================================
   # lambda0 = initial value of \lambda
   # NumFS   = the number of iterations in FS algorithm
   # ================ Output =================================
   # TH      = storing the approximation MLEs of \lambda
   # ===========================================================
   options(width=68, digits=6)
   x <- 1:5;  nx <- c(32, 16, 6, 1, 0)
   n <- sum(nx)
   xbar <- sum(x*nx)/n
   lambda <- lambda0
   TH <- matrix(0, NumFS, 1)
   for (tt in 1:NumFS) {
        a <- exp(-lambda)
        b <- 1 - a - lambda*a
        lambda <- lambda + (1-a)*(xbar*(1-a)-lambda)/b
        TH[tt, 1] <- lambda
   }
   return(TH)
}
```

12.2• Output with initial value $\lambda^{(0)} = 2$

— Let $\lambda^{(0)} = 2$ and NumFS $= 4$, by implementing the above R program, we obtain

—
```
> ZTP.mode.FS(2, 4)
          [,1]
[1,] 1.056751
[2,] 0.972979
[3,] 0.972178
[4,] 0.972178
```

— Obviously, if the error is set to be 10^{-6}, the Fisher scoring algorithm converged to 0.972178 in 4 iterations.

12.3• Output with initial value $\lambda^{(0)} = 0.4$

— Let $\lambda^{(0)} = 0.4$ and NumFS $= 4$, by implementing the above R program, we obtain

—
```
> ZTP.mode.FS(0.4, 4)
          [,1]
[1,] 1.018630
[2,] 0.972423
[3,] 0.972178
[4,] 0.972178
```

— Obviously, if the error is set to be 10^{-6}, the Fisher scoring algorithm still converged to 0.972178 in 4 iterations, while the NR algorithm converged to the wrong value of zero.

— If we take $\lambda^{(0)} = 0.01$, we have

—
```
> ZTP.mode.FS(0.01, 5)
          [,1]
[1,] 1.123544
[2,] 0.974684
[3,] 0.972179
[4,] 0.972178
[5,] 0.972178
```