Chapter 1

Elements of Statistics

In this chapter, the statistical methods used in the proceeding chapters are summarized. Mood, Graybill and Bose (1974), Hogg and Craig (1995) and Stuart and Ord (1991, 1994) are good references in Sections 1.1 – 1.8, while Judge, Hill, Griffiths and Lee (1980) and Greene (1993, 1997) are representative textbooks in Section 1.9.

1.1 Event and Probability

1.1.1 Event

We consider an **experiment** whose outcome is not known in advance, which is sometimes called a **random experiment**. The **sample space** of an experiment is the set of all possible outcomes. Each element of a sample space is called an **element** of the sample space or a **sample point**, which represents each outcome obtained by the experiment. An **event** is any collection of outcomes contained in the sample space, or equivalently a subset of the sample space. A **simple event** consists of exactly one element and a **compound event** consists of more than one element. Sample space is denoted by Ω and sample point is given by ω.

Suppose that event A is a subset of sample space Ω. Let ω be a sample point in event A. Then, we say that a sample point ω is contained in a sample space A, which is denoted by $\omega \in A$.

A set of the sample points which do not belong to event A is called the **complementary event**, which is denoted by A^c. An event which do not have any sample point is called the **empty event**, denoted by $\emptyset$. Conversely, an event which

includes all possible sample points is called the **whole event**, represented by Ω.

Next, consider two events A and B. A set consisting of the whole sample points which belong to either event A or event B is called the **sum event**, which is denoted by $A \cap B$. A set consisting of the whole sample points which belong to both event A and event B is called the **product event**, denoted by $A \cap B$. When $A \cap B = \emptyset$, we say that events A and B are **mutually exclusive**.

Example 1.1: Consider an experiment of casting a die. We have six sample points, which are denoted by $\omega_1 = \{1\}$, $\omega_2 = \{2\}$, $\omega_3 = \{3\}$, $\omega_4 = \{4\}$, $\omega_5 = \{5\}$ and $\omega_6 = \{6\}$, where ω_i represents the sample point that we have i. In this experiment, the sample space is given by $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6\}$. Let A be the event that we have even numbers and B be the event that we have multiples of three. Then, we can write as $A = \{\omega_2, \omega_4, \omega_6\}$ and $B = \{\omega_3, \omega_6\}$. The complementary event of A is given by $A^c = \{\omega_1, \omega_3, \omega_5\}$, which is the event that we have odd numbers. The sum event of A and B is written as $A \cup B = \{\omega_2, \omega_3, \omega_4, \omega_6\}$, while the product event is $A \cap B = \{\omega_6\}$. Since $A \cap A^c = \emptyset$, we have the fact that A and A^c are mutually exclusive.

Example 1.2: Consider an experiment that consists in flipping a coin three times. In this case, we have the following eight sample points:

$$\begin{array}{llll} \omega_1 = \text{(H,H,H)}, & \omega_2 = \text{(H,H,T)}, & \omega_3 = \text{(H,T,H)}, & \omega_4 = \text{(H,T,T)}, \\ \omega_5 = \text{(T,H,H)}, & \omega_6 = \text{(T,H,T)}, & \omega_7 = \text{(T,T,H)}, & \omega_8 = \text{(T,T,T)}, \end{array}$$

where H represents head while T indicates tail. For example, (H,T,H) means that the first flip lands head, the second flip is tail and the third one is head. Therefore, the sample space of this experiments can be written as:

$$\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6, \omega_7, \omega_8\}.$$

Let A be an event that we have two heads, B be an event that we obtain at least one tail, C be an event that we have head in the second flip, and D be an event that we obtain tail in the third flip. Then, the events A, B and C are give by:

$$\begin{aligned} A &= \{\omega_2, \omega_3, \omega_5\}, \\ B &= \{\omega_2, \omega_3, \omega_4, \omega_5, \omega_6, \omega_7, \omega_8\}, \\ C &= \{\omega_1, \omega_2, \omega_5, \omega_6\}, \\ D &= \{\omega_2, \omega_4, \omega_6, \omega_8\}. \end{aligned}$$

Since A is a subset of B, denoted by $A \subset B$, a sum event is $A \cup B = B$, while a product event is $A \cap B = A$. Moreover, we obtain $C \cap D = \{\omega_2, \omega_6\}$ and $C \cup D = \{\omega_1, \omega_2, \omega_4, \omega_5, \omega_6, \omega_8\}$.

1.1.2 Probability

Let $n(A)$ be the number of sample points in A. We have $n(A) \leq n(B)$ when $A \subseteq B$. Each sample point is equally likely to occur. In the case of Example 1.1 (Section 1.1.1), each of the six possible outcomes has probability 1/6 and in Example 1.2 (Section 1.1.1), each of the eight possible outcomes has probability 1/8. Thus, the probability which the event A occurs is defined as:

$$P(A) = \frac{n(A)}{n(\Omega)}.$$

In Example 1.1, $P(A) = 3/6$ and $P(A \cap B) = 1/6$ are obtained, because $n(\Omega) = 6$, $n(A) = 3$ and $n(A \cap B) = 1$. Similarly, in Example 1.2, we have $P(C) = 4/8$, $P(A \cap B) = P(A) = 3/8$ and so on. Note that we obtain $P(A) \leq P(B)$ when $A \subseteq B$.

It is known that we have the following three properties on probability:

$$0 \leq P(A) \leq 1, \tag{1.1}$$

$$P(\Omega) = 1, \tag{1.2}$$

$$P(\emptyset) = 0. \tag{1.3}$$

$\emptyset \subseteq A \subseteq \Omega$ implies $n(\emptyset) \leq n(A) \leq n(\Omega)$. Therefore, we have:

$$\frac{n(\emptyset)}{n(\Omega)} \leq \frac{n(A)}{n(\Omega)} \leq \frac{n(\Omega)}{n(\Omega)} = 1.$$

Dividing by $n(\Omega)$, we obtain:

$$P(\emptyset) \leq P(A) \leq P(\Omega) = 1.$$

Because $\emptyset$ has no sample point, the number of the sample point is given by $n(\emptyset) = 0$ and accordingly we have $P(\emptyset) = 0$. Therefore, $0 \leq P(A) \leq 1$ is obtained as in (1.1). Thus, (1.1) – (1.3) are obtained.

When events A and B are mutually exclusive, i.e., when $A \cap B = \emptyset$, then $P(A \cup B) = P(A) + P(B)$ holds. Moreover, since A and A^c are mutually exclusive, $P(A^c) = 1 - P(A)$ is obtained. Note that $P(A \cup A^c) = 1$ holds. Generally, unless A and B are not exclusive, we have the following formula:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B),$$

which is known as the **addition rule**. In Example 1.1, each probability is given by $P(A \cup B) = 2/3$, $P(A) = 1/2$, $P(B) = 1/3$ and $P(A \cap B) = 1/6$. Thus, in the example we can verify that the above addition rule holds.

The probability which event A occurs, given that event B has occurred, is called the **conditional probability**, i.e.,

$$P(A|B) = \frac{n(A \cap B)}{n(B)} = \frac{P(A \cap B)}{P(B)},$$

or equivalently,

$$P(A \cap B) = P(A|B)P(B),$$

which is called the **multiplication rule**. When event A is **independent** of event B, we have $P(A \cap B) = P(A)P(B)$, which implies that $P(A|B) = P(A)$. Conversely, $P(A \cap B) = P(A)P(B)$ implies that A is independent of B. In Example 1.2, because of $P(A \cap C) = 1/4$ and $P(C) = 1/2$, the conditional probability $P(A|C) = 1/2$ is obtained. From $P(A) = 3/8$, we have $P(A \cap C) \neq P(A)P(C)$. Therefore, A is not independent of C. As for C and D, since we have $P(C) = 1/2$, $P(D) = 1/2$ and $P(C \cap D) = 1/4$, we can show that C is independent of D.

1.2 Random Variable and Distribution

1.2.1 Univariate Random Variable and Distribution

The **random variable** X is defined as the real value function on sample space Ω. Since X is a function of a sample point ω, it is written as $X = X(\omega)$. Suppose that $X(\omega)$ takes a real value on the interval I. That is, X depends on a set of the sample point ω, i.e., $\{\omega; X(\omega) \in I\}$, which is simply written as $\{X \in I\}$.

In Example 1.1 (Section 1.1.1), suppose that X is a random variable which takes the number of spots up on the die. Then, X is a function of ω and takes the following values:

$$X(\omega_1) = 1, \quad X(\omega_2) = 2, \quad X(\omega_3) = 3, \quad X(\omega_4) = 4,$$
$$X(\omega_5) = 5, \quad X(\omega_6) = 6.$$

In Example 1.2 (Section 1.1.1), suppose that X is a random variable which takes the number of heads. Depending on the sample point ω_i, X takes the following values:

$$X(\omega_1) = 3, \quad X(\omega_2) = 2, \quad X(\omega_3) = 2, \quad X(\omega_4) = 1,$$
$$X(\omega_5) = 2, \quad X(\omega_6) = 1, \quad X(\omega_7) = 1, \quad X(\omega_8) = 0.$$

Thus, the random variable depends on a sample point.

There are two kinds of random variables. One is a **discrete random variable**, while another is a **continuous random variable**.

Discrete random variable and Probability function: Suppose that the discrete random variable X takes $x_1, x_2, \cdots$, where $x_1 < x_2 < \cdots$ is assumed. Consider the probability that X takes x_i, i.e., $P(X = x_i) = p_i$, which is a function of x_i. That is, a function of x_i, say $f(x_i)$, is associated with $P(X = x_i) = p_i$. The function $f(x_i)$ represents the probability in the case where X takes x_i. Therefore, we have the following relation:

$$P(X = x_i) = p_i = f(x_i), \quad i = 1, 2, \cdots,$$

where $f(x_i)$ is called the **probability function** of X.

More formally, the function $f(x_i)$ which has the following properties is defined as the probability function.

$$f(x_i) \geq 0, \quad i = 1, 2, \cdots,$$
$$\sum_i f(x_i) = 1.$$

Furthermore, for an event A, we have the following equation:

$$P(X \in A) = \sum_{x_i \in A} f(x_i).$$

Several functional forms of $f(x_i)$ are shown in Section 2.4.

In Example 1.2 (Section 1.1.1), all the possible values of X are 0, 1, 2 and 3. That is, $x_1 = 0$, $x_2 = 1$, $x_3 = 2$ and $x_4 = 3$ are assigned in this case. The probability that X takes x_1, x_2, x_3 or x_4 is given by:

$$P(X = 0) = f(0) = P(\{\omega_8\}) = \frac{1}{8},$$
$$P(X = 1) = f(1) = P(\{\omega_4, \omega_6, \omega_7\}) = P(\{\omega_4\}) + P(\{\omega_6\}) + P(\{\omega_7\}) = \frac{3}{8},$$
$$P(X = 2) = f(2) = P(\{\omega_2, \omega_3, \omega_5\}) = P(\{\omega_2\}) + P(\{\omega_3\}) + P(\{\omega_5\}) = \frac{3}{8},$$
$$P(X = 3) = f(3) = P(\{\omega_1\}) = \frac{1}{8},$$

which can be written as:

$$P(X = x) = f(x) = \frac{3!}{x!(3-x)!}(\frac{1}{2})^3, \quad x = 0, 1, 2, 3.$$

For $P(X = 1)$ and $P(X = 2)$, note that each sample point is mutually exclusive. The above probability function is called the **binomial distribution** discussed in Section 2.4.5. Thus, we can check $f(x) \geq 0$ and $\sum_x f(x) = 1$ in Example 1.2.

Continuous random variable and Probability density function: Whereas a discrete random variable assumes at most a countable set of possible values, a continuous random variable X takes any real number within an interval I. For the interval I, the probability which X is contained in A is defined as:

$$P(X \in I) = \int_I f(x)\mathrm{d}x.$$

For example, let I be the interval between a and b for $b > a$. Then, we can rewrite $P(X \in I)$ as follows:

$$P(a < X < b) = \int_a^b f(x)\mathrm{d}x,$$

where $f(x)$ is called the **probability density function** of X, or simply the **density function** of X.

In order for $f(x)$ to be a probability density function, $f(x)$ has to satisfy the following properties:

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

Some functional forms of $f(x)$ are discussed in Sections 2.1 – 2.3.

For a continuous random variable, note as follows:

$$P(X = x) = \int_x^x f(t)\mathrm{d}t = 0.$$

In the case of discrete random variables, $P(X = x_i)$ represents the probability which X takes x_i, i.e., $p_i = f(x_i)$. Thus, the probability function $f(x_i)$ itself implies probability. However, in the case of continuous random variables, $P(a < X < b)$ indicates the probability which X lies on the interval (a, b).

Example 1.3: As an example, consider the following function:

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

Clearly, since $f(x) \geq 0$ for $-\infty < x < \infty$ and $\int_{-\infty}^{\infty} f(x)\mathrm{d}x = \int_0^1 f(x)\mathrm{d}x = [x]_0^1 = 1$, the above function can be a probability density function. In fact, it is called a **uniform distribution**. See Section 2.1 for the uniform distribution.

Example 1.4: As another example, consider the following function:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2},$$

for $-\infty < x < \infty$. Clearly, we have $f(x) \geq 0$ for all x. We check whether $\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1$. Define $I = \int_{-\infty}^{\infty} f(x)\mathrm{d}x$.

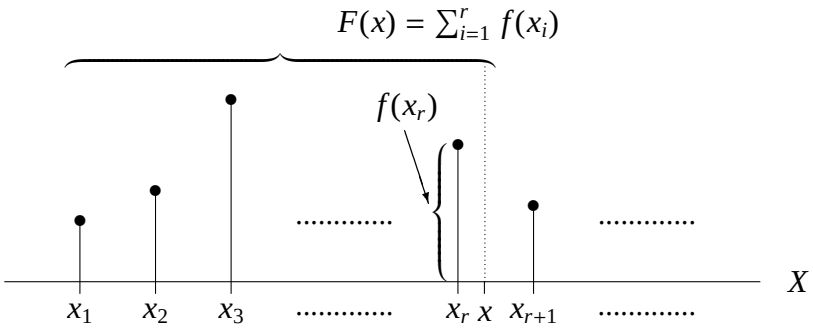
To prove $I = 1$, we may prove $I^2 = 1$ because of $f(x) > 0$ for all x, which is shown as follows:

$$\begin{aligned} I^2 &= \Big(\int_{-\infty}^{\infty} f(x)\mathrm{d}x\Big)^2 = \Big(\int_{-\infty}^{\infty} f(x)\mathrm{d}x\Big)\Big(\int_{-\infty}^{\infty} f(y)\mathrm{d}y\Big) \\ &= \Big(\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}x^2)\mathrm{d}x\Big)\Big(\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}y^2)\mathrm{d}y\Big) \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\Big(-\frac{1}{2}(x^2+y^2)\Big)\mathrm{d}x\mathrm{d}y \\ &= \frac{1}{2\pi}\Big(\int_0^{2\pi} \mathrm{d}\theta\Big)\Big(\int_0^{\infty} \exp(-\frac{1}{2}r^2) r\mathrm{d}r\Big) \\ &= \frac{1}{2\pi}\Big(\int_0^{2\pi} \mathrm{d}\theta\Big)\Big(\int_0^{\infty} \exp(-s)\mathrm{d}s\Big) = \frac{1}{2\pi} 2\pi[-\exp(-s)]_0^{\infty} = 1. \end{aligned}$$

In the fifth equality, integration by substitution is used. See Appendix 1.1 for the integration by substitution. $x = r\cos\theta$ and $y = r\sin\theta$ are taken for transformation, which is a one-to-one transformation from (x, y) to (r, θ). Note that $0 < r < +\infty$ and $0 < \theta < 2\pi$. The Jacobian is given by:

$$J = \begin{vmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{vmatrix} = \begin{vmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{vmatrix} = r.$$

Figure 1.1: Probability Function $f(x)$ and Distribution Function $F(x)$
— Discrete Random Variable —



Note that r is the integer which satisfies $x_r \leq x < x_{r+1}$.

In the second integration of the sixth equality, again, integration by substitution is utilized, where transformation is $s = \frac{1}{2}r^2$.

Thus, we obtain the result $I^2 = 1$ and accordingly we have $I = 1$ because of $f(x) \geq 0$. Therefore, $f(x) = e^{-\frac{1}{2}x^2}/\sqrt{2\pi}$ is also taken as a probability density function, which is called the **standard normal probability density function**, discussed in Section 2.2.1.

Distribution Function: The **distribution function** (or the **cumulative distribution function**), denoted by $F(x)$, is defined as:

$$P(X \leq x) = F(x).$$

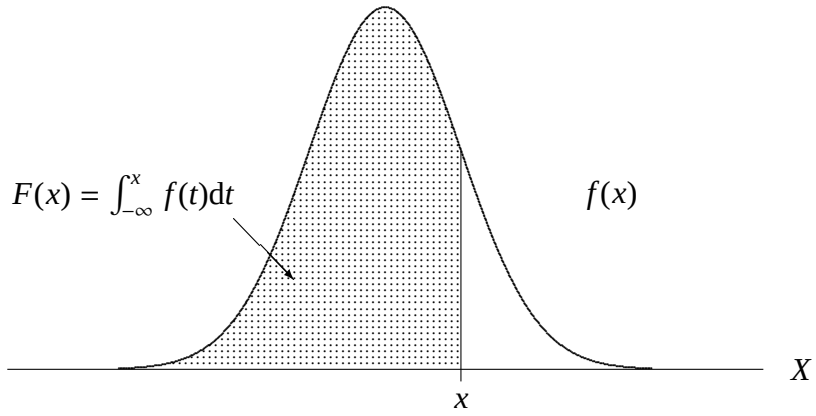
The properties of the distribution function $F(x)$ are represented by:

$$\begin{aligned} &F(x_1) \leq F(x_2), \quad \text{for } x_1 < x_2, \\ &P(a < X \leq b) = F(b) - F(a), \\ &F(-\infty) = 0, \quad F(+\infty) = 1. \end{aligned}$$

The difference between the discrete and continuous random variables is given by:

1. Discrete random variable (Figure 1.1):

Figure 1.2: Density Function $f(x)$ and Distribution Function $F(x)$
— Continuous Random Variable —



- $F(x) = \sum_{i=1}^{r} f(x_i) = \sum_{i=1}^{r} p_i$, where r denotes the integer which satisfies $x_r \leq x < x_{r+1}$.
- $F(x_i) - F(x_i - \epsilon) = f(x_i) = p_i$, where ϵ is a small positive number less than $x_i - x_{i-1}$.

2. Continuous random variable (Figure 1.2):
   - $F(x) = \int_{-\infty}^{x} f(t)\mathrm{d}t$,
   - $F'(x) = f(x)$.

$f(x)$ and $F(x)$ are displayed in Figure 1.1 for a discrete random variable and Figure 1.2 for a continuous random variable.

1.2.2 Multivariate Random Variable and Distribution

We consider two random variables X and Y in this section. It is easy to extend to more than two random variables.

Discrete Random Variables: Suppose that discrete random variables X and Y take $x_1, x_2, \cdots$ and $y_1, y_2, \cdots$, respectively. The probability which event $\{\omega; X(\omega) =$

x_i and $Y(\omega) = y_j\}$ occurs is given by:

$$P(X = x_i, Y = y_j) = f_{xy}(x_i, y_j),$$

where $f_{xy}(x_i, y_j)$ represents the **joint probability function** of X and Y. In order for $f_{xy}(x_i, y_j)$ to be a joint probability function, $f_{xy}(x_i, y_j)$ has to satisfies the following properties:

$$f_{xy}(x_i, y_j) \geq 0, \quad i, j = 1, 2, \cdots$$
$$\sum_{i,j} f_{xy}(x_i, y_j) = 1.$$

Define $f_x(x_i)$ and $f_y(y_j)$ as:

$$f_x(x_i) = \sum_j f_{xy}(x_i, y_j), \quad i = 1, 2, \cdots,$$
$$f_y(y_j) = \sum_i f_{xy}(x_i, y_j), \quad j = 1, 2, \cdots.$$

Then, $f_x(x_i)$ and $f_y(y_j)$ are called the **marginal probability functions** of X and Y. $f_x(x_i)$ and $f_y(y_j)$ also have the properties of the probability functions, i.e., $f_x(x_i) \geq 0$ and $\sum_i f_x(x_i) = 1$, and $f_y(y_j) \geq 0$ and $\sum_i f_y(y_j) = 1$.

Continuous Random Variables: Consider two continuous random variables X and Y. For a domain D, the probability which event $\{\omega; (X(\omega), Y(\omega)) \in D\}$ occurs is given by:

$$P((X, Y) \in D) = \iint_D f_{xy}(x, y)\mathrm{d}x\mathrm{d}y,$$

where $f_{xy}(x, y)$ is called the **joint probability density function** of X and Y or the **joint density function** of X and Y. $f_{xy}(x, y)$ has to satisfy the following properties:

$$f_{xy}(x, y) \geq 0,$$
$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{xy}(x, y)\mathrm{d}x\mathrm{d}y = 1.$$

Define $f_x(x)$ and $f_y(y)$ as:

$$f_x(x) = \int_{-\infty}^{\infty} f_{xy}(x, y)\mathrm{d}y,$$
$$f_y(y) = \int_{-\infty}^{\infty} f_{xy}(x, y)\mathrm{d}x,$$

where $f_x(x)$ and $f_y(y)$ are called the **marginal probability density functions** of X and Y or the **marginal density functions** of X and Y.

For example, consider the event $\{\omega; a < X(\omega) < b,\ c < Y(\omega) < d\}$, which is the specific case of the domain D. Then, the probability that we have the event $\{\omega; a < X(\omega) < b,\ c < Y(\omega) < d\}$ is written as:

$$P(a < X < b,\ c < Y < d) = \int_a^b \int_c^d f_{xy}(x, y)\mathrm{d}x\mathrm{d}y.$$

The mixture of discrete and continuous random variables is also possible. For example, Let X be a discrete random variable and Y be a continuous random variable. X takes $x_1, x_2, \cdots$. The probability which both X takes x_i and Y takes real numbers within the interval I is given by:

$$P(X = x_i, Y \in I) = \int_I f_{xy}(x_i, y)\mathrm{d}y.$$

Then, we have the following properties:

$$f_{xy}(x_i, y) \geq 0, \quad i = 1, 2, \cdots,$$
$$\sum_i \int_{-\infty}^{\infty} f_{xy}(x_i, y)\mathrm{d}y = 1.$$

The marginal probability function of X is given by:

$$f_x(x_i) = \int_{-\infty}^{\infty} f_{xy}(x_i, y)\mathrm{d}y.$$

The marginal probability density function of Y is:

$$f_y(y) = \sum_i f_{xy}(x_i, y).$$

1.2.3 Conditional Distribution

Discrete Random Variable: The **conditional probability function** of X given $Y = y_j$ is represented as:

$$P(X = x_i | Y = y_j) = f_{x|y}(x_i|y_j) = \frac{f_{xy}(x_i, y_j)}{f_y(y_j)} = \frac{f_{xy}(x_i, y_j)}{\sum_i f_{xy}(x_i, y_j)}.$$

The second equality indicates the definition of the conditional probability, which is shown in Section 1.1.2. The features of the conditional probability function $f_{x|y}(x_i|y_j)$ are:

$$\begin{aligned} &f_{x|y}(x_i|y_j) \geq 0, \quad i = 1, 2, \cdots, \\ &\sum_i f_{x|y}(x_i|y_j) = 1, \quad \text{for any } j. \end{aligned}$$

Continuous Random Variable: The **conditional probability density function** of X given $Y = y$ (or the **conditional density function** of X given $Y = y$) is:

$$f_{x|y}(x|y) = \frac{f_{xy}(x,y)}{f_y(y)} = \frac{f_{xy}(x,y)}{\int_{-\infty}^{\infty} f_{xy}(x,y)\mathrm{d}x}.$$

The properties of the conditional probability density function $f_{x|y}(x|y)$ are given by:

$$\begin{aligned} &f_{x|y}(x|y) \geq 0, \\ &\int_{-\infty}^{\infty} f_{x|y}(x|y)\mathrm{d}x = 1, \quad \text{for any } Y = y. \end{aligned}$$

Independence of Random Variables: For discrete random variables X and Y, we say that X is **independent** (or **stochastically independent**) of Y if and only if $f_{xy}(x_i, y_j) = f_x(x_i) f_y(y_j)$. Similarly, for continuous random variables X and Y, we say that X is independent of Y if and only if $f_{xy}(x, y) = f_x(x) f_y(y)$.

When X and Y are stochastically independent, $g(X)$ and $h(Y)$ are also stochastically independent, where $g(X)$ and $h(Y)$ are functions of X and Y.

1.3 Mathematical Expectation

1.3.1 Univariate Random Variable

Definition of Mathematical Expectation: Let $g(X)$ be a function of random variable X. The mathematical expectation of $g(X)$, denoted by $\mathrm{E}(g(X))$, is defined as follows:

$$\mathrm{E}(g(X)) = \begin{cases} \sum_i g(x_i)p_i = \sum_i g(x_i) f(x_i), & \text{(Discrete Random Variable)}, \\ \int_{-\infty}^{\infty} g(x) f(x)\mathrm{d}x, & \text{(Continuous Random Variable)}. \end{cases}$$

The following three functional forms of $g(X)$ are important.

1. $\boldsymbol{g(X) = X}$.

   The expectation of X, $\mathrm{E}(X)$, is known as **mean** of random variable X.

$$\begin{aligned}\mathrm{E}(X) &= \begin{cases} \displaystyle\sum_i x_i f(x_i), & \text{(Discrete Random Variable)}, \\ \displaystyle\int_{-\infty}^{\infty} x f(x)\mathrm{d}x, & \text{(Continuous Random Variable)}, \end{cases} \\ &= \mu, \quad (\text{or } \mu_x).\end{aligned}$$

   When a distribution of X is symmetric, mean indicates the center of the distribution.

2. $\boldsymbol{g(X) = (X - \mu)^2}$.

   The expectation of $(X - \mu)^2$ is known as **variance** of random variable X, which is denoted by $\mathrm{V}(X)$.

$$\begin{aligned}\mathrm{V}(X) &= \mathrm{E}((X-\mu)^2) \\ &= \begin{cases} \displaystyle\sum_i (x_i - \mu)^2 f(x_i), & \text{(Discrete Random Variable)}, \\ \displaystyle\int_{-\infty}^{\infty} (x-\mu)^2 f(x)\mathrm{d}x, & \text{(Continuous Random Variable)}, \end{cases} \\ &= \sigma^2, \quad (\text{or } \sigma_x^2).\end{aligned}$$

   If X is broadly distributed, $\sigma^2 = \mathrm{V}(X)$ becomes large. Conversely, if the distribution is concentrated on the center, σ^2 becomes small. Note that $\sigma = \sqrt{V(X)}$ is called the **standard deviation**.

3. $\boldsymbol{g(X) = e^{\theta X}}$.

   The expectation of $e^{\theta X}$ is called the **moment-generating function**, which is denoted by $\phi(\theta)$.

$$\begin{aligned}\phi(\theta) &= \mathrm{E}(e^{\theta X}) \\ &= \begin{cases} \displaystyle\sum_i e^{\theta x_i} f(x_i), & \text{(Discrete Random Variable)}, \\ \displaystyle\int_{-\infty}^{\infty} e^{\theta x} f(x)\mathrm{d}x, & \text{(Continuous Random Variable)}. \end{cases}\end{aligned}$$

Note that the definition of e is given by:

$$e = \lim_{x\to 0}(1+x)^{\frac{1}{x}} = \lim_{h\to\infty}\Big(1+\frac{1}{h}\Big)^h = 2.71828182845905.$$

The moment-generating function plays an important roll in statistics, which is discussed in Section 1.5.

In Examples 1.5 – 1.8, mean, variance and the moment-generating function are computed.

Example 1.5: In Example 1.2 of flipping a coin three times (Section 1.1.1), we see in Section 1.2.1 that the probability function is written as the following binomial distribution:

$$P(X=x) = f(x) = \frac{n!}{x!(n-x)!}p^x(1-p)^{n-x}, \quad \text{for } x = 0,1,2,\cdots,n,$$

where $n = 3$ and $p = 1/2$. When X has the binomial distribution above, we obtain $\mathrm{E}(X)$, $\mathrm{V}(X)$ and $\phi(\theta)$ as follows.

First, $\mathrm{E}(X)$ is computed as:

$$\begin{aligned}\mu = \mathrm{E}(X) &= \sum_x x f(x) = \sum_x x\frac{n!}{x!(n-x)!}p^x(1-p)^{n-x}\\ &= \sum_x \frac{n!}{(x-1)!(n-x)!}p^x(1-p)^{n-x} = np\sum_x \frac{(n-1)!}{(x-1)!(n-x)!}p^{x-1}(1-p)^{n-x}\\ &= np\sum_{x'}\frac{n'!}{x'!(n'-x')!}p^{x'}(1-p)^{n'-x'} = np,\end{aligned}$$

where $n' = n-1$ and $x' = x-1$ are set.

Second, in order to obtain $\mathrm{V}(X)$, we rewrite $\mathrm{V}(X)$ as:

$$\sigma^2 = \mathrm{V}(X) = \mathrm{E}(X^2) - \mu^2 = \mathrm{E}(X(X-1)) + \mu - \mu^2.$$

$\mathrm{E}(X(X-1))$ is given by:

$$\mathrm{E}(X(X-1)) = \sum_x x(x-1)f(x) = \sum_x x(x-1)\frac{n!}{x!(n-x)!}p^x(1-p)^{n-x}$$

$$= \sum_x \frac{n!}{(x-2)!(n-x)!} p^x (1-p)^{n-x}$$
$$= n(n-1)p^2 \sum_x \frac{(n-2)!}{(x-2)!(n-x)!} p^{x-2}(1-p)^{n-x}$$
$$= n(n-1)p^2 \sum_{x'} \frac{n'!}{x'!(n'-x')!} p^{x'}(1-p)^{n'-x'} = n(n-1)p^2,$$

where $n' = n-2$ and $x' = x-2$ are re-defined. Therefore, $\mathrm{V}(X)$ is obtained as:

$$\sigma^2 = \mathrm{V}(X) = \mathrm{E}(X(X-1)) + \mu - \mu^2$$
$$= n(n-1)p^2 + np - n^2p^2 = -np^2 + np = np(1-p).$$

Finally, the moment-generating function $\phi(\theta)$ is represented as:

$$\phi(\theta) = \mathrm{E}(e^{\theta X}) = \sum_x e^{\theta x} \frac{n!}{x!(n-x)!} p^x (1-p)^{n-p}$$
$$= \sum_x \frac{n!}{x!(n-x)!} (pe^\theta)^x (1-p)^{n-p} = (pe^\theta + 1 - p)^n.$$

In the last equality, we utilize the following formula:

$$(a+b)^n = \sum_{x=0}^{n} \frac{n!}{x!(n-x)!} a^x b^{n-x},$$

which is called the **binomial theorem**.

Example 1.6: As an example of continuous random variables, in Section 1.2.1 the uniform distribution is introduced, which is given by:

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

When X has the uniform distribution above, $\mathrm{E}(X)$, $\mathrm{V}(X)$ and $\phi(\theta)$ are computed as follows:

$$\mu = \mathrm{E}(X) = \int_{-\infty}^{\infty} x f(x) \mathrm{d}x = \int_0^1 x \mathrm{d}x = [\frac{1}{2}x^2]_0^1 = \frac{1}{2},$$

$$\sigma^2 = \mathrm{V}(X) = \mathrm{E}(X^2) - \mu^2$$
$$= \int_{-\infty}^{\infty} x^2 f(x) \mathrm{d}x - \mu^2 = \int_0^1 x^2 \mathrm{d}x - \mu^2 = [\frac{1}{3}x^3]_0^1 - (\frac{1}{2})^2 = \frac{1}{12},$$

$$\phi(\theta) = \mathrm{E}(e^{\theta X}) = \int_{-\infty}^{\infty} e^{\theta x} f(x)\mathrm{d}x = \int_{0}^{1} e^{\theta x}\mathrm{d}x = [\frac{1}{\theta}e^{\theta x}]_0^1 = \frac{1}{\theta}(e^{\theta} - 1).$$

Example 1.7: As another example of continuous random variables, we take the standard normal distribution:

$$f(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}, \quad \text{for } -\infty < x < \infty,$$

which is discussed in Section 2.2.1. When X has a standard normal distribution, i.e., when $X \sim N(0,1)$, $\mathrm{E}(X)$, $\mathrm{V}(X)$ and $\phi(\theta)$ are as follows.

$\mathrm{E}(X)$ is obtained as:

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x f(x)\mathrm{d}x = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} xe^{-\frac{1}{2}x^2}\mathrm{d}x = \frac{1}{\sqrt{2\pi}}\left[-e^{-\frac{1}{2}x^2}\right]_{-\infty}^{\infty} = 0,$$

because $\lim_{x\to\pm\infty} -e^{-\frac{1}{2}x^2} = 0$.

$\mathrm{V}(X)$ is computed as follows:

$$\begin{aligned}\mathrm{V}(X) = \mathrm{E}(X^2) &= \int_{-\infty}^{\infty} x^2 f(x)\mathrm{d}x = \int_{-\infty}^{\infty} x^2\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}\mathrm{d}x = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x\frac{\mathrm{d}(-e^{-\frac{1}{2}x^2})}{\mathrm{d}x}\mathrm{d}x \\ &= \frac{1}{\sqrt{2\pi}}\left[x(-e^{-\frac{1}{2}x^2})\right]_{-\infty}^{\infty} + \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2}\mathrm{d}x = \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}\mathrm{d}x = 1.\end{aligned}$$

The first equality holds because of $\mathrm{E}(X) = 0$. In the fifth equality, use the following integration formula, called the **integration by parts**:

$$\int_a^b h(x)g'(x)\mathrm{d}x = \left[h(x)g(x)\right]_a^b - \int_a^b h'(x)g(x)\mathrm{d}x,$$

where we take $h(x) = x$ and $g(x) = -e^{-\frac{1}{2}x^2}$ in this case. See Appendix 1.2 for the integration by parts. In the sixth equality, $\lim_{x\to\pm\infty} -xe^{-\frac{1}{2}x^2} = 0$ is utilized. The last equality is because the integration of the standard normal probability density function is equal to one (see p.7 in Section 1.2.1 for the integration of the standard normal probability density function).

$\phi(\theta)$ is derived as follows:

$$\begin{aligned}\phi(\theta) &= \int_{-\infty}^{\infty} e^{\theta x} f(x)\mathrm{d}x = \int_{-\infty}^{\infty} e^{\theta x}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}\mathrm{d}x = \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2+\theta x}\mathrm{d}x \\ &= \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\left((x-\theta)^2-\theta^2\right)}\mathrm{d}x = e^{\frac{1}{2}\theta^2}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x-\theta)^2}\mathrm{d}x = e^{\frac{1}{2}\theta^2}.\end{aligned}$$

The last equality holds because the integration indicates the normal density with mean θ and variance one. See Section 2.2.2 for the normal density.

Example 1.8: When the moment-generating function of X is given by $\phi_x(\theta) = e^{\frac{1}{2}\theta^2}$ (i.e., X has a standard normal distribution), we want to obtain the moment-generating function of $Y = \mu + \sigma X$.

Let $\phi_x(\theta)$ and $\phi_y(\theta)$ be the moment-generating functions of X and Y, respectively. Then, the moment-generating function of Y is obtained as follows:

$$\begin{aligned}\phi_y(\theta) &= \mathrm{E}(e^{\theta Y}) = \mathrm{E}(e^{\theta(\mu+\sigma X)}) = e^{\theta\mu}\mathrm{E}(e^{\theta\sigma X}) = e^{\theta\mu}\phi_x(\theta\sigma) = e^{\theta\mu}e^{\frac{1}{2}\sigma^2\theta^2} \\ &= \exp(\mu\theta + \frac{1}{2}\sigma^2\theta^2).\end{aligned}$$

Some Formulas of Mean and Variance:

1. **Theorem:** $\mathrm{E}(aX + b) = a\mathrm{E}(X) + b$, where a and b are constant.

   Proof:

   When X is a discrete random variable,

   $$\begin{aligned}\mathrm{E}(aX + b) &= \sum_i (ax_i + b)f(x_i) = a\sum_i x_i f(x_i) + b\sum_i f(x_i) \\ &= a\mathrm{E}(X) + b.\end{aligned}$$

   Note that we have $\sum_i x_i f(x_i) = \mathrm{E}(X)$ from the definition of mean and $\sum_i f(x_i) = 1$ because $f(x_i)$ is a probability function.

   If X is a continuous random variable,

   $$\begin{aligned}\mathrm{E}(aX + b) &= \int_{-\infty}^{\infty}(ax + b)f(x)\mathrm{d}x = a\int_{-\infty}^{\infty} xf(x)\mathrm{d}x + b\int_{-\infty}^{\infty} f(x)\mathrm{d}x \\ &= a\mathrm{E}(X) + b\end{aligned}$$

   Similarly, we have $\int_{-\infty}^{\infty} xf(x)\mathrm{d}x = \mathrm{E}(X)$ from the definition of mean and $\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1$ because $f(x)$ is a probability density function.

2. **Theorem:** $\mathrm{V}(X) = \mathrm{E}(X^2) - \mu^2$, where $\mu = \mathrm{E}(X)$.

   Proof:

   $\mathrm{V}(X)$ is rewritten as follows:

$$\begin{aligned}\mathrm{V}(X) &= \mathrm{E}((X-\mu)^2) = \mathrm{E}(X^2 - 2\mu X - \mu^2) \\ &= \mathrm{E}(X^2) - 2\mu\mathrm{E}(X) + \mu^2 = \mathrm{E}(X^2) - \mu^2.\end{aligned}$$

   The first equality is due to the definition of variance.

3. **Theorem:** $\mathrm{V}(aX + b) = a^2\mathrm{V}(X)$, where a and b are constant.

   Proof:

   From the definition of the mathematical expectation, $\mathrm{V}(aX + b)$ is represented as:

$$\begin{aligned}\mathrm{V}(aX+b) &= \mathrm{E}\Big(((aX+b) - \mathrm{E}(aX+b))^2\Big) = \mathrm{E}\Big((aX - a\mu)^2\Big) \\ &= \mathrm{E}(a^2(X-\mu)^2) = a^2\mathrm{E}((X-\mu)^2) = a^2\mathrm{V}(X)\end{aligned}$$

   The first and the fifth equalities are from the definition of variance. We use $\mathrm{E}(aX + b) = a\mu + b$ in the second equality.

4. **Theorem:** The random variable X is assumed to be distributed with mean $\mathrm{E}(X) = \mu$ and variance $\mathrm{V}(X) = \sigma^2$. Define $Z = \dfrac{X-\mu}{\sigma}$. Then, we have $\mathrm{E}(Z) = 0$ and $\mathrm{V}(Z) = 1$.

   Proof:

   $\mathrm{E}(X)$ and $\mathrm{V}(X)$ are obtained as:

$$\mathrm{E}(Z) = \mathrm{E}\Big(\frac{X-\mu}{\sigma}\Big) = \frac{\mathrm{E}(X)-\mu}{\sigma} = 0,$$

$$\mathrm{V}(Z) = \mathrm{V}\Big(\frac{1}{\sigma}X - \frac{\mu}{\sigma}\Big) = \frac{1}{\sigma^2}\mathrm{V}(X) = 1.$$

   The transformation from X to Z is known as normalization or standardization.

1.3.2 Bivariate Random Variable

Definition: Let $g(X, Y)$ be a function of random variables X and Y. The mathematical expectation of $g(X, Y)$, denoted by $\mathrm{E}(g(X, Y))$, is defined as:

$$\mathrm{E}(g(X,Y)) = \begin{cases} \sum_i \sum_j g(x_i, y_j) f(x_i, y_j), & \text{(Discrete Random Variables)}, \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f(x,y) \mathrm{d}x\mathrm{d}y, & \text{(Continuous Random Variables)}. \end{cases}$$

The following four functional forms are important, i.e., mean, variance, covariance and the moment-generating function.

1. $\boldsymbol{g(X, Y) = X}$:

   The expectation of random variable X, i.e., $\mathrm{E}(X)$, is given by:

   $$\begin{aligned} \mathrm{E}(X) &= \begin{cases} \sum_i \sum_j x_i f(x_i, y_j), & \text{(Discrete Random Variables)}, \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f(x,y) \mathrm{d}x\mathrm{d}y, & \text{(Continuous Random Variables)}, \end{cases} \\ &= \mu_x. \end{aligned}$$

   The case of $g(X, Y) = Y$ is exactly the same formulation as above, i.e., $\mathrm{E}(Y) = \mu_y$.

2. $\boldsymbol{g(X, Y) = (X - \mu_x)^2}$:

   The expectation of $(X - \mu_x)^2$ is known as variance of random variable X, which is denoted by $\mathrm{V}(X)$ and represented as follows:

   $$\begin{aligned} \mathrm{V}(X) &= \mathrm{E}((X - \mu_x)^2) \\ &= \begin{cases} \sum_i \sum_j (x_i - \mu_x)^2 f(x_i, y_j), & \text{(Discrete Cases)}, \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_x)^2 f(x,y) \mathrm{d}x\mathrm{d}y, & \text{(Continuous Cases)}, \end{cases} \\ &= \sigma_x^2. \end{aligned}$$

   The variance of Y is also obtained in the same fashion, i.e., $\mathrm{V}(Y) = \sigma_y^2$.

3. $\boldsymbol{g(X, Y) = (X - \mu_x)(Y - \mu_y)}$:

The expectation of $(X-\mu_x)(Y-\mu_y)$ is known as **covariance** of X and Y, which is denoted by $\mathrm{Cov}(X,Y)$ and written as:

$$\begin{aligned}\mathrm{Cov}(X,Y) &= \mathrm{E}((X-\mu_x)(Y-\mu_y)) \\ &= \begin{cases} \sum_i \sum_j (x_i-\mu_x)(y_j-\mu_y) f(x_i,y_j), & \text{(Discrete Cases)}, \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-\mu_x)(y-\mu_y) f(x,y)\mathrm{d}x\mathrm{d}y, & \text{(Continuous Cases)}. \end{cases}\end{aligned}$$

Thus, covariance is defined in the case of bivariate random variables.

4. $\boldsymbol{g(X,Y) = e^{\theta_1 X+\theta_2 Y}}$:

   The mathematical expectation of $e^{\theta_1 X+\theta_2 Y}$ is called the moment-generating function, which is denoted by $\phi(\theta_1,\theta_2)$ and written as:

$$\begin{aligned}\phi(\theta_1,\theta_2) &= \mathrm{E}(e^{\theta_1 X+\theta_2 Y}) \\ &= \begin{cases} \sum_i \sum_j e^{\theta_1 x_i+\theta_2 y_j} f(x_i,y_j), & \text{(Discrete Cases)}, \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{\theta_1 x+\theta_2 y} f(x,y)\mathrm{d}x\mathrm{d}y, & \text{(Continuous Cases)}. \end{cases}\end{aligned}$$

   In Section 1.5, the moment-generating function in the multivariate cases is discussed in more detail.

Some Formulas of Mean and Variance: We consider two random variables X and Y.

1. **Theorem:** $\mathrm{E}(X+Y) = \mathrm{E}(X) + \mathrm{E}(Y)$.

   Proof:

   For discrete random variables X and Y, it is shown as follows:

$$\begin{aligned}\mathrm{E}(X+Y) &= \sum_i \sum_j (x_i+y_j) f_{xy}(x_i,y_j) \\ &= \sum_i \sum_j x_i f_{xy}(x_i,y_j) + \sum_i \sum_j y_j f_{xy}(x_i,y_j) \\ &= \mathrm{E}(X) + \mathrm{E}(Y).\end{aligned}$$

For continuous random variables X and Y, we can show:

$$\begin{aligned}\mathrm{E}(X+Y) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x+y)f_{xy}(x,y)\mathrm{d}x\mathrm{d}y \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xf_{xy}(x,y)\mathrm{d}x\mathrm{d}y + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}yf_{xy}(x,y)\mathrm{d}x\mathrm{d}y \\ &= \mathrm{E}(X)+\mathrm{E}(Y).\end{aligned}$$

2. **Theorem:** $\mathrm{E}(XY) = \mathrm{E}(X)\mathrm{E}(Y)$, when X is independent of Y.

   Proof:

   For discrete random variables X and Y,

$$\begin{aligned}\mathrm{E}(XY) &= \sum_i\sum_j x_i y_j f_{xy}(x_i,y_j) = \sum_i\sum_j x_i y_j f_x(x_i)f_y(y_j) \\ &= \Big(\sum_i x_i f_x(x_i)\Big)\Big(\sum_j y_j f_y(y_j)\Big) = \mathrm{E}(X)\mathrm{E}(Y).\end{aligned}$$

   For continuous random variables X and Y,

$$\begin{aligned}\mathrm{E}(XY) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xyf_{xy}(x,y)\mathrm{d}x\mathrm{d}y \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xyf_x(x)f_y(y)\mathrm{d}x\mathrm{d}y \\ &= \Big(\int_{-\infty}^{\infty}xf_x(x)\mathrm{d}x\Big)\Big(\int_{-\infty}^{\infty}yf_y(y)\mathrm{d}y\Big) = \mathrm{E}(X)\mathrm{E}(Y)\end{aligned}$$

3. **Theorem:** $\mathrm{Cov}(X,Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)$.

   Proof:

   For both discrete and continuous random variables, we can rewrite as follows:

$$\begin{aligned}\mathrm{Cov}(X,Y) &= \mathrm{E}((X-\mu_x)(Y-\mu_y)) = \mathrm{E}(XY-\mu_xY-\mu_yX+\mu_x\mu_y) \\ &= \mathrm{E}(XY)-\mathrm{E}(\mu_xY)-\mathrm{E}(\mu_yX)+\mu_x\mu_y \\ &= \mathrm{E}(XY)-\mu_x\mathrm{E}(Y)-\mu_y\mathrm{E}(X)+\mu_x\mu_y \\ &= \mathrm{E}(XY)-\mu_x\mu_y-\mu_y\mu_x+\mu_x\mu_y = \mathrm{E}(XY)-\mu_x\mu_y \\ &= \mathrm{E}(XY)-\mathrm{E}(X)\mathrm{E}(Y).\end{aligned}$$

   In the fourth equality, the theorem in Section 1.3.1 is used.

4. **Theorem:** $\mathrm{Cov}(X, Y) = 0$, when X is independent of Y.

   Proof:

   From the above two theorems, we have $\mathrm{E}(XY) = \mathrm{E}(X)\mathrm{E}(Y)$ when X is independent of Y and $\mathrm{Cov}(X, Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)$. Therefore, $\mathrm{Cov}(X, Y) = 0$ is obtained when X is independent of Y.

5. **Definition:** The **correlation coefficient** between X and Y, denoted by ρ_{xy}, is defined as:

$$\rho_{xy} = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{V}(X)}\sqrt{\mathrm{V}(Y)}} = \frac{\mathrm{Cov}(X, Y)}{\sigma_x \sigma_y}.$$

   When $\rho_{xy} > 0$, we say that there is a **positive correlation** between X and Y. As ρ_{xy} approaches 1, we say that there is a **strong positive correlation** between X and Y. When $\rho_{xy} < 0$, we say that there is a **negative correlation** between X and Y. As ρ_{xy} approaches -1, we say that there is a **strong negative correlation** between X and Y.

6. **Theorem:** $\rho_{xy} = 0$, when X is independent of Y.

   Proof:

   When X is independent of Y, we have $\mathrm{Cov}(X, Y) = 0$. Therefore, we can obtain the result $\rho_{xy} = \dfrac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{V}(X)}\sqrt{\mathrm{V}(Y)}} = 0$. However, note that $\rho_{xy} = 0$ does not mean the independence between X and Y.

7. **Theorem:** $\mathrm{V}(X \pm Y) = \mathrm{V}(X) \pm 2\mathrm{Cov}(X, Y) + \mathrm{V}(Y)$.

   Proof:

   For both discrete and continuous random variables, $\mathrm{V}(X \pm Y)$ is rewritten as follow:

$$\begin{aligned}
\mathrm{V}(X \pm Y) &= \mathrm{E}\Big(((X \pm Y) - \mathrm{E}(X \pm Y))^2\Big) = \mathrm{E}\Big(((X - \mu_x) \pm (Y - \mu_y))^2\Big) \\
&= \mathrm{E}((X - \mu_x)^2 \pm 2(X - \mu_x)(Y - \mu_y) + (Y - \mu_y)^2) \\
&= \mathrm{E}((X - \mu_x)^2) \pm 2\mathrm{E}((X - \mu_x)(Y - \mu_y)) + \mathrm{E}((Y - \mu_y)^2) \\
&= \mathrm{V}(X) \pm 2\mathrm{Cov}(X, Y) + \mathrm{V}(Y).
\end{aligned}$$

8. **Theorem:** $-1 \le \rho_{xy} \le 1$.

   Proof:

Consider the following function of t: $f(t) = \mathrm{V}(Xt - Y)$, which is greater than zero because of the definition of variance. Therefore, for all t, we have $f(t) \geq 0$. $f(t)$ is rewritten as follows:

$$\begin{aligned} f(t) &= \mathrm{V}(Xt - Y) = \mathrm{V}(Xt) - 2\mathrm{Cov}(Xt, Y) + \mathrm{V}(Y) \\ &= t^2\mathrm{V}(X) - 2t\mathrm{Cov}(X, Y) + \mathrm{V}(Y) \\ &= \mathrm{V}(X)\Big(t - \frac{\mathrm{Cov}(X, Y)}{\mathrm{V}(X)}\Big)^2 + \mathrm{V}(Y) - \frac{(\mathrm{Cov}(X, Y))^2}{\mathrm{V}(X)}. \end{aligned}$$

In order to have $f(t) \geq 0$ for all t, we need the following condition:

$$\mathrm{V}(Y) - \frac{(\mathrm{Cov}(X, Y))^2}{\mathrm{V}(X)} \geq 0,$$

which implies:

$$\frac{(\mathrm{Cov}(X, Y))^2}{\mathrm{V}(X)\mathrm{V}(Y)} \leq 1.$$

Therefore, we have:

$$-1 \leq \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{V}(X)}\sqrt{\mathrm{V}(Y)}} \leq 1.$$

From the definition of correlation coefficient, i.e., $\rho_{xy} = \dfrac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{V}(X)}\sqrt{\mathrm{V}(Y)}}$, we obtain the result: $-1 \leq \rho_{xy} \leq 1$.

9. **Theorem:** $\mathrm{V}(X \pm Y) = \mathrm{V}(X) + \mathrm{V}(Y)$, when X is independent of Y.

   Proof:

   From the theorem above, $\mathrm{V}(X \pm Y) = \mathrm{V}(X) \pm 2\mathrm{Cov}(X, Y) + \mathrm{V}(Y)$ generally holds. When random variables X and Y are independent, we have $\mathrm{Cov}(X, Y) = 0$. Therefore, $\mathrm{V}(X + Y) = \mathrm{V}(X) + \mathrm{V}(Y)$ holds, when X is independent of Y.

10. **Theorem:** For n random variables $X_1, X_2, \cdots, X_n$,

$$\begin{aligned} \mathrm{E}(\sum_i a_i X_i) &= \sum_i a_i \mu_i, \\ \mathrm{V}(\sum_i a_i X_i) &= \sum_i \sum_j a_i a_j \mathrm{Cov}(X_i, X_j), \end{aligned}$$

where $\mathrm{E}(X_i) = \mu_i$ and a_i is a constant value. Especially, when $X_1, X_2, \cdots, X_n$ are mutually independent, we have the following:

$$\mathrm{V}(\sum_i a_i X_i) = \sum_i a_i^2 \mathrm{V}(X_i).$$

Proof:

For mean of $\sum_i a_i X_i$, the following representation is obtained.

$$\mathrm{E}(\sum_i a_i X_i) = \sum_i a_i \mathrm{E}(X_i) = \sum_i a_i \mu_i.$$

For variance of $\sum_i a_i X_i$, we can rewrite as follows:

$$\begin{aligned}
\mathrm{V}(\sum_i a_i X_i) &= \mathrm{E}\Big(\sum_i a_i (X_i - \mu_i)\Big)^2 = \mathrm{E}\Big(\sum_i a_i (X_i - \mu_i)\Big)\Big(\sum_j a_j (X_j - \mu_j)\Big) \\
&= \mathrm{E}\Big(\sum_i \sum_j a_i a_j (X_i - \mu_i)(X_j - \mu_j)\Big) \\
&= \sum_i \sum_j a_i a_j \mathrm{E}\Big((X_i - \mu_i)(X_j - \mu_j)\Big) = \sum_i \sum_j a_i a_j \mathrm{Cov}(X_i, X_j).
\end{aligned}$$

When $X_1, X_2, \cdots, X_n$ are mutually independent, we obtain $\mathrm{Cov}(X_i, X_j) = 0$ for $i \neq j$ from the previous theorem. Therefore, we obtain:

$$\mathrm{V}(\sum_i a_i X_i) = \sum_i a_i^2 \mathrm{V}(X_i).$$

Note that $\mathrm{Cov}(X_i, X_i) = \mathrm{E}((X_i - \mu)^2) = \mathrm{V}(X_i)$.

11. **Theorem:** n random variables $X_1, X_2, \cdots, X_n$ are mutually independently and identically distributed with mean μ and variance σ^2. That is, for all $i = 1, 2, \cdots, n$, $\mathrm{E}(X_i) = \mu$ and $\mathrm{V}(X_i) = \sigma^2$ are assumed. Consider arithmetic average $\overline{X} = (1/n) \sum_{i=1}^n X_i$. Then, we have:

$$\mathrm{E}(\overline{X}) = \mu, \qquad \mathrm{V}(\overline{X}) = \frac{\sigma^2}{n}.$$

Proof:

The mathematical expectation of $\overline{X}$ is given by:

$$\mathrm{E}(\overline{X}) = \mathrm{E}(\frac{1}{n} \sum_{i=1}^n X_i) = \frac{1}{n} \mathrm{E}(\sum_{i=1}^n X_i) = \frac{1}{n} \sum_{i=1}^n \mathrm{E}(X_i) = \frac{1}{n} \sum_{i=1}^n \mu = \frac{1}{n} n\mu = \mu.$$

$\mathrm{E}(aX) = a\mathrm{E}(X)$ in the second equality and $\mathrm{E}(X+Y) = \mathrm{E}(X) + \mathrm{E}(Y)$ in the third equality are utilized, where X and Y are random variables and a is a constant value. For these formulas, see p.17 in Section 1.3.1 and p.20 in this section.

The variance of $\overline{X}$ is computed as follows:

$$\mathrm{V}(\overline{X}) = \mathrm{V}(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\mathrm{V}(\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{V}(X_i) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}.$$

We use $\mathrm{V}(aX) = a^2\mathrm{V}(X)$ in the second equality and $\mathrm{V}(X+Y) = \mathrm{V}(X)+\mathrm{V}(Y)$ for X independent of Y in the third equality, where X and Y denote random variables and a is a constant value. For these formulas, see p.18 in Section 1.3.1 and p.23 in this section.

1.4 Transformation of Variables

Transformation of variables is used in the case of continuous random variables. Based on a distribution of a random variable, a distribution of the transformed random variable is derived. In other words, when a distribution of X is known, we can find a distribution of Y using the transformation of variables, where Y is a function of X.

1.4.1 Univariate Cases

Distribution of $Y = \psi^{-1}(X)$: Let $f_x(x)$ be the probability density function of continuous random variable X and $X = \psi(Y)$ be a one-to-one transformation. Then, the probability density function of Y, i.e., $f_y(y)$, is given by:

$$f_y(y) = |\psi'(y)| f_x(\psi(y)).$$

We can derive the above transformation of variables from X to Y as follows. Let $f_x(x)$ and $F_x(x)$ be the probability density function and the distribution function of X, respectively. Note that $F_x(x) = P(X \leq x)$ and $f_x(x) = F_x'(x)$.

Suppose that $X = \psi(Y)$ implies $Y = h(X)$. That is, we have $h^{-1}(Y) = \psi(Y)$. When $X = \psi(Y)$, we want to obtain the probability density function of Y. Let $f_y(y)$

and $F_y(y)$ be the probability density function and the distribution function of Y, respectively.

In the case of $\psi'(X) > 0$, the distribution function of Y, $F_y(y)$, is rewritten as follows:

$$F_y(y) = P(Y \leq y) = P\Big(h(X) \leq y\Big) = P\Big(X \leq h^{-1}(y)\Big) = P\Big(X \leq \psi(y)\Big) = F_x\Big(\psi(y)\Big).$$

Therefore, differentiating $F_y(y)$ with respect to y, we can obtain the following expression:

$$f_y(y) = F'_y(y) = \psi'(y)F'_x\Big(\psi(y)\Big) = \psi'(y)\, f_x\Big(\psi(y)\Big). \tag{1.4}$$

Next, in the case of $\psi'(X) < 0$, the distribution function of Y, $F_y(y)$, is rewritten as follows:

$$\begin{aligned} F_y(y) &= P(Y \leq y) = P\Big(h(X) \leq y\Big) = P\Big(X \geq h^{-1}(y)\Big) = P\Big(X \geq \psi(y)\Big) \\ &= 1 - P\Big(X < \psi(y)\Big) = 1 - F_x\Big(\psi(y)\Big). \end{aligned}$$

Thus, in the case of $\psi'(X) < 0$, pay attention to the third equality. Differentiating $F_y(y)$ with respect to y, we obtain the following result:

$$f_y(y) = F'_y(y) = -\psi'(y)F'_x\Big(\psi(y)\Big) = -\psi'(y)\, f_x\Big(\psi(y)\Big). \tag{1.5}$$

Note that $-\psi'(y) > 0$.

Thus, summarizing the above two cases, i.e., $\psi'(X) > 0$ and $\psi'(X) < 0$, equations (1.4) and (1.5) indicate the following result:

$$f_y(y) = |\psi'(y)|\, f_x\Big(\psi(y)\Big),$$

which is called the **transformation of variables**.

Example 1.9: When X has a standard normal density function, i.e., when $X \sim N(0, 1)$, we derive the probability density function of Y, where $Y = \mu + \sigma X$.

Since we have:

$$X = \psi(Y) = \frac{Y - \mu}{\sigma},$$

$\psi'(y) = 1/\sigma$ is obtained. Therefore, the density function of Y, $f_y(y)$, is given by:

$$f_y(y) = |\psi'(y)|\, f_x\Big(\psi(y)\Big) = \frac{1}{\sigma\sqrt{2\pi}} \exp\Big(-\frac{1}{2\sigma^2}(y - \mu)^2\Big),$$

which indicates the normal distribution with mean μ and variance σ^2, denoted by $N(\mu, \sigma^2)$.

On Distribution of $Y = X^2$: As an example, when we know the distribution function of X as $F_x(x)$, we want to obtain the distribution function of Y, $F_y(y)$, where $Y = X^2$. Using $F_x(x)$, $F_y(y)$ is rewritten as follows:

$$\begin{aligned} F_y(y) &= P(Y \le y) = P(X^2 \le y) = P(-\sqrt{y} \le X \le \sqrt{y}) \\ &= F_x(\sqrt{y}) - F_x(-\sqrt{y}). \end{aligned}$$

Therefore, when we have $f_x(x)$ and $Y = X^2$, the probability density function of Y is obtained as follows:

$$f_y(y) = F'_y(y) = \frac{1}{2\sqrt{y}}\Big(f_x(\sqrt{y}) + f_x(-\sqrt{y})\Big).$$

1.4.2 Multivariate Cases

Bivariate Case: Let $f_{xy}(x, y)$ be a joint probability density function of X and Y. Let $X = \psi_1(U, V)$ and $Y = \psi_2(U, V)$ be a one-to-one transformation from (X, Y) to (U, V). Then, we obtain a joint probability density function of U and V, denoted by $f_{uv}(u, v)$, as follows:

$$f_{uv}(u, v) = |J| f_{xy}(\psi_1(u, v), \psi_2(u, v)),$$

where J is called the **Jacobian** of the transformation, which is defined as:

$$J = \begin{vmatrix} \dfrac{\partial x}{\partial u} & \dfrac{\partial x}{\partial v} \\ \dfrac{\partial y}{\partial u} & \dfrac{\partial y}{\partial v} \end{vmatrix}.$$

Multivariate Case: Let $f_x(x_1, x_2, \cdots, x_n)$ be a joint probability density function of $X_1, X_2, \cdots X_n$. Suppose that the one-to-one transformation from $(X_1, X_2, \cdots, X_n)$ to $(Y_1, Y_2, \cdots, Y_n)$ is given by:

$$\begin{aligned} X_1 &= \psi_1(Y_1, Y_2, \cdots, Y_n), \\ X_2 &= \psi_2(Y_1, Y_2, \cdots, Y_n), \\ &\vdots \\ X_n &= \psi_n(Y_1, Y_2, \cdots, Y_n). \end{aligned}$$

Then, we obtain a joint probability density function of Y_1, Y_2, $\cdots$, Y_n, denoted by $f_y(y_1, y_2, \cdots, y_n)$, as follows:

$$f_y(y_1, y_2, \cdots, y_n) = |J| f_x\Big(\psi_1(y_1, \cdots, y_n), \psi_2(y_1, \cdots, y_n), \cdots, \psi_n(y_1, \cdots, y_n)\Big),$$

where J is called the Jacobian of the transformation, which is defined as:

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \cdots & \frac{\partial x_1}{\partial y_n} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} & \cdots & \frac{\partial x_2}{\partial y_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial x_n}{\partial y_1} & \frac{\partial x_n}{\partial y_2} & \cdots & \frac{\partial x_n}{\partial y_n} \end{vmatrix}.$$

1.5 Moment-Generating Function

1.5.1 Univariate Cases

As discussed in Section 1.3.1, the moment-generating function is defined as $\phi(\theta) = \mathrm{E}(e^{\theta X})$. In this section, the important theorems and remarks of the moment-generating function are summarized.

For a random variable X, $\mu'_n \equiv \mathrm{E}(X^n)$ is called the ***n*-th moment** of X. Then, we have the following first theorem.

1. **Theorem:** $\phi^{(n)}(0) = \mu'_n \equiv \mathrm{E}(X^n)$.

   Proof:

   First, from the definition of the moment-generating function, $\phi(\theta)$ is written as:

   $$\phi(\theta) = \mathrm{E}(e^{\theta X}) = \int_{-\infty}^{\infty} e^{\theta x} f(x) \mathrm{d}x.$$

   The n-th derivative of $\phi(\theta)$, denoted by $\phi^{(n)}(\theta)$, is:

   $$\phi^{(n)}(\theta) = \int_{-\infty}^{\infty} x^n e^{\theta x} f(x) \mathrm{d}x.$$

   Evaluating $\phi^{(n)}(\theta)$ at $\theta = 0$, we obtain:

   $$\phi^{(n)}(0) = \int_{-\infty}^{\infty} x^n f(x) \mathrm{d}x = \mathrm{E}(X^n) \equiv \mu'_n,$$

   where the second equality comes from the definition of the mathematical expectation.

2. **Remark:** Consider two random variables X and Y. When the moment-generating function of X is equivalent to that of Y, we have the fact that X has the same distribution as Y.

3. **Theorem:** Let $\phi(\theta)$ be the moment-generating function of X. Then, the moment-generating function of Y, where $Y = aX + b$, is given by $e^{b\theta}\phi(a\theta)$.

   Proof:

   Let $\phi_y(\theta)$ be the moment-generating function of Y. Then, $\phi_y(\theta)$ is rewritten as follows:

   $$\phi_y(\theta) = \mathrm{E}(e^{\theta Y}) = \mathrm{E}(e^{\theta(aX+b)}) = e^{b\theta}\mathrm{E}(e^{a\theta X}) = e^{b\theta}\phi(a\theta).$$

4. **Theorem:** Let $\phi_1(\theta)$, $\phi_2(\theta)$, $\cdots$, $\phi_n(\theta)$ be the moment-generating functions of X_1, X_2, $\cdots$, X_n, which are mutually independently distributed random variables. Define $Y = X_1 + X_2 + \cdots + X_n$. Then, the moment-generating function of Y is given by $\phi_1(\theta)\phi_2(\theta)\cdots\phi_n(\theta)$, i.e.,

   $$\phi_y(\theta) = \mathrm{E}(e^{\theta Y}) = \phi_1(\theta)\phi_2(\theta)\cdots\phi_n(\theta),$$

   where $\phi_y(\theta)$ represents the moment-generating function of Y.

   Proof:

   The moment-generating function of Y, i.e., $\phi_y(\theta)$, is:

   $$\begin{aligned}\phi_y(\theta) &= \mathrm{E}(e^{\theta Y}) = \mathrm{E}(e^{\theta(X_1+X_2+\cdots+X_n)}) = \mathrm{E}(e^{\theta X_1})\mathrm{E}(e^{\theta X_2})\cdots\mathrm{E}(e^{\theta X_n}) \\ &= \phi_1(\theta)\phi_2(\theta)\cdots\phi_n(\theta).\end{aligned}$$

   The third equality holds because X_1, X_2, $\cdots$, X_n are mutually independently distributed random variables.

5. **Theorem:** When X_1, X_2, $\cdots$, X_n are mutually independently and identically distributed and the moment-generating function of X_i is given by $\phi(\theta)$ for all i, the moment-generating function of Y is represented by $\left(\phi(\theta)\right)^n$, where $Y = X_1 + X_2 + \cdots + X_n$.

   Proof:

   Using the above theorem, we have the following:

   $$\phi_y(\theta) = \phi_1(\theta)\phi_2(\theta)\cdots\phi_n(\theta) = \phi(\theta)\phi(\theta)\cdots\phi(\theta) = \left(\phi(\theta)\right)^n.$$

   Note that $\phi_i(\theta) = \phi(\theta)$ for all i.

6. **Theorem:** When $X_1, X_2, \cdots, X_n$ are mutually independently and identically distributed and the moment-generating function of X_i is given by $\phi(\theta)$ for all i, the moment-generating function of $\overline{X}$ is represented by $\left(\phi(\frac{\theta}{n})\right)^n$, where $\overline{X} = (1/n)\sum_{i=1}^{n} X_i$.

   Proof:

   Let $\phi_{\overline{x}}(\theta)$ be the moment-generating function of $\overline{X}$.

$$\phi_{\overline{x}}(\theta) = \mathrm{E}(e^{\theta\overline{X}}) = \mathrm{E}(e^{\frac{\theta}{n}\sum_{i=1}^{n} X_i}) = \prod_{i=1}^{n} \mathrm{E}(e^{\frac{\theta}{n}X_i}) = \prod_{i=1}^{n} \phi(\frac{\theta}{n}) = \left(\phi(\frac{\theta}{n})\right)^n$$

Example 1.10: For the binomial random variable, the moment-generating function $\phi(\theta)$ is shown as:

$$\phi(\theta) = (pe^{\theta} + 1 - p)^n,$$

which is discussed in Example 1.5 (Section 1.3.1). Using the moment-generating function, we check whether $\mathrm{E}(X) = np$ and $\mathrm{V}(X) = np(1-p)$ are obtained when X is a binomial random variable.

The first- and the second-derivatives with respect to θ are given by:

$$\phi'(\theta) = npe^{\theta}(pe^{\theta} + 1 - p)^{n-1},$$
$$\phi''(\theta) = npe^{\theta}(pe^{\theta} + 1 - p)^{n-1} + n(n-1)p^2e^{2\theta}(pe^{\theta} + 1 - p)^{n-2}.$$

Evaluating at $\theta = 0$, we have:

$$\mathrm{E}(X) = \phi'(0) = np, \quad \mathrm{E}(X^2) = \phi''(0) = np + n(n-1)p^2.$$

Therefore, $\mathrm{V}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2 = np(1-p)$ can be derived. Thus, we can make sure that $\mathrm{E}(X)$ and $\mathrm{V}(X)$ are obtained from $\phi(\theta)$.

1.5.2 Multivariate Cases

Bivariate Case: As discussed in Section 1.3.2, for two random variables X and Y, the moment-generating function is defined as $\phi(\theta_1, \theta_2) = \mathrm{E}(e^{\theta_1 X + \theta_2 Y})$. Some useful and important theorems and remarks are shown as follows.

1. **Theorem:** Consider two random variables X and Y. Let $\phi(\theta_1, \theta_2)$ be the moment-generating function of X and Y. Then, we have the following result:

$$\frac{\partial^{j+k}\phi(0,0)}{\partial\theta_1^j\partial\theta_2^k} = \mathrm{E}(X^jY^k).$$

Proof:

Let $f_{xy}(x,y)$ be the probability density function of X and Y. From the definition, $\phi(\theta_1, \theta_2)$ is written as:

$$\phi(\theta_1, \theta_2) = \mathrm{E}(e^{\theta_1 X+\theta_2 Y}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{\theta_1 x+\theta_2 y} f_{xy}(x,y)\mathrm{d}x\mathrm{d}y.$$

Taking the j-th derivative of $\phi(\theta_1, \theta_2)$ with respect to θ_1 and at the same time the k-th derivative with respect to θ_2, we have the following expression:

$$\frac{\partial^{j+k}\phi(\theta_1,\theta_2)}{\partial\theta_1^j\partial\theta_2^k} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^jy^k e^{\theta_1 x+\theta_2 y} f_{xy}(x,y)\mathrm{d}x\mathrm{d}y.$$

Evaluating the above equation at $(\theta_1, \theta_2) = (0, 0)$, we can easily obtain:

$$\frac{\partial^{j+k}\phi(0,0)}{\partial\theta_1^j\partial\theta_2^k} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^jy^k f_{xy}(x,y)\mathrm{d}x\mathrm{d}y \equiv \mathrm{E}(X^jY^k).$$

2. **Remark:** Let (X_i, Y_i) be a pair of random variables. Suppose that the moment-generating function of (X_1, Y_1) is equivalent to that of (X_2, Y_2). Then, (X_1, Y_1) has the same distribution function as (X_2, Y_2).

3. **Theorem:** Let $\phi(\theta_1, \theta_2)$ be the moment-generating function of (X, Y). The moment-generating function of X is given by $\phi_1(\theta_1)$ and that of Y is $\phi_2(\theta_2)$. Then, we have the following facts:

$$\phi_1(\theta_1) = \phi(\theta_1, 0), \quad \phi_2(\theta_2) = \phi(0, \theta_2).$$

Proof:

Again, the definition of the moment-generating function of X and Y is represented as:

$$\phi(\theta_1, \theta_2) = \mathrm{E}(e^{\theta_1 X+\theta_2 Y}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{\theta_1 x+\theta_2 y} f_{xy}(x,y)\mathrm{d}x\mathrm{d}y.$$

When $\phi(\theta_1, \theta_2)$ is evaluated at $\theta_2 = 0$, $\phi(\theta_1, 0)$ is rewritten as follows:

$$\begin{aligned}\phi(\theta_1, 0) = \mathrm{E}(e^{\theta_1 X}) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{\theta_1 x} f_{xy}(x, y)\mathrm{d}x\mathrm{d}y \\ &= \int_{-\infty}^{\infty} e^{\theta_1 x}\Big(\int_{-\infty}^{\infty} f_{xy}(x, y)\mathrm{d}y\Big)\mathrm{d}x \\ &= \int_{-\infty}^{\infty} e^{\theta_1 x} f_x(x)\mathrm{d}x = \mathrm{E}(e^{\theta_1 X}) = \phi_1(\theta_1).\end{aligned}$$

Thus, we obtain the result: $\phi(\theta_1, 0) = \phi_1(\theta_1)$. Similarly, $\phi(0, \theta_2) = \phi_2(\theta_2)$ can be derived.

4. **Theorem:** The moment-generating function of (X, Y) is given by $\phi(\theta_1, \theta_2)$. Let $\phi_1(\theta_1)$ and $\phi_2(\theta_2)$ be the moment-generating functions of X and Y, respectively. If X is independent of Y, we have:

$$\phi(\theta_1, \theta_2) = \phi_1(\theta_1)\phi_2(\theta_2).$$

Proof:

From the definition of $\phi(\theta_1, \theta_2)$, the moment-generating function of X and Y is rewritten as follows:

$$\phi(\theta_1, \theta_2) = \mathrm{E}(e^{\theta_1 X + \theta_2 Y}) = \mathrm{E}(e^{\theta_1 X})\mathrm{E}(e^{\theta_2 Y}) = \phi_1(\theta_1)\phi_2(\theta_2).$$

The second equality holds because X is independent of Y.

Multivariate Case: For multivariate random variables X_1, X_2, $\cdots$, X_n, the moment-generating function is defined as:

$$\phi(\theta_1, \theta_2, \cdots, \theta_n) = \mathrm{E}(e^{\theta_1 X_1 + \theta_2 X_2 + \cdots + \theta_n X_n}).$$

1. **Theorem:** If the multivariate random variables X_1, X_2, $\cdots$, X_n are mutually independent, the moment-generating function of X_1, X_2, $\cdots$, X_n, denoted by $\phi(\theta_1, \theta_2, \cdots, \theta_n)$, is given by:

$$\phi(\theta_1, \theta_2, \cdots, \theta_n) = \phi_1(\theta_1)\phi_2(\theta_2)\cdots\phi_n(\theta_n),$$

where $\phi_i(\theta) = \mathrm{E}(e^{\theta X_i})$.

Proof:

From the definition of the moment-generating function in the multivariate cases, we obtain the following:

$$\begin{aligned}\phi(\theta_1, \theta_2, \cdots, \theta_n) &= \mathrm{E}(e^{\theta_1 X_1 + \theta_2 X_2 + \cdots + \theta_n X_n}) \\ &= \mathrm{E}(e^{\theta_1 X_1})\mathrm{E}(e^{\theta_2 X_2}) \cdots \mathrm{E}(e^{\theta_n X_n}) \\ &= \phi_1(\theta_1)\phi_2(\theta_2) \cdots \phi_n(\theta_n).\end{aligned}$$

2. **Theorem:** Suppose that the multivariate random variables X_1, X_2, $\cdots$, X_n are mutually independently and identically distributed. X_i has a normal distribution with mean μ and variance σ^2, i.e., $X_i \sim N(\mu, \sigma^2)$. Let us define $\hat{\mu} = \sum_{i=1}^n a_i X_i$, where a_i, $i = 1, 2, \cdots, n$, are assumed to be known. Then, $\hat{\mu}$ has a normal distribution with mean $\mu \sum_{i=1}^n a_i$ and variance $\sigma^2 \sum_{i=1}^n a_i^2$, i.e., $\hat{\mu} \sim N(\mu \sum_{i=1}^n a_i, \sigma^2 \sum_{i=1}^n a_i^2)$.

Proof:

From Example 1.8 (p.17) and Example 1.9 (p.26), it is shown that the moment-generating function of X is given by: $\phi_x(\theta) = \exp(\mu\theta + \frac{1}{2}\sigma^2\theta^2)$, when X is normally distributed as $X \sim N(\mu, \sigma^2)$.

Let $\phi_{\hat{\mu}}$ be the moment-generating function of $\hat{\mu}$.

$$\begin{aligned}\phi_{\hat{\mu}}(\theta) = \mathrm{E}(e^{\theta\hat{\mu}}) = \mathrm{E}(e^{\theta \sum_{i=1}^n a_i X_i}) = \prod_{i=1}^n \mathrm{E}(e^{\theta a_i X_i}) = \prod_{i=1}^n \phi_x(a_i\theta) \\ = \prod_{i=1}^n \exp(\mu a_i\theta + \frac{1}{2}\sigma^2 a_i^2\theta^2) = \exp(\mu \sum_{i=1}^n a_i\theta + \frac{1}{2}\sigma^2 \sum_{i=1}^n a_i^2\theta^2)\end{aligned}$$

which is equivalent to the moment-generating function of the normal distribution with mean $\mu \sum_{i=1}^n a_i$ and variance $\sigma^2 \sum_{i=1}^n a_i^2$, where μ and σ^2 in $\phi_x(\theta)$ is simply replaced by $\mu \sum_{i=1}^n a_i$ and $\sigma^2 \sum_{i=1}^n a_i^2$ in $\phi_{\hat{\mu}}(\theta)$, respectively.

Moreover, note as follows. When $a_i = 1/n$ is taken for all $i = 1, 2, \cdots, n$, i.e., when $\hat{\mu} = \overline{X}$ is taken, $\hat{\mu} = \overline{X}$ is normally distributed as: $\overline{X} \sim N(\mu, \sigma^2/n)$.

1.6 Law of Large Numbers and Central Limit Theorem

1.6.1 Chebyshev's Inequality

In this section, we introduce Chebyshev's inequality, which enables us to find upper and lower bounds given a certain probability.

Theorem: Let $g(X)$ be a nonnegative function of the random variable X, i.e., $g(X) \geq 0$. If $\mathrm{E}(g(X))$ exists, then we have:

$$P(g(X) \geq k) \leq \frac{\mathrm{E}(g(X))}{k},$$

for a positive constant value k.

Proof:

We define the discrete random variable U as follows:

$$U = \begin{cases} 1, & \text{if } g(X) \geq k, \\ 0, & \text{if } g(X) < k. \end{cases}$$

Thus, the discrete random variable U takes 0 or 1. Suppose that the probability function of U is given by:

$$f(u) = P(U = u),$$

where $P(U = u)$ is represented as:

$$P(U = 1) = P(g(X) \geq k),$$
$$P(U = 0) = P(g(X) < k).$$

Then, in spite of the value which U takes, the following equation always holds:

$$g(X) \geq kU.$$

Therefore, taking the expectation on both sides, we obtain:

$$\mathrm{E}(g(X)) \geq k\mathrm{E}(U), \tag{1.6}$$

where $\mathrm{E}(U)$ is given by:

$$\begin{aligned} \mathrm{E}(U) &= \sum_{u=0}^{1} uP(U = u) = 1 \times P(U = 1) + 0 \times P(U = 0) = P(U = 1) \\ &= P(g(X) \geq k). \end{aligned} \tag{1.7}$$

Accordingly, substituting equation (1.7) into equation (1.6), we have the following inequality:

$$P(g(X) \geq k) \leq \frac{\mathrm{E}(g(X))}{k}.$$

Chebyshev's Inequality: Assume that $\mathrm{E}(X) = \mu$, $\mathrm{V}(X) = \sigma^2$, and λ is a positive constant value. Then, we have the following inequality:

$$P(|X - \mu| \geq \lambda\sigma) \leq \frac{1}{\lambda^2},$$

or equivalently,

$$P(|X - \mu| < \lambda\sigma) \geq 1 - \frac{1}{\lambda^2},$$

which is called **Chebyshev's inequality**.

Proof:

Take $g(X) = (X - \mu)^2$ and $k = \lambda^2\sigma^2$. Then, we have:

$$P((X - \mu)^2 \geq \lambda^2\sigma^2) \leq \frac{\mathrm{E}(X - \mu)^2}{\lambda^2\sigma^2},$$

which implies

$$P(|X - \mu| \geq \lambda\sigma) \leq \frac{1}{\lambda^2}.$$

Note that $\mathrm{E}(X - \mu)^2 = \mathrm{V}(X) = \sigma^2$.

Since we have $P(|X - \mu| \geq \lambda\sigma) + P(|X - \mu| < \lambda\sigma) = 1$, we can derive the following inequality:

$$P(|X - \mu| < \lambda\sigma) \geq 1 - \frac{1}{\lambda^2}. \tag{1.8}$$

An Interpretation of Chebyshev's inequality: The number $1/\lambda^2$ is an upper bound for the probability $P(|X - \mu| \geq \lambda\sigma)$. Equation (1.8) is rewritten as:

$$P(\mu - \lambda\sigma < X < \mu + \lambda\sigma) \geq 1 - \frac{1}{\lambda^2}.$$

That is, the probability that X falls within $\lambda\sigma$ units of μ is greater than or equal to $1 - 1/\lambda^2$. Taking an example of $\lambda = 2$, the probability that X falls within two standard deviations of its mean is at least 0.75.

Finally, note as follows. Taking $\epsilon = \lambda\sigma$, we obtain as follows:

$$P(|X - \mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2},$$

i.e.,

$$P(|X - \mathrm{E}(X)| \geq \epsilon) \leq \frac{\mathrm{V}(X)}{\epsilon^2}, \tag{1.9}$$

which inequality is used in the next section.

1.6.2 Law of Large Numbers (Convergence in probability)

Law of Large Numbers: Assume that X_1, X_2, $\cdots$, X_n are mutually independently and identically distributed with mean $\mathrm{E}(X_i) = \mu$ and variance $\mathrm{V}(X_i) = \sigma^2 < \infty$ for all i. Then, for any positive value ϵ, as $n \longrightarrow \infty$, we have the following result:

$$P(|\overline{X}_n - \mu| > \epsilon) \longrightarrow 0,$$

where $\overline{X}_n = (1/n)\sum_{i=1}^{n} X_i$. We say that $\overline{X}_n$ converges to μ in probability.

Proof:

Using (1.9), Chebyshev's inequality represents as follows:

$$P(|\overline{X}_n - \mathrm{E}(\overline{X}_n)| > \epsilon) \leq \frac{\mathrm{V}(\overline{X}_n)}{\epsilon^2},$$

where X in (1.9) is replaced by $\overline{X}_n$. As in Section 1.3.2 (p.24), we have $\mathrm{E}(\overline{X}_n) = \mu$ and $\mathrm{V}(\overline{X}_n) = \sigma^2/n$, which are substituted into the above inequality. Then, we obtain:

$$P(|\overline{X}_n - \mu| > \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}.$$

Accordingly, when $n \longrightarrow \infty$, the following equation holds:

$$P(|\overline{X}_n - \mu| > \epsilon) \leq \frac{\sigma^2}{n\epsilon^2} \longrightarrow 0.$$

That is, $\overline{X}_n \longrightarrow \mu$ is obtained as $n \longrightarrow \infty$, which is written as: $\mathrm{plim}\, \overline{X}_n = \mu$. This theorem is called the **law of large numbers**.

The condition $P(|\overline{X}_n - \mu| > \epsilon) \longrightarrow 0$ or equivalently $P(|\overline{X}_n - \mu| < \epsilon) \longrightarrow 1$ is used as the definition of **convergence in probability**. In this case, we say that $\overline{X}_n$ converges to μ in probability.

Theorem: In the case where $X_1, X_2, \cdots, X_n$ are not identically distributed and they are not mutually independently distributed, we assume that

$$m_n = \mathrm{E}(\sum_{i=1}^{n} X_i) < \infty,$$
$$V_n = \mathrm{V}(\sum_{i=1}^{n} X_i) < \infty,$$
$$\frac{V_n}{n^2} \longrightarrow 0, \quad \text{as } n \longrightarrow \infty.$$

Then, we obtain the following result:

$$\frac{\sum_{i=1}^{n} X_i - m_n}{n} \longrightarrow 0.$$

That is, $\overline{X}_n$ converges to m_n/n in probability. This theorem is also called the law of large numbers.

1.6.3 Central Limit Theorem

Central Limit Theorem: $X_1, X_2, \cdots, X_n$ are mutually independently and identically distributed with $\mathrm{E}(X_i) = \mu$ and $\mathrm{V}(X_i) = \sigma^2$ for all i. Both μ and σ^2 are finite. Under the above assumptions, when $n \longrightarrow \infty$, we have:

$$P\Big(\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} < x\Big) \longrightarrow \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} \mathrm{d}u,$$

which is called the **central limit theorem**.

Proof:

Define $Y_i = \dfrac{X_i - \mu}{\sigma}$. We can rewrite as follows:

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i.$$

Since $Y_1, Y_2, \cdots, Y_n$ are mutually independently and identically distributed, the moment-generating function of Y_i is identical for all i, which is denoted by $\phi(\theta)$.

Using $\mathrm{E}(Y_i) = 0$ and $\mathrm{V}(Y_i) = 1$, the moment-generating function of Y_i, $\phi(\theta)$, is rewritten as:

$$\begin{aligned}\phi(\theta) = \mathrm{E}(e^{Y_i\theta}) &= \mathrm{E}\Big(1 + Y_i\theta + \frac{1}{2}Y_i^2\theta^2 + \frac{1}{3!}Y_i^3\theta^3\cdots\Big) \\ &= 1 + \frac{1}{2}\theta^2 + O(\theta^3).\end{aligned}$$

In the second equality, $e^{Y_i\theta}$ is approximated by the Taylor series expansion around $Y_i = 0$. See Appendix 1.3 for the Taylor series expansion.

Define Z as:

$$Z = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y_i.$$

Then, the moment-generating function of Z, i.e., $\phi_z(\theta)$, is given by:

$$\begin{aligned}\phi_z(\theta) = \mathrm{E}(e^{Z\theta}) = \mathrm{E}\Big(e^{\frac{\theta}{\sqrt{n}}\sum_{i=1}^{n} Y_i}\Big) = \prod_{i=1}^{n}\mathrm{E}\Big(e^{\frac{\theta}{\sqrt{n}}Y_i}\Big) &= \Big(\phi(\frac{\theta}{\sqrt{n}})\Big)^n \\ = \Big(1 + \frac{1}{2}\frac{\theta^2}{n} + O(\frac{\theta^3}{n^{\frac{3}{2}}})\Big)^n &= \Big(1 + \frac{1}{2}\frac{\theta^2}{n} + O(n^{-\frac{3}{2}})\Big)^n.\end{aligned}$$

Moreover, set $x = \dfrac{1}{2}\dfrac{\theta^2}{n} + O(n^{-\frac{3}{2}})$. Multiply n/x on both sides. Substitute $n = \dfrac{1}{x}\Big(\dfrac{1}{2}\theta^2 + O(n^{-\frac{1}{2}})\Big)$ into the moment-generating function of Z, i.e., $\phi_z(\theta)$. Then, we obtain:

$$\begin{aligned}\phi_z(\theta) = \Big(1 + \frac{1}{2}\frac{\theta^2}{n} + O(n^{-\frac{3}{2}})\Big)^n &= (1+x)^{\frac{1}{x}(\frac{\theta^2}{2} + O(n^{-\frac{1}{2}}))} \\ = \Big((1+x)^{\frac{1}{x}}\Big)^{\frac{\theta^2}{2} + O(n^{-\frac{1}{2}})} &\longrightarrow e^{\frac{\theta^2}{2}}.\end{aligned}$$

Note that $x \longrightarrow 0$ when $n \longrightarrow \infty$ and that $\lim_{x\to 0}(1+x)^{1/x} = e$ as in Section 1.2.3 (p.14).

Since $\phi_z(\theta) = e^{\frac{\theta^2}{2}}$ is the moment-generating function of the standard normal distribution (see p.16 in Section 1.3.1 for the moment-generating function of the standard normal probability density), we have:

$$P\Big(\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} < x\Big) \longrightarrow \int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}u^2}\mathrm{d}u,$$

or equivalently,

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \longrightarrow N(0,1).$$

The following expression is also possible:

$$\sqrt{n}(\overline{X}_n - \mu) \longrightarrow N(0,\sigma^2). \tag{1.10}$$

Corollary 1: When $\mathrm{E}(X_i) = \mu$, $\mathrm{V}(X_i) = \sigma^2$ and $\overline{X}_n = (1/n)\sum_{i=1}^n X_i$, note that

$$\frac{\overline{X}_n - \mathrm{E}(\overline{X}_n)}{\sqrt{\mathrm{V}(\overline{X}_n)}} = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}.$$

Therefore, we can rewrite the above theorem as:

$$P\Big(\frac{\overline{X}_n - \mathrm{E}(\overline{X}_n)}{\sqrt{\mathrm{V}(\overline{X}_n)}} < x\Big) \longrightarrow \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}\mathrm{d}u.$$

Corollary 2: Consider the case where $X_1, X_2, \cdots, X_n$ are not identically distributed and they are not mutually independently distributed. Assume that

$$\lim_{n\to\infty} n\mathrm{V}(\overline{X_n}) = \sigma^2 < \infty,$$

where $\overline{X_n} = (1/n)\sum_{i=1}^n X_i$. Then, when $n \longrightarrow \infty$, we have:

$$P\Big(\frac{\overline{X_n} - \mathrm{E}(\overline{X_n})}{\sqrt{\mathrm{V}(\overline{X_n})}} < x\Big) = P\Big(\frac{\sum_{i=1}^n X_i - \mathrm{E}(\sum_{i=1}^n X_i)}{\sqrt{\mathrm{V}(\sum_{i=1}^n X_i)}} < x\Big) \longrightarrow \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}\mathrm{d}u.$$

1.7 Statistical Inference

1.7.1 Point Estimation

Suppose that the functional form of the underlying distribution on population is known but the parameter θ included in the distribution is not known. The distribution function of population is given by $f(x;\theta)$. Let $x_1, x_2, \cdots, x_n$ be the n observed data drawn from population. Consider estimating the parameter θ using the n observed data. Let $\hat{\theta}_n(x_1, x_2, \cdots, x_n)$ be a function of the observed data $x_1, x_2, \cdots, x_n$.

Suppose that $\hat{\theta}_n(x_1, x_2, \cdots, x_n)$ is constructed from the purpose of estimating the parameter θ. That is, $\hat{\theta}_n(x_1, x_2, \cdots, x_n)$ takes a certain value given the n observed data. In this case, $\hat{\theta}_n(x_1, x_2, \cdots, x_n)$ is called the **point estimate** of θ, or simply the **estimate** of θ.

Example 1.11: Consider the case of $\theta = (\mu, \sigma^2)$, where the unknown parameters contained in population is given by mean and variance. A point estimate of population mean μ is given by:

$$\hat{\mu}_n(x_1, x_2, \cdots, x_n) \equiv \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

A point estimate of population variance σ^2 is:

$$\hat{\sigma}_n^2(x_1, x_2, \cdots, x_n) \equiv s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

An alternative point estimate of population variance σ^2 is:

$$\widetilde{\sigma}_n^2(x_1, x_2, \cdots, x_n) \equiv s^{**2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

1.7.2 Statistic, Estimate and Estimator

The underlying distribution of population is assumed to be known, but the parameter θ, which characterizes the underlying distribution, is unknown. The probability density function of population is given by $f(x;\theta)$. Let $X_1, X_2, \cdots, X_n$ be a subset of population, which are regarded as the random variables and are assumed to be mutually independent. $x_1, x_2, \cdots, x_n$ are taken as the experimental values of the random variables $X_1, X_2, \cdots, X_n$. In statistics, we consider that n-variate random variables $X_1, X_2, \cdots, X_n$ takes the experiments values $x_1, x_2, \cdots, x_n$ by chance. There, the experiments values and the actually observed data series are used in the same meaning.

As discussed in Section 1.7.1, $\hat{\theta}_n(x_1, x_2, \cdots, x_n)$ denotes the point estimate of θ. In the case where the observed data $x_1, x_2, \cdots, x_n$ are replaced by the corresponding random variables $X_1, X_2, \cdots, X_n$, a function of $X_1, X_2, \cdots, X_n$, i.e., $\hat{\theta}(X_1, X_2, \cdots, X_n)$, is called the **estimator** of θ, which should be distinguished from the **estimate** of θ, i.e., $\hat{\theta}(x_1, x_2, \cdots, x_n)$.

Example 1.12: Let $X_1, X_2, \cdots, X_n$ denote a random sample of n from a given distribution $f(x;\theta)$. Consider the case of $\theta = (\mu, \sigma^2)$.

The estimator of μ is given by $\overline{X} = (1/n)\sum_{i=1}^{n} X_i$, while the estimate of μ is $\overline{x} = (1/n)\sum_{i=1}^{n} x_i$. The estimator of σ^2 is $S^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2/(n-1)$ and the estimate of σ^2 is $s^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2/(n-1)$.

There are numerous estimators and estimates of θ. All of $(1/n)\sum_{i=1}^{n} X_i$, $(X_1 + X_n)/2$, median of $(X_1, X_2, \cdots, X_n)$ and so on are taken as the estimators of μ. Of course, they are called the estimates of θ when X_i is replaced by x_i for all i. Similarly, both $S^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2/(n-1)$ and $S^{*2} = \sum_{i=1}^{2}(X_i - \overline{X})^2/n$ are the estimators of σ^2. We need to choose one out of the numerous estimators of θ. The problem of choosing an optimal estimator out of the numerous estimators is discussed in Sections 1.7.4 and 1.7.5.

Finally, note as follows. A function of random variables is called a **statistic**. The statistic for estimation of the parameter is called an estimator. Therefore, an estimator is a family of a statistic.

1.7.3 Estimation of Mean and Variance

Suppose that the population distribution is given by $f(x;\theta)$. The random sample $X_1, X_2, \cdots, X_n$ are assumed to be drawn from the population distribution $f(x;\theta)$, where $\theta = (\mu, \sigma^2)$. Therefore, we can assume that $X_1, X_2, \cdots, X_n$ are mutually independently and identically distributed, where "identically" implies $\mathrm{E}(X_i) = \mu$ and $\mathrm{V}(X_i) = \sigma^2$ for all i.

Consider the estimators of $\theta = (\mu, \sigma^2)$ as follows.

1. The estimator of population mean μ is:
   - $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$.
2. The estimators of population variance σ^2 are:
   - $S^{*2} = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$, when μ is known,
   - $S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$,
   - $S^{**2} = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$,

Properties of $\overline{X}$: From Theorem on p.24, mean and variance of $\overline{X}$ are obtained as follows:

$$\mathrm{E}(\overline{X}) = \mu, \quad \mathrm{V}(\overline{X}) = \frac{\sigma^2}{n}.$$

Properties of S^{*2}, S^2 and S^{**2}: The expectation of S^{*2} is:

$$\begin{aligned}\mathrm{E}(S^{*2}) &= \mathrm{E}\Big(\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2\Big) = \frac{1}{n}\mathrm{E}\Big(\sum_{i=1}^{n}(X_i-\mu)^2\Big) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\Big((X_i-\mu)^2\Big) \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathrm{V}(X_i) = \frac{1}{n}\sum_{i=1}^{n}\sigma^2 = \frac{1}{n}n\sigma^2 = \sigma^2,\end{aligned}$$

where $\mathrm{E}((X_i-\mu)^2) = \mathrm{V}(X_i) = \sigma^2$ is used in the fourth and the fifth equalities.

Next, the expectation of S^2 is given by:

$$\begin{aligned}\mathrm{E}(S^2) &= \mathrm{E}\Big(\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2\Big) = \frac{1}{n-1}\mathrm{E}\Big(\sum_{i=1}^{n}(X_i-\overline{X})^2\Big) \\ &= \frac{1}{n-1}\mathrm{E}\Big(\sum_{i=1}^{n}((X_i-\mu)-(\overline{X}-\mu))^2\Big) \\ &= \frac{1}{n-1}\mathrm{E}\Big(\sum_{i=1}^{n}((X_i-\mu)^2-2(X_i-\mu)(\overline{X}-\mu)+(\overline{X}-\mu)^2)\Big) \\ &= \frac{1}{n-1}\mathrm{E}\Big(\sum_{i=1}^{n}(X_i-\mu)^2-2(\overline{X}-\mu)\sum_{i=1}^{n}(X_i-\mu)+n(\overline{X}-\mu)^2\Big) \\ &= \frac{1}{n-1}\mathrm{E}\Big(\sum_{i=1}^{n}(X_i-\mu)^2-n(\overline{X}-\mu)^2\Big) \\ &= \frac{n}{n-1}\mathrm{E}\Big(\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2\Big)-\frac{n}{n-1}\mathrm{E}((\overline{X}-\mu)^2) \\ &= \frac{n}{n-1}\sigma^2-\frac{n}{n-1}\frac{\sigma^2}{n} = \sigma^2.\end{aligned}$$

$\sum_{i=1}^{n}(X_i-\mu) = n(\overline{X}-\mu)$ is used in the sixth equality. $\mathrm{E}\Big((1/n)\sum_{i=1}^{n}(X_i-\mu)^2\Big) = \mathrm{E}(S^{*2}) = \sigma^2$ and $\mathrm{E}((\overline{X}-\mu)^2) = \mathrm{V}(\overline{X}) = \sigma^2/n$ are required in the eighth equality.

Finally, the mathematical expectation of S^{**2} is represented by:

$$\mathrm{E}(S^{**2}) = \mathrm{E}\Big(\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\Big) = \mathrm{E}\Big(\frac{n-1}{n}\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2\Big)$$
$$= \mathrm{E}\Big(\frac{n-1}{n}S^2\Big) = \frac{n-1}{n}\mathrm{E}(S^2) = \frac{n-1}{n}\sigma^2 \neq \sigma^2.$$

Summarizing the above results, we obtain as follows:

$$\mathrm{E}(S^{*2}) = \sigma^2, \quad \mathrm{E}(S^2) = \sigma^2, \quad \mathrm{E}(S^{**2}) = \frac{n-1}{n}\sigma^2 \neq \sigma^2.$$

1.7.4 Point Estimation: Optimality

As mentioned in the previous sections, θ denotes the parameter to be estimated. $\hat{\theta}_n(X_1, X_2, \cdots, X_n)$ represents the estimator of θ, while $\hat{\theta}_n(x_1, x_2, \cdots, x_n)$ indicates the estimate of θ. Hereafter, in the case of no confusion, $\hat{\theta}_n(X_1, X_2, \cdots, X_n)$ is simply written as $\hat{\theta}_n$.

As discussed above, there are numerous candidates of the estimator $\hat{\theta}_n$. The desired properties which $\hat{\theta}_n$ have to satisfy include unbiasedness, efficiency and consistency.

Unbiasedness: One of the desirable features that the estimator of the parameter should have is given by:

$$\mathrm{E}(\hat{\theta}_n) = \theta, \tag{1.11}$$

which implies that $\hat{\theta}_n$ is distributed around θ. When the condition (1.11) holds, $\hat{\theta}_n$ is called the **unbiased estimator** of θ. $\mathrm{E}(\hat{\theta}_n) - \theta$ is defined as **bias**.

As an example of unbiasedness, consider the case of $\theta = (\mu, \sigma^2)$. Suppose that $X_1, X_2, \cdots, X_n$ are mutually independently and identically distributed with mean μ and variance σ^2. Consider the following estimators of μ and σ^2.

1. The estimator of μ is:
   - $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$.
2. The estimators of σ^2 are:
   - $S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$,

- $S^{**2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$.

Since we have obtained $\mathrm{E}(\overline{X}) = \mu$ and $\mathrm{E}(S^2) = \sigma^2$ in Section 1.7.3, $\overline{X}$ and S^2 are unbiased estimators of μ and σ^2. However, we have obtained the result $\mathrm{E}(S^{**2}) \neq \sigma^2$ in Section 1.7.3 and therefore S^{**2} is not an unbiased estimator of σ^2. Thus, according to the criterion of unbiasedness, S^2 is preferred to S^{**2} for estimation of σ^2.

Efficiency: Consider two estimators, i.e., $\hat{\theta}_n$ and $\widetilde{\theta}_n$. Both are assumed to be unbiased. That is, we have the following condition: $\mathrm{E}(\hat{\theta}_n) = \theta$ and $\mathrm{E}(\widetilde{\theta}_n) = \theta$. When $\mathrm{V}(\hat{\theta}_n) < \mathrm{V}(\widetilde{\theta}_n)$, we say that $\hat{\theta}_n$ is more efficient than $\widetilde{\theta}_n$. The estimator which is widely distributed is not preferred.

Consider as many unbiased estimators as possible. The unbiased estimator with the least variance is known as the efficient estimator. We have the case where an **efficient estimator** does not exist.

In order to obtain the efficient estimator, we utilize Cramer-Rao inequality. Suppose that X_i has the probability density function $f(x_i;\theta)$ for all i, i.e., X_1, X_2, $\cdots$, X_n are mutually independently and identically distributed. For any unbiased estimator of θ, denoted by $\hat{\theta}_n$, it is known that we have the following inequality:

$$\mathrm{V}(\hat{\theta}_n) \geq \frac{\sigma^2(\theta)}{n}, \tag{1.12}$$

where

$$\sigma^2(\theta) = \frac{1}{\mathrm{E}\left(\left(\frac{\partial \log f(X;\theta)}{\partial\theta}\right)^2\right)} = \frac{1}{\mathrm{V}\left(\left(\frac{\partial \log f(X;\theta)}{\partial\theta}\right)\right)} = -\frac{1}{\mathrm{E}\left(\frac{\partial^2 \log f(X;\theta)}{\partial\theta^2}\right)}, \tag{1.13}$$

which is known as the **Cramer-Rao inequality**. See Appendix 1.4 for proof of the Cramer-Rao inequality.

When there exists the unbiased estimator $\hat{\theta}_n$ such that the equality in (1.12) holds, $\hat{\theta}_n$ becomes the unbiased estimator with minimum variance, which is the efficient estimator. $\sigma^2(\theta)/n$ is called the **Cramer-Rao lower bound**.

Example 1.13 (Efficient Estimator): Suppose that X_1, X_2, $\cdots$, X_n are mutually independently, identically and normally distributed with mean μ and variance σ^2. Then, we show that $\overline{X}$ is an efficient estimator of μ.

When $\sigma^2 < \infty$, from Theorem on p.24, $\mathrm{V}(\overline{X})$ is given by σ^2/n in spite of the distribution of X_i, $i = 1, 2, \cdots, n$. .....(A)

On the other hand, because we assume that X_i is normally distributed with mean μ and variance σ^2, the probability density function of X_i is given by:

$$f(x;\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{1}{2\sigma^2}(x-\mu)^2\Big).$$

The Cramer-Rao inequality is represented as:

$$\mathrm{V}(\overline{X}) \geq \frac{1}{n\mathrm{E}\left(\left(\frac{\partial \log f(X;\mu)}{\partial \mu}\right)^2\right)},$$

where the logarithm of $f(X;\mu)$ is written as:

$$\log f(X;\mu) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(X-\mu)^2.$$

Therefore, the partial derivative of $f(X;\mu)$ with respect to μ is:

$$\frac{\partial \log f(X;\mu)}{\partial \mu} = \frac{1}{\sigma^2}(X-\mu).$$

Accordingly, the Cramer-Rao inequality in this case is written as:

$$\mathrm{V}(\overline{X}) \geq \frac{1}{n\mathrm{E}\left(\left(\frac{1}{\sigma^2}(X-\mu)\right)^2\right)} = \frac{1}{n\frac{1}{\sigma^4}\mathrm{E}((X-\mu)^2)} = \frac{\sigma^2}{n}. \quad \text{.....(B)}$$

From (A) and (B), The variance of $\overline{X}$ is equal to the lower bound of Cramer-Rao inequality, i.e., $\mathrm{V}(\overline{X}) = \frac{\sigma^2}{n}$, which implies that the equality included in the Cramer-Rao inequality holds. Therefore, we can conclude that the sample mean $\overline{X}$ is an efficient estimator of μ.

Example 1.14 (Linear Unbiased Minimum Variance Estimator): Suppose that X_1, X_2, $\cdots$, X_n are mutually independently and identically distributed with mean μ and variance σ^2 (note that the normality assumption is excluded from Example 1.13). Consider the following linear estimator: $\hat{\mu} = \sum_{i=1}^{n} a_i X_i$. Then, we

want to show $\hat{\mu}$ (i.e., $\overline{X}$) is a **linear unbiased minimum variance estimator** if $a_i = 1/n$ for all i, i.e., if $\hat{\mu} = \overline{X}$.

Utilizing Theorem on p.23, when $\mathrm{E}(X_i) = \mu$ and $\mathrm{V}(X_i) = \sigma^2$ for all i, we have: $\mathrm{E}(\hat{\mu}) = \mu \sum_{i=1}^n a_i$ and $\mathrm{V}(\hat{\mu}) = \sigma^2 \sum_{i=1}^n a_i^2$.

Since $\hat{\mu}$ is linear in X_i, $\hat{\mu}$ is called a **linear estimator** of μ. In order for $\hat{\mu}$ to be unbiased, we need to have the condition: $\mathrm{E}(\hat{\mu}) = \mu \sum_{i=1}^n a_i = \mu$. That is, if $\sum_{i=1}^n a_i = 1$ is satisfied, $\hat{\mu}$ gives us a **linear unbiased estimator**. Thus, as mentioned in Example 1.12 of Section 1.7.2, there are numerous unbiased estimators.

The variance of $\hat{\mu}$ is given by $\sigma^2 \sum_{i=1}^n a_i^2$. We obtain the value of a_i which minimizes $\sum_{i=1}^n a_i^2$ with the constraint $\sum_{i=1}^n a_i = 1$. Construct the Lagrange function as follows:

$$L = \frac{1}{2} \sum_{i=1}^n a_i^2 + \lambda(1 - \sum_{i=1}^n a_i),$$

where λ denotes the Lagrange multiplier. The $\frac{1}{2}$ in front of the first term appears to make life easier later on and does not affect the outcome. To determine the optimum values, we set the partial derivatives of L with respect to a_i and λ equal to zero, i.e.,

$$\frac{\partial L}{\partial a_i} = a_i - \lambda = 0, \qquad i = 1, 2, \cdots, n,$$
$$\frac{\partial L}{\partial \lambda} = 1 - \sum_{i=1}^n a_i = 0.$$

Solving the above equations, $a_i = \lambda = 1/n$ is obtained. Therefore, when $a_i = 1/n$ for all i, $\hat{\mu}$ has minimum variance in a class of linear unbiased estimators. That is, $\overline{X}$ is a **linear unbiased minimum variance estimator**.

The linear unbiased minimum variance estimator should be distinguished from the efficient estimator discussed in Example 1.13. The former does not requires the assumption on the underlying distribution. The latter gives us the unbiased estimator which variance is equal to the Cramer-Rao lower bound, which is not restricted to a class of the linear unbiased estimators. Under the assumption of normal population, the linear unbiased minimum variance estimator leads to the efficient estimator. However, both are different in general. In addition, note that the efficient estimator does not necessarily exist.

Consistency: Let $\hat{\theta}_n$ be an estimator of θ. Suppose that for any $\epsilon > 0$ we have the following:

$$P(|\hat{\theta}_n - \theta| > \epsilon) \longrightarrow 0, \qquad \text{as} \quad n \longrightarrow \infty,$$

which implies that $\hat{\theta} \longrightarrow \theta$ as $n \longrightarrow \infty$. Then, we say that $\hat{\theta}_n$ is a **consistent estimator** of θ. That is, the estimator which approaches the true parameter value as the sample size is large is called the consistent estimator of the parameter.

Example 1.15: Suppose that $X_1, X_2, \cdots, X_n$ are mutually independently and identically distributed with mean μ and variance σ^2. Assume that σ^2 is known. Then, it is shown that $\overline{X}$ is a consistent estimator of μ.

From (1.9), Chebyshev's inequality is given by:

$$P(|X - \mathrm{E}(X)| > \epsilon) \leq \frac{\mathrm{V}(X)}{\epsilon^2},$$

for a random variable X. Here, replacing X by $\overline{X}$, we obtain $\mathrm{E}(\overline{X})$ and $\mathrm{V}(\overline{X})$ as follows:

$$\mathrm{E}(\overline{X}) = \mu, \quad \mathrm{V}(\overline{X}) = \frac{\sigma^2}{n},$$

because $\mathrm{E}(X_i) = \mu$ and $\mathrm{V}(X_i) = \sigma^2 < \infty$ are assumed for all i.

Then, when $n \longrightarrow \infty$, we obtain the following result:

$$P(|\overline{X} - \mu| > \epsilon) \leq \frac{\sigma^2}{n\epsilon^2} \longrightarrow 0,$$

which implies that $\overline{X} \longrightarrow \mu$ as $n \longrightarrow \infty$. Therefore, we can conclude that $\overline{X}$ is a consistent estimator of μ.

Summarizing the results up to now, $\overline{X}$ is an unbiased, minimum variance and consistent estimator of population mean μ. When the distribution of X_i is assumed to be normal for all i, $\overline{X}$ leads to an unbiased, efficient and consistent estimator of μ.

1.7.5 Maximum Likelihood Estimator

In Section 1.7.4, the properties of the estimators $\overline{X}$ and S^2 are discussed. It is shown that $\overline{X}$ is an unbiased, efficient and consistent estimator of μ under normality assumption and that S^2 is an unbiased estimator of σ^2. Note that S^2 is not efficient but consistent (we do not check these features of S^2 in this book).

The population parameter θ depends on a functional form of the population distribution $f(x;\theta)$. It corresponds to (μ, σ^2) in the case of the normal distribution and β in the case of the exponential distribution (Section 2.2.4). Now, in more general cases, we want to consider how to estimate θ. The maximum likelihood estimator gives us one of the solutions.

Let $X_1, X_2, \cdots, X_n$ be mutually independently and identically distributed random samples. X_i has the probability density function $f(x;\theta)$. Under these assumptions, the joint density function of $X_1, X_2, \cdots, X_n$ is given by:

$$f(x_1, x_2, \cdots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta),$$

where θ denotes the unknown parameter.

Given the actually observed data $(x_1, x_2, \cdots, x_n)$, the joint density $f(x_1, x_2, \cdots, x_n; \theta)$ is regarded as a function of θ, i.e.,

$$l(\theta) = l(\theta; x) = l(\theta; x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} f(x_i; \theta).$$

$l(\theta)$ is called the **likelihood function**.

Let $\hat{\theta}_n$ be the θ which maximizes the likelihood function. Replacing $x_1, x_2, \cdots, x_n$ by $X_1, X_2, \cdots, X_n$, $\hat{\theta}_n = \hat{\theta}_n(X_1, X_2, \cdots, X_n)$ is called the **maximum likelihood estimator**, while $\hat{\theta}_n(x_1, x_2, \cdots, x_n)$ is called the **maximum likelihood estimate**.

That is, solving the following equation:

$$\frac{\partial l(\theta)}{\partial \theta} = 0,$$

the maximum likelihood estimator $\hat{\theta}_n \equiv \hat{\theta}_n(X_1, X_2, \cdots, X_n)$ is obtained.

Example 1.16: Suppose that $X_1, X_2, \cdots, X_n$ are mutually independently, identically and normally distributed with mean μ and variance σ^2. We derive the maximum likelihood estimators of μ and σ^2. The joint density (or the likelihood function) of $X_1, X_2, \cdots, X_n$ is written as:

$$\begin{aligned} f(x_1, x_2, \cdots, x_n; \mu, \sigma^2) &= \prod_{i=1}^{n} f(x_i; \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{1}{2\sigma^2}(x_i - \mu)^2\Big) \\ &= (2\pi\sigma^2)^{-n/2} \exp\Big(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2\Big) = l(\mu, \sigma^2). \end{aligned}$$

The logarithm of the likelihood function is given by:

$$\log l(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2,$$

which is called the **log-likelihood function**. For maximization of the likelihood function, differentiating the log-likelihood function $\log l(\mu, \sigma^2)$ with respect to μ and σ^2, the first derivatives should be equal to zero, i.e.,

$$\frac{\partial \log l(\mu, \sigma^2)}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = 0,$$
$$\frac{\partial \log l(\mu, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2 = 0.$$

Let $\hat{\mu}$ and $\hat{\sigma}^2$ be the solution which satisfies the above two equations. Solving the two equations, we obtain the maximum likelihood estimates as follows:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x},$$
$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = s^{**2}.$$

Replacing x_i by X_i for $i = 1, 2, \cdots, n$, the maximum likelihood estimators of μ and σ^2 are given by $\overline{X}$ and S^{**2}, respectively. Since $\mathrm{E}(\overline{X}) = \mu$, the maximum likelihood estimator of μ, $\overline{X}$, is an unbiased estimator. However, because of $\mathrm{E}(S^{**2}) = \dfrac{n-1}{n}\sigma^2 \neq \sigma^2$ as shown in Section 1.7.3, the maximum likelihood estimator of σ^2, S^{**2}, is not an unbiased estimator.

Properties of Maximum Likelihood Estimator: For small sample, the maximum likelihood estimator has the following properties.

- The maximum likelihood estimator is not unbiased in general, but we often have the case where we can construct the unbiased estimator by an appropriate transformation.

  For instance, in Example 1.16, we find that the maximum likelihood estimator of σ^2, S^{**2}, is not unbiased. However, $\dfrac{n}{n-1}S^{**2}$ is an unbiased estimator of σ^2.

- If the efficient estimator exists, i.e., if there exists the estimator which satisfies the equality in the Cramer-Rao inequality, the maximum likelihood estimator is efficient.

For large sample, as $n \longrightarrow \infty$, the maximum likelihood estimator of θ, $\hat{\theta}_n$, has the following property:

$$\sqrt{n}(\hat{\theta}_n - \theta) \longrightarrow N(0, \sigma^2(\theta)), \tag{1.14}$$

where

$$\sigma^2(\theta) = \frac{1}{\mathrm{E}\left(\left(\frac{\partial \log f(X;\theta)}{\partial \theta}\right)^2\right)}.$$

(1.14) indicates that the maximum likelihood estimator has consistency, asymptotic unbiasedness, asymptotic efficiency and asymptotic normality. Asymptotic normality of the maximum likelihood estimator comes from the central limit theorem discussed in Section 1.6.3. Even though the underlying distribution is not normal, i.e., even though $f(x;\theta)$ is not normal, the maximum likelihood estimator is asymptotically normally distributed. Note that the properties of $n \longrightarrow \infty$ are called the asymptotic properties, which include consistency, asymptotic normality and so on.

That is, by normalizing, as $n \longrightarrow \infty$, we obtain as follows:

$$\frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\sigma(\theta)} = \frac{\hat{\theta}_n - \theta}{\sigma(\theta)/\sqrt{n}} \longrightarrow N(0,1).$$

As another representation, when n is large, we can approximate as follows:

$$\hat{\theta}_n \sim N(\theta, \frac{\sigma^2(\theta)}{n}).$$

This implies that when $n \longrightarrow \infty$, $\hat{\theta}_n$ approaches the lower bound of Cramer-Rao inequality: $\frac{\sigma^2(\theta)}{n}$, which property is called an asymptotic efficiency.

Moreover, replacing θ in variance $\sigma^2(\theta)$ by $\hat{\theta}_n$, when $n \longrightarrow \infty$, we have the following property:

$$\frac{\hat{\theta}_n - \theta}{\sigma(\hat{\theta}_n)/\sqrt{n}} \longrightarrow N(0,1), \tag{1.15}$$

which also comes from the central limit theorem.

Practically, when n is large, we approximately use as follows:

$$\hat{\theta}_n \sim N(\theta, \frac{\sigma^2(\hat{\theta}_n)}{n}). \tag{1.16}$$

Proof of (1.14): By the central limit theorem (1.10),

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial \log f(X_i;\theta)}{\partial \theta} \longrightarrow N(0, \frac{1}{\sigma^2(\theta)}), \tag{1.17}$$

where $\sigma^2(\theta)$ is defined in (1.13), i.e., $\mathrm{V}(\partial \log f(X_i;\theta)/\partial\theta) = 1/\sigma^2(\theta)$. As shown in (1.45) of Appendix 1.4, note that $\mathrm{E}(\partial \log f(X_i;\theta)/\partial\theta) = 0$. We can apply the central limit theorem, taking $\partial \log f(X_i;\theta)/\partial\theta$ as the i-th random variable.

By performing the first-order Taylor series expansion around $\hat{\theta}_n = \theta$, we have the following approximation:

$$\begin{aligned} 0 &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial \log f(X_i;\hat{\theta}_n)}{\partial \theta} \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial \log f(X_i;\theta)}{\partial \theta} + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial^2 \log f(X_i;\theta)}{\partial \theta^2} (\hat{\theta}_n - \theta) + \cdots. \end{aligned}$$

Therefore, the following approximation also holds:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial \log f(X_i;\theta)}{\partial \theta} \approx -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial^2 \log f(X_i;\theta)}{\partial \theta^2} (\hat{\theta}_n - \theta).$$

From (1.17) and the above equation, we obtain:

$$-\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \log f(X_i;\theta)}{\partial \theta^2} \sqrt{n}(\hat{\theta}_n - \theta) \longrightarrow N(0, \frac{1}{\sigma^2(\theta)}).$$

The law of large numbers indicates as follows:

$$-\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \log f(X_i;\theta)}{\partial \theta^2} \longrightarrow -\mathrm{E}\Big(\frac{\partial^2 \log f(X_i;\theta)}{\partial \theta^2}\Big) = \frac{1}{\sigma^2(\theta)},$$

where the last equality is from (1.13). Thus, we have the following relation:

$$-\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \log f(X_i;\theta)}{\partial \theta^2} \sqrt{n}(\hat{\theta}_n - \theta) \longrightarrow \frac{1}{\sigma^2(\theta)} \sqrt{n}(\hat{\theta}_n - \theta) \longrightarrow N(0, \frac{1}{\sigma^2(\theta)})$$

Therefore, the asymptotic normality of the maximum likelihood estimator is obtained as follows:

$$\sqrt{n}(\hat{\theta}_n - \theta) \longrightarrow N(0, \sigma^2(\theta)).$$

Thus, (1.14) is obtained.

1.7.6 Interval Estimation

In Sections 1.7.1 – 1.7.5, the point estimation is discussed. It is important to know where the true parameter value of θ is likely to lie.

Suppose that the population distribution is given by $f(x;\theta)$. Using the random sample $X_1, X_2, \cdots, X_n$ drawn from the population distribution, we construct the two statistics, say, $\hat{\theta}_U(X_1, X_2, \cdots, X_n; \theta^*)$ and $\hat{\theta}_L(X_1, X_2, \cdots, X_n; \theta^{**})$, where θ^* and θ^{**} denote the constant values which satisfy:

$$P\Big(\theta^* < \hat{\theta}_n < \theta^{**}\Big) = 1 - \alpha, \tag{1.18}$$

for $\theta^{**} > \theta^*$. Note that $\hat{\theta}_n$ depends on $X_1, X_2, \cdots, X_n$ as well as θ, i.e., $\hat{\theta}_n \equiv \hat{\theta}_n(X_1, X_2, \cdots, X_n; \theta)$. Now we assume that we can solve (1.18) with respect to θ, which is rewritten as follows:

$$P\Big(\hat{\theta}_L(X_1, X_2, \cdots, X_n; \theta^*) < \theta < \hat{\theta}_U(X_1, X_2, \cdots, X_n; \theta^{**})\Big) = 1 - \alpha. \tag{1.19}$$

(1.19) implies that θ lies on the interval $\Big(\hat{\theta}_L(X_1, X_2, \cdots, X_n; \theta^*), \hat{\theta}_U(X_1, X_2, \cdots, X_n; \theta^{**})\Big)$ with probability $1 - \alpha$. Depending on a functional form of $\hat{\theta}_n(X_1, X_2, \cdots, X_n; \theta)$, we possibly have the situation that θ^* and θ^{**} are switched with each other.

Now, we replace the random variables $X_1, X_2, \cdots, X_n$ by the experimental values $x_1, x_2, \cdots, x_n$. Then, we say that the interval:

$$\Big(\hat{\theta}_L(x_1, x_2, \cdots, x_n; \theta^*),\ \hat{\theta}_U(x_1, x_2, \cdots, x_n; \theta^{**})\Big)$$

is called the $100 \times (1 - \alpha)\%$ **confidence interval** of θ. Thus, estimating the interval is known as the **interval estimation**, which is distinguished from the point estimation. In the interval, $\hat{\theta}_L(x_1, x_2, \cdots, x_n; \theta^*)$ is known as the **lower bound** of the confidence interval, while $\hat{\theta}_U(x_1, x_2, \cdots, x_n; \theta^{**})$ is the **upper bound** of the confidence interval.

Given probability α, the $\hat{\theta}_L(X_1, X_2, \cdots, X_n; \theta^*)$ and $\hat{\theta}_U(X_1, X_2, \cdots, X_n; \theta^{**})$ which satisfies equation (1.19) are not unique. For estimation of the unknown parameter θ, it is more optimal to minimize the width of the confidence interval. Therefore, we should choose θ^* and θ^{**} which minimizes the width $\hat{\theta}_U(X_1, X_2, \cdots, X_n; \theta^{**}) - \hat{\theta}_L(X_1, X_2, \cdots, X_n; \theta^*)$.

Interval Estimation of $\overline{X}$: Let $X_1, X_2, \cdots, X_n$ be mutually independently and identically distributed random variables. X_i has a distribution with mean μ and variance σ^2. From the central limit theorem,

$$\frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \longrightarrow N(0, 1).$$

Replacing σ^2 by its estimator S^2 (or S^{**2}),

$$\frac{\overline{X} - \mu}{S / \sqrt{n}} \longrightarrow N(0, 1).$$

Therefore, when n is large enough,

$$P(z^* < \frac{\overline{X} - \mu}{S / \sqrt{n}} < z^{**}) = 1 - \alpha,$$

where z^* and z^{**} ($z^* < z^{**}$) are percent points from the standard normal density function. Solving the inequality above with respect to μ, the following expression is obtained.

$$P(\overline{X} - z^{**} \frac{S}{\sqrt{n}} < \mu < \overline{X} - z^* \frac{S}{\sqrt{n}}) = 1 - \alpha,$$

where $\hat{\theta}_L$ and $\hat{\theta}_U$ correspond to $\overline{X} - z^{**} S / \sqrt{n}$ and $\overline{X} - z^* S / \sqrt{n}$, respectively.

The length of the confidence interval is given by:

$$\hat{\theta}_U - \hat{\theta}_L = \frac{S}{\sqrt{n}}(z^{**} - z^*),$$

which should be minimized subject to:

$$\int_{z^*}^{z^{**}} f(x)\mathrm{d}x = 1 - \alpha,$$

i.e.,

$$F(z^{**}) - F(z^*) = 1 - \alpha,$$

where $F(\cdot)$ denotes the standard normal cumulative distribution function.

Solving the minimization problem above, we can obtain the conditions that $f(z^*) = f(z^{**})$ for $z^* < z^{**}$ and that $f(x)$ is symmetric. Therefore, we have:

$$-z^* = z^{**} = z_{\alpha/2},$$

where $z_{\alpha/2}$ denotes the $100 \times \alpha/2$ percent point from the standard normal density function.

Accordingly, replacing the estimators $\overline{X}$ and S^2 by their estimates $\overline{x}$ and s^2, the $100 \times (1-\alpha)\%$ confidence interval of μ is approximately represented as:

$$\Big(\overline{x} - z_{\alpha/2}\frac{s}{\sqrt{n}},\ \overline{x} + z_{\alpha/2}\frac{s}{\sqrt{n}}\Big),$$

for large n.

For now, we do not impose any assumptions on the distribution of X_i. If we assume that X_i is normal, $\sqrt{n}(\overline{X}-\mu)/S$ has a t distribution with $n-1$ degrees of freedom for any n. Therefore, $100\times(1-\alpha)\%$ confidence interval of μ is given by:

$$\Big(\overline{x} - t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}},\ \overline{x} + t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}\Big),$$

where $t_{\alpha/2}(n-1)$ denotes the $100 \times \alpha/2$ percent points of the t distribution with $n-1$ degrees of freedom. See Section 2.2.10 for the t distribution.

Interval Estimation of $\hat{\theta}_n$: Let $X_1, X_2, \cdots, X_n$ be mutually independently and identically distributed random variables. X_i has the probability density function $f(x_i;\theta)$. Suppose that $\hat{\theta}_n$ represents the maximum likelihood estimator of θ.

From (1.16), we can approximate the $100 \times (1-\alpha)\%$ confidence interval of θ as follows:

$$\Big(\hat{\theta}_n - z_{\alpha/2}\frac{\sigma(\hat{\theta}_n)}{\sqrt{n}},\ \hat{\theta}_n + z_{\alpha/2}\frac{\sigma(\hat{\theta}_n)}{\sqrt{n}}\Big).$$

1.8 Testing Hypothesis

1.8.1 Basic Concepts in Testing Hypothesis

Given the population distribution $f(x;\theta)$, we want to judge from the observed values $(x_1, x_2, \cdots, x_n)$ whether the hypothesis on the parameter θ, e.g. $\theta = \theta_0$, is correct or not. The hypothesis that we want to test is called the **null hypothesis**, which is denoted by $H_0 :\ \theta = \theta_0$. The hypothesis against the null hypothesis, e.g. $\theta \neq \theta_0$, is called the **alternative hypothesis**, which is denoted by $H_1 :\ \theta \neq \theta_0$.

Table 1.1: Type I and Type II Errors

	H_0 is true.	H_0 is false.
Acceptance of H_0	Correct judgment	Type II Error (Probability β)
Rejection of H_0	Type I Error (Probability α = Significance Level)	Correct judgment ($1-\beta$ = Power)

Type I and Type II Errors: When we test the null hypothesis H_0, as shown in Table 1.1 we have four cases, i.e., (i) we accept H_0 when H_0 is true, (ii) we reject H_0 when H_0 is true, (iii) we accept H_0 when H_0 is false, and (iv) we reject H_0 when H_0 is false. (i) and (iv) are correct judgments while (ii) and (iii) are not correct. (ii) is called a **type I error**. and (iii) is called a **type II error**. The probability of committing a type I error is called the **significance level**, which is denoted by α, and the probability of committing a type II error is denoted by β. Probability of (iv) is called the **power** or the **power function**, because it is a function of the parameter θ.

Testing Procedures: The testing procedure is summarized as follows.

1. Construct the null hypothesis (H_0) on the parameter.
2. Consider an appropriate statistic, which is called a **test statistic**. Derive a distribution function of the test statistic when H_0 is true.
3. From the observed data, compute the observed value of the test statistic.
4. Compare the distribution and the observed value of the test statistic. The observed value of the test statistic is in the tails of the distribution, we consider that H_0 is not likely to occur and we reject H_0.

The region that H_0 is unlikely to occur and accordingly H_0 is rejected is called the **rejection region** or the **critical region**, denoted by R. Conversely, the region that H_0 is likely to occur and accordingly H_0 is accepted is called the **acceptance region**, denoted by A.

Using the rejection region R and the acceptance region A, the type I and II errors and the power are formulated as follows. Suppose that the test statistic is

give by $T = T(X_1, X_2, \cdots, X_n)$. The probability of committing a type I error, i.e., the significance level α is given by:

$$P(T(X_1, X_2, \cdots, X_n) \in R | H_0 \text{ is true}) = \alpha,$$

which is the probability that rejects H_0 when H_0 is true. Conventionally, the significance level $\alpha = 0.1, 0.05, 0.001$ is chosen in practice. The probability of committing a type II error, i.e., β is represented as:

$$P(T(X_1, X_2, \cdots, X_n) \in A | H_0 \text{ is not true}) = \beta,$$

which corresponds to the probability that accepts H_0 when H_0 is not true. The power is defined as $1 - \beta$, i.e.,

$$P(T(X_1, X_2, \cdots, X_n) \in R | H_0 \text{ is not true}) = 1 - \beta,$$

which is the probability that rejects H_0 when H_0 is not true.

1.8.2 Power Function

Let $X_1, X_2, \cdots, X_n$ be mutually independently, identically and normally distributed with mean μ and variance σ^2. Assume that σ^2 is known.

In Figure 1.3, we consider the hypothesis on the population mean μ, i.e., the null hypothesis $H_0 : \mu = \mu_0$ against the alternative hypothesis $H_1 : \mu = \mu_1$, where $\mu_1 > \mu_0$ is taken. The dark shadow area corresponds to the probability of committing a type I error, i.e., the significance level, while the light shadow area indicates the probability of committing a type II error. The probability of the right-hand side of f^* in the distribution under H_1 represents the power of the test, i.e., $1 - \beta$.

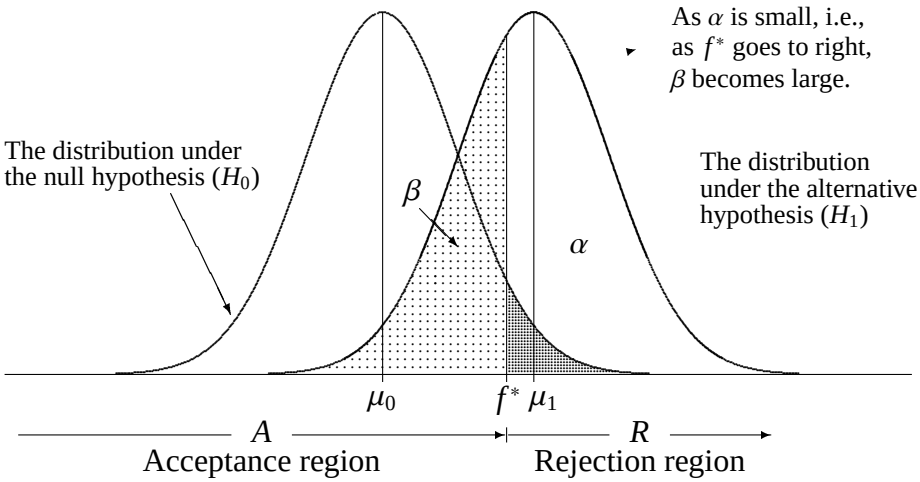
In the case of normal population, the distribution of sample mean $\overline{X}$ is given by:

$$\overline{X} \sim N(\mu, \frac{\sigma^2}{n}).$$

For the distribution of $\overline{X}$, see the moment-generating function of $\overline{X}$ in Theorem on p.33. By normalization, we have:

$$\frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \sim N(0, 1).$$

Figure 1.3: Type I Error (α) and Type II Error (β)



Therefore, under the null hypothesis $H_0 : \mu = \mu_0$, we obtain:

$$\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1),$$

where μ is replaced by μ_0. Since the significance level α is the probability which rejects H_0 when H_0 is true, it is given by:

$$\alpha = P(\overline{X} > \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}),$$

where z_α denotes $100 \times \alpha$ percent points of the standard normal density function. Therefore, the rejection region is given by: $\overline{X} > \mu_0 + z_\alpha \sigma/\sqrt{n}$.

Since the power $1-\beta$ is the probability which rejects H_0 when H_1 is true, it is given by:

$$1 - \beta = P(\overline{X} > \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}) = P(\frac{\overline{X} - \mu_1}{\sigma/\sqrt{n}} > \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} + z_\alpha)$$

$$= 1 - P(\frac{\overline{X} - \mu_1}{\sigma/\sqrt{n}} < \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} + z_\alpha) = 1 - F(\frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} + z_\alpha),$$

where $F(\cdot)$ represents the standard normal cumulative distribution function, i.e., $F(x) = \int_{-\infty}^{x} (2\pi)^{-1/2} \exp(-\frac{1}{2}t^2)\mathrm{d}t$. The power function is a function of μ_1, given μ_0 and α.

1.8.3 Testing Hypothesis on Population Mean

Let $X_1, X_2, \cdots, X_n$ be mutually independently, identically and normally distributed with mean μ and variance σ^2. Assume that σ^2 is known.

Consider testing the null hypothesis $H_0 : \mu = \mu_0$. When the null hypothesis H_0 is true, the distribution of $\overline{X}$ is given by:

$$\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1).$$

Therefore, the test statistic is given by: $\sqrt{n}(\overline{X} - \mu_0)/\sigma$, while the test statistic value is: $\sqrt{n}(\overline{x} - \mu_0)/\sigma$, where the sample mean $\overline{X}$ is replaced by the observed value $\overline{x}$.

Depending on the alternative hypothesis, we have the following three cases.

1. **The alternative hypothesis $H_1 : \mu < \mu_0$** (one-sided test):

   We have: $P\Big(\dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} < -z_\alpha\Big) = \alpha$. Therefore, when $\dfrac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} < -z_\alpha$, we reject the null hypothesis $H_0 : \mu = \mu_0$ at the significance level α.

2. **The alternative hypothesis $H_1 : \mu > \mu_0$** (one-sided test):

   We have: $P\Big(\dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} > z_\alpha\Big) = \alpha$. Therefore, when $\dfrac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} > z_\alpha$, we reject the null hypothesis $H_0 : \mu = \mu_0$ at the significance level α.

3. **The alternative hypothesis $H_1 : \mu \neq \mu_0$** (two-sided test):

   We have: $P\Big(\left|\dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}\right| > z_{\alpha/2}\Big) = \alpha$. Therefore, when $\left|\dfrac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}\right| > z_{\alpha/2}$, we reject the null hypothesis $H_0 : \mu = \mu_0$ at the significance level α.

When the sample size n is large enough, the testing procedure above can be applied to the cases: (i) the distribution of X_i is not known and (ii) σ^2 is replaced by its estimator S^2 (in the case where σ^2 is not known).

1.8.4 Wald Test

From (1.15), under the null hypothesis $H_0 : \theta = \theta_0$, as $n \longrightarrow \infty$, the maximum likelihood estimator $\hat{\theta}_n$ is distributed as follows:

$$\frac{\hat{\theta}_n - \theta_0}{\sigma(\hat{\theta}_n)/\sqrt{n}} \sim N(0,1).$$

For $H_0 : \theta = \theta_0$ and $H_1 : \theta \neq \theta_0$, replacing $X_1, X_2, \cdots, X_n$ in $\hat{\theta}_n$ by the observed values $x_1, x_2, \cdots, x_n$, we obtain the following testing procedure:

1. If we have:
$$\left| \frac{\hat{\theta}_n - \theta_0}{\sigma(\hat{\theta}_n)/\sqrt{n}} \right| > z_{\alpha/2},$$
we reject the null hypothesis H_0 at the significance level α, because the probability which H_0 occurs is small enough.

2. As for $H_0 : \theta = \theta_0$ and $H_1 : \theta > \theta_0$, if we have:
$$\frac{\hat{\theta}_n - \theta_0}{\sigma(\hat{\theta}_n)/\sqrt{n}} > z_{\alpha},$$
we reject H_0 at the significance level α.

3. For $H_0 : \theta = \theta_0$ and $H_1 : \theta < \theta_0$, when we have the following:
$$\frac{\hat{\theta}_n - \theta_0}{\sigma(\hat{\theta}_n)/\sqrt{n}} < -z_{\alpha},$$
we reject H_0 at the significance level α.

The testing procedure introduced here is called the **Wald test**.

Example 1.17: $X_1, X_2, \cdots, X_n$ are mutually independently, identically and exponentially distributed. Consider the following exponential probability density function:
$$f(x;\gamma) = \gamma e^{-\gamma x},$$
for $0 < x < \infty$.

Using the Wald test, we want to test the null hypothesis $H_0 : \gamma = \gamma_0$ against the alternative hypothesis $H_1 : \gamma \neq \gamma_0$.

Generally, as $n \longrightarrow \infty$, the distribution of the maximum likelihood estimator of the parameter γ, $\hat{\gamma}_n$, is asymptotically represented as:
$$\frac{\hat{\gamma}_n - \gamma}{\sigma(\hat{\gamma}_n)/\sqrt{n}} \sim N(0,1),$$
where
$$\sigma^2(\gamma) = \left(\mathrm{E}\left(\left(\frac{\mathrm{d} \log f(X;\gamma)}{\mathrm{d}\gamma} \right)^2 \right) \right)^{-1} = -\left(\mathrm{E}\left(\frac{\mathrm{d}^2 \log f(X;\gamma)}{\mathrm{d}\gamma^2} \right) \right)^{-1}.$$

See (1.13) and (1.15) for the above properties on the maximum likelihood estimator.

Therefore, under the null hypothesis $H_0 : \gamma = \gamma_0$, when n is large enough, we have the following distribution:

$$\frac{\hat{\gamma}_n - \gamma_0}{\sigma(\hat{\gamma}_n)/\sqrt{n}} \sim N(0,1).$$

As for the null hypothesis $H_0 : \gamma = \gamma_0$ against the alternative hypothesis $H_1 : \gamma \neq \gamma_0$, if we have:

$$\left| \frac{\hat{\gamma}_n - \gamma_0}{\sigma(\hat{\gamma}_n)/\sqrt{n}} \right| > z_{\alpha/2},$$

we can reject H_0 at the significance level α.

We need to derive $\sigma^2(\gamma)$ and $\hat{\gamma}_n$ to perform the testing procedure. First, $\sigma^2(\gamma)$ is given by:

$$\sigma^2(\gamma) = -\left(\mathrm{E}\Big(\frac{\mathrm{d}^2 \log f(X;\gamma)}{\mathrm{d}\gamma^2} \Big) \right)^{-1} = \gamma^2.$$

Note that the first- and the second-derivatives of $\log f(X;\gamma)$ with respect to γ are given by:

$$\frac{\mathrm{d} \log f(X;\gamma)}{\mathrm{d}\gamma} = \frac{1}{\gamma} - X, \qquad \frac{\mathrm{d}^2 \log f(X;\gamma)}{\mathrm{d}\gamma^2} = -\frac{1}{\gamma^2}.$$

Next, the maximum likelihood estimator of γ, i.e., $\hat{\gamma}_n$, is obtained as follows. Since X_1, $X_2 \cdots$, X_n are mutually independently and identically distributed, the likelihood function $l(\gamma)$ is given by:

$$l(\gamma) = \prod_{i=1}^{n} f(x_i;\gamma) = \prod_{i=1}^{n} \gamma e^{-\gamma x_i} = \gamma^n e^{-\gamma \sum x_i}.$$

Therefore, the log-likelihood function is written as:

$$\log l(\gamma) = n \log(\gamma) - \gamma \sum_{i=1}^{n} x_i.$$

We obtain the value of γ which maximizes $\log l(\gamma)$. Solving the following equation:

$$\frac{\mathrm{d} \log l(\gamma)}{\mathrm{d}\gamma} = \frac{n}{\gamma} - \sum_{i=1}^{n} x_i = 0,$$

the maximum likelihood estimator of γ, i.e., $\hat{\gamma}_n$ is represented as:

$$\hat{\gamma}_n = \frac{n}{\sum_{i=1}^{n} X_i} = \frac{1}{\overline{X}}.$$

Then, we have the following:

$$\frac{\hat{\gamma}_n - \gamma}{\sigma(\hat{\gamma}_n)/\sqrt{n}} = \frac{\hat{\gamma}_n - \gamma}{\hat{\gamma}_n/\sqrt{n}} \longrightarrow N(0,1),$$

where $\hat{\gamma}_n$ is given by $1/\overline{X}$.

For $H_0 : \gamma = \gamma_0$ and $H_1 : \gamma \neq \gamma_0$, if we have:

$$\left| \frac{\hat{\gamma}_n - \gamma_0}{\hat{\gamma}_n/\sqrt{n}} \right| > z_{\alpha/2},$$

we reject H_0 at the significance level α.

1.8.5 Likelihood Ratio Test

Suppose that the population distribution is given by $f(x;\theta)$, where $\theta = (\theta_1, \theta_2)$. Consider testing the null hypothesis $\theta_1 = \theta_1^*$ against the alternative hypothesis $H_1 : \theta_1 \neq \theta_1^*$, using the observed values $(x_1, x_2, \cdots, x_n)$ corresponding to the random sample $(X_1, X_2, \cdots, X_n)$.

Let θ_1 and θ_2 be $1 \times k_1$ and $1 \times k_2$ vectors, respectively. Therefore, $\theta = (\theta_1, \theta_2)$ denotes a $1 \times (k_1 + k_2)$ vector. Since we take the null hypothesis as $H_0 : \theta_1 = \theta_1^*$, the number of restrictions is given by k_1, which is equal to the dimension of θ_1.

The likelihood function is written as:

$$l(\theta_1, \theta_2) = \prod_{i=1}^{n} f(x_i; \theta_1, \theta_2).$$

Let $(\widetilde{\theta}_1, \widetilde{\theta}_2)$ be the maximum likelihood estimator of (θ_1, θ_2). That is, $(\widetilde{\theta}_1, \widetilde{\theta}_2)$ indicates the solution of (θ_1, θ_2), obtained from the following equations:

$$\frac{\partial l(\theta_1, \theta_2)}{\partial \theta_1} = 0, \quad \frac{\partial l(\theta_1, \theta_2)}{\partial \theta_2} = 0.$$

The solution $(\widetilde{\theta}_1, \widetilde{\theta}_2)$ is called the **unconstrained maximum likelihood estimator**, because the null hypothesis $H_0 : \theta_1 = \theta_1^*$ is not taken into account.

Let $\hat{\theta}_2$ be the maximum likelihood estimator of θ_2 under the null hypothesis $H_0 : \theta_1 = \theta_1^*$. That is, $\hat{\theta}_2$ is a solution of the following equation:

$$\frac{\partial l(\theta_1^*, \theta_2)}{\partial \theta_2} = 0.$$

The solution $\hat{\theta}_2$ is called the **constrained maximum likelihood estimator** of θ_2, because the likelihood function is maximized with respect to θ_2 subject to the constraint $\theta_1 = \theta_1^*$.

Define λ as follows:

$$\lambda = \frac{l(\theta_1^*, \hat{\theta}_2)}{l(\widetilde{\theta}_1, \widetilde{\theta}_2)},$$

which is called the **likelihood ratio**.

As n goes to infinity, it is known that we have:

$$-2\log(\lambda) \sim \chi^2(k_1),$$

where k_1 denotes the number of the constraints.

Let $\chi^2_\alpha(k_1)$ be the α percent point from the chi-square distribution with k_1 degrees of freedom. When $-2\log(\lambda) > \chi^2_\alpha(k_1)$, we reject the null hypothesis $H_0 : \theta_1 = \theta_1^*$ at the significance level α. If $-2\log(\lambda)$ is close to zero, we accept the null hypothesis. When $(\theta_1^*, \hat{\theta}_2)$ is close to $(\widetilde{\theta}_1, \widetilde{\theta}_2)$, $-2\log(\lambda)$ approaches zero.

The likelihood ratio test is useful in the case where it is not easy to derive the distribution of $(\widetilde{\theta}_1, \widetilde{\theta}_2)$.

Example 1.18: $X_1, X_2, \cdots, X_n$ are mutually independently, identically and exponentially distributed. Consider the following exponential probability density function:

$$f(x; \gamma) = \gamma e^{-\gamma x},$$

for $0 < x < \infty$.

Using the likelihood ratio test, we want to test the null hypothesis $H_0 : \gamma = \gamma_0$ against the alternative hypothesis $H_1 : \gamma \neq \gamma_0$. Remember that in Example 1.17 we test the hypothesis with the Wald test.

In this case, the likelihood ratio is given by:

$$\lambda = \frac{l(\gamma_0)}{l(\hat{\gamma}_n)},$$

where $\hat{\gamma}_n$ is derived in Example 1.17, i.e.,

$$\hat{\gamma}_n = \frac{n}{\sum_{i=1}^{n} X_i} = \frac{1}{\overline{X}}.$$

Since the number of the constraint is equal to one, as the sample size n goes to infinity we have the following asymptotic distribution:

$$-2\log\lambda \longrightarrow \chi^2(1).$$

The likelihood ratio is computed as follows:

$$\lambda = \frac{l(\gamma_0)}{l(\hat{\gamma}_n)} = \frac{\gamma_0^n e^{-\gamma_0 \sum X_i}}{\hat{\gamma}_n^n e^{-n}}.$$

If $-2\log\lambda > \chi^2_\alpha(1)$, we reject the null hypothesis $H_0 : \mu = \mu_0$ at the significance level α. Note that $\chi^2_\alpha(1)$ denotes the $100 \times \alpha$ percent point from the chi-square distribution with one degree of freedom.

Example 1.19: Suppose that $X_1, X_2, \cdots, X_n$ are mutually independently, identically and normally distributed with mean zero and variance σ^2.

The normal probability density function with mean μ and variance σ^2 is given by:

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}.$$

By the likelihood ratio test, we want to test the null hypothesis $H_0 : \mu = \mu_0$ against the alternative hypothesis $H_1 : \mu \neq \mu_0$.

The likelihood ratio is given by:

$$\lambda = \frac{l(\mu_0, \widetilde{\sigma}^2)}{l(\hat{\mu}, \hat{\sigma}^2)},$$

where $\widetilde{\sigma}^2$ is the constrained maximum likelihood estimator with the constraint $\mu = \mu_0$, while $(\hat{\mu}, \hat{\sigma}^2)$ denotes the unconstrained maximum likelihood estimator. In this case, since the number of the constraint is one, the asymptotic distribution is as follows:

$$-2\log\lambda \longrightarrow \chi^2(1).$$

Now, we derive $l(\mu_0, \widetilde{\sigma}^2)$ and $l(\hat{\mu}, \hat{\sigma}^2)$. $l(\mu, \sigma^2)$ is written as:

$$\begin{aligned} l(\mu, \sigma^2) &= f(x_1, x_2, \cdots, x_n; \mu, \sigma^2) = \prod_{i=1}^{n} f(x_i; \mu, \sigma^2) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{1}{2\sigma^2}(x_i - \mu)^2\Big) \\ &= (2\pi\sigma^2)^{-n/2} \exp\Big(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2\Big). \end{aligned}$$

The log-likelihood function $\log l(\mu, \sigma^2)$ is represented as:

$$\log l(\mu, \sigma^2) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2.$$

For the numerator of the likelihood ratio, under the constraint $\mu = \mu_0$, maximize $\log l(\mu_0, \sigma^2)$ with respect to σ^2. Since we obtain the first-derivative:

$$\frac{\partial \log l(\mu_0, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2} \frac{1}{\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \mu_0)^2 = 0,$$

the constrained maximum likelihood estimator $\widetilde{\sigma}^2$ is given by:

$$\widetilde{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_0)^2.$$

Therefore, replacing σ^2 by $\widetilde{\sigma}^2$, $l(\mu_0, \widetilde{\sigma}^2)$ is written as:

$$l(\mu_0, \widetilde{\sigma}^2) = (2\pi\widetilde{\sigma}^2)^{-n/2} \exp\Big(-\frac{1}{2\widetilde{\sigma}^2} \sum_{i=1}^{n} (x_i - \mu_0)^2\Big) = (2\pi\widetilde{\sigma}^2)^{-n/2} \exp\Big(-\frac{n}{2}\Big).$$

For the denominator of the likelihood ratio, because the unconstrained maximum likelihood estimators are obtained as:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i, \qquad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2,$$

$l(\hat{\mu}, \hat{\sigma}^2)$ is written as:

$$l(\hat{\mu}, \hat{\sigma}^2) = (2\pi\hat{\sigma}^2)^{-n/2} \exp\Big(-\frac{1}{2\hat{\sigma}^2} \sum_{i=1}^{n} (x_i - \hat{\mu})^2\Big) = (2\pi\hat{\sigma}^2)^{-n/2} \exp\Big(-\frac{n}{2}\Big).$$

Thus, the likelihood ratio is given by:

$$\lambda = \frac{l(\mu_0, \widetilde{\sigma}^2)}{l(\hat{\mu}, \hat{\sigma}^2)} = \frac{(2\pi\widetilde{\sigma}^2)^{-n/2} \exp\left(-\frac{n}{2}\right)}{(2\pi\hat{\sigma}^2)^{-n/2} \exp\left(-\frac{n}{2}\right)} = \left(\frac{\widetilde{\sigma}^2}{\hat{\sigma}^2}\right)^{-n/2}.$$

Asymptotically, we have:

$$-2\log\lambda = n(\log\widetilde{\sigma}^2 - \log\hat{\sigma}^2) \sim \chi^2(1).$$

When $-2\log\lambda > \chi^2_\alpha(1)$, we reject the null hypothesis $H_0 : \mu = \mu_0$ at the significance level α.

1.9 Regression Analysis

1.9.1 Setup of the Model

When (X_1, Y_1), (X_2, Y_2), $\cdots$, (X_n, Y_n) are available, suppose that there is a linear relationship between Y and X, i.e.,

$$Y_i = \beta_1 + \beta_2 X_i + u_i, \tag{1.20}$$

for $i = 1, 2, \cdots, n$.

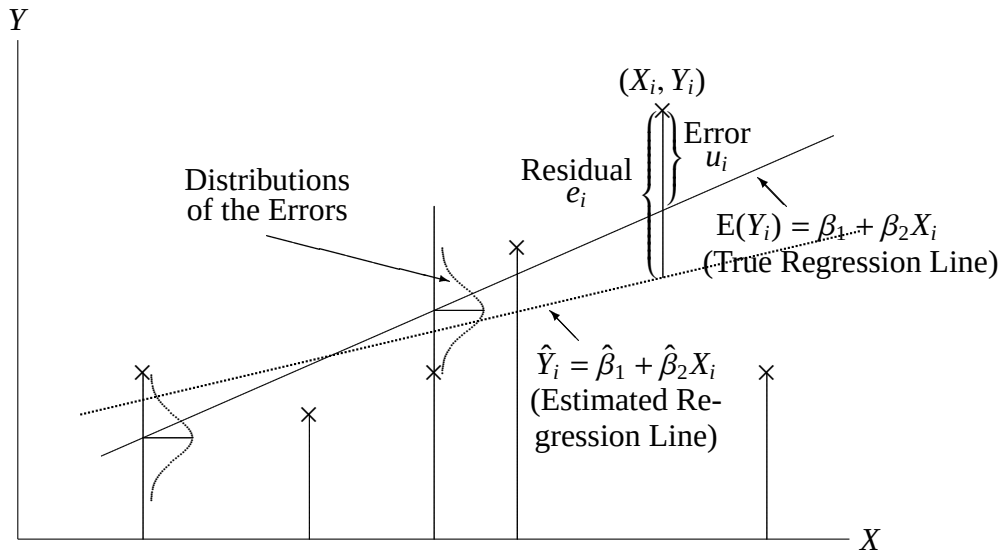
X_i and Y_i denote the i-th observations. Y_i is called the **dependent variable** or the **unexplanatory variable**, while X_i is known as the **independent variable** or the **explanatory variable**. β_1 and β_2 are unknown **parameter**s to be estimated. u_i is the unobserved **error term** assumed to be a random variable with mean zero and variance σ^2. β_1 and β_2 are called the **regression coefficient**s.

X_i is assumed to be nonstochastic, but Y_i is stochastic because Y_i depends on the error u_i. The error terms u_1, u_2, $\cdots$, u_n are assumed to be mutually independently and identically distributed. it is assumed that u_i has a distribution with mean zero, i.e., $\mathrm{E}(u_i) = 0$ is assumed. Taking the expectation on both sides of equation (1.20), the expectation of Y_i is represented as:

$$\begin{aligned} \mathrm{E}(Y_i) &= \mathrm{E}(\beta_1 + \beta_2 X_i + u_i) = \beta_1 + \beta_2 X_i + \mathrm{E}(u_i) \\ &= \beta_1 + \beta_2 X_i, \end{aligned} \tag{1.21}$$

for $i = 1, 2, \cdots, n$. Using $\mathrm{E}(Y_i)$ we can rewrite (1.20) as $Y_i = \mathrm{E}(Y_i) + u_i$. Equation (1.21) represents the true regression line.

Figure 1.4: True and Estimated Regression Lines



Let $\hat{\beta}_1$ and $\hat{\beta}_2$ be estimators of β_1 and β_2. Replacing (β_1, β_2) by $(\hat{\beta}_1, \hat{\beta}_2)$, equation (1.21) turns out to be:

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i, \tag{1.22}$$

for $i = 1, 2, \cdots, n$, where e_i is called the **residual**. The residual e_i is taken as the experimental value of u_i.

We define $\hat{Y}_i$ as follows:

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i, \tag{1.23}$$

for $i = 1, 2, \cdots, n$, which is interpreted as the predicted value of Y_i. Equation (1.23) indicates the estimated regression line, which is different from equation (1.21). Moreover, using $\hat{Y}_i$ we can rewrite (1.22) as $Y_i = \hat{Y}_i + e_i$.

Equations (1.21) and (1.23) are displayed in Figure 1.4. Consider the case of $n = 6$ for simplicity. $\times$ indicates the observed data series. The true regression line (1.21) is represented by the solid line, while the estimated regression line (1.23) is drawn with the dotted line. Based on the observed data, β_1 and β_2 are estimated as: $\hat{\beta}_1$ and $\hat{\beta}_2$.

In the next section, we consider how to obtain the estimates of β_1 and β_2, i.e., $\hat{\beta}_1$ and $\hat{\beta}_2$.

1.9.2 Ordinary Least Squares Estimation

Suppose that (X_1, Y_1), (X_2, Y_2), $\cdots$, (X_n, Y_n) are available. For the regression model (1.20), we consider estimating β_1, β_2 and σ^2. Replacing β_1 and β_2 by their estimates $\hat{\beta}_1$ and $\hat{\beta}_2$, remember that the residual e_i is given by:

$$e_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i.$$

The sum of squared residuals is defined as follows:

$$S(\hat{\beta}_1, \hat{\beta}_2) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2.$$

It might be plausible to choose the $\hat{\beta}_1$ and $\hat{\beta}_2$ which minimize the sum of squared residuals, i.e., $S(\hat{\beta}_1, \hat{\beta}_2)$. This method is called the **ordinary least squares (OLS) estimation**. To minimize $S(\hat{\beta}_1, \hat{\beta}_2)$ with respect to $\hat{\beta}_1$ and $\hat{\beta}_2$, we set the partial derivatives equal to zero:

$$\frac{\partial S(\hat{\beta}_1, \hat{\beta}_2)}{\partial \hat{\beta}_1} = -2 \sum_{i=1}^{n} (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) = 0,$$
$$\frac{\partial S(\hat{\beta}_1, \hat{\beta}_2)}{\partial \hat{\beta}_2} = -2 \sum_{i=1}^{n} X_i (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) = 0,$$

which yields the following two equations:

$$\overline{Y} = \hat{\beta}_1 + \hat{\beta}_2 \overline{X}, \tag{1.24}$$
$$\sum_{i=1}^{n} X_i Y_i = n\overline{X}\hat{\beta}_1 + \hat{\beta}_2 \sum_{i=1}^{n} X_i^2, \tag{1.25}$$

where $\overline{Y} = (1/n)\sum_{i=1}^{n} Y_i$ and $\overline{X} = (1/n)\sum_{i=1}^{n} X_i$. Multiplying (1.24) by $n\overline{X}$ and subtracting (1.25), we can derive $\hat{\beta}_2$ as follows:

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n} X_i Y_i - n\overline{XY}}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2} = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (X_i - \overline{X})^2}. \tag{1.26}$$

From equation (1.24), $\hat{\beta}_1$ is directly obtained as follows:

$$\hat{\beta}_1 = \overline{Y} - \hat{\beta}_2 \overline{X}. \tag{1.27}$$

When the observed values are taken for Y_i and X_i for $i = 1, 2, \cdots, n$, we say that $\hat{\beta}_1$ and $\hat{\beta}_2$ are called the **ordinary least squares estimate**s (or simply the **least squares estimate**s) of β_1 and β_2. When Y_i for $i = 1, 2, \cdots, n$ are regarded as the random sample, we say that $\hat{\beta}_1$ and $\hat{\beta}_2$ are called the **ordinary least squares estimator**s (or the **least squares estimator**s) of β_1 and β_2.

1.9.3 Properties of Least Squares Estimator

Equation (1.26) is rewritten as:

$$\begin{aligned}\hat{\beta}_2 &= \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2} = \frac{\sum_{i=1}^n (X_i - \overline{X}) Y_i}{\sum_{i=1}^n (X_i - \overline{X})^2} - \frac{\overline{Y} \sum_{i=1}^n (X_i - \overline{X})}{\sum_{i=1}^n (X_i - \overline{X})^2} \\ &= \sum_{i=1}^n \frac{X_i - \overline{X}}{\sum_{i=1}^n (X_i - \overline{X})^2} Y_i = \sum_{i=1}^n \omega_i Y_i. \end{aligned} \tag{1.28}$$

In the third equality, $\sum_{i=1}^n (X_i - \overline{X}) = 0$ is utilized because of $\overline{X} = (1/n) \sum_{i=1}^n X_i$. In the fourth equality, ω_i is defined as:

$$\omega_i = \frac{X_i - \overline{X}}{\sum_{i=1}^n (X_i - \overline{X})^2}.$$

ω_i is nonstochastic because X_i is assumed to be nonstochastic. ω_i has the following properties:

$$\sum_{i=1}^n \omega_i = \sum_{i=1}^n \frac{X_i - \overline{X}}{\sum_{i=1}^n (X_i - \overline{X})^2} = \frac{\sum_{i=1}^n (X_i - \overline{X})}{\sum_{i=1}^n (X_i - \overline{X})^2} = 0, \tag{1.29}$$

$$\sum_{i=1}^n \omega_i X_i = \sum_{i=1}^n \omega_i (X_i - \overline{X}) = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{\sum_{i=1}^n (X_i - \overline{X})^2} = 1, \tag{1.30}$$

$$\begin{aligned}\sum_{i=1}^n \omega_i^2 &= \sum_{i=1}^n \left(\frac{X_i - \overline{X}}{\sum_{i=1}^n (X_i - \overline{X})^2} \right)^2 = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{\left(\sum_{i=1}^n (X_i - \overline{X})^2 \right)^2} \\ &= \frac{1}{\sum_{i=1}^n (X_i - \overline{X})^2}. \end{aligned} \tag{1.31}$$

The first equality of equation (1.30) comes from equation (1.29).

From now on, we focus only on $\hat{\beta}_2$, because usually β_2 is more important than β_1 in the regression model (1.20). In order to obtain the properties of the least squares estimator $\hat{\beta}_2$, we rewrite equation (1.28) as:

$$\begin{aligned}\hat{\beta}_2 &= \sum_{i=1}^{n} \omega_i Y_i = \sum_{i=1}^{n} \omega_i(\beta_1 + \beta_2 X_i + u_i) \\ &= \beta_1 \sum_{i=1}^{n} \omega_i + \beta_2 \sum_{i=1}^{n} \omega_i X_i + \sum_{i=1}^{n} \omega_i u_i \\ &= \beta_2 + \sum_{i=1}^{n} \omega_i u_i. \end{aligned} \tag{1.32}$$

In the fourth equality of (1.32), equations (1.29) and (1.30) are utilized.

Mean and Variance of $\hat{\beta}_2$: $u_1, u_2, \cdots, u_n$ are assumed to be mutually independently and identically distributed with mean zero and variance σ^2, but they are not necessarily normal. Remember that we do not need normality assumption to obtain mean and variance but the normality assumption is required to test a hypothesis.

From equation (1.32), the expectation of $\hat{\beta}_2$ is derived as follows:

$$\begin{aligned}\mathrm{E}(\hat{\beta}_2) &= \mathrm{E}(\beta_2 + \sum_{i=1}^{n} \omega_i u_i) = \beta_2 + \mathrm{E}(\sum_{i=1}^{n} \omega_i u_i) \\ &= \beta_2 + \sum_{i=1}^{n} \omega_i \mathrm{E}(u_i) = \beta_2. \end{aligned} \tag{1.33}$$

It is shown from (1.33) that the ordinary least squares estimator $\hat{\beta}_2$ are the unbiased estimator of β_2.

From (1.32), the variance of $\hat{\beta}_2$ is computed as:

$$\begin{aligned}\mathrm{V}(\hat{\beta}_2) &= \mathrm{V}(\beta_2 + \sum_{i=1}^{n} \omega_i u_i) = \mathrm{V}(\sum_{i=1}^{n} \omega_i u_i) = \sum_{i=1}^{n} \mathrm{V}(\omega_i u_i) = \sum_{i=1}^{n} \omega_i^2 \mathrm{V}(u_i) \\ &= \sigma^2 \sum_{i=1}^{n} \omega_i^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}. \end{aligned} \tag{1.34}$$

From Theorem on p.18, the second and the fourth equalities hold. The third equality holds because $u_1, u_2, \cdots, u_n$ are mutually independent (see the theorem on p.23). The last equality comes from equation (1.31).

Thus, $\mathrm{E}(\hat{\beta}_2)$ and $\mathrm{V}(\hat{\beta}_2)$ are given by (1.33) and (1.34).

Gauss-Markov Theorem: It has been discussed above that $\hat{\beta}_2$ is represented as (1.28), which implies that $\hat{\beta}_2$ is a linear estimator, i.e., linear in Y_i. In addition, (1.33) indicates that $\hat{\beta}_2$ is an unbiased estimator. Therefore, summarizing these two facts, it is shown that $\hat{\beta}_2$ is a **linear unbiased estimator**. Furthermore, here we show that $\hat{\beta}_2$ has minimum variance within a class of the linear unbiased estimators.

Consider the alternative linear unbiased estimator $\widetilde{\beta}_2$ as follows:

$$\widetilde{\beta}_2 = \sum_{i=1}^{n} c_i Y_i = \sum_{i=1}^{n} (\omega_i + d_i) Y_i,$$

where $c_i = \omega_i + d_i$ is defined and d_i is nonstochastic. Then, $\widetilde{\beta}_2$ is transformed into:

$$\begin{aligned}
\widetilde{\beta}_2 &= \sum_{i=1}^{n} c_i Y_i = \sum_{i=1}^{n} (\omega_i + d_i)(\beta_1 + \beta_2 X_i + u_i) \\
&= \beta_1 \sum_{i=1}^{n} \omega_i + \beta_2 \sum_{i=1}^{n} \omega_i X_i + \sum_{i=1}^{n} \omega_i u_i + \beta_1 \sum_{i=1}^{n} d_i + \beta_2 \sum_{i=1}^{n} d_i X_i + \sum_{i=1}^{n} d_i u_i \\
&= \beta_2 + \beta_1 \sum_{i=1}^{n} d_i + \beta_2 \sum_{i=1}^{n} d_i X_i + \sum_{i=1}^{n} \omega_i u_i + \sum_{i=1}^{n} d_i u_i.
\end{aligned}$$

Equations (1.29) and (1.30) are used in the forth equality. Taking the expectation on both sides of the above equation, we obtain:

$$\begin{aligned}
\mathrm{E}(\widetilde{\beta}_2) &= \beta_2 + \beta_1 \sum_{i=1}^{n} d_i + \beta_2 \sum_{i=1}^{n} d_i X_i + \sum_{i=1}^{n} \omega_i \mathrm{E}(u_i) + \sum_{i=1}^{n} d_i \mathrm{E}(u_i) \\
&= \beta_2 + \beta_1 \sum_{i=1}^{n} d_i + \beta_2 \sum_{i=1}^{n} d_i X_i.
\end{aligned}$$

Note that d_i is not a random variable and that $\mathrm{E}(u_i) = 0$. Since $\widetilde{\beta}_2$ is assumed to be unbiased, we need the following conditions:

$$\sum_{i=1}^{n} d_i = 0, \qquad \sum_{i=1}^{n} d_i X_i = 0.$$

When these conditions hold, we can rewrite $\widetilde{\beta}_2$ as:

$$\widetilde{\beta}_2 = \beta_2 + \sum_{i=1}^{n} (\omega_i + d_i) u_i.$$

The variance of $\widetilde{\beta}_2$ is derived as:

$$\begin{aligned}
\mathrm{V}(\widetilde{\beta}_2) &= \mathrm{V}\Big(\beta_2 + \sum_{i=1}^{n}(\omega_i + d_i)u_i\Big) = \mathrm{V}\Big(\sum_{i=1}^{n}(\omega_i + d_i)u_i\Big) = \sum_{i=1}^{n}\mathrm{V}\Big((\omega_i + d_i)u_i\Big) \\
&= \sum_{i=1}^{n}(\omega_i + d_i)^2\mathrm{V}(u_i) = \sigma^2(\sum_{i=1}^{n}\omega_i^2 + 2\sum_{i=1}^{n}\omega_i d_i + \sum_{i=1}^{n}d_i^2) \\
&= \sigma^2(\sum_{i=1}^{n}\omega_i^2 + \sum_{i=1}^{n}d_i^2).
\end{aligned}$$

From unbiasedness of $\widetilde{\beta}_2$, using $\sum_{i=1}^{n} d_i = 0$ and $\sum_{i=1}^{n} d_i X_i = 0$, we obtain:

$$\sum_{i=1}^{n}\omega_i d_i = \frac{\sum_{i=1}^{n}(X_i - \overline{X})d_i}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{\sum_{i=1}^{n}X_i d_i - \overline{X}\sum_{i=1}^{n}d_i}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = 0,$$

which is utilized to obtain the variance of $\widetilde{\beta}_2$. From (1.34), the variance of $\hat{\beta}_2$ is given by: $\mathrm{V}(\hat{\beta}_2) = \sigma^2 \sum_{i=1}^{n}\omega_i^2$. Therefore, we have:

$$\mathrm{V}(\widetilde{\beta}_2) \geq \mathrm{V}(\hat{\beta}_2),$$

because $\sum_{i=1}^{n} d_i^2 \geq 0$. When $\sum_{i=1}^{n} d_i^2 = 0$, i.e., when $d_1 = d_2 = \cdots = d_n = 0$, we have the equality: $\mathrm{V}(\widetilde{\beta}_2) = \mathrm{V}(\hat{\beta}_2)$. In the case of $d_1 = d_2 = \cdots = d_n = 0$, $\hat{\beta}_2$ is equivalent to $\widetilde{\beta}_2$.

Thus, the least squares estimator $\hat{\beta}_2$ gives us the **linear unbiased minimum variance estimator**, or equivalently the **best linear unbiased estimator (BLUE)**, which is called the **Gauss-Markov theorem**.

Asymptotic Properties of $\hat{\beta}_2$: We assume that as n goes to infinity we have the following:

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \longrightarrow M < \infty,$$

where M is a constant value. From (1.31), we obtain:

$$n\sum_{i=1}^{n}\omega_i^2 = \frac{1}{(1/n)\sum_{i=1}^{n}(X_i - \overline{X})} \longrightarrow \frac{1}{M}.$$

Note that $f(x_n) \longrightarrow f(m)$ when $x_n \longrightarrow m$, where m is a constant value and $f(\cdot)$ is a function.

Here, we show both consistency of $\hat{\beta}_2$ and asymptotic normality of $\sqrt{n}(\hat{\beta}_2-\beta_2)$. First, we prove that $\hat{\beta}_2$ is a consistent estimator of β_2. As in (1.9), Chebyshev's inequality is given by:

$$P(|X-\mu|>\epsilon) \leq \frac{\sigma^2}{\epsilon^2},$$

where $\mu = \mathrm{E}(X)$ and $\sigma^2 = \mathrm{V}(X)$. Here, we replace X, $\mathrm{E}(X)$ and $\mathrm{V}(X)$ by $\hat{\beta}_2$,

$$\mathrm{E}(\hat{\beta}_2) = \beta_2, \quad \mathrm{V}(\hat{\beta}_2) = \sigma^2 \sum_{i=1}^{n} \omega_i^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(X_i-\overline{X})},$$

respectively. Then, when $n \longrightarrow \infty$, we obtain the following result:

$$P(|\hat{\beta}_2-\beta_2|>\epsilon) \leq \frac{\sigma^2 \sum_{i=1}^{n}\omega_i^2}{\epsilon^2} = \frac{\sigma^2}{\epsilon^2 \sum_{i=1}^{n}(X_i-\overline{X})} \longrightarrow 0,$$

where $\sum_{i=1}^{n}\omega_i^2 \longrightarrow 0$ because $n\sum_{i=1}^{n}\omega_i^2 \longrightarrow 1/M$ from the assumption. Thus, we obtain the result that $\hat{\beta}_2 \longrightarrow \beta_2$ as $n \longrightarrow \infty$. Therefore, we can conclude that $\hat{\beta}_2$ is a consistent estimator of β_2.

Next, we want to show that $\sqrt{n}(\hat{\beta}_2-\beta_2)$ is asymptotically normal. Noting that $\hat{\beta}_2 = \beta_2 + \sum_{i=1}^{n}\omega_i u_i$ as in (1.32) from Corollary 2 on p.39 (central limit theorem), asymptotic normality is shown as follows:

$$\frac{\sum_{i=1}^{n}\omega_i u_i - \mathrm{E}(\sum_{i=1}^{n}\omega_i u_i)}{\sqrt{\mathrm{V}(\sum_{i=1}^{n}\omega_i u_i)}} = \frac{\sum_{i=1}^{n}\omega_i u_i}{\sigma\sqrt{\sum_{i=1}^{n}\omega_i^2}} = \frac{\hat{\beta}_2-\beta_2}{\sigma/\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})^2}} \longrightarrow N(0,1),$$

where $\mathrm{E}(\sum_{i=1}^{n}\omega_i u_i) = 0$, $\mathrm{V}(\sum_{i=1}^{n}\omega_i u_i) = \sigma^2\sum_{i=1}^{n}\omega_i^2$ and $\sum_{i=1}^{n}\omega_i u_i = \hat{\beta}_2-\beta_2$ are substituted in the second equality. Moreover, we can rewrite as follows:

$$\frac{\hat{\beta}_2-\beta_2}{\sigma/\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})^2}} = \frac{\sqrt{n}(\hat{\beta}_2-\beta_2)}{\sigma/\sqrt{(1/n)\sum_{i=1}^{n}(X_i-\overline{X})^2}} \longrightarrow \frac{\sqrt{n}(\hat{\beta}_2-\beta_2)}{\sigma/\sqrt{M}} \longrightarrow N(0,1),$$

or equivalently,

$$\sqrt{n}(\hat{\beta}_2-\beta_2) \longrightarrow N(0, \frac{\sigma^2}{M}).$$

Thus, asymptotic normality of $\sqrt{n}(\hat{\beta}_2-\beta_2)$ is shown.

Finally, replacing σ^2 by its consistent estimator s^2, it is known as follows:

$$\frac{\hat{\beta}_2-\beta_2}{s\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})^2}} \longrightarrow N(0,1), \tag{1.35}$$

where s^2 is defined as:

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2, \tag{1.36}$$

which is a consistent and unbiased estimator of σ^2.

Thus, using (1.35), in large sample we can construct the confidence interval discussed in Section (1.7.6) and test the hypothesis discussed in Section 1.8.

Exact Distribution of $\hat{\beta}_2$: We have shown asymptotic normality of $\sqrt{n}(\hat{\beta}_2 - \beta_2)$, which is one of the large sample properties. Now, we discuss the small sample properties of $\hat{\beta}_2$. In order to obtain the distribution of $\hat{\beta}_2$ in small sample, the distribution of the error term has to be assumed. Therefore, the extra assumption is that $u_i \sim N(0, \sigma^2)$. Writing equation (1.32), again, $\hat{\beta}_2$ is represented as:

$$\hat{\beta}_2 = \beta_2 + \sum_{i=1}^{n} \omega_i u_i.$$

First, we obtain the distribution of the second term in the above equation. From Theorem on p.33, $\sum_{i=1}^{n} \omega_i u_i$ is distributed as:

$$\sum_{i=1}^{n} \omega_i u_i \sim N(0, \sigma^2 \sum_{i=1}^{n} \omega_i^2).$$

Therefore, from Example 1.9 on p.26, $\hat{\beta}_2$ is distributed as:

$$\hat{\beta}_2 = \beta_2 + \sum_{i=1}^{n} \omega_i u_i \sim N(\beta_2, \sigma^2 \sum_{i=1}^{n} \omega_i^2),$$

or equivalently,

$$\frac{\hat{\beta}_2 - \beta_2}{\sigma\sqrt{\sum_{i=1}^{n} \omega_i^2}} = \frac{\hat{\beta}_2 - \beta_2}{\sigma/\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}} \sim N(0,1),$$

for any n.

Moreover, replacing σ^2 by its estimator s^2 defined in (1.36), it is known that we have:

$$\frac{\hat{\beta}_2 - \beta_2}{s/\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}} \sim t(n-2),$$

where $t(n-2)$ denotes t distribution with $n-2$ degrees of freedom. See Section 2.2.10 for derivation of the t distribution. Thus, under normality assumption on the error term u_i, the $t(n-2)$ distribution is used for the confidence interval and the testing hypothesis in small sample.

1.9.4 Multiple Regression Model

In Sections 1.9.1 – 1.9.3, only one independent variable, i.e., X_i, is taken into the regression model. In this section, we extend it to more independent variables, which is called the **multiple regression**. We consider the following regression model:

$$\begin{aligned} Y_i &= \beta_1 X_{i,1} + \beta_2 X_{i,2} + \cdots + \beta_k X_{i,k} + u_i \\ &= X_i \beta + u_i, \end{aligned}$$

for $i = 1, 2, \cdots, n$, where X_i and β denote a $1 \times k$ vector of the independent variables and a $k \times 1$ vector of the unknown parameters to be estimated, which are represented as:

$$X_i = (X_{i,1}, X_{i,2}, \cdots, X_{i,k}), \qquad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}.$$

$X_{i,j}$ denotes the i-th observation of the j-th independent variable. The case of $k = 2$ and $X_{i,1} = 1$ for all i is exactly equivalent to (1.20). Therefore, the matrix form above is a generalization of (1.20). Writing all the equations for $i = 1, 2, \cdots, n$, we have:

$$\begin{aligned} Y_1 &= \beta_1 X_{1,1} + \beta_2 X_{1,2} + \cdots + \beta_k X_{1,k} + u_1, \\ Y_2 &= \beta_1 X_{2,1} + \beta_2 X_{2,2} + \cdots + \beta_k X_{2,k} + u_2, \\ &\vdots \\ Y_n &= \beta_1 X_{n,1} + \beta_2 X_{n,2} + \cdots + \beta_k X_{n,k} + u_n, \end{aligned}$$

which is rewritten as:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} X_{1,1} & X_{1,2} & \cdots & X_{1,k} \\ X_{2,1} & X_{2,2} & \cdots & X_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n,1} & X_{n,2} & \cdots & X_{n,k} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_k \end{pmatrix}.$$

Again, the above equation is compactly rewritten as:

$$Y = X\beta + u. \tag{1.37}$$

where Y, X and u are denoted by:

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \qquad X = \begin{pmatrix} X_{1,1} & X_{1,2} & \cdots & X_{1,k} \\ X_{2,1} & X_{2,2} & \cdots & X_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n,1} & X_{n,2} & \cdots & X_{n,k} \end{pmatrix}, \qquad u = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_k \end{pmatrix}.$$

Utilizing the matrix form (1.37), we derive the ordinary least squares estimator of β, denoted by $\hat{\beta}$. In equation (1.37), replacing β by $\hat{\beta}$, we have the following equation:

$$Y = X\hat{\beta} + e,$$

where e denotes a $1 \times n$ vector of the residuals. The i-th element of e is given by e_i. The sum of squared residuals is written as follows:

$$\begin{aligned} S(\hat{\beta}) &= \sum_{i=1}^{n} e_i^2 = e'e = (Y - X\hat{\beta})'(Y - X\hat{\beta}) = (Y' - \hat{\beta}'X')(Y - X\hat{\beta}) \\ &= Y'Y - Y'X\hat{\beta} - \hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta} = Y'Y - 2Y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta}. \end{aligned}$$

See Appendix 1.5 for the transpose in the fourth equality. In the last equality, note that $\hat{\beta}'X'Y = Y'X\hat{\beta}$ because both are scalars. To minimize $S(\hat{\beta})$ with respect to $\hat{\beta}$, we set the first derivative of $S(\hat{\beta})$ equal to zero, i.e.,

$$\frac{\partial S(\hat{\beta})}{\partial \hat{\beta}} = -2X'Y + 2X'X\hat{\beta} = 0.$$

See Appendix 1.5 for the derivatives of matrices. Solving the equation above with respect to $\hat{\beta}$, the ordinary least squares estimator of β is given by:

$$\hat{\beta} = (X'X)^{-1}X'Y. \tag{1.38}$$

See Appendix 1.5 for the inverse of the matrix. Thus, the ordinary least squares estimator is derived in the matrix form.

Now, in order to obtain the properties of $\hat{\beta}$ such as mean, variance, distribution and so on, (1.38) is rewritten as follows:

$$\begin{aligned} \hat{\beta} &= (X'X)^{-1}X'Y = (X'X)^{-1}X'(X\beta + u) = (X'X)^{-1}X'X\beta + (X'X)^{-1}X'u \\ &= \beta + (X'X)^{-1}X'u. \end{aligned} \tag{1.39}$$

Taking the expectation on both sides of equation (1.39), we have the following:

$$\mathrm{E}(\hat{\beta}) = \mathrm{E}(\beta + (X'X)^{-1}X'u) = \beta + (X'X)^{-1}X'\mathrm{E}(u) = \beta,$$

because of $\mathrm{E}(u) = 0$ by the assumption of the error term u_i. Thus, unbiasedness of $\hat{\beta}$ is shown.

The variance of $\hat{\beta}$ is obtained as:

$$\begin{aligned}\mathrm{V}(\hat{\beta}) &= \mathrm{E}((\hat{\beta}-\beta)(\hat{\beta}-\beta)') = \mathrm{E}\Big((X'X)^{-1}X'u((X'X)^{-1}X'u)'\Big) \\ &= \mathrm{E}((X'X)^{-1}X'uu'X(X'X)^{-1}) = (X'X)^{-1}X'\mathrm{E}(uu')X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}X'X(X'X)^{-1} = \sigma^2(X'X)^{-1}.\end{aligned}$$

The first equality is the definition of variance in the case of vector. In the fifth equality, $\mathrm{E}(uu') = \sigma^2 I_n$ is used, which implies that $\mathrm{E}(u_i^2) = \sigma^2$ for all i and $\mathrm{E}(u_i u_j) = 0$ for $i \neq j$. Remember that $u_1, u_2, \cdots, u_n$ are assumed to be mutually independently and identically distributed with mean zero and variance σ^2.

Under normality assumption on the error term u, it is known that the distribution of $\hat{\beta}$ is given by:

$$\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1}).$$

Taking the j-th element of $\hat{\beta}$, its distribution is given by:

$$\hat{\beta}_j \sim N(\beta_j, \sigma^2 a_{jj}), \qquad \text{i.e.,} \qquad \frac{\hat{\beta}_j - \beta_j}{\sigma\sqrt{a_{jj}}} \sim N(0,1),$$

where a_{jj} denotes the j-th diagonal element of $(X'X)^{-1}$.

Replacing σ^2 by its estimator s^2, we have the following t distribution:

$$\frac{\hat{\beta}_j - \beta_j}{s\sqrt{a_{jj}}} \sim t(n-k),$$

where $t(n-k)$ denotes the t distribution with $n-k$ degrees of freedom. s^2 is taken as follows:

$$s^2 = \frac{1}{n-k}\sum_{i=1}^{n} e_i^2 = \frac{1}{n-k}e'e = \frac{1}{n-k}(Y - X\hat{\beta})'(Y - X\hat{\beta}),$$

which leads to an unbiased estimator of σ^2.

Using the central limit theorem, without normality assumption we can show that as $n \longrightarrow \infty$, under the condition of $(1/n)X'X \longrightarrow M$ we have the following result:

$$\frac{\hat{\beta}_j - \beta_j}{s\sqrt{a_{jj}}} \longrightarrow N(0,1),$$

where M denotes a $k \times k$ constant matrix.

Thus, we can construct the confidence interval and the testing procedure, using the t distribution under the normality assumption or the normal distribution without the normality assumption.

Appendix 1.1: Integration by Substitution

Univariate Case: For a function of x, $f(x)$, we perform integration by substitution, using $x = \psi(y)$. Then, it is easy to obtain the following formula:

$$\int f(x)\mathrm{d}x = \int \psi'(y) f(\psi(y))\mathrm{d}y,$$

which formula is called the **integration by substitution**.

Proof:

Let $F(x)$ be the integration of $f(x)$, i.e.,

$$F(x) = \int_{-\infty}^{x} f(t)\mathrm{d}t,$$

which implies that $F'(x) = f(x)$.

Differentiating $F(x) = F(\psi(y))$ with respect to y, we have:

$$f(x) \equiv \frac{\mathrm{d}F(\psi(y))}{\mathrm{d}y} = \frac{\mathrm{d}F(x)}{\mathrm{d}x}\frac{\mathrm{d}x}{\mathrm{d}y} = f(x)\psi'(y) = f(\psi(y))\psi'(y).$$

Bivariate Case: For $f(x,y)$, define $x = \psi_1(u,v)$ and $y = \psi_2(u,v)$.

$$\iint f(x,y)\mathrm{d}x\mathrm{d}y = \iint J f(\psi_1(u,v),\psi_2(u,v))\mathrm{d}u\mathrm{d}v,$$

where J is called the **Jacobian**, which represents the following determinant:

$$J = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial x}{\partial v}\frac{\partial y}{\partial u}.$$

Appendix 1.2: Integration by Parts

Let $h(x)$ and $g(x)$ be functions of x. Then, we have the following formula:

$$\int h(x)g'(x)\mathrm{d}x = h(x)g(x) - \int h'(x)g(x)\mathrm{d}x,$$

which is called the **integration by parts**.

Proof:

Consider the derivative of $f(x)g(x)$ with respect to x, i.e.,

$$\Big(f(x)g(x)\Big)' = f'(x)g(x) + f(x)g'(x).$$

Integrating the above equation on both sides, we have:

$$\int \Big(f(x)g(x)\Big)'\mathrm{d}x = \int f'(x)g(x)\mathrm{d}x + \int f(x)g'(x)\mathrm{d}x.$$

Therefore, we obtain:

$$f(x)g(x) = \int f'(x)g(x)\mathrm{d}x + \int f(x)g'(x)\mathrm{d}x.$$

Thus, the following result is derived.

$$\int f(x)g'(x)\mathrm{d}x = f(x)g(x) - \int f'(x)g(x)\mathrm{d}x.$$

When we want to integrate $f(x)g'(x)$ within the range between a and b for $a < b$, the above formula is modified as:

$$\int_a^b f(x)g'(x)\mathrm{d}x = \Big[f(x)g(x)\Big]_a^b - \int_a^b f'(x)g(x)\mathrm{d}x.$$

Appendix 1.3: Taylor Series Expansion

Consider approximating $f(x)$ around $x = x_0$ by the Taylor series expansion. Then, $f(x)$ is approximated as follows:

$$\begin{aligned} f(x) &= f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2!}f''(x_0)(x - x_0)^2 + \frac{1}{3!}f'''(x_0)(x - x_0)^3 + \cdots \\ &= \sum_{n=0}^{\infty} \frac{1}{n!} f^{(n)}(x_0)(x - x_0)^n, \end{aligned}$$

where $f^{(n)}(x_0)$ denotes the n-th derivative of $f(x)$ evaluated at $x = x_0$. Note that $f^{(0)}(x_0) = f(x_0)$ and $0! = 1$.

In addition, the following approximation is called the k-order Taylor series expansion:

$$f(x) \approx \sum_{n=0}^{k} \frac{1}{n!} f^{(n)}(x_0)(x - x_0)^n.$$

Appendix 1.4: Cramer-Rao Inequality

As seen in (1.12) and (1.13), the Cramer-Rao inequality is given by:

$$\mathrm{V}(\hat{\theta}_n) \geq \frac{\sigma^2(\theta)}{n},$$

where

$$\sigma^2(\theta) = \frac{1}{\mathrm{E}\left(\left(\dfrac{\partial \log f(X;\theta)}{\partial \theta}\right)^2\right)} = \frac{1}{\mathrm{V}\left(\left(\dfrac{\partial \log f(X;\theta)}{\partial \theta}\right)\right)} = -\frac{1}{\mathrm{E}\left(\dfrac{\partial^2 \log f(X;\theta)}{\partial \theta^2}\right)}.$$

Here, we prove the above inequality and the equalities in $\sigma^2(\theta)$.

Proof:

The likelihood function $l(\theta; x) = l(\theta; x_1, x_2, \cdots, x_n)$ is a joint density of $X_1, X_2, \cdots, X_n$. Therefore, the integration of $l(\theta; x_1, x_2, \cdots, x_n)$ with respect to $x_1, x_2, \cdots, x_n$ is equal to one. See Section 1.7.5 for the likelihood function.

That is, we have the following equation:

$$1 = \int l(\theta; x)\mathrm{d}x, \tag{1.40}$$

where the likelihood function $l(\theta; x)$ is given by $l(\theta; x) = \prod_{i=1}^{n} f(x_i; \theta)$ and $\int \cdots \mathrm{d}x$ implies n-tuple integral.

Differentiating both sides of equation (1.40) with respect to θ, we obtain the following equation:

$$\begin{aligned} 0 &= \int \frac{\partial l(\theta; x)}{\partial \theta}\mathrm{d}x = \int \frac{1}{l(\theta; x)} \frac{\partial l(\theta; x)}{\partial \theta} l(\theta; x)\mathrm{d}x \\ &= \int \frac{\partial \log l(\theta; x)}{\partial \theta} l(\theta; x)\mathrm{d}x = \mathrm{E}\Big(\frac{\partial \log l(\theta; X)}{\partial \theta}\Big), \end{aligned} \tag{1.41}$$

which implies that the expectation of $\partial \log l(\theta; X)/\partial\theta$ is equal to zero. In the third equality, note that $\mathrm{d}\log x/\mathrm{d}x = 1/x$.

Now, let $\hat{\theta}_n$ be an estimator of θ. The definition of the mathematical expectation of the estimator $\hat{\theta}_n$ is represented as:

$$\mathrm{E}(\hat{\theta}_n) = \int \hat{\theta}_n l(\theta; x)\mathrm{d}x. \tag{1.42}$$

Differentiating equation (1.42) with respect to θ on both sides, we can rewrite as follows:

$$\begin{aligned}\frac{\partial \mathrm{E}(\hat{\theta}_n)}{\partial\theta} &= \int \hat{\theta}_n \frac{\partial l(\theta; x)}{\partial\theta}\mathrm{d}x = \int \hat{\theta}_n \frac{\partial \log l(\theta; x)}{\partial\theta} l(\theta; x)\mathrm{d}x \\ &= \int \Big(\hat{\theta}_n - \mathrm{E}(\hat{\theta}_n)\Big)\Big(\frac{\partial \log l(\theta; x)}{\partial\theta} - \mathrm{E}(\frac{\partial \log l(\theta; x)}{\partial\theta})\Big) l(\theta; x)\mathrm{d}x \\ &= \mathrm{Cov}\Big(\hat{\theta}_n, \frac{\partial \log l(\theta; X)}{\partial\theta}\Big). \end{aligned} \tag{1.43}$$

In the second equality, $\mathrm{d}\log x/\mathrm{d}x = 1/x$ is utilized. The third equality holds because of $\mathrm{E}(\partial \log l(\theta; x)/\partial\theta)$ from equation (1.41).

For simplicity of discussion, suppose that θ is a scalar. Taking the square on both sides of equation (1.43), we obtain the following expression:

$$\begin{aligned}\Big(\frac{\partial \mathrm{E}(\hat{\theta}_n)}{\partial\theta}\Big)^2 &= \left(\mathrm{Cov}\Big(\hat{\theta}_n, \frac{\partial \log l(\theta; X)}{\partial\theta}\Big)\right)^2 = \rho^2 \mathrm{V}(\hat{\theta}_n)\mathrm{V}\Big(\frac{\partial \log l(\theta; X)}{\partial\theta}\Big) \\ &\leq \mathrm{V}(\hat{\theta}_n)\mathrm{V}\left(\frac{\partial \log l(\theta; X)}{\partial\theta}\right),\end{aligned}$$

where ρ denotes the correlation coefficient between $\hat{\theta}_n$ and $\partial \log l(\theta; X)/\partial\theta$. That is, we have the definition of ρ is given by:

$$\rho = \frac{\mathrm{Cov}\Big(\hat{\theta}_n, \dfrac{\partial \log l(\theta; X)}{\partial\theta}\Big)}{\sqrt{\mathrm{V}(\hat{\theta}_n)}\sqrt{\mathrm{V}\Big(\dfrac{\partial \log l(\theta; X)}{\partial\theta}\Big)}}.$$

Moreover, we have $-1 \leq \rho \leq 1$ (i.e., $\rho^2 \leq 1$). Then, the following inequality is obtained.

$$\Big(\frac{\partial \mathrm{E}(\hat{\theta}_n)}{\partial\theta}\Big)^2 \leq \mathrm{V}(\hat{\theta}_n)\ \mathrm{V}\Big(\frac{\partial \log l(\theta; X)}{\partial\theta}\Big),$$

which is rewritten as:

$$\mathrm{V}(\hat{\theta}_n) \geq \frac{\left(\frac{\partial \mathrm{E}(\hat{\theta}_n)}{\partial \theta}\right)^2}{\mathrm{V}\left(\frac{\partial \log l(\theta; X)}{\partial \theta}\right)}. \tag{1.44}$$

When $\mathrm{E}(\hat{\theta}_n) = \theta$, i.e., when $\hat{\theta}_n$ is an unbiased estimator of θ, the numerator in the right-hand side of equation (1.44) is equal to one. Therefore, we have the following inequality:

$$\mathrm{V}(\hat{\theta}_n) \geq \frac{1}{\mathrm{V}\left(\frac{\partial \log l(\theta; X)}{\partial \theta}\right)} = \frac{1}{\mathrm{E}\left(\left(\frac{\partial \log l(\theta; X)}{\partial \theta}\right)^2\right)}.$$

Note that we have $\mathrm{V}(\partial \log l(\theta; X)/\partial\theta) = \mathrm{E}((\partial \log l(\theta; X)/\partial\theta)^2)$ in the equality above, because of $\mathrm{E}(\partial \log l(\theta; X)/\partial\theta) = 0$.

Moreover, the denominator in the right-hand side of the above inequality is rewritten as follows:

$$\begin{aligned}\mathrm{E}\left(\left(\frac{\partial \log l(\theta; X)}{\partial \theta}\right)^2\right) &= \mathrm{E}\left(\left(\sum_{i=1}^{n} \frac{\partial \log f(X_i; \theta)}{\partial \theta}\right)^2\right) = \sum_{i=1}^{n} \mathrm{E}\left(\left(\frac{\partial \log f(X_i; \theta)}{\partial \theta}\right)^2\right) \\ &= n\mathrm{E}\left(\left(\frac{\partial \log f(X; \theta)}{\partial \theta}\right)^2\right) = n \int_{-\infty}^{\infty} \left(\frac{\partial \log f(x; \theta)}{\partial \theta}\right)^2 f(x; \theta) \mathrm{d}x.\end{aligned}$$

In the first equality, $\log l(\theta; X) = \sum_{i=1}^{n} \log f(X_i; \theta)$ is utilized. Since X_i, $i = 1, 2, \cdots, n$, are mutually independent, the second equality holds. The third equality holds because X_1, X_2, $\cdots$, X_n are identically distributed.

Therefore, we obtain the following inequality:

$$\mathrm{V}(\hat{\theta}_n) \geq \frac{1}{\mathrm{E}\left(\left(\frac{\partial \log l(\theta; X)}{\partial \theta}\right)^2\right)} = \frac{1}{n\mathrm{E}\left(\left(\frac{\partial \log f(X; \theta)}{\partial \theta}\right)^2\right)} = \frac{\sigma^2(\theta)}{n},$$

which is equivalent to (1.12).

Next, we prove the equalities in (1.13), i.e.,

$$-\mathrm{E}\left(\frac{\partial^2 \log f(X; \theta)}{\partial \theta^2}\right) = \mathrm{E}\left(\left(\frac{\partial \log f(X; \theta)}{\partial \theta}\right)^2\right) = \mathrm{V}\left(\frac{\partial \log f(X; \theta)}{\partial \theta}\right).$$

Differentiating $\int f(x;\theta)\mathrm{d}x = 1$ with respect to θ, we obtain as follows:

$$\int \frac{\partial f(x;\theta)}{\partial \theta}\mathrm{d}x = 0.$$

We assume that the range of x does not depend on the parameter θ and that $\partial f(x;\theta)/\partial\theta$ exists. The above equation is rewritten as:

$$\int \frac{\partial \log f(x;\theta)}{\partial \theta} f(x;\theta)\mathrm{d}x = 0,$$

or equivalently,

$$\mathrm{E}\Big(\frac{\partial \log f(x;\theta)}{\partial \theta}\Big) = 0. \tag{1.45}$$

Again, differentiating with respect to θ,

$$\int \frac{\partial^2 \log f(x;\theta)}{\partial \theta^2} f(x;\theta)\mathrm{d}x + \int \frac{\partial \log f(x;\theta)}{\partial \theta}\frac{\partial f(x;\theta)}{\partial \theta}\mathrm{d}x = 0,$$

i.e.,

$$\int \frac{\partial^2 \log f(x;\theta)}{\partial \theta^2} f(x;\theta)\mathrm{d}x + \int \Big(\frac{\partial \log f(x;\theta)}{\partial \theta}\Big)^2 f(x;\theta)\mathrm{d}x = 0,$$

i.e.,

$$\mathrm{E}\Big(\frac{\partial^2 \log f(x;\theta)}{\partial \theta^2}\Big) + \mathrm{E}\left(\Big(\frac{\partial \log f(x;\theta)}{\partial \theta}\Big)^2\right) = 0.$$

Thus, we obtain:

$$-\mathrm{E}\Big(\frac{\partial^2 \log f(x;\theta)}{\partial \theta^2}\Big) = \mathrm{E}\left(\Big(\frac{\partial \log f(x;\theta)}{\partial \theta}\Big)^2\right).$$

Moreover, from equation (1.45), the following equation is derived.

$$\mathrm{E}\left(\Big(\frac{\partial \log f(x;\theta)}{\partial \theta}\Big)^2\right) = \mathrm{V}\Big(\frac{\partial \log f(x;\theta)}{\partial \theta}\Big).$$

Therefore, we have:

$$-\mathrm{E}\Big(\frac{\partial^2 \log f(X;\theta)}{\partial \theta^2}\Big) = \mathrm{E}\left(\Big(\frac{\partial \log f(X;\theta)}{\partial \theta}\Big)^2\right) = \mathrm{V}\Big(\frac{\partial \log f(X;\theta)}{\partial \theta}\Big).$$

Thus, the Cramer-Rao inequality is derived as:

$$\sigma^2(\theta) = \frac{1}{\mathrm{E}\left(\Big(\frac{\partial \log f(X;\theta)}{\partial \theta}\Big)^2\right)} = \frac{1}{\mathrm{V}\left(\Big(\frac{\partial \log f(X;\theta)}{\partial \theta}\Big)\right)} = -\frac{1}{\mathrm{E}\Big(\frac{\partial^2 \log f(X;\theta)}{\partial \theta^2}\Big)}.$$

Appendix 1.5: Some Formulas of Matrix Algebra

1. Let $A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{l1} & a_{l2} & \cdots & a_{lk} \end{pmatrix} = [a_{ij}]$, which is a $l \times k$ matrix, where a_{ij} denotes i-th row and j-th column of A. The **transpose** of A, denoted by A', is defined as:

   $$A' = \begin{pmatrix} a_{11} & a_{21} & \cdots & a_{l1} \\ a_{12} & a_{22} & \cdots & a_{l2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1k} & a_{2k} & \cdots & a_{lk} \end{pmatrix} = [a_{ji}],$$

   where the i-th row of A' is the i-th column of A.

2. $(Ax)' = x'A'$,

   where A and x are a $l \times k$ matrix and a $k \times 1$ vector, respectively.

3. $a' = a$,

   where a denotes a scalar.

4. $\dfrac{\partial a'x}{\partial x} = a$,

   where a and x are $k \times 1$ vectors.

5. $\dfrac{\partial x'Ax}{\partial x} = (A + A')x$,

   where A and x are a $k \times k$ matrix and a $k \times 1$ vector, respectively.

   Especially, when A is symmetric,

   $\dfrac{\partial x'Ax}{\partial x} = 2Ax$.

6. Let A and B be $k \times k$ matrices, and I_k be a $k \times k$ **identity matrix** (one in the diagonal elements and zero in the other elements).

   When $AB = I_k$, B is called the **inverse** of A, denoted by $B = A^{-1}$.

   That is, $AA^{-1} = A^{-1}A = I_k$.

7. Let A be a $k \times k$ matrix and x be a $k \times 1$ vector.

If A is a **positive definite matrix**, for any x we have:

$$x'Ax > 0.$$

If A is a **positive semidefinite matrix**, for any x we have:

$$x'Ax \geq 0.$$

If A is a **negative definite matrix**, for any x we have:

$$x'Ax < 0.$$

If A is a **negative semidefinite matrix**, for any x we have:

$$x'Ax \leq 0.$$

References

Greene, W.H., 1993, *Econometric Analysis* (Second Edition), Prentice Hall.

Greene, W.H., 1997, *Econometric Analysis* (Third Edition), Prentice-Hall.

Hogg, R.V. and Craig, A.T., 1995, *Introduction to Mathematical Statistics* (Fifth Edition), Prentice Hall.

Judge, G., Hill, C., Griffiths, W. and Lee, T., 1980, *The Theory and Practice of Econometrics*, John Wiley & Sons.

Mood, A.M., Graybill, F.A. and Boes, D.C., 1974, *Introduction to the Theory of Statistics* (Third Edition), McGraw-Hill.

Stuart, A. and Ord, J.K., 1991, *Kendall's Advanced Theory of Statistics, Vol.2* (Fifth Edition), Edward Arnold.

Stuart, A. and Ord, J.K., 1994, *Kendall's Advanced Theory of Statistics, Vol.1* (Sixth Edition), Edward Arnold.