# Econometrics I



## English Class from This Year!!

## Too bad!! (for you and me)

Econometrics I —> Statistics
Econometrics II —> Econometrics

TA session: Tue, 3rd class (13:00 – 14:30), Room #4, 4/17 –,
by Mr. Kinoshita (2nd year of the doctor course)



You can get this lecture note from:

`www2.econ.osaka-u.ac.jp/~tanizaki/class/2012`

## Some Textbooks

• 『確率統計演習 1 確率』(国沢清典編，1966，培風館)
• 『確率統計演習 2 統計』(国沢清典編，1966，培風館)
• H. Tanizaki, 2004, *Computational Methods in Statistics and Econometrics* (STATISTICS: textbooks and monographs, Vol.172), Mercel Dekker.



• R.V. Hogg, J.W. McKean and A.T. Craig, 2005, *Introduction to Mathematical Statistics* (Sixth edition), Pearson Prentice Hall.

More elementary statistics:
・Undergraduate, Tue., 3rd class, Room #5, Prof. Oya
・Graduate, Tue., 6th class, Room #1, Prof. Fukushige



# 1 Event and Probability

## 1.1 Event (事象)

We consider an **experiment (実験)** whose outcome is not known in advance but an event occurs with probability, which is sometimes called a **random experiment (無作為実験)**.
The **sample space (標本空間)** of an experiment is a set of all possible outcomes.



Each outcome of a sample space is called an **element (要素，元)** of the sample space or a **sample point (標本点)**, which represents each outcome obtained by the experiment.
An **event (事象)** is any collection of outcomes contained in the sample space, or equivalently a subset of the sample space.

An **elementary event (根元事象)** consists of exactly one element and a **compound event (複合事象)** consists of more than one element.
**Sample space** is denoted by $\Omega$ and **sample point** is given by $\omega$.

Suppose that event $A$ is a subset of sample space $\Omega$.
Let $\omega$ be a sample point in event $A$.
Then, we say that a sample point $\omega$ is contained in a sample space $A$, which is denoted by $\omega \in A$.
The event which does not belong to event $A$ is called the **complementary event (余事象)** of $A$, which is denoted by $A^c$.



The event which does not have any sample point is called the **empty event (空事象)**, denoted by $\phi$.
Conversely, the event which includes all possible sample points is called the **whole event (全事象)**, represented by $\Omega$, which is equivalent to a **sample space (標本空間)**.



Next, consider two events $A$ and $B$.
The event which belongs to either event $A$ or event $B$ is called the **sum event (和事象)**, which is denoted by $A \cup B$.
The event which belongs to both event $A$ and event $B$ is called the **product event (積事象)**, denoted by $A \cap B$.
When $A \cap B = \phi$, we say that events $A$ and $B$ are **exclusive (排反)**.



**Example 1.1:** Consider an experiment of casting a die (サイコロ).
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 $\omega_i$ 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 $A \cap A^c = \phi$, we have the fact that $A$ and $A^c$ are exclusive.



**Example 1.2:** Cast a coin three times. In this case, we have the following eight sample points:

$$\begin{aligned}
&\omega_1 = (\mathrm{H,H,H}), \quad \omega_2 = (\mathrm{H,H,T}), \quad \omega_3 = (\mathrm{H,T,H}), \\
&\omega_4 = (\mathrm{H,T,T}), \quad \omega_5 = (\mathrm{T,H,H}), \quad \omega_6 = (\mathrm{T,H,T}), \\
&\omega_7 = (\mathrm{T,T,H}), \quad \omega_8 = (\mathrm{T,T,T})
\end{aligned}$$

where H represents head (表) while T indicates tail (裏).
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 experiment can be written as:

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

Let $A$ be the event that we have two heads, $B$ be the event that we obtain at least one tail, $C$ be the 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 $A$, $B$, $C$ and $D$ are give by:

$$A = \{\omega_2, \omega_3, \omega_5\}, \qquad B = \{\omega_2, \omega_3, \omega_4, \omega_5, \omega_6, \omega_7, \omega_8\},$$

$$C = \{\omega_1, \omega_2, \omega_5, \omega_6\}, \qquad D = \{\omega_2, \omega_4, \omega_6, \omega_8\}.$$



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 $C \cup D = \{\omega_1, \omega_2, \omega_4, \omega_5, \omega_6, \omega_8\}$.



## 1.2 Probability

Let $n(A)$ be the number of sample points in $A$.
We have $n(A) \le n(B)$ when $A \subseteq B$.
Each sample point is equally likely to occur.
In the case of Example 1.1 (Section 1.1), each of the six possible outcomes has probability 1/6 and in Example 1.2 (Section 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 $P(A) \le P(B)$ because of $A \subseteq B$.



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

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

$$P(\Omega) = 1, \tag{2}$$

$$P(\phi) = 0. \tag{3}$$



$\phi \subseteq A \subseteq \Omega$ implies $n(\phi) \le n(A) \le n(\Omega)$.
Therefore, we have:

$$\frac{n(\phi)}{n(\Omega)} \le \frac{n(A)}{n(\Omega)} \le \frac{n(\Omega)}{n(\Omega)} = 1.$$

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

$$P(\phi) \le P(A) \le P(\Omega) = 1.$$

Because $\phi$ has no sample point, the number of the sample point is given by $n(\phi) = 0$ and accordingly we have $P(\phi) = 0$.
Therefore, $0 \le P(A) \le 1$ is obtained as in (1).

When events $A$ and $B$ are exclusive, i.e., when $A \cap B = \phi$, then $P(A \cup B) = P(A) + P(B)$ holds.

Moreover, since $A$ and $A^c$ are exclusive, $P(A^c) = 1 - P(A)$ is obtained.

Note that $P(A \cup A^c) = P(\Omega) = 1$ 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 (加法定理)**.

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 (条件付確率)**, 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 (乗法定理)**.



When event $A$ is **independent (独立)** 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 $D$.



# 2 Random Variable and Distribution

## 2.1 Univariate Random Variable and Distribution

The **random variable (確率変数)** $X$ is defined as the real value function on sample space $\Omega$.

Since $X$ is a function of a sample point $\omega$, 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 $\omega$, i.e., $\{\omega; X(\omega) \in I\}$, which is simply written as $\{X \in I\}$.

In Example 1.1 (Section 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 $\omega$ and takes the following values:

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



Next, suppose that $X$ is a random variable which takes 1 for odd numbers and 0 for even numbers on the die.

Then, $X$ is a function of $\omega$ and takes the following values:

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



In Example 1.2 (Section 1.1), suppose that $X$ is a random variable which takes the number of heads.

Depending on the sample point $\omega_i$, $X$ 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 (離散型確率変数)**, while another is a **continuous random variable (連続型確率変数)**.



**Discrete Random Variable (離散型確率変数) and Probability Function (確率関数):** 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 (確率関数)** 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 a set $A$, we can write a probability as the following equation:

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



Several functional forms of $f(x_i)$ are as follows.

**Discrete uniform distribution (離散型一様分布):**

$$f(x) = \begin{cases} \dfrac{1}{N}, & x = 1, 2, \cdots, N \\ 0, & \text{otherwise} \end{cases}$$

where $N = 1, 2, \cdots$.

**Bernoulli distribution (ベルヌイ分布):**

$$f(x) = \begin{cases} p^x(1-p)^{1-x}, & x = 0, 1 \\ 0, & \text{otherwise} \end{cases}$$

where $0 \leq p \leq 1$.



**Binomial distribution (二項分布):**

$$f(x) = \begin{cases} \dbinom{n}{x} p^x(1-p)^{n-x}, & x = 0, 1, 2, \cdots, n \\ 0, & \text{otherwise,} \end{cases}$$

where $0 \leq p \leq 1$ and $n = 1, 2, \cdots$.

$$(a+b)^n = \sum_{x=0}^{n} {}_nC_x a^x b^{n-x} \quad \longrightarrow \quad \text{Binomial Theorem (二項定理)}$$

$${}_nC_x = \binom{n}{x} = \frac{n!}{x!(n-x)!} \qquad n! = 1 \cdot 2 \cdots n \ \ (\text{factorial of } n)$$

$X \sim B(n, p)$



**Poisson distribution (ポアソン分布):**

$$f(x) = \begin{cases} \dfrac{e^{-\lambda}\lambda^x}{x!}, & x = 0, 1, \cdots \\ 0, & \text{otherwise,} \end{cases}$$

where $\lambda > 0$.



$<$ **Review** $>$ $e$

Note that the definition of $e$ is given by:

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

Notation

$$\exp(x) = e^x$$



$<$ **Review** $>$ Taylor series expansion about $x_0$

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2!}(x-x_0)^2 + \cdots + \frac{f^{(k)}(x_0)}{k!}(x-x_0)^k + \cdots$$

where the $k$th derivative of $f(x)$ is $f^{(k)}(x)$.

Taylor series expansion of $f(x) = e^x$ about $x = 0$

$$e^x = 1 + x + \frac{1}{2!}x^2 + \frac{1}{3!}x^3 + \cdots = \sum_{k=0}^{\infty} \frac{x^k}{k!}$$

where $f^{(k)}(x) = e^x$. Set $x = \lambda$ and $k = x$.

$$1 = \sum_{k=0}^{\infty} \frac{e^{-\lambda}\lambda^x}{x!}$$

**Geometric distribution (幾何分布):**

$$f(x) = \begin{cases} p(1-p)^x, & x = 0, 1, \cdots \\ 0, & \text{otherwise,} \end{cases}$$

where $0 < p \leq 1$.

**Negative binomial distribution (負の二項分布):**

$$f(x) = \begin{cases} \begin{pmatrix} r+x-1 \\ x \end{pmatrix} p^r (1-p)^x, & x = 0, 1, \cdots \\ 0, & \text{otherwise,} \end{cases}$$

where $0 < p \leq 1$ and $r > 0$.



Taylor series expansion of $f(x) = (1-x)^{-r}$ about $x = 0$

$$f(x) = 1 + rx + \frac{r(r+1)}{2!}x^2 + \frac{r(r+1)(r+2)}{3!}x^3 + \cdots + \begin{pmatrix} r+k-1 \\ k \end{pmatrix} x^k + \cdots$$

where $f^{(k)}(x) = \begin{pmatrix} r+k-1 \\ k \end{pmatrix} x^k$.

$$(1-x)^{-r} = \sum_{k=0}^{\infty} \begin{pmatrix} r+k-1 \\ k \end{pmatrix} x^k$$

Set $x = 1 - p$ and $k = x$

$$p^{-r} = \sum_{x=0}^{\infty} \begin{pmatrix} r+x-1 \\ x \end{pmatrix} (1-p)^x, \quad \text{i.e.,} \quad 1 = \sum_{x=0}^{\infty} \begin{pmatrix} r+x-1 \\ x \end{pmatrix} p^r (1-p)^x$$



**Hypergeometric distribution (超幾何分布):**

$$f(x) = \begin{cases} \dfrac{\begin{pmatrix} K \\ x \end{pmatrix}\begin{pmatrix} M-K \\ n-x \end{pmatrix}}{\begin{pmatrix} M \\ n \end{pmatrix}}, & x = 0, 1, \cdots, n \\ 0, & \text{otherwise,} \end{cases}$$

where $M = 1, 2, \cdots$, $K = 0, 1, \cdots, M$, and $n = 1, 2, \cdots, M$.



In Example 1.2 (Section 1.1), all the possible values of $X$ are 0, 1, 2 and 3. (note that $X$ denotes the number of heads when a coin is cast three times).

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:

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

which can be written as:



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

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 (二項分布)**.

Thus, it is easy to check $f(x) \geq 0$ and $\sum_x f(x) = 1$ in Example 1.2.

**Continuous Random Variable (連続型確率変数) and Probability Density Function (確率密度関数):** 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 $a < b$.
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 (確率密度関数)** of $X$, or simply the **density function (密度関数)** 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 as follows:

**Uniform distribution (一様分布):**

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & a < x < b \\ 0, & \text{otherwise,} \end{cases}$$

where $-\infty < a < b < \infty$.
$X \sim U(a, b)$



**Normmal distribution (正規分布):**

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{1}{2\sigma^2}(x-\mu)^2\Big), \qquad -\infty < x < \infty$$

where $-\infty < \mu < \infty$ and $\sigma > 0$.
$X \sim N(\mu, \sigma^2)$
$N(0, 1)$ = Standard normal distribution



**Exponential distribution (指数分布):**

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0, & \text{otherwise,} \end{cases}$$

where $\lambda > 0$

**Gamma distribution (ガンマ分布):**

$$f(x) = \begin{cases} \dfrac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x}, & x > 0 \\ 0, & \text{otherwise,} \end{cases}$$

where $\lambda > 0$ and $r > 0$.

Gamma dist. with $r = 1 \iff$ Exponential dist.

Gamma function: $\Gamma(a) = \displaystyle\int_0^\infty x^{a-1} e^{-x} \mathrm{d}x, \quad a > 0$

$\Gamma(a+1) = a\Gamma(a)$ —> Use integration by parts (部分積分)

$\Gamma(n+1) = n!$ for integer $n$

$\Gamma(n+\frac{1}{2}) = \dfrac{1 \cdot 3 \cdot 5 \cdots (2n-1)}{2^n} \sqrt{\pi}, \quad \Gamma(\frac{1}{2}) = 2\Gamma(\frac{3}{2}) = \sqrt{\pi}$



**Beta distribution (ベータ分布):**

$$f(x) = \begin{cases} \dfrac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}, & 0 < x < 1 \\ 0, & \text{otherwise,} \end{cases}$$

where $a > 0$ and $b > 0$.

Beta function:

$$\begin{aligned} B(a,b) &= \int_0^1 x^{a-1}(1-x)^{b-1} \mathrm{d}x \\ &= \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \\ B(a,b) &= B(b,a) \end{aligned}$$



**Cauchy distribution (コーシー分布):**

$$f(x) = \frac{1}{\pi\beta(1 + (x-\alpha)^2/\beta^2)}, \qquad -\infty < x < \infty$$

where $-\infty < \alpha < \infty$ and $\beta > 0$.

**Log-normal distribution (対数正規分布):**

$$f(x) = \begin{cases} \dfrac{1}{x\sqrt{2\pi\sigma^2}} \exp\Big(-\dfrac{1}{2\sigma^2}(\ln x - \mu)^2\Big), & 0 < x < \infty \\ 0, & \text{otherwise,} \end{cases}$$

where $-\infty < \mu < \infty$ and $\sigma > 0$.



**Double exponentioal distribution (二重指数分布), or Laplace distribution (ラプラス分布):**

$$f(x) = \frac{1}{2\beta} \exp\Big(-\frac{|x-\alpha|}{\beta}\Big), \qquad -\infty < x < \infty$$

where $-\infty < \alpha < \infty$ and $\beta > 0$.

**Weibull distribution (ワイブル分布):**

$$f(x) = \begin{cases} abx^{b-1} \exp(-ax^b), & x > 0 \\ 0, & \text{otherwise,} \end{cases}$$

where $a > 0$ and $b > 0$.



**Logistic distribution (ロジスティック分布):**

$$F(x) = (1 + e^{-(x-\alpha)/\beta})^{-1}, \qquad -\infty < x < \infty$$

where $-\infty < \alpha < \infty$ and $\beta > 0$.

**Pareto distribution (パレート分布):**

$$f(x) = \begin{cases} \dfrac{\theta x_0^\theta}{x^{\theta+1}}, & x > x_0 \\ 0, & \text{otherwise,} \end{cases}$$

where $x_0 > 0$ and $\theta > 0$.



**Gumbel distribution (ガンベル分布), or Extreme value distribution (極値分布):**

$$F(x) = \exp(-e^{-(x-\alpha)/\beta}), \qquad -\infty < x < \infty$$

where $-\infty < \alpha < \infty$ and $\beta > 0$.

***t* distribution (*t* 分布):**

$$f(x) = \frac{\Gamma(\frac{k+1}{2})}{\Gamma(\frac{k}{2})} \frac{1}{\sqrt{k\pi}} \Big(1 + \frac{x^2}{k}\Big)^{-(k+1)/2}, \qquad -\infty < x < \infty$$

where $k > 0$.

$X \sim t(k)$ —> $t$ dist. with $k$ degrees of freedom (自由度)

$t(1) \iff$ Cauchy dist.

$t(\infty) \iff N(0,1)$



***F* distribution (*F* 分布):**

$$f(x) = \begin{cases} \dfrac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \Big(\dfrac{m}{n}\Big)^{m/2} \dfrac{x^{m/2-1}}{(1 + \frac{m}{n}x)^{(m+n)/2}}, & x > 0 \\ 0, & \text{otherwise,} \end{cases}$$

where $m, n = 1, 2, \cdots$.

$X \sim F(m, n)$ —> $F$ dist. with $(m, n)$ degrees of freedom



**Chi-square distribution (カイ二乗分布):**

$$f(x) = \begin{cases} \dfrac{1}{\Gamma(\frac{k}{2}) 2^{k/2}} x^{k/2-1} e^{-x/2}, & x > 0 \\ 0, & \text{otherwise,} \end{cases}$$

where $k = 1, 2, \cdots$.

$X \sim \chi^2(k)$ —> $\chi^2$ dist. with $k$ degrees of freedom



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**.



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

First of all, we define $I$ as $I = \int_{-\infty}^{\infty} f(x)\, \mathrm{d}x$.

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

$$
\begin{aligned}
I^2 &= \Big(\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x\Big)^2 = \Big(\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x\Big)\Big(\int_{-\infty}^{\infty} f(y)\,\mathrm{d}y\Big) \\
&= \Big(\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}x^2)\,\mathrm{d}x\Big)\Big(\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}y^2)\,\mathrm{d}y\Big) \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\Big(-\frac{1}{2}(x^2+y^2)\Big)\,\mathrm{d}x\,\mathrm{d}y \\
&= \frac{1}{2\pi}\int_{0}^{2\pi}\int_{0}^{\infty} \exp(-\frac{1}{2}r^2)r\,\mathrm{d}r\,\mathrm{d}\theta \\
&= \frac{1}{2\pi}\int_{0}^{2\pi}\int_{0}^{\infty} \exp(-s)\,\mathrm{d}s\,\mathrm{d}\theta = \frac{1}{2\pi}2\pi[-\exp(-s)]_0^{\infty} = 1.
\end{aligned}
$$



## < Review > Integration by Substitution (置換積分):

**Univariate (1 変数) 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 (2 変数) 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 (ヤコビアン)**, which represents the following determinant (行列式):

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

**< End of Review >**



**< Go back to the Integration >**

In the fifth equality, integration by substitution (置換積分) is used.

The polar coordinate transformation (極座標変換) is used as $x = r\cos\theta$ and $y = r\sin\theta$.

Note that $0 \le r < +\infty$ and $0 \le \theta < 2\pi$.

The Jacobian is given by:

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



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

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.

Actually, this density function is called the **standard normal probability density function (標準正規分布)**.

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

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

which represents the probability less than $x$.



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

$$F(x_1) \leq F(x_2), \quad \text{for } x_1 < x_2, \quad \text{—>} \ \text{nondecreasing function}$$

$$P(a < X \leq b) = F(b) - F(a), \quad \text{for } a < b,$$

$$F(-\infty) = 0, \quad F(+\infty) = 1.$$

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



1. Discrete random variable (Figure 1):
   - $F(x) = \sum_{i=1}^{r} f(x_i) = \sum_{i=1}^{r} p_i$,
   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 $\epsilon$ is a small positive number less than $x_i - x_{i-1}$.

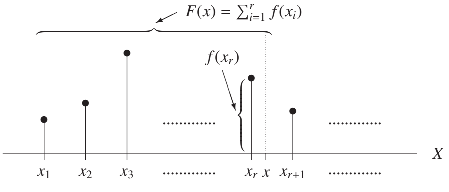


2. Continuous random variable (Figure 2):
   - $F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t$,
   - $F'(x) = f(x)$.

$f(x)$ and $F(x)$ are displayed in Figure 1 for a discrete random variable and Figure 2 for a continuous random variable.



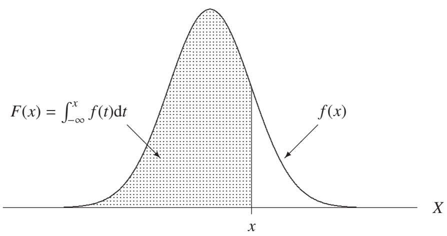
Figure 1: Probability Function $f(x)$ and Distribution Function $F(x)$— Discrete Case




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



Figure 2: Density Function $f(x)$ and Distribution Function $F(x)$ — Continuous Case



## 2.2 Multivariate Random Variable (多変量確率変数) 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 $\{\omega; X(\omega) = x_i \text{ and } Y(\omega) = y_j\}$ 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 (結合確率関数)** 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:

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

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 (周辺確率関数)** 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 $\sum_j f_y(y_j) = 1$.



**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 (結合確率密度関数)** 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:

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

where $f_x(x)$ and $f_y(y)$ are called the **marginal probability density functions (周辺確率密度関数)** of $X$ and $Y$ or the **marginal density functions (周辺密度関数)** of $X$ and $Y$.



For example, consider the event $\{\omega; a < X(\omega) < b,\ c < Y(\omega) < d\}$, which is a 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 RVs is also possible. For example, let $X$ be a discrete RV and $Y$ be a continuous RV. $X$ 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:

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



The marginal probability function of $X$ is given by:

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

for $i = 1, 2, \cdots$. The marginal probability density function of $Y$ is:

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



## 2.3 Conditional Distribution

**Discrete Random Variable:** The **conditional probability function (条件付確率関数)** 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.



The features of the conditional probability function $f_{x|y}(x_i|y_j)$ are:

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



**Continuous Random Variable:** The **conditional probability density function (条件付確率密度関数)** of $X$ given $Y = y$ (or the **conditional density function (条件付密度関数)** 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:

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

**Independence of Random Variables:** For discrete random variables $X$ and $Y$, we say that $X$ is **independent (独立)** (or **stochastically independent (確率的に独立)**) 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 $Y$.



# 3 Mathematical Expectation

## 3.1 Univariate Random Variable

**Definition of Mathematical Expectation (数学的期待値):** 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:



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

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 (平均)** of random variable $X$.

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

   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 (分散)** of random variable $X$, which is denoted by $\mathrm{V}(X)$.

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



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

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

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

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



**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,

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

Similarly, note that we have $\int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x = \mathrm{E}(X)$ 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:**

E($X$) and 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 (正規化) or standardization (標準化).



**Example 1.5:** In Example 1.2 of flipping a coin three times (Section 1.1), we see in Section 2.1 that the probability function is written as the following binomial distribution:

$$P(X = x) = f(x) = \frac{n!}{x!(n-x)!}p^x(1-p)^{n-x},$$

$$\text{for } x = 0, 1, 2, \cdots, n,$$

where $n = 3$ and $p = 1/2$.

When $X$ has the binomial distribution above, we obtain E($X$), V($X$) and $\phi(\theta)$ as follows.



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

$$\begin{aligned}
\mu = \mathrm{E}(X) &= \sum_{x=0}^{n} xf(x) = \sum_{x=1}^{n} xf(x) = \sum_{x=1}^{n} x\frac{n!}{x!(n-x)!}p^x(1-p)^{n-x} \\
&= \sum_{x=1}^{n} \frac{n!}{(x-1)!(n-x)!}p^x(1-p)^{n-x} \\
&= np\sum_{x=1}^{n} \frac{(n-1)!}{(x-1)!(n-x)!}p^{x-1}(1-p)^{n-x} \\
&= np\sum_{x'=0}^{n'} \frac{n'!}{x'!(n'-x')!}p^{x'}(1-p)^{n'-x'} = np,
\end{aligned}$$

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



Second, in order to obtain $\sigma^2 = \mathrm{V}(X)$, we rewrite V($X$) as:

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

E($X(X-1)$) is given by:

$$\begin{aligned}
\mathrm{E}(X(X-1)) &= \sum_{x=0}^{n} x(x-1)f(x) = \sum_{x=2}^{n} x(x-1)f(x) \\
&= \sum_{x=2}^{n} x(x-1)\frac{n!}{x!(n-x)!}p^x(1-p)^{n-x} \\
&= \sum_{x=2}^{n} \frac{n!}{(x-2)!(n-x)!}p^x(1-p)^{n-x}
\end{aligned}$$



$$\begin{aligned}
&= n(n-1)p^2\sum_{x=2}^{n} \frac{(n-2)!}{(x-2)!(n-x)!}p^{x-2}(1-p)^{n-x} \\
&= n(n-1)p^2\sum_{x'=0}^{n'} \frac{n'!}{x'!(n'-x')!}p^{x'}(1-p)^{n'-x'} \\
&= n(n-1)p^2,
\end{aligned}$$

where $n' = n-2$ and $x' = x-2$ are re-defined.



Therefore, $\sigma^2 = \mathrm{V}(X)$ is obtained as:

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

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

$$\begin{aligned}\phi(\theta) = \mathrm{E}(e^{\theta X}) &= \sum_{x=0}^{n} e^{\theta x} \frac{n!}{x!(n-x)!} p^x (1-p)^{n-p} \\ &= \sum_{x=0}^{n} \frac{n!}{x!(n-x)!} (pe^{\theta})^x (1-p)^{n-p} = (pe^{\theta} + 1 - p)^n.\end{aligned}$$

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 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:

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



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

$$\begin{aligned}\phi(\theta) = \mathrm{E}(e^{\theta X}) &= \int_{-\infty}^{\infty} e^{\theta x} f(x)\,\mathrm{d}x = \int_0^1 e^{\theta x}\,\mathrm{d}x \\ &= [\frac{1}{\theta}e^{\theta x}]_0^1 = \frac{1}{\theta}(e^{\theta} - 1).\end{aligned}$$



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

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

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:

$$\begin{aligned}\mathrm{E}(X) &= \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x e^{-\frac{1}{2}x^2}\,\mathrm{d}x \\ &= \frac{1}{\sqrt{2\pi}} \Big[-e^{-\frac{1}{2}x^2}\Big]_{-\infty}^{\infty} = 0,\end{aligned}$$

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



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

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

The first equality holds because of $\mathrm{E}(X) = 0$.

In the fifth equality, use the following integration formula, called the **integration by parts**:

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

where we take $h(x) = x$ and $g(x) = -e^{-\frac{1}{2}x^2}$ in this case.

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.



$\phi(\theta)$ 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 $\theta$ and variance one.



**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:

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



**Example 1.8(b):** When $X \sim N(\mu, \sigma^2)$, what is the moment-generating function of $X$?

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



## 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:

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

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:

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

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 $X$.

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

The variance of $Y$ is also obtained in the same way, 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 (共分散)** of $X$ and $Y$, which is denoted by $\mathrm{Cov}(X,Y)$ and written as:



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

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:

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

In Section 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 given by:

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



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

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

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

**Proof:**

For discrete random variables $X$ and $Y$,

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

If $X$ is independent of $Y$, the second equality holds, i.e., $f_{xy}(x_i, y_j) = f_x(x_i) f_y(y_j)$.



For continuous random variables $X$ and $Y$,

$$\begin{aligned}\mathrm{E}(XY) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy f_{xy}(x, y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy f_x(x) f_y(y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \Big(\int_{-\infty}^{\infty} x f_x(x)\,\mathrm{d}x\Big)\Big(\int_{-\infty}^{\infty} y f_y(y)\,\mathrm{d}y\Big) = \mathrm{E}(X)\mathrm{E}(Y).\end{aligned}$$

When $X$ is independent of $Y$, we have $f_{xy}(x, y) = f_x(x) f_y(y)$ in the second equality.



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:

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



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

In the fourth equality, the theorem in Section 3.1 is used, i.e., $\mathrm{E}(\mu_x Y) = \mu_x \mathrm{E}(Y)$ and $\mathrm{E}(\mu_y X) = \mu_y \mathrm{E}(X)$.



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 (相関係数)** between $X$ and $Y$, denoted by $\rho_{xy}$, is defined as:

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

$\rho_{xy} > 0 \Longrightarrow$ **positive correlation** between $X$ and $Y$

$\rho_{xy} \longrightarrow 1 \Longrightarrow$ **strong positive correlation**

$\rho_{xy} < 0 \Longrightarrow$ **negative correlation** between $X$ and $Y$

$\rho_{xy} \longrightarrow -1 \Longrightarrow$ **strong negative correlation**

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

**Proof:**

When $X$ is independent of $Y$, we have $\mathrm{Cov}(X, Y) = 0$.

We 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 follows:

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



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



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

**Proof:**

Consider the following function of $t$: $f(t) = \mathrm{V}(Xt - Y)$, which is always greater than or equal to zero because of the definition of variance. Therefore, for all $t$, we have $f(t) \ge 0$. $f(t)$ 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 $f(t) \ge 0$ for all $t$, we need the following condition:

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

because the first term in the last equality is nonnegative, which implies:

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



Therefore, we have:

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

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: $-1 \le \rho_{xy} \le 1$.

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 $X_1, X_2, \cdots, X_n$,

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



$$\mathrm{V}(\sum_i a_i X_i) = \sum_i \sum_j a_i a_j \mathrm{Cov}(X_i, X_j),$$

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.

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



The first and second equalities come from the previous theorems on mean.



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 all $i \neq j$ 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 $\mu$ and variance $\sigma^2$. 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, mean and variance of $\overline{X}$ are given by:

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

**Proof:**

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

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

$\mathrm{E}(aX) = a\mathrm{E}(X)$ 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.



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

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

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.



# 4 Transformation of Variables (変数変換)

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 $X$.



## 4.1 Univariate Case

**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 (一対一) transformation. Then, the probability density function of $Y$, i.e., $f_y(y)$, is given by:

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

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 $F_x(x) = P(X \le x)$ and $f_x(x) = F'_x(x)$.



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:

$$\begin{aligned}F_y(y) &= P(Y \le y) = P\big(\psi(Y) \le \psi(y)\big) \\ &= P\big(X \le \psi(y)\big) = F_x\big(\psi(y)\big).\end{aligned}$$

The first equality is the definition of the cumulative distribution function. The second equality holds because of $\psi'(Y) > 0$. Therefore, differentiating $F_y(y)$ with



respect to $y$, we can obtain the following expression:

$$f_y(y) = F'_y(y) = \psi'(y)F'_x\big(\psi(y)\big) = \psi'(y)f_x\big(\psi(y)\big). \tag{4}$$

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

$$\begin{aligned} F_y(y) &= P(Y \le y) = P\Big(\psi(Y) \ge \psi(y)\Big) = P\Big(X \ge \psi(y)\Big) \\ &= 1 - P\Big(X < \psi(y)\Big) = 1 - F_x\Big(\psi(y)\Big). \end{aligned}$$

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

$$f_y(y) = F_y'(y) = -\psi'(y)F_x'\Big(\psi(y)\Big) = -\psi'(y)f_x\Big(\psi(y)\Big). \qquad (5)$$



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

Thus, summarizing the above two cases, i.e., $\psi'(X) > 0$ and $\psi'(X) < 0$, equations (4) and (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 $X \sim N(0,1)$, we derive the probability density function of $Y = \mu + \sigma X$.

Since we have:

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

$\psi'(y) = 1/\sigma$ is obtained. Therefore, $f_y(y)$ 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 $\mu$ and variance $\sigma^2$, 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:

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

The probability density function of $Y$ is obtained as follows:

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



**Example: $\chi^2(1)$ Distribution:** Define $Y = X^2$, where $X \sim N(0,1)$. Then, $Y \sim \chi^2(1)$.

**proof:**

$$\begin{aligned} f_y(y) &= \frac{1}{2\sqrt{y}}\Big(f_x(\sqrt{y}) + f_x(-\sqrt{y})\Big) \\ &= \frac{1}{\sqrt{2\pi}} y^{-1/2} \exp\Big(-\frac{1}{2}y\Big) \end{aligned}$$

which is $\chi^2(1)$ distribution, where

$$f_x(x) = \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{1}{2}x^2\Big)$$

$$f_x(\sqrt{y}) = f_x(-\sqrt{y}) = \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{1}{2}y\Big)$$



Note that the $\chi^2(n)$ distribution is:

$$f_x(x) = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} \exp(-\frac{x}{2}), \qquad x > 0,$$

where $\Gamma(\frac{n}{2}) = \sqrt{\pi}$

## 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:

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



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 a 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:

$$\begin{aligned} &f_y(y_1, y_2, \cdots, y_n) \\ &\quad = |J| f_x\Big(\psi_1(y_1, \cdots, y_n), \psi_2(y_1, \cdots, y_n), \cdots, \psi_n(y_1, \cdots, y_n)\Big), \end{aligned}$$



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

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



**Example: Normal Distribution:** $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$. $X$ is independent of $Y$.

Then, $X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$

**Proof:**

The density functions of $X$ and $Y$ are:

$$f_x(x) = \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{1}{2\sigma_1^2}(x - \mu_1)^2\right)$$

$$f_y(y) = \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left(-\frac{1}{2\sigma_2^2}(y - \mu_2)^2\right)$$

The joint density of $X$ and $Y$ is:

$$
\begin{aligned}
f_{xy}(x,y) &= f_x(x)f_y(y) \\
&= \frac{1}{2\pi\sigma_1\sigma_2}\exp\left(-\frac{1}{2\sigma_1^2}(x-\mu_1)^2 - \frac{1}{2\sigma_2^2}(y-\mu_2)^2\right)
\end{aligned}
$$

Define $U = X + Y$ and $V = Y$. We obtain the joint distribution of $U$ and $V$.

Using $X = U - V$ and $Y = V$, the Jacobian is:

$$
J = \begin{vmatrix} \dfrac{\partial x}{\partial u} & \dfrac{\partial x}{\partial v} \\ \dfrac{\partial y}{\partial u} & \dfrac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} 1 & -1 \\ 0 & 1 \end{vmatrix} = 1
$$



The joint density function of $U$ and $V$, $f_{uv}(u,v)$, is given by:

$$
\begin{aligned}
f_{uv}(u,v) &= |J| f_{xy}(u-v, v) \\
&= \frac{1}{2\pi\sigma_1\sigma_2}\exp\left(-\frac{1}{2\sigma_1^2}(u-v-\mu_1)^2 - \frac{1}{2\sigma_2^2}(v-\mu_2)^2\right)
\end{aligned}
$$

The marginal density function of $U$ is:

$$
\begin{aligned}
f_u(u) &= \int f_{uv}(u,v)\mathrm{d}v \\
&= \int \frac{1}{2\pi\sigma_1\sigma_2}\exp\Big(-\frac{1}{2\sigma_1^2}(u-v-\mu_1)^2 - \frac{1}{2\sigma_2^2}(v-\mu_2)^2\Big)\mathrm{d}v
\end{aligned}
$$



$$
\begin{aligned}
&= \int \frac{1}{2\pi\sigma_1\sigma_2}\exp\bigg(-\frac{1}{2\sigma_1^2}((v-\mu_2)-(u-\mu_1-\mu_2))^2 \\
&\qquad\qquad -\frac{1}{2\sigma_2^2}(v-\mu_2)^2\bigg)\mathrm{d}v \\
&= \int \frac{1}{2\pi\sigma_1\sigma_2}\exp\bigg(-\frac{1}{2/(1/\sigma_1^2+1/\sigma_2^2)}((v-\mu_2) \\
&\qquad -\frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2}(u-\mu_1-\mu_2))^2 - \frac{1}{2(\sigma_1^2+\sigma_2^2)}(u-\mu_1-\mu_2)^2\bigg)\mathrm{d}v
\end{aligned}
$$



$$
\begin{aligned}
&= \int \frac{1}{\sqrt{2\pi/(1/\sigma_1^2+1/\sigma_2^2)}} \\
&\times\exp\bigg(-\frac{1}{2/(1/\sigma_1^2+1/\sigma_2^2)}((v-\mu_2)-\frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2}(u-\mu_1-\mu_2))^2\bigg) \\
&\quad\times\frac{1}{\sqrt{2\pi(\sigma_1^2+\sigma_2^2)}}\exp\bigg(-\frac{1}{2(\sigma_1^2+\sigma_2^2)}(u-\mu_1-\mu_2)^2\bigg)\mathrm{d}v
\end{aligned}
$$



$$
\begin{aligned}
&= \int \frac{1}{\sqrt{2\pi/(1/\sigma_1^2+1/\sigma_2^2)}} \\
&\quad\times\exp\bigg(-\frac{1}{2/(1/\sigma_1^2+1/\sigma_2^2)}((v-\mu_2) \\
&\qquad -\frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2}(u-\mu_1-\mu_2))^2\bigg)\mathrm{d}v \\
&\times\frac{1}{\sqrt{2\pi(\sigma_1^2+\sigma_2^2)}}\exp\bigg(-\frac{1}{2(\sigma_1^2+\sigma_2^2)}(u-\mu_1-\mu_2)^2\bigg) \\
&= \frac{1}{\sqrt{2\pi(\sigma_1^2+\sigma_2^2)}}\exp\bigg(-\frac{1}{2(\sigma_1^2+\sigma_2^2)}(u-\mu_1-\mu_2)^2\bigg)
\end{aligned}
$$



**Example: $\chi^2$ Distribution:** $X \sim \chi^2(n)$ and $Y \sim \chi^2(m)$. $X$ is independent of $Y$.

Then, $X + Y \sim \chi^2(n+m)$.

**Proof:**

The density functions of $X$ and $Y$ are:

$$
\begin{aligned}
f_x(x) &= \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}x^{\frac{n}{2}-1}\exp(-\frac{x}{2}), \quad x > 0 \\
f_y(y) &= \frac{1}{2^{\frac{m}{2}}\Gamma(\frac{m}{2})}y^{\frac{m}{2}-1}\exp(-\frac{y}{2}), \quad y > 0
\end{aligned}
$$

The joint density function of $X$ and $Y$ is:

$$
f_{xy}(x,y) = f_x(x)f_y(y)
$$

$$
\begin{aligned}
&= \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} \exp(-\frac{x}{2}) \frac{1}{2^{\frac{m}{2}}\Gamma(\frac{m}{2})} y^{\frac{m}{2}-1} \exp(-\frac{y}{2}) \\
&= C x^{\frac{n}{2}-1} y^{\frac{m}{2}-1} \exp(-\frac{x+y}{2})
\end{aligned}
$$

where $C = \dfrac{1}{2^{\frac{n+m}{2}}\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})}$.

From $X = U - V$ and $Y = V$, the Jacobian is:

$$
J = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} 1 & -1 \\ 0 & 1 \end{vmatrix} = 1
$$



The joint density function of $U$ and $V$, $f_{uv}(u, v)$, is given by:

$$
\begin{aligned}
f_{uv}(u, v) &= |J| f_{xy}(u - v, v) \\
&= C(u - v)^{\frac{n}{2}-1} v^{\frac{m}{2}-1} \exp(-\frac{u}{2})
\end{aligned}
$$

The marginal density function of $U$ is:

$$
\begin{aligned}
f_u(u) &= \int f_{uv}(u, v) \mathrm{d}v \\
&= C \exp(-\frac{u}{2}) \int_0^\infty (u - v)^{\frac{n}{2}-1} v^{\frac{m}{2}-1} \mathrm{d}v \\
&= C \exp(-\frac{u}{2}) \int_0^\infty (u - uw)^{\frac{n}{2}-1} (uw)^{\frac{m}{2}-1} u \mathrm{d}w
\end{aligned}
$$



$$
\begin{aligned}
&= C u^{\frac{n+m}{2}-1} \exp(-\frac{u}{2}) \int_0^1 (1 - w)^{\frac{n}{2}-1} w^{\frac{m}{2}-1} \mathrm{d}w \\
&= C B(\frac{n}{2}, \frac{m}{2}) u^{\frac{n+m}{2}-1} \exp(-\frac{u}{2}) \\
&= \frac{1}{2^{\frac{n+m}{2}}\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})} \frac{\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})}{\Gamma(\frac{n+m}{2})} u^{\frac{n+m}{2}-1} \exp(-\frac{u}{2}) \\
&= \frac{1}{2^{\frac{n+m}{2}}\Gamma(\frac{n+m}{2})} u^{\frac{n+m}{2}-1} \exp(-\frac{u}{2})
\end{aligned}
$$

Beta function $B(n, m)$ is:

$$
B(n, m) = \int_0^\infty (1 - x)^{n-1} x^{m-1} \mathrm{d}x = \frac{\Gamma(n)\Gamma(m)}{\Gamma(n + m)}
$$



**Example: *t* Distribution:** $X \sim N(0, 1)$ and $Y \sim \chi^2(n)$. $X$ is independent of $Y$.

Then, $U = \dfrac{X}{\sqrt{Y/n}} \sim t(n)$

Note that the density function of $t(n)$ is:

$$
f_u(x) = \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \frac{1}{\sqrt{n\pi}} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}
$$



**Proof:** The density functions of $X$ and $Y$ are:

$$
\begin{aligned}
f_x(x) &= \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}x^2), \qquad -\infty < x < \infty \\
f_y(y) &= \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} y^{\frac{n}{2}-1} \exp(-\frac{y}{2}), \qquad y > 0
\end{aligned}
$$

The joint density functions of $X$ and $Y$, $f_{xy}(x, y)$, is:

$$
\begin{aligned}
f_{xy}(x, y) &= f_x(x) f_y(y) \\
&= \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}x^2) \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} y^{\frac{n}{2}-1} \exp(-\frac{y}{2}) \\
&= \frac{1}{\sqrt{2\pi}} \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} y^{\frac{n}{2}-1} \exp(-\frac{y}{2} - \frac{1}{2}x^2)
\end{aligned}
$$



From $X = U\sqrt{\dfrac{V}{n}}$ and $Y = V$, the Jacobian is:

$$
J = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} \sqrt{\frac{v}{n}} & \frac{u}{2\sqrt{nv}} \\ 0 & 1 \end{vmatrix} = \sqrt{\frac{v}{n}}
$$

The joint density function of $U$ and $V$, $f_{uv}(u, v)$, is:

$$
\begin{aligned}
f_{uv}(u, v) &= |J| f_{xy}(u\sqrt{\frac{v}{n}}, v) \\
&= \sqrt{\frac{v}{n}} \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}\frac{u^2 v}{n}) \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} v^{\frac{n}{2}-1} \exp(-\frac{v}{2}) \\
&= C v^{\frac{n-1}{2}} \exp\left(-\frac{v}{2}(1 + \frac{u^2}{n})\right)
\end{aligned}
$$

where $C = \dfrac{1}{\sqrt{n}} \dfrac{1}{\sqrt{\pi}} \dfrac{1}{2^{\frac{n+1}{2}}\Gamma(\frac{n}{2})}$.

The marginal density function of $U$ is:

$$
\begin{aligned}
f_u(u) &= \int f_{uv}(u, v)\mathrm{d}v \\
&= C \int v^{\frac{n-1}{2}} \exp\Big( -\frac{v}{2}(1 + \frac{u^2}{n})\Big)\mathrm{d}v \\
&= C \int \Big( w(1 + \frac{u^2}{n})^{-1}\Big)^{\frac{n-1}{2}} \exp(-\frac{1}{2}w)\Big(1 + \frac{u^2}{n}\Big)^{-1} \mathrm{d}w \\
&= C\Big(1 + \frac{u^2}{n}\Big)^{-\frac{n+1}{2}} \int w^{\frac{n+1}{2}-1} \exp(-\frac{1}{2}w)\mathrm{d}w
\end{aligned}
$$



$$
\begin{aligned}
&= C\Big(1 + \frac{u^2}{n}\Big)^{-\frac{n+1}{2}} 2^{\frac{n+1}{2}}\Gamma(\frac{n+1}{2}) \\
&\times \int \frac{1}{2^{\frac{n+1}{2}}\Gamma(\frac{n+1}{2})} w^{\frac{n+1}{2}-1} \exp(-\frac{1}{2}w)\mathrm{d}w \\
&= C\Big(1 + \frac{u^2}{n}\Big)^{-\frac{n+1}{2}} 2^{\frac{n+1}{2}}\Gamma(\frac{n+1}{2}) \\
&= \frac{1}{\sqrt{\pi}} \frac{1}{2^{\frac{n+1}{2}}\Gamma(\frac{n}{2})} \frac{1}{\sqrt{n}} 2^{\frac{n+1}{2}}\Gamma(\frac{n+1}{2})\Big(1 + \frac{u^2}{n}\Big)^{-\frac{n+1}{2}} \\
&= \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \frac{1}{\sqrt{n\pi}}\Big(1 + \frac{u^2}{n}\Big)^{-\frac{n+1}{2}}
\end{aligned}
$$



Use integration by substitution by $w = v(1 + \dfrac{u^2}{n})$.
Note that $f(w) = \dfrac{1}{2^{\frac{n+1}{2}}\Gamma(\frac{n+1}{2})} w^{\frac{n+1}{2}-1} \exp(-\dfrac{1}{2}w)$ is the density function of $\chi^2(n + 1)$.



**Example: Cauchy Distribution:** $X \sim N(0, 1)$ and $Y \sim N(0, 1)$. $X$ is independent of $Y$.
Then, $U = \dfrac{X}{Y}$ is Cauchy.

Note that the density function of $U$, $f_u(u)$, is:

$$
f(u) = \frac{1}{\pi(1 + u^2)}
$$



**Proof:** The density functions of $X$ and $Y$ are:

$$
\begin{aligned}
f_x(x) &= \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}x^2), \qquad -\infty < x < \infty \\
f_y(y) &= \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}y^2), \qquad -\infty < y < \infty
\end{aligned}
$$

The joint density function of $X$ and $Y$ is:

$$
\begin{aligned}
f_{xy}(x, y) &= f_x(x)f_y(y) \\
&= \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}x^2) \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}y^2) \\
&= \frac{1}{2\pi} \exp(-\frac{1}{2}(x^2 + y^2))
\end{aligned}
$$



Transformation of variables by $u = \dfrac{x}{y}$ and $v = y$.
From $x = uv$ and $y = v$, the Jacobian is:

$$
J = \begin{vmatrix} \dfrac{\partial x}{\partial u} & \dfrac{\partial x}{\partial v} \\ \dfrac{\partial y}{\partial u} & \dfrac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} v & u \\ 0 & 1 \end{vmatrix} = v
$$

The joint density function of $U$ and $V$, $f_{uv}(u, v)$, is:

$$
\begin{aligned}
f_{uv}(u, v) &= |J| f_{xy}(uv, v) \\
&= |v| \frac{1}{2\pi} \exp(-\frac{1}{2}v^2(1 + u^2))
\end{aligned}
$$

The marginal density function of $U$ is:

$$\begin{aligned}
f_u(u) &= \int f_{uv}(u, v)\mathrm{d}v \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty} |v| \exp(-\frac{1}{2}v^2(1 + u^2))\mathrm{d}v \\
&= \frac{1}{\pi}\int_{0}^{\infty} v \exp(-\frac{1}{2}v^2(1 + u^2))\mathrm{d}v \\
&= \frac{1}{\pi}\Big[-\frac{1}{1+u^2}\exp(-\frac{1}{2}v^2(1 + u^2))\Big]_{v=0}^{\infty} \\
&= \frac{1}{\pi(1+u^2)}
\end{aligned}$$



# 5 Moment-Generating Function (積率母関数)

## 5.1 Univariate Case

As discussed in Section 3.1, the moment-generating function is defined as $\phi(\theta) = \mathrm{E}(e^{\theta X})$.

For a random variable $X$, $\mu'_n \equiv \mathrm{E}(X^n)$ is called the ***n*th moment (*n* 次の積率)** of $X$.



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:** The moment-generating function is a weighted sum of all the moments.

$$\begin{aligned}
\phi(\theta) = \mathrm{E}(e^{\theta X}) &= \mathrm{E}\Big(\sum_{n=0}^{\infty}\frac{1}{n!}f^{(n)}(0)\theta^n\Big) \\
&= \mathrm{E}\Big(\sum_{n=0}^{\infty}\frac{1}{n!}X^n\theta^n\Big) = \sum_{n=0}^{\infty}\frac{1}{n!}\mathrm{E}(X^n)\theta^n
\end{aligned}$$

   where $f(\theta) = e^{\theta X}$. $\quad f(\theta) = \sum_{n=0}^{\infty}\frac{1}{n!}f^{(n)}(0)\theta^n$

   Note that $f^{(n)}(\theta) = X^n e^{\theta X}$.



3. **Remark:** $\phi(\theta)$ does not exist, if $\mathrm{E}(X^n)$ for some $n$ does not exist.

   $\phi(\theta)$ is finite. $\iff$ All the moments exist.

4. **Remark:** Let $X$ and $Y$ be two random variables. Suppose that both moment-generating functions exist. When the moment-generating function of $X$ is equivalent to that of $Y$, we have the fact that $X$ has the same distribution as $Y$.

$$\begin{aligned}\phi_x(\theta) = \phi_y(\theta) \qquad &\Longleftrightarrow \qquad \mathrm{E}(X^n) = \mathrm{E}(Y^n) \text{ for all } n \\ &\Longleftrightarrow \qquad f_x(t) = f_y(t)\end{aligned}$$



5. **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:

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

$\phi(\theta)$ represents the moment-generating function of $X$.



6. **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 rewritten as:

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

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



7. **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 $\big(\phi(\theta)\big)^n$, where $Y = X_1 + X_2 + \cdots + X_n$.



**Proof:**

Using the above theorem, we have the following:

$$\begin{aligned}\phi_y(\theta) &= \phi_1(\theta)\phi_2(\theta)\cdots\phi_n(\theta) \\ &= \phi(\theta)\phi(\theta)\cdots\phi(\theta) = \big(\phi(\theta)\big)^n.\end{aligned}$$

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

8. **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 $\Big(\phi(\dfrac{\theta}{n})\Big)^n$, where $\overline{X} = (1/n)\sum_{i=1}^n X_i$.



**Proof:**

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

$$\begin{aligned}\phi_{\overline{x}}(\theta) &= \mathrm{E}(e^{\theta\overline{X}}) = \mathrm{E}(e^{\frac{\theta}{n}\sum_{i=1}^n X_i}) = \prod_{i=1}^n \mathrm{E}(e^{\frac{\theta}{n}X_i}) \\ &= \prod_{i=1}^n \phi(\frac{\theta}{n}) = \Big(\phi(\frac{\theta}{n})\Big)^n.\end{aligned}$$



**Bernoulli Distribution:** The probability function of Bernoulli random variable $X$ is:

$$f(x) = p^x(1-p)^{1-x}, \qquad x = 0, 1$$

The moment-generating function of $X$ is:

$$\phi(\theta) = pe^\theta + 1 - p$$

Mean: $\mathrm{E}(X) = \phi'(0) = p$

Variance: $\mathrm{V}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2 = \phi''(0) - p^2 = p(1-p)$



**Binomial Distribution:** For the binomial random variable, the moment-generating function $\phi(\theta)$ is known as:

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

which is discussed in Example 1.5 (Section 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 $\theta$ 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 $\phi(\theta)$.



**Poisson Distribution:** The probability function of Poisson random variable $X$ is:

$$f(x) = e^{-\lambda}\frac{\lambda^x}{x!}, \qquad x = 0, 1, 2, \cdots$$

The moment-generating function of $X$ is:

$$\begin{aligned}\phi(\theta) &= \sum_{x=0}^\infty e^{\theta x}e^{-\lambda}\frac{\lambda^x}{x!} \\ &= \sum_{x=0}^\infty e^{-\lambda}e^{e^\theta\lambda}e^{-e^\theta\lambda}\frac{(e^\theta\lambda)^x}{x!} \\ &= \exp(\lambda(e^\theta - 1))\end{aligned}$$

**Normal Distribution:** When $X \sim N(\mu, \sigma^2)$, the moment-generating function of $X$ is given by: $\phi(\theta) = \exp(\mu\theta + \frac{1}{2}\sigma^2\theta^2)$ from the previous example.

Obtain $\mathrm{E}(X)$ and $\mathrm{V}(X)$, using $\phi(\theta)$.

- $\mathrm{E}(X) = \phi'(0) = \mu$

from $\phi'(\theta) = (\mu + \sigma^2\theta)\exp(\mu\theta + \frac{1}{2}\sigma^2\theta^2)$.

- $\mathrm{E}(X^2) = \phi''(0) = \sigma^2 + \mu^2$

from $\phi''(\theta) = \sigma^2\exp(\mu\theta + \frac{1}{2}\sigma^2\theta^2) + (\mu + \sigma^2\theta)^2\exp(\mu\theta + \frac{1}{2}\sigma^2\theta^2)$.

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



**Cauchy Distribution:** Cauchy distribution: $f(x) = \dfrac{1}{\pi(1+x^2)}$ for $-\infty < x < \infty$.

$$\begin{aligned}\mathrm{E}(X) &= \int x f(x)\mathrm{d}x = \int \frac{x}{\pi(1+x^2)}\mathrm{d}x \\ &= \frac{1}{2\pi}[\log(1+x^2)]_{-\infty}^{\infty}\end{aligned}$$

$\Longrightarrow$ $\phi(\theta)$ does not exists.

$t(k)$ distrubution $\Longrightarrow$ $\mathrm{E}(X^k)$ does not exists.



**Uniform Distribution:** The density function is:

$$f(x) = \frac{1}{b-a}, \quad a < x < b$$

The moment-generating function is:

$$\begin{aligned}\phi(\theta) &= \int_{-\infty}^{\infty} e^{\theta x} f(x)\mathrm{d}x = \int_a^b e^{\theta x}\frac{1}{b-a}\mathrm{d}x \\ &= \left[\frac{e^{\theta x}}{\theta(b-a)}\right]_a^b = \frac{e^{\theta b} - e^{\theta a}}{\theta(b-a)}\end{aligned}$$



$$\phi'(\theta) = \frac{\theta(be^{\theta b} - ae^{\theta a}) - (e^{\theta b} - e^{\theta a})}{\theta^2(b-a)}$$

Mean:

$$\begin{aligned}\mathrm{E}(X) &= \phi'(0) \longleftarrow \text{ Use L'Hospital's rule.} \\ &= \frac{a+b}{2}\end{aligned}$$

(*) $f(\theta) = \theta(be^{\theta b} - ae^{\theta a}) - (e^{\theta b} - e^{\theta a})$, $\quad g(\theta) = \theta^2(b-a)$

$f'(\theta) = \theta(b^2e^{\theta b} - a^2e^{\theta a})$, $\quad g'(\theta) = 2\theta(b-a)$

$$\lim_{\theta\to 0}\frac{f(\theta)}{g(\theta)} = \lim_{\theta\to 0}\frac{f'(\theta)}{g'(\theta)} = \lim_{\theta\to 0}\frac{\theta(b^2e^{\theta b} - a^2e^{\theta a})}{2\theta(b-a)} = \frac{a+b}{2}$$



(*) L'Hospital's rule

For two continuous functions $f(x)$ and $g(x)$,

$$\lim_{x\to\infty}\frac{f(x)}{g(x)} = \lim_{x\to\infty}\frac{f'(x)}{g'(x)}, \qquad \text{or} \qquad \lim_{x\to 0}\frac{f(x)}{g(x)} = \lim_{x\to 0}\frac{f'(x)}{g'(x)},$$

L'Hospital's rule is used when we have:

$$\lim_{x\to\infty}\frac{f(x)}{g(x)} = \frac{\infty}{\infty} \text{ or } \frac{0}{0},$$

or

$$\lim_{x\to 0}\frac{f(x)}{g(x)} = \frac{\infty}{\infty} \text{ or } \frac{0}{0}.$$



Variance: $\quad \mathrm{V}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2$

$$\begin{aligned}\mathrm{E}(X^2) &= \phi''(0) \\ &= \frac{\theta^2(b^2e^{\theta b} - a^2e^{\theta a}) - 2\theta(be^{\theta b} - ae^{\theta a}) + 2(e^{\theta b} - e^{\theta a})}{\theta^3(b-a)}\end{aligned}$$

$$\begin{cases} f(\theta) = \theta^2(b^2e^{\theta b} - a^2e^{\theta a}) - 2\theta(be^{\theta b} - ae^{\theta a}) + 2(e^{\theta b} - e^{\theta a}) \\ g(\theta) = \theta^3(b-a)\end{cases}$$

$$\begin{cases} f'(\theta) = \theta^2(b^3e^{\theta b} - a^3e^{\theta a}) \\ g'(\theta) = 3\theta^2(b-a)\end{cases}$$

$$\phi''(0) = \lim_{\theta\to 0}\frac{f(\theta)}{g(\theta)} = \lim_{\theta\to 0}\frac{f'(\theta)}{g'(\theta)}$$
$$= \lim_{\theta\to 0}\frac{\theta^2(b^3e^{\theta b} - a^3e^{\theta a})}{3\theta^2(b-a)} = \frac{b^2+ba+a^2}{3}$$

$$\mathrm{V}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2$$
$$= \phi''(0) - (\phi'(0))^2 \longleftarrow \text{L'Hospital's rule}$$
$$= \frac{b^2+ba+a^2}{3} - \left(\frac{a+b}{2}\right)^2 = \frac{(b-a)^2}{12}$$



**Exponential Distribution:** The exponential distribution is:

$$f(x) = \lambda e^{-\lambda x}, \quad 0 < x$$

The moment-generating function is:

$$\phi(\theta) = \int_{-\infty}^{\infty} e^{\theta x} f(x)\mathrm{d}x = \int_0^{\infty} e^{\theta x}\lambda e^{-\lambda x}\mathrm{d}x$$
$$= \frac{\lambda}{\lambda-\theta}\int_0^{\infty}(\lambda-\theta)e^{-(\lambda-\theta)x}\mathrm{d}x = \frac{\lambda}{\lambda-\theta}$$

Use the exponential distribution with parameter $\lambda - \theta$ in the integration.



1. Mean: $\mathrm{E}(X) = \phi'(0)$

   $$\phi'(\theta) = \frac{\lambda}{(\lambda-\theta)^2}$$
   $$\mathrm{E}(X) = \phi'(0) = \frac{1}{\lambda}$$

2. Variance: $\mathrm{V}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2$

   $\mathrm{E}(X^2) = \phi''(0) \qquad \phi''(\theta) = 2\dfrac{\lambda}{(\lambda-\theta)^3}$
   $$\mathrm{V}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2 = \phi''(0) - (\phi'(0))^2$$
   $$= \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$



**$\chi^2$ Distribution:** The density function is:

$$f(x) = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}x^{\frac{n}{2}-1}\exp(-\frac{x}{2}), \quad 0 < x$$

The moment-generating function is:

$$\phi(\theta) = \int_{-\infty}^{\infty} e^{\theta x} f(x)\mathrm{d}x$$
$$= \int_0^{\infty} e^{\theta x}\frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}x^{\frac{n}{2}-1}\exp(-\frac{x}{2})\mathrm{d}x$$
$$= \int_0^{\infty}\frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}x^{\frac{n}{2}-1}\exp\left(-\frac{1}{2}(1-2\theta)x\right)\mathrm{d}x$$
$$= \int_0^{\infty}\frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}\left(\frac{y}{1-2\theta}\right)^{\frac{n}{2}-1}\exp(-\frac{1}{2}y)\frac{1}{1-2\theta}\mathrm{d}y$$



$$= \left(\frac{1}{1-2\theta}\right)^{\frac{n}{2}}\int_0^{\infty}\frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}y^{\frac{n}{2}-1}\exp(-\frac{1}{2}y)\mathrm{d}y$$
$$= \left(\frac{1}{1-2\theta}\right)^{\frac{n}{2}}$$

Use integration by substitution by $y = (1-2\theta)x$

$\dfrac{\mathrm{d}x}{\mathrm{d}y} = (1-2\theta)^{-1}$

Use the $\chi^2(n)$ distribution in the integration.



1. Mean: $\mathrm{E}(X) = \phi'(0)$

   $\phi'(\theta) = (-\frac{n}{2})(-2)(1-2\theta)^{-\frac{n}{2}-1}$
   $$\mathrm{E}(X) = \phi'(0) = n$$

2. Variance: $\mathrm{V}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2$

   $\mathrm{E}(X^2) = \phi''(0)$

   $\phi''(\theta) = (-\frac{n}{2})(-\frac{n}{2}-1)(-2)^2(1-2\theta)^{-\frac{n}{2}-1}$
   $$\mathrm{V}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2 = \phi''(0) - (\phi'(0))^2$$
   $$= n(n+2) - n^2 = 2n$$

**Sum of Bernoulli Random Variables:** $X_1, X_2, \cdots, X_n$ are mutually independently and identically distributed as Bernoulli random variable with parameter $p$. Then, the probability function of $Y = X_1 + X_2 + \cdots + X_n$ is $B(n, p)$.

**Proof:** The moment-generating function of $X_i$, $\phi_i(\theta)$, is:

$$\phi_i(\theta) = pe^{\theta} + 1 - p$$



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

$$\begin{aligned}\phi_y(\theta) &= \mathrm{E}(e^{\theta Y}) = \mathrm{E}(e^{\theta(X_1+X_2+\cdots+X_n)}) \\ &= \mathrm{E}(e^{\theta X_1})\mathrm{E}(e^{\theta X_2})\cdots\mathrm{E}(e^{\theta X_n}) = \phi_1(\theta)\phi_2(\theta)\cdots\phi_n(\theta) \\ &= \Big(\phi(\theta)\Big)^n = (pe^{\theta} + 1 - p)^n,\end{aligned}$$

which is the moment-generating function of $B(n, p)$.

Note:

In the third equality, $X_1, X_2, \cdots, X_n$ are mutually independent.

In the fifth equality, $X_1, X_2, \cdots, X_n$ are identically distributed.



**Sum of Two Normal Random Variables:** $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$. $X$ is independent of $Y$.

Then, $aX + bY \sim N(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2)$, where $a$ are $b$ are constant.

**Proof:** Suppose tha the moment-generating functions of $X$ and $Y$ are given by $\phi_x(\theta)$ and $\phi_y(\theta)$.

$$\begin{aligned}\phi_x(\theta) &= \exp\Big(\mu_1\theta + \frac{1}{2}\sigma_1^2\theta^2\Big) \\ \phi_y(\theta) &= \exp\Big(\mu_2\theta + \frac{1}{2}\sigma_2^2\theta^2\Big)\end{aligned}$$



The moment-generating function of $W = aX + bY$ is:

$$\begin{aligned}\phi_w(\theta) &= \mathrm{E}(e^{\theta W}) = \mathrm{E}(e^{\theta(aX+bY)}) = \mathrm{E}(e^{a\theta X})\mathrm{E}(e^{b\theta Y}) = \phi_x(a\theta)\phi_y(b\theta) \\ &= \exp\Big(\mu_1(a\theta) + \frac{1}{2}\sigma_1^2(a\theta)^2\Big) \times \exp\Big(\mu_2(b\theta) + \frac{1}{2}\sigma_2^2(b\theta)^2\Big) \\ &= \exp\Big((a\mu_1 + b\mu_2)\theta + \frac{1}{2}(a^2\sigma_1^2 + b^2\sigma_2^2)\theta^2\Big)\end{aligned}$$

which is the moment-generating function of normal distribution with mean $a\mu_1+b\mu_2$ and variance $a^2\sigma^2 + b^2\sigma_2^2$.

Therefore, $aX + bY \sim N(a\mu_1 + b\mu_2, a^2\sigma^2 + b^2\sigma_2^2)$



**Sum of Two $\chi^2$ Random Variables:** $X \sim \chi^2(n)$ and $Y \sim \chi^2(m)$. $X$ is independent of $Y$.

Then, $Z = X + Y \sim \chi^2(n + m)$

**Proof:**

Let $\phi_x(\theta)$ and $\phi_y(\theta)$ be the moment-generating functions of $X$ and $Y$.

$\phi_x(\theta)$ and $\phi_y(\theta)$ are given by:

$$\phi_x(\theta) = \Big(\frac{1}{1-2\theta}\Big)^{\frac{n}{2}}, \qquad \phi_y(\theta) = \Big(\frac{1}{1-2\theta}\Big)^{\frac{m}{2}}.$$



The moment-generating function of $Z = X + Y$ is:

$$\begin{aligned}\phi_z(\theta) &\equiv \mathrm{E}(e^{\theta Z}) = \mathrm{E}(e^{\theta(X+Y)}) = \mathrm{E}(e^{\theta X})\mathrm{E}(e^{\theta Y}) = \phi_x(\theta)\phi_y(\theta) \\ &= \Big(\frac{1}{1-2\theta}\Big)^{\frac{n}{2}}\Big(\frac{1}{1-2\theta}\Big)^{\frac{m}{2}} = \Big(\frac{1}{1-2\theta}\Big)^{\frac{n+m}{2}}\end{aligned}$$

which is the moment-generating function of $\chi^2(n + m)$ distribution. Therefore, $Z \sim \chi^2(n + m)$.

Note:

In the third equality, $X$ and $Y$ are independent.

## 5.2 Multivariate Cases

**Bivariate Case:** As discussed in Section 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:

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

**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 $\theta_1$ and at the same time the $k$th derivative with respect to $\theta_2$, we have the following expression:

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

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

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



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:

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

$$= \int_{-\infty}^{\infty} e^{\theta_1 x} f_x(x)\,\mathrm{d}x = \mathrm{E}(e^{\theta_1 X}) = \phi_1(\theta_1).$$

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.

   Suppose that $X_i \sim N(\mu, \sigma^2)$.

   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, $\hat{\mu} \sim N(\mu \sum_{i=1}^n a_i, \sigma^2 \sum_{i=1}^n a_i^2)$.



**Proof:**

From Example 1.8 (p.111) and Example 1.9 (p.147), 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 $\mu$ and $\sigma^2$ 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: $\overline{X} \sim N(\mu, \sigma^2/n)$.



# 6 Law of Large Numbers (対数の法則) and Central Limit Theorem (中心極限定理)

## 6.1 Chebyshev's Inequality (チェビシェフの不等式)



**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:

$$P(g(X) \geq k) \leq \frac{\mathrm{E}(g(X))}{k}, \tag{6}$$

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:

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

which implies that we have $g(X) \geq k$ when $U = 1$ and $g(X) \geq 0$ when $U = 0$, where $k$ is a positive constant value.

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

$$\mathrm{E}(g(X)) \geq k\mathrm{E}(U), \tag{7}$$



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

$$\begin{aligned} \mathrm{E}(U) &= \sum_{u=0}^{1} uP(U = u) = 1 \times P(U = 1) + 0 \times P(U = 0) \\ &= P(U = 1) = P(g(X) \geq k). \end{aligned} \tag{8}$$

Accordingly, substituting equation (8) into equation (7), 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 $\lambda$ 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 $P(|X - \mu| \geq \lambda\sigma) \leq \dfrac{1}{\lambda^2}$.

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:

$$P(|X - \mu| < \lambda\sigma) \geq 1 - \frac{1}{\lambda^2}. \tag{9}$$



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

Equation (9) 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 $\mu$ 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.

Furthermore, note as follows.

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

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

i.e.,

$$P(|X-\mathrm{E}(X)| \geq \epsilon) \leq \frac{\mathrm{V}(X)}{\epsilon^2}, \tag{10}$$

which inequality is used in the next section.



**Remark:** Equation (10) can be derived when we take $g(X) = (X-\mu)^2$, $\mu = \mathrm{E}(X)$ and $k = \epsilon^2$ in equation (6).

Even when we have $\mu \neq \mathrm{E}(X)$, the following inequality still hold:

$$P(|X-\mu| \geq \epsilon) \leq \frac{\mathrm{E}((X-\mu)^2)}{\epsilon^2}.$$

Note that $\mathrm{E}((X-\mu)^2)$ represents the mean square error (MSE).

When $\mu = \mathrm{E}(X)$, the mean square error reduces to the variance.



## 6.2 Law of Large Numbers (対数の法則) and Convergence in Probability (確率収束)

**Law of Large Numbers 1:** Assume that $X_1, X_2, \cdots, X_n$ are mutually independently and identically distributed with mean $\mathrm{E}(X_i) = \mu$ for all $i$.

Supopose that the moment-generating function of $X_i$ is finite.

Define $\overline{X}_n = \dfrac{1}{n}\sum_{i=1}^n X_i$.

Then, $\overline{X}_n \longrightarrow \mu$ as $n \longrightarrow \infty$.



**Proof:** The moment-generating function is written as:

$$\begin{aligned}\phi(\theta) &= 1 + \mu_1'\theta + \frac{1}{2!}\mu_2'\theta^2 + \frac{1}{3!}\mu_3'\theta^3 + \cdots \\ &= 1 + \mu_1'\theta + O(\theta^2)\end{aligned}$$

where $\mu_k' = \mathrm{E}(X^k)$ for all $k$. That is, all the moments exist.

$$\begin{aligned}\phi_{\overline{x}}(\theta) &= \Big(\phi\Big(\frac{\theta}{n}\Big)\Big)^n = \Big(1 + \mu_1'\frac{\theta}{n} + O(\frac{\theta^2}{n^2})\Big)^n \\ &= \Big(1 + \mu_1'\frac{\theta}{n} + O(\frac{1}{n^2})\Big)^n = \Big((1+x)^{\frac{1}{x}}\Big)^{\mu\theta + O(n^{-1})} \\ &\longrightarrow \exp(\mu\theta) \quad \text{as } x \longrightarrow 0,\end{aligned}$$



which is the following probability function:

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

$$\phi(\theta) = \sum e^{\theta x} f(x) = e^{\theta\mu} f(\mu) = e^{\theta\mu}$$



**Law of Large Numbers 2:** 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 $\epsilon$, as $n \longrightarrow \infty$, we have the following result:

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

where $\overline{X}_n = \dfrac{1}{n}\sum_{i=1}^n X_i$.

We say that $\overline{X}_n$ converges in probability to $\mu$.

**Proof:**

Using (10), Chebyshev's inequality is represented as follows:

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

where $X$ in (10) is replaced by $\overline{X}_n$.

We know $\mathrm{E}(\overline{X}_n) = \mu$ and $\mathrm{V}(\overline{X}_n) = \dfrac{\sigma^2}{n}$, which are substituted into the above inequality.

Then, we obtain:

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



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

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

That is, $\overline{X}_n \longrightarrow \mu$ is obtained as $n \longrightarrow \infty$, which is written as: $\operatorname{plim} \overline{X}_n = \mu$.

This theorem is called the **law of large numbers**.

The condition $P(|\overline{X}_n - \mu| \geq \epsilon) \longrightarrow 0$ or equivalently $P(|\overline{X}_n - \mu| < \epsilon) \longrightarrow 1$ is used as the definition of **convergence in probability (確率収束)**.

In this case, we say that $\overline{X}_n$ converges in probability to $\mu$.



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

$$m_n = \mathrm{E}(\sum_{i=1}^{n} X_i), \qquad V_n = \mathrm{V}(\sum_{i=1}^{n} X_i),$$

and assume that

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



Then, we obtain the following result:

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

That is, $\overline{X}_n$ converges in probability to $\lim_{n\to\infty} \dfrac{m_n}{n}$.

This theorem is also called the law of large numbers.



**Proof:**

Remember Chebyshev's inequality:

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

Replace $X$, $\mathrm{E}(X)$ and $\mathrm{V}(X)$
by $\overline{X}_n$, $\mathrm{E}(\overline{X}_n) = \dfrac{m_n}{n}$ and $\mathrm{V}(\overline{X}_n) = \dfrac{V_n}{n^2}$.

Then, we obtain:

$$P(|\overline{X}_n - \frac{m_n}{n}| \geq \epsilon) \leq \frac{V_n}{n^2\epsilon^2}.$$



As $n$ goes to infinity,

$$P(|\overline{X}_n - \frac{m_n}{n}| \geq \epsilon) \leq \frac{V_n}{n^2\epsilon^2} \longrightarrow 0.$$

Therefore, $\overline{X}_n \longrightarrow \lim_{n\to\infty} \dfrac{m_n}{n}$ as $n \longrightarrow \infty$.

## 6.3 Central Limit Theorem (中心極限定理) and Convergence in Distribution (分布収束)

**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 $\mu$ and $\sigma^2$ 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 $\theta = 0$.



**(*) Remark:**

$O(x)$ implies that it is a polynomial function of $x$ and the higher-order terms but it is dominated by $x$.

In this case, $O(\theta^3)$ is a function of $\theta^3, \theta^4, \cdots$.

Since the moment-generating function is conventionally evaluated at $\theta = 0$, $\theta^3$ is the largest value of $\theta^3, \theta^4, \cdots$ and accordingly $O(\theta^3)$ is dominated by $\theta^3$ (in other words, $\theta^4, \theta^5, \cdots$ are small enough, compared with $\theta^3$).



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:

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

We consider that $n$ goes to infinity.

Therefore, $O(\frac{\theta^3}{n^{\frac{3}{2}}})$ indicates a function of $n^{-\frac{3}{2}}$.



Moreover, consider $x = \dfrac{1}{2}\dfrac{\theta^2}{n} + O(n^{-\frac{3}{2}})$.

Multiply $n/x$ on both sides of $x = \dfrac{1}{2}\dfrac{\theta^2}{n} + O(n^{-\frac{3}{2}})$.

Then, we obtain $n = \dfrac{1}{x}\Big(\dfrac{1}{2}\theta^2 + O(n^{-\frac{1}{2}})\Big)$.

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 2.3 (p.35).
Furthermore, we have $O(n^{-\frac{1}{2}}) \longrightarrow 0$ as $n \longrightarrow \infty$.
Since $\phi_z(\theta) = e^{\frac{\theta^2}{2}}$ is the moment-generating function of the standard normal distribution (see p.110 in Section 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,

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



We say that $\dfrac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$ converges in distribution to $N(0,1)$.
$\Longrightarrow$ **Convergence in distribution (分布収束)**

The following expression is also possible:

$$\sqrt{n}(\overline{X}_n - \mu) \longrightarrow N(0, \sigma^2). \tag{11}$$



**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

$$\lim_{n\to\infty} n\mathrm{V}(\overline{X_n}) = \sigma^2 < \infty, \qquad \text{where } \overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Then, when $n \longrightarrow \infty$, we have:

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



**Summary:** Let $X_1, X_2, \cdots, X_n, \cdots$ be a sequence of random variables. Let $X$ be a random variable. Let $F_n$ be the distribution function of $X_n$ and $F$ be that of $X$.

• $X_n$ converges in probability to $X$ if $\lim_{n\to\infty} P(|X_n - X| \geq \epsilon) = 0$ or $\lim_{n\to\infty} P(|X_n - X| < \epsilon) = 1$ for all $\epsilon > 0$.
Equivalently, we write $X_n \xrightarrow{P} X$.

• $X_n$ converges in distribution to $X$ (or $F$) if $\lim_{n\to\infty} F_n(x) = F(x)$ for all $x$.
Equivalently, we write $X_n \xrightarrow{D} X$ or $X_n \xrightarrow{D} F$.



# 7 Statistical Inference

## 7.1 Point Estimation (点推定)

Suppose that the underlying distribution is known but the parameter $\theta$ 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 the population distribution.

Consider estimating the parameter $\theta$ using the $n$ observed data.

Let $\hat\theta_n(x_1, x_2, \cdots, x_n)$ be a function of the observed data $x_1, x_2, \cdots, x_n$.

$\hat\theta_n(x_1, x_2, \cdots, x_n)$ is constructed to estimate the parameter $\theta$.

$\hat\theta_n(x_1, x_2, \cdots, x_n)$ takes a certain value given the $n$ observed data.

$\hat\theta_n(x_1, x_2, \cdots, x_n)$ is called the **point estimate** of $\theta$, or simply the **estimate** of $\theta$.



**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 $\mu$ 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 $\sigma^2$ 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 $\sigma^2$ is:

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



## 7.2 Statistic, Estimate and Estimator (統計量，推定値，推定量)

The underlying distribution of population is assumed to be known, but the parameter $\theta$, 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 $X_1, X_2, \cdots, X_n$ take the experimental values $x_1, x_2, \cdots, x_n$ by chance.



There, the experimental values and the actually observed data series are used in the same meaning.

$\hat\theta_n(x_1, x_2, \cdots, x_n)$ denotes the point estimate of $\theta$.

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 (推定量)** of $\theta$, which should be distinguished from the **estimate (推定値)** of $\theta$, 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 $\mu$ is given by $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$, while the estimate of $\mu$ is $\overline{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$.

The estimator of $\sigma^2$ is $S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ and the estimate of $\sigma^2$ is $s^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$.



There are numerous estimators and estimates of $\theta$.

All of $\dfrac{1}{n}\sum_{i=1}^{n} X_i$, $\dfrac{X_1 + X_n}{2}$, median of $(X_1, X_2, \cdots, X_n)$ and so on are taken as the estimators of $\mu$.

Of course, they are called the estimates of $\theta$ when $X_i$ is replaced by $x_i$ for all $i$.

Both $S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ and $S^{*2} = \dfrac{1}{n}\sum_{i=1}^{2}(X_i - \overline{X})^2$ are the estimators of $\sigma^2$.



We need to choose one out of the numerous estimators of $\theta$.

The problem of choosing an optimal estimator out of the numerous estimators is discussed in Sections 7.4 and 7.5.1.

In addition, note as follows.
A function of random variables is called a **statistic (統計量)**. The statistic for estimation of the parameter is called an estimator.

Therefore, an estimator is a family of a statistic.



## 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 $\mu$ is:
   - $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$.
2. The estimators of population variance $\sigma^2$ are:
   - $S^{*2} = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$, when $\mu$ 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.138, mean and variance of $\overline{X}$ are obtained as follows:

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

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

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

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

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



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

$\sum_{i=1}^n (X_i-\mu) = n(\overline{X}-\mu)$ is used in the sixth equality.

$\mathrm{E}\Big(\frac{1}{n}\sum_{i=1}^n (X_i-\mu)^2\Big) = \mathrm{E}(S^{*2}) = \sigma^2$ and

$\mathrm{E}((\overline{X}-\mu)^2) = \mathrm{V}(\overline{X}) = \dfrac{\sigma^2}{n}$ are required in the eighth equality.



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

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

Summarizing the above results, we obtain as follows:

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



## 7.4 Point Estimation: Optimality

$\theta$ denotes the parameter to be estimated.

$\hat{\theta}_n(X_1, X_2, \cdots, X_n)$ represents the estimator of $\theta$, while $\hat{\theta}_n(x_1, x_2, \cdots, x_n)$ indicates the estimate of $\theta$.

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 of $\hat{\theta}_n$ are:

- **unbiasedness (不偏性)**,
- **efficiency (有効性)**.
- **consistency (一致性)** and
- **sufficiency (十分性)**. ⟵ Not discussed in this class.



**Unbiasedness (不偏性):** One of the desirable features that the estimator of the parameter should have is given by:

$$\mathrm{E}(\hat{\theta}_n) = \theta, \tag{12}$$

which implies that $\hat{\theta}_n$ is distributed around $\theta$.

When (12) holds, $\hat{\theta}_n$ is called the **unbiased estimator (不偏推定量)** of $\theta$.

$\mathrm{E}(\hat{\theta}_n) - \theta$ is defined as **bias (偏り)**.

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 $\mu$ and variance $\sigma^2$.

Consider the following estimators of $\mu$ and $\sigma^2$.

1. The estimator of $\mu$ is:
   - $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$.



2. The estimators of $\sigma^2$ are:

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

Since we have obtained $\mathrm{E}(\overline{X}) = \mu$ and $\mathrm{E}(S^2) = \sigma^2$, $\overline{X}$ and $S^2$ are unbiased estimators of $\mu$ and $\sigma^2$.

We have obtained the result $\mathrm{E}(S^{**2}) \neq \sigma^2$ and therefore $S^{**2}$ is not an unbiased estimator of $\sigma^2$.

According to the criterion of unbiasedness, $S^2$ is preferred to $S^{**2}$ for estimation of $\sigma^2$.



**Efficiency (有効性):** Consider two estimators, $\hat{\theta}_n$ and $\widetilde{\theta}_n$.

Both are assumed to be unbiased.

That is, $\mathrm{E}(\hat{\theta}_n) = \theta$ 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 unbiased estimator with the least variance is known as the **efficient estimator (有効推定量)**.

We have the case where an efficient estimator does not exist.

In order to find the efficient estimator, we utilize **Cramer-Rao inequality (クラメール・ラオの不等式)**.



Suppose that $X_1, X_2, \cdots, X_n$ are mutually independently and identically distributed and the distribution of $X_i$ is $f(x_i;\theta)$.

For any unbiased estimator of $\theta$, denoted by $\hat{\theta}_n$, it is known that we have the following inequality:

$$\mathrm{V}(\hat{\theta}_n) \geq \frac{\sigma^2(\theta)}{n}, \tag{13}$$

$$\text{where } \sigma^2(\theta) = \frac{1}{\mathrm{E}\Big(\big(\frac{\partial \log f(X;\theta)}{\partial\theta}\big)^2\Big)} = \frac{1}{\mathrm{V}\Big(\big(\frac{\partial \log f(X;\theta)}{\partial\theta}\big)\Big)}$$

$$= -\frac{1}{\mathrm{E}\big(\frac{\partial^2 \log f(X;\theta)}{\partial\theta^2}\big)}, \tag{14}$$



which is known as the **Cramer-Rao inequality (クラメール・ラオの不等式)**.

When there exists the unbiased estimator $\hat{\theta}_n$ such that the equality in (13) holds, $\hat{\theta}_n$ becomes the unbiased estimator with minimum variance, which is the **efficient estimator (有効推定量)**.

$\dfrac{\sigma^2(\theta)}{n}$ is called the **Cramer-Rao lower bound (クラメール・ラオの下限)**.



**Proof of the Cramer-Rao inequality:** We prove the above inequality and the equalities in $\sigma^2(\theta)$.

The **likelihood function (尤度関数)** $l(\theta; x) = l(\theta; x_1, x_2, \cdots, x_n)$ is a joint density of $X_1, X_2, \cdots, X_n$.

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

See Section 7.5.1 for the **likelihood function (尤度関数)**.

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. That is, we have the following equation:

$$1 = \int l(\theta; x)\,\mathrm{d}x, \tag{15}$$

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 (15) with respect to $\theta$, we obtain the following equation:

$$\begin{aligned} 0 &= \int \frac{\partial l(\theta; x)}{\partial \theta}\,\mathrm{d}x = \int \frac{1}{l(\theta; x)} \frac{\partial l(\theta; x)}{\partial \theta} l(\theta; x)\,\mathrm{d}x \\ &= \int \frac{\partial \log l(\theta; x)}{\partial \theta} l(\theta; x)\,\mathrm{d}x = \mathrm{E}\Big(\frac{\partial \log l(\theta; X)}{\partial \theta}\Big), \end{aligned} \tag{16}$$

which implies that the expectation of $\dfrac{\partial \log l(\theta; X)}{\partial \theta}$ is equal to zero.

In the third equality, note that $\dfrac{\mathrm{d} \log x}{\mathrm{d}x} = \dfrac{1}{x}$.



Now, let $\hat{\theta}_n$ be an estimator of $\theta$. The definition of the mathematical expectation of the estimator $\hat{\theta}_n$ is represented as:

$$\mathrm{E}(\hat{\theta}_n) = \int \hat{\theta}_n l(\theta; x)\,\mathrm{d}x. \tag{17}$$

Differentiating equation (17) with respect to $\theta$ on both sides, we can rewrite as follows:

$$\begin{aligned} \frac{\partial \mathrm{E}(\hat{\theta}_n)}{\partial \theta} &= \int \hat{\theta}_n \frac{\partial l(\theta; x)}{\partial \theta}\,\mathrm{d}x = \int \hat{\theta}_n \frac{\partial \log l(\theta; x)}{\partial \theta} l(\theta; x)\,\mathrm{d}x \\ &= \int \Big(\hat{\theta}_n - \mathrm{E}(\hat{\theta}_n)\Big)\Big(\frac{\partial \log l(\theta; x)}{\partial \theta} - \mathrm{E}(\frac{\partial \log l(\theta; x)}{\partial \theta})\Big) l(\theta; x)\,\mathrm{d}x \\ &= \mathrm{Cov}\Big(\hat{\theta}_n, \frac{\partial \log l(\theta; X)}{\partial \theta}\Big). \end{aligned} \tag{18}$$



In the second equality, $\dfrac{\mathrm{d} \log x}{\mathrm{d}x} = \dfrac{1}{x}$ is utilized.

The third equality holds because of $\mathrm{E}(\dfrac{\partial \log l(\theta; X)}{\partial \theta}) = 0$ from equation (16).

For simplicity of discussion, suppose that $\theta$ is a scalar.



Taking the square on both sides of equation (18), we obtain the following expression:

$$\begin{aligned} \Big(\frac{\partial \mathrm{E}(\hat{\theta}_n)}{\partial \theta}\Big)^2 &= \left(\mathrm{Cov}\Big(\hat{\theta}_n, \frac{\partial \log l(\theta; X)}{\partial \theta}\Big)\right)^2 = \rho^2 \mathrm{V}(\hat{\theta}_n) \mathrm{V}\Big(\frac{\partial \log l(\theta; X)}{\partial \theta}\Big) \\ &\leq \mathrm{V}(\hat{\theta}_n) \mathrm{V}\left(\frac{\partial \log l(\theta; X)}{\partial \theta}\right), \end{aligned} \tag{19}$$

where $\rho$ denotes the correlation coefficient between $\hat{\theta}_n$ and $\dfrac{\partial \log l(\theta; X)}{\partial \theta}$.



Note that we have the definition of $\rho$ 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 inequality (19) is obtained, which is rewritten as:

$$\mathrm{V}(\hat{\theta}_n) \geq \frac{\Big(\dfrac{\partial \mathrm{E}(\hat{\theta}_n)}{\partial \theta}\Big)^2}{\mathrm{V}\Big(\dfrac{\partial \log l(\theta; X)}{\partial \theta}\Big)}. \tag{20}$$

When $\mathrm{E}(\hat\theta_n) = \theta$, i.e., when $\hat\theta_n$ is an unbiased estimator of $\theta$, the numerator in the right-hand side of equation (20) is equal to one.

Therefore, we have the following result:

$$\mathrm{V}(\hat\theta_n) \geq \frac{1}{\mathrm{V}\Big(\dfrac{\partial \log l(\theta;X)}{\partial\theta}\Big)} = \frac{1}{\mathrm{E}\Big(\Big(\dfrac{\partial \log l(\theta;X)}{\partial\theta}\Big)^2\Big)}.$$

Note that we have $\mathrm{V}\Big(\dfrac{\partial \log l(\theta;X)}{\partial\theta}\Big) = \mathrm{E}\Big((\dfrac{\partial \log l(\theta;X)}{\partial\theta})^2\Big)$ in the equality above, because of $\mathrm{E}\Big(\dfrac{\partial \log l(\theta;X)}{\partial\theta}\Big) = 0$.



Moreover, the denominator in the right-hand side of the above inequality is rewritten as follows:

$$\begin{aligned}
&\mathrm{E}\Big(\Big(\frac{\partial \log l(\theta;X)}{\partial\theta}\Big)^2\Big) \\
&= \mathrm{E}\Big(\Big(\sum_{i=1}^n \frac{\partial \log f(X_i;\theta)}{\partial\theta}\Big)^2\Big) = \sum_{i=1}^n \mathrm{E}\Big(\Big(\frac{\partial \log f(X_i;\theta)}{\partial\theta}\Big)^2\Big) \\
&= n\mathrm{E}\Big(\Big(\frac{\partial \log f(X;\theta)}{\partial\theta}\Big)^2\Big) = n\int_{-\infty}^{\infty}\Big(\frac{\partial \log f(x;\theta)}{\partial\theta}\Big)^2 f(x;\theta)\,\mathrm{d}x.
\end{aligned}$$

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:

$$\mathrm{V}(\hat\theta_n) \geq \frac{1}{\mathrm{E}\Big(\Big(\dfrac{\partial \log l(\theta;X)}{\partial\theta}\Big)^2\Big)} = \frac{1}{n\mathrm{E}\Big(\Big(\dfrac{\partial \log f(X;\theta)}{\partial\theta}\Big)^2\Big)} = \frac{\sigma^2(\theta)}{n},$$

which is equivalent to (13).

Next, we prove the equalities in (14), i.e.,

$$-\mathrm{E}\Big(\frac{\partial^2 \log f(X;\theta)}{\partial\theta^2}\Big) = \mathrm{E}\Big(\Big(\frac{\partial \log f(X;\theta)}{\partial\theta}\Big)^2\Big)$$



$$= \mathrm{V}\Big(\frac{\partial \log f(X;\theta)}{\partial\theta}\Big).$$

Differentiating $\int f(x;\theta)\,\mathrm{d}x = 1$ with respect to $\theta$, 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 $\theta$ and that $\dfrac{\partial f(x;\theta)}{\partial\theta}$ exists.

The above equation is rewritten as:

$$\int \frac{\partial \log f(x;\theta)}{\partial\theta} f(x;\theta)\,\mathrm{d}x = 0, \tag{21}$$



or equivalently,

$$\mathrm{E}\Big(\frac{\partial \log f(X;\theta)}{\partial\theta}\Big) = 0. \tag{22}$$

Again, differentiating equation (21) with respect to $\theta$,

$$\int \frac{\partial^2 \log f(x;\theta)}{\partial\theta^2} f(x;\theta)\,\mathrm{d}x + \int \frac{\partial \log f(x;\theta)}{\partial\theta}\frac{\partial f(x;\theta)}{\partial\theta}\,\mathrm{d}x = 0,$$

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.,

$$\mathrm{E}\Big(\frac{\partial^2 \log f(x;\theta)}{\partial\theta^2}\Big) + \mathrm{E}\Big(\Big(\frac{\partial \log f(x;\theta)}{\partial\theta}\Big)^2\Big) = 0.$$



Thus, we obtain:

$$-\mathrm{E}\Big(\frac{\partial^2 \log f(x;\theta)}{\partial\theta^2}\Big) = \mathrm{E}\Big(\Big(\frac{\partial \log f(x;\theta)}{\partial\theta}\Big)^2\Big).$$

Moreover, from equation (22), the following equation is derived.

$$\mathrm{E}\Big(\Big(\frac{\partial \log f(x;\theta)}{\partial\theta}\Big)^2\Big) = \mathrm{V}\Big(\frac{\partial \log f(x;\theta)}{\partial\theta}\Big).$$

Therefore, we have:

$$-\mathrm{E}\Big(\frac{\partial^2 \log f(X;\theta)}{\partial\theta^2}\Big) = \mathrm{E}\Big(\Big(\frac{\partial \log f(X;\theta)}{\partial\theta}\Big)^2\Big) = \mathrm{V}\Big(\frac{\partial \log f(X;\theta)}{\partial\theta}\Big).$$

Thus, the Cramer-Rao inequality is derived as:

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

where

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



**Example 1.13a (Efficient Estimator of $\mu$):** Suppose that $X_1, X_2, \cdots, X_n$ are mutually independently, identically and normally distributed with mean $\mu$ and variance $\sigma^2$.

Then, we show that $\overline{X}$ is an efficient estimator of $\mu$.

$\mathrm{V}(\overline{X})$ is given by $\dfrac{\sigma^2}{n}$, which does not depend on the distribution of $X_i$, $i = 1, 2, \cdots, n$. (A)



Because $X_i$ is normally distributed with mean $\mu$ and variance $\sigma^2$, the 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:

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

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

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



The partial derivative of $f(X;\mu)$ with respect to $\mu$ is:

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

The Cramer-Rao inequality in this case is written as:

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



From (A) and (B), variance of $\overline{X}$ is equal to the lower bound of Cramer-Rao inequality, i.e., $\mathrm{V}(\overline{X}) = \dfrac{\sigma^2}{n}$, 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 $\mu$.



**Example 1.13b (Efficient Estimator of $\sigma^2$):** Suppose that $X_1, X_2, \cdots, X_n$ are mutually independently, identically and normally distributed with mean $\mu$ and variance $\sigma^2$.

Is $S^2$ is an efficient estimator of $\sigma^2$?

$\mathrm{E}(S^2) = \sigma^2$ ....... Unbiased estimator

Under normality assumption, $\mathrm{V}(S^2)$ is given by $\dfrac{2\sigma^4}{n-1}$, because $\mathrm{V}(U) = 2(n-1)$ from $U = \dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$.

$\cdots\cdots$ (A)

Because $X_i$ is normally distributed with mean $\mu$ and variance $\sigma^2$, the density function of $X_i$ is given by:

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

The Cramer-Rao inequality is represented as:

$$\mathrm{V}(S^2) \geq \frac{1}{n\mathrm{E}\Big(\Big(\dfrac{\partial \log f(X;\sigma^2)}{\partial \sigma^2}\Big)^2\Big)} = \frac{1}{-n\mathrm{E}\Big(\dfrac{\partial^2 \log f(X;\sigma^2)}{\partial (\sigma^2)^2}\Big)},$$

where the logarithm of $f(X;\sigma^2)$ is written as:

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



The partial derivative of $f(X;\sigma^2)$ with respect to $\sigma^2$ is:

$$\frac{\partial \log f(X;\sigma^2)}{\partial \sigma^2} = -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}(X-\mu)^2.$$

The 2nd partial derivative of $f(X;\sigma^2)$ with respect to $\sigma^2$ is:

$$\frac{\partial^2 \log f(X;\sigma^2)}{\partial (\sigma^2)^2} = \frac{1}{2\sigma^4} - \frac{1}{\sigma^6}(X-\mu)^2.$$

The Cramer-Rao inequality in this case is written as:

$$\mathrm{V}(S^2) \geq \frac{1}{-n\mathrm{E}\Big(\dfrac{1}{2\sigma^4} - \dfrac{1}{\sigma^6}(X-\mu)^2.\Big)} = \frac{2\sigma^4}{n}. \qquad \text{(B)}$$



From (A) and (B), variance of $S^2$ is not equal to the lower bound of Cramer-Rao inequality, i.e., $\mathrm{V}(S^2) = \dfrac{2\sigma^4}{n-1} > \dfrac{2\sigma^4}{n}$.

Therefore, we can conclude that the sample unbiased variance $S^2$ is not an efficient estimator of $\sigma^2$.



**Example 1.14: Minimum Variance Linear Unbiased Estimator (最小分散線形不偏推定量):** Suppose that $X_1, X_2, \cdots, X_n$ are mutually independently and identically distributed with mean $\mu$ and variance $\sigma^2$ (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 **minimum variance linear unbiased estimator** if $a_i = \dfrac{1}{n}$ for all $i$, i.e., if $\hat{\mu} = \overline{X}$.



Utilizing Theorem on p.133, 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 (線形推定量)** of $\mu$.

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 (線形不偏推定量)**.

Thus, as mentioned in Example 1.12 of Section 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 $\lambda$ denotes the Lagrange multiplier.

The $\frac{1}{2}$ in the first term makes computation easier.



For minimization, the partial derivatives of $L$ with respect to $a_i$ and $\lambda$ are 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, $a_i = \lambda = \dfrac{1}{n}$ is obtained.
When $a_i = \dfrac{1}{n}$ for all $i$, $\hat{\mu}$ has minimum variance in a class of linear unbiased estimators.
$\overline{X}$ is a **minimum variance linear unbiased estimator**.



The **minimum variance linear unbiased estimator** is different from the **efficient estimator**.
The former does not requires the normality assumption.
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 normality assumption, he linear unbiased minimum variance estimator leads to the efficient estimator.
Note that the efficient estimator does not necessarily exist.



**Consistency (一致性):** Let $\hat{\theta}_n$ be an estimator of $\theta$.

Suppose that for any $\epsilon > 0$ we have the following:

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

which implies that $\hat{\theta} \longrightarrow \theta$ as $n \longrightarrow \infty$.

We say that $\hat{\theta}_n$ is a **consistent estimator (一致推定量)** of $\theta$.



**Example 1.15:** Suppose that $X_1, X_2, \cdots, X_n$ are mutually independently and identically distributed with mean $\mu$ and variance $\sigma^2$.
Assume that $\sigma^2$ is known.
Then, it is shown that $\overline{X}$ is a consistent estimator of $\mu$.
For RV $X$, Chebyshev's inequality is given by:

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

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 $\mathrm{V}(X_i) = \sigma^2 < \infty$ for all $i$.
Then, when $n \longrightarrow \infty$, we obtain the following result:

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

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

**Summary:**

When the distribution of $X_i$ is **not** assumed for all $i$, $\overline{X}$ is an **minimum variance linear unbiased and consistent estimator** of $\mu$.

When the distribution of $X_i$ is assumed to be **normal** for all $i$, $\overline{X}$ leads to an **efficient and consistent estimator** of $\mu$.



**Example 1.16a:** Suppose that $X_1, X_2, \cdots, X_n$ are mutually independently, identically and normally distributed with mean $\mu$ and variance $\sigma^2$.

Consider $S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$, which is an unbiased estimator of $\sigma^2$.

We obtain the following Chebyshev's inequality:

$$P(|S^2 - \sigma^2| \geq \epsilon) \leq \frac{\mathrm{E}((S^2 - \sigma^2)^2)}{\epsilon^2}.$$

We compute $\mathrm{E}((S^2 - \sigma^2)^2) \equiv \mathrm{V}(S^2)$.
$U = \dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$.



$\mathrm{E}(U) = n-1$ and $\mathrm{V}(U) = 2(n-1)$.
$\mathrm{V}(U) = \mathrm{V}(\dfrac{(n-1)S^2}{\sigma^2}) = 2(n-1)$
$\dfrac{(n-1)^2}{\sigma^4}\mathrm{V}(S^2) = 2(n-1)$
$\mathrm{V}(S^2) = \dfrac{2\sigma^2}{n-1}$

$$P(|S^2 - \sigma^2| \geq \epsilon) \leq \frac{\mathrm{E}((S^2 - \sigma^2)^2)}{\epsilon^2} = \frac{2\sigma^2}{(n-1)\epsilon^2} \longrightarrow 0,$$

which implies that $S^2 \longrightarrow \sigma^2$ as $n \longrightarrow \infty$.
Threfore, $S^2$ is a consistent estimator of $\sigma^2$.



**Example 1.16b:** Suppose that $X_1, X_2, \cdots, X_n$ are mutually independently, identically and normally distributed with mean $\mu$ and variance $\sigma^2$.

Consider $S^{**2} = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$, which is an estimate of $\sigma^2$.

We obtain the following Chebyshev's inequality:

$$P(|S^{**2} - \sigma^2| \geq \epsilon) \leq \frac{\mathrm{E}((S^{**2} - \sigma^2)^2)}{\epsilon^2}.$$

We compute $\mathrm{E}((S^{**2} - \sigma^2)^2)$.

Define $S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ as an estimator $\sigma^2$.



From $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$, we obtain $\mathrm{E}\Big(\dfrac{(n-1)S^2}{\sigma^2}\Big) = n-1$ and $\mathrm{V}\Big(\dfrac{(n-1)S^2}{\sigma^2}\Big) = 2(n-1)$.

Therefore, $\mathrm{E}(S^2) = \sigma^2$ and $\mathrm{V}(S^2) = \dfrac{2\sigma^4}{n-1}$ can be derived.



Using $S^{**2} = \dfrac{n-1}{n}S^2$, we have the following:

$$\begin{aligned}
\mathrm{E}((S^{**2} - \sigma^2)^2) &= \mathrm{E}\Big(\Big(\frac{n-1}{n}S^2 - \sigma^2\Big)^2\Big) \\
&= \mathrm{E}\Big(\Big(\frac{n-1}{n}(S^2 - \sigma^2) - \frac{\sigma^2}{n}\Big)^2\Big) \\
&= \frac{(n-1)^2}{n^2}\mathrm{E}((S^2 - \sigma^2)^2) + \frac{\sigma^4}{n^2} \\
&= \frac{(n-1)^2}{n^2}\mathrm{V}(S^2) + \frac{\sigma^4}{n^2} = \frac{(2n-1)}{n^2}\sigma^4.
\end{aligned}$$

Therefore, as $n \longrightarrow \infty$, we obtain:

$$P(|S^{**2} - \sigma^2| \geq \epsilon) \leq \frac{1}{\epsilon^2}\frac{(2n-1)}{n^2}\sigma^4 \longrightarrow 0.$$

Because $S^{**2} \longrightarrow \sigma^2$, $S^{**2}$ is a consistent estimator of $\sigma^2$.

$S^{**2}$ is biased (see Section 7.3, p.273), but is is consistent.



## 7.5 Estimation Methods

• **Maximum Likelihood Estimation Method (最尤推定法)**

• **Least Squares Estimation Method (最小二乗法)**

• **Method of Moment (積率法)**



### 7.5.1 Maximum Likelihood Estimator (最尤推定量)

In Section 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 $\mu$ under normality assumption and that $S^2$ is an unbiased and consistent estimator of $\sigma^2$.

The parameter $\theta$ is included in the underlying distribution $f(x;\theta)$.



$\theta = (\mu, \sigma^2)$ in the case of the normal distribution.

Now, in more general cases, we want to consider how to estimate $\theta$.

The **maximum likelihood estimator (最尤推定量)** 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 $f(x;\theta)$.



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 $\theta$ 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 $\theta$, 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 (尤度関数)**.



Let $\hat{\theta}_n$ be the $\theta$ which maximizes the likelihood function.
Given data $x_1$, $x_2$, $\cdots$, $x_n$, $\hat{\theta}_n(x_1, x_2, \cdots, x_n)$ is called the **maximum likelihood estimate (MLE, 最尤推定値)**.
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 (MLE, 最尤推定量)**.
That is, solving the following equation:

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

MLE $\hat{\theta}_n \equiv \hat{\theta}_n(X_1, X_2, \cdots, X_n)$ is obtained.

**Example 1.17a:** Suppose that $X_1$, $X_2$, $\cdots$, $X_n$ are mutually independently, identically and normally distributed with mean $\mu$ and variance $\sigma^2$.

We derive the maximum likelihood estimators of $\mu$ and $\sigma^2$.

The joint density (or the likelihood function) of $X_1$, $X_2$, $\cdots$, $X_n$ is:

$$f(x_1, x_2, \cdots, x_n; \mu, \sigma^2) = \prod_{i=1}^{n} f(x_i; \mu, \sigma^2)$$
$$= \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\Big(-\frac{1}{2\sigma^2}(x_i - \mu)^2\Big)$$



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

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 (対数尤度関数)**.



For maximization of the likelihood function, differentiating the log-likelihood function $\log l(\mu, \sigma^2)$ with respect to $\mu$ and $\sigma^2$, 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 $\mu$ and $\sigma^2$ are given by $\overline{X}$ and $S^{**2}$.



Since $\mathrm{E}(\overline{X}) = \mu$, the maximum likelihood estimator of $\mu$, $\overline{X}$, is an unbiased estimator.

We have checked that $\overline{X}$ is efficient and consistent.

However, because of $\mathrm{E}(S^{**2}) = \dfrac{n-1}{n}\sigma^2 \neq \sigma^2$ as shown in Section 7.3, the maximum likelihood estimator of $\sigma^2$, $S^{**2}$, is not an unbiased estimator.

We have checked that $S^{**2}$ is inefficient but consistent.



**Example 1.17b:** Suppose that $X_1$, $X_2$, $\cdots$, $X_n$ are mutually independently and identically distributed as Bernoulli random variables with parameter $p$.

We derive the maximum likelihood estimators of $p$.

The joint density (or the likelihood function) of $X_1$, $X_2$, $\cdots$, $X_n$ is:

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

The log-likelihood function is given by:

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

For maximization of the likelihood function, differentiating the log-likelihood function $\log l(p)$ with respect to $p$ , the first derivatives should be equal to zero, i.e.,

$$\begin{aligned} \frac{\mathrm{d} \log l(p)}{\mathrm{d}p} &= \frac{1}{p} \sum_{i=1}^{n} x_i - \frac{1}{1-p}(n - \sum_{i=1}^{n} x_i) \\ &= \frac{n}{p}\overline{x} - \frac{n}{1-p}(1-\overline{x}) = 0 \end{aligned}$$

Let $\hat{p}$ be the solution which satisfies the above equation.



We obtain the maximum likelihood estimates as follows:

$$\hat{p} = \overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i,$$

Replacing $x_i$ by $X_i$ for $i = 1, 2, \cdots, n$, the maximum likelihood estimator of $p$ is given by $\hat{p} = \overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$.



● We check whether $\hat{p}$ is unbiased.

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

Remember that $\mathrm{E}(X_i) = \sum_{x_i=0}^{1} x_i p^{x_i}(1-p)^{1-x_i} = p$, where $x_i$ takes 0 or 1.

Thus, $\hat{p}$ is an unbiased estimator of $p$.



● Next, we check whether $\hat{p}$ is efficient.

From Cramer-Rao inequality,

$$\mathrm{V}(\hat{p}) \geq -\frac{1}{n\mathrm{E}\Big(\frac{\mathrm{d}^2 \log f(X;p)}{\mathrm{d}p^2}\Big)}.$$

$$f(X;p) = p^X(1-p)^{1-X}$$

$$\begin{aligned} \log f(X;p) &= X \log(p) + (1-X) \log(1-p) \\ \frac{\mathrm{d} \log f(X;p)}{\mathrm{d}p} &= \frac{X}{p} - \frac{1-X}{1-p} \\ \frac{\mathrm{d}^2 \log f(X;p)}{\mathrm{d}p^2} &= -\frac{X}{p^2} - \frac{1-X}{(1-p)^2} \end{aligned}$$



We need to check whether the equality holds.

$$\begin{aligned} \mathrm{V}(\hat{p}) &= \mathrm{V}(\frac{1}{n} \sum_{i=1}^{n} X_i) = \frac{1}{n^2} \mathrm{V}(\sum_{i=1}^{n} X_i) = \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{V}(X_i) \\ &= \frac{1}{n^2} \sum_{i=1}^{n} p(1-p) = \frac{p(1-p)}{n}, \end{aligned}$$

Note as follows:

$\mathrm{V}(X_i) = \mathrm{E}((X_i - p)^2) = \sum_{x_i=0}^{1} (x_i - p)^2 p^{x_i}(1-p)^{1-x_i} = p(1-p).$



The Cramer-Rao lower bound is:

$$\begin{aligned} -\frac{1}{n\mathrm{E}\Big(\frac{\mathrm{d}^2 \log f(X;p)}{\mathrm{d}p^2}\Big)} &= -\frac{1}{n\mathrm{E}\Big(-\frac{X}{p^2} - \frac{1-X}{(1-p)^2}\Big)} \\ &= -\frac{1}{n\Big(-\frac{\mathrm{E}(X)}{p^2} - \frac{1-\mathrm{E}(X)}{(1-p)^2}\Big)} = \frac{1}{n\Big(\frac{1}{p} + \frac{1}{1-p}\Big)} = \frac{p(1-p)}{n}, \end{aligned}$$

which is equal to $\mathrm{V}(\hat{p})$.

Thus, $\hat{p}$ is an efficient estimator of $p$.

● We check whether $\hat{p}$ is consistent.

From Chebyshev's inequality,

$$P(|\hat{p} - p| \geq \epsilon) \leq \frac{\mathrm{E}((\hat{p} - p)^2)}{\epsilon^2} = \frac{p(1-p)}{n\epsilon^2}.$$

As $n \longrightarrow \infty$, $P(|\hat{p} - p| \geq \epsilon) \longrightarrow 0$.

That is, $\hat{p}$ converges in probability to $p$.

Thus, $\hat{p}$ is a consistent estimator of $p$.



**Properties of Maximum Likelihood Estimator:** For **small sample (小標本)**, the MLE has the following properties.

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

  For instance, the MLE of $\sigma^2$, $S^{**2}$, is not unbiased.

  However, $\dfrac{n}{n-1}S^{**2} = S^2$ is an unbiased estimator of $\sigma^2$.

- If the efficient estimator exists, the maximum likelihood estimator is efficient.



Efficient estimator $\Longleftrightarrow$ The variance of the estimator is equal to the Cramer-Rao lower bound.

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

$$\sqrt{n}(\hat{\theta}_n - \theta) \longrightarrow N(0, \sigma^2(\theta)), \tag{23}$$

where

$$\sigma^2(\theta) = \frac{1}{\mathrm{E}\bigg(\Big(\dfrac{\partial \log f(X;\theta)}{\partial \theta}\Big)^2\bigg)} = -\frac{1}{\mathrm{E}\bigg(\dfrac{\partial^2 \log f(X;\theta)}{\partial \theta^2}\bigg)}.$$



(23) indicates that the MLE has consistency, **asymptotic unbiasedness (漸近不偏性)**, **asymptotic efficiency (漸近有効性)** and **asymptotic normality (漸近正規性)**.

Asymptotic normality of the MLE comes from the central limit theorem discussed in Section 6.3.

Even though the underlying distribution is not normal, i.e., even though $f(x;\theta)$ is not normal, the MLE 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.

By normalizing, as $n \longrightarrow \infty$, we obtain as follows:

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

$\sqrt{n}(\hat{\theta}_n - \theta)$ has the distribution, which does not depend on $n$.

$\sqrt{n}(\hat{\theta}_n - \theta) = O(1)$ is written, where $O()$ is a function $n$.

That is, $\hat{\theta}_n - \theta = n^{-1/2} \times O(1) = O(n^{-1/2})$.



As another representation, when $n$ is large, we can approximate the distribution of $\hat{\theta}_n$ as follows:

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

This implies that when $n \longrightarrow \infty$, $\hat{\theta}_n$ approaches the lower bound of Cramer-Rao inequality: $\dfrac{\sigma^2(\theta)}{n}$.

This property is called an asymptotic efficiency.

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

$$\frac{\hat\theta_n - \theta}{\sigma(\hat\theta_n)/\sqrt{n}} \longrightarrow N(0,1). \tag{24}$$

Practically, when $n$ is large, we approximately use:

$$\hat\theta_n \sim N\Big(\theta, \frac{\sigma^2(\hat\theta_n)}{n}\Big). \tag{25}$$



**Proof of (23):** By the central limit theorem (11) on p.254,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \log f(X_i;\theta)}{\partial\theta} \longrightarrow N\Big(0, \frac{1}{\sigma^2(\theta)}\Big), \tag{26}$$

where $\sigma^2(\theta)$ is defined in (14), i.e., $\mathrm{V}\Big(\frac{\partial \log f(X_i;\theta)}{\partial\theta}\Big) = \frac{1}{\sigma^2(\theta)}$.

Note that $\mathrm{E}\Big(\frac{\partial \log f(X_i;\theta)}{\partial\theta}\Big) = 0$.

Apply the central limit theorem, taking $\frac{\partial \log f(X_i;\theta)}{\partial\theta}$ as the $i$th random variable.



By the Taylor series expansion around $\hat\theta_n = \theta$,

$$\begin{aligned}
0 &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \log f(X_i;\hat\theta_n)}{\partial\theta} \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \log f(X_i;\theta)}{\partial\theta} + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial^2 \log f(X_i;\theta)}{\partial\theta^2}(\hat\theta_n - \theta) \\
&\quad + \frac{1}{2!}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial^3 \log f(X_i;\theta)}{\partial\theta^3}(\hat\theta_n - \theta)^2 + \cdots
\end{aligned}$$



The third and above terms in the right-hand side are:

$$\frac{1}{2!}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial^3 \log f(X_i;\theta)}{\partial\theta^3}(\hat\theta_n - \theta)^2 + \cdots \longrightarrow 0.$$

It can be shown that the sum of the above terms is equal to $O(n^{-1/2})$.
Note that $\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^3 \log f(X_i;\theta)}{\partial\theta^3} \longrightarrow \mathrm{E}\Big(\frac{\partial^3 \log f(X_i;\theta)}{\partial\theta^3}\Big)$ from Chebyshev's inequality.
In addition, for now, we consider $\sqrt{n}(\hat\theta_n - \theta)^2 \longrightarrow 0$ as $n \longrightarrow \infty$. Actually, we obtain $\sqrt{n}(\hat\theta_n - \theta)^2 = O(n^{-1/2})$ from $\hat\theta_n - \theta = O(n^{-1/2})$.



Therefore,

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

which implies that the asy. dist. of $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \log f(X_i;\theta)}{\partial\theta}$ is equivalent to that of $-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial^2 \log f(X_i;\theta)}{\partial\theta^2}(\hat\theta_n - \theta)$.



From (26) and the above equations, we obtain:

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

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 comes from (14).

Thus, we have the following relationship:

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

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, (23) is obtained.



### 7.5.2 Least Squares Estimation Method (最小二乗法)

$X_1, X_2, \cdots, X_n$ are mutually independently distributed with mean $\mu$.
$x_1, x_2, \cdots, x_n$ are generated from $X_1, X_2, \cdots, X_n$, respectively.
Solve the following problem:

$$\min_{\mu} S(\mu), \qquad \text{where}\quad S(\mu)=\sum_{i=1}^{n}(x_i-\mu)^2.$$

Let $\hat\mu$ be the least squares estimate of $\mu$.

$$\frac{\mathrm{d}S(\mu)}{\mathrm{d}\mu}=0 \quad\Longrightarrow\quad \hat\mu=\frac{1}{n}\sum_{i=1}^{n}x_i$$



The least squares estimator is given by:

$$\hat\mu=\frac{1}{n}\sum_{i=1}^{n}X_i,$$

which is equivalent to MLE.



### 7.5.3 Method of Moment (積率法)

The distribution of $X_i$ is $f(x;\theta)$.
Let $\mu'_k$ be the $k$th moment.
From the definition of the $k$th moment,

$$\mathrm{E}(X^k)=\mu'_k$$

where $\mu'_k$ depends on $\theta$.
Let $\hat\mu'_k$ be the estimate of the $k$th moment.

$$\mathrm{E}(X^k)\approx\frac{1}{n}\sum_{i=1}^{n}x_i^k=\hat\mu'_k$$



The estimator of $\mu'_k$ is:

$$\hat\mu'_k=\frac{1}{n}\sum_{i=1}^{n}X_i^k$$

**Example: $\theta=(\mu,\sigma^2)$:** Because we have two parameters, we use the 1st and 2nd moments.
$\mu'_1=\mathrm{E}(X)=\mu$
$\mu'_2=\mathrm{E}(X^2)=\mathrm{V}(X)+(\mathrm{E}(X))^2=\sigma^2+\mu^2$



Estimates:

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

Estimators:

$$\hat\mu=\frac{1}{n}\sum_{i=1}^{n}X_i=\overline{X} \qquad \hat\sigma^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2$$

## 7.6 Interval Estimation

In Sections 7.1 – 7.5.1, the point estimation is discussed.

It is important to know where the true parameter value of $\theta$ 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, $\theta_U(X_1, X_2, \cdots, X_n)$ and $\theta_L(X_1, X_2, \cdots, X_n)$, where

$$P(\theta_L(X_1, X_2, \cdots, X_n) < \theta < \theta_U(X_1, X_2, \cdots, X_n)) = 1 - \alpha. \tag{27}$$

(27) implies that $\theta$ lies on the interval $\Big(\theta_L(X_1, X_2, \cdots, X_n), \theta_U(X_1, X_2, \cdots, X_n)\Big)$ with probability $1 - \alpha$.



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:

$$\Big(\theta_L(x_1, x_2, \cdots, x_n),\ \theta_U(x_1, x_2, \cdots, x_n)\Big)$$

is called the $100 \times (1 - \alpha)\%$ **confidence interval (信頼区間)** of $\theta$.

Thus, estimating the interval is known as the **interval estimation (区間推定)**, which is distinguished from the point estimation.

In the interval, $\theta_L(x_1, x_2, \cdots, x_n)$ is known as the **lower bound** of the confidence interval, while $\theta_U(x_1, x_2, \cdots, x_n)$ is the **upper bound** of the confidence interval.



Given probability $\alpha$, the $\theta_L(X_1, X_2, \cdots, X_n)$ and $\theta_U(X_1, X_2, \cdots, X_n)$ which satisfies equation (27) are not unique.

For estimation of the unknown parameter $\theta$, it is more optimal to minimize the width of the confidence interval.

Therefore, we should choose $\theta_L$ and $\theta_U$ which minimizes the width $\theta_U(X_1, X_2, \cdots, X_n) - \theta_L(X_1, X_2, \cdots, X_n)$.



**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 $\mu$ and variance $\sigma^2$.

From the central limit theorem,

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

Replacing $\sigma^2$ 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 $\mu$, the following expression is obtained.

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

where $\hat{\theta}_L$ and $\hat{\theta}_U$ correspond to $\overline{X} - z^{**}\dfrac{S}{\sqrt{n}}$ and $\overline{X} - z^*\dfrac{S}{\sqrt{n}}$, 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 $\mu$ 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, $\dfrac{\overline{X} - \mu}{S/\sqrt{n}}$ has a $t$ distribution with $n - 1$ degrees of freedom for any $n$.
Therefore, $100 \times (1 - \alpha)\%$ confidence interval of $\mu$ 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 point of the $t$ distribution with $n-1$ degrees of freedom.



**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 $\theta$.
From (25), we can approximate the $100 \times (1 - \alpha)\%$ confidence interval of $\theta$ as follows:

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



# 8 Testing Hypothesis (仮説検定)

## 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 $\theta$, e.g. $\theta = \theta_0$, is correct or not.
The hypothesis that we want to test is called the **null hypothesis (帰無仮説)**, 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 (対立仮説)**, which is denoted by $H_1 : \theta \neq \theta_0$.

**Type I and Type II Errors (第一種の誤り，第二種の誤り):** When we test the null hypothesis $H_0$, as shown in Table 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 (第一種の誤り)** and (iii) is called a **type II error (第二種の誤り)**.



The probability which a type I error occurs is called the **significance level (有意水準)**, which is denoted by $\alpha$, and the probability of committing a type II error is denoted by $\beta$.

Probability of (iv) is called the **power (検出力)** or the **power function (検出力関数)**, because it is a function of the parameter $\theta$.



Table 1: Type I and Type II Errors

| | $H_0$ is true. | $H_0$ is false. |
|---|---|---|
| Acceptance of $H_0$ | Correct judgment | **Type II Error**<br>第二種の誤り<br>(Probability $\beta$) |
| Rejection of $H_0$ | **Type I Error**<br>第一種の誤り<br>(Probability $\alpha$<br>= **Significance Level**)<br>有意水準 | Correct judgment<br>($1-\beta$ = **Power**)<br>検出力 |



**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 (検定等計量)**.

   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.

   When 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 (棄却域)** 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 (採択域)**, 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 (第一種の誤り)**, i.e., the **significance level (有意水準)** $\alpha$, 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 $\alpha = 0.1, 0.05, 0.01$ is chosen in practice.



The probability of committing a **type II error (第二種の誤り)**, i.e., $\beta$, 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 (検出力，または，検定力)** is defined as $1 - \beta$,

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

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



## 8.2 Power Function (検出力関数)

Let $X_1, X_2, \cdots, X_n$ be mutually independently, identically and normally distributed with mean $\mu$ and variance $\sigma^2$.

Assume that $\sigma^2$ is known.

In Figure 3, we consider:

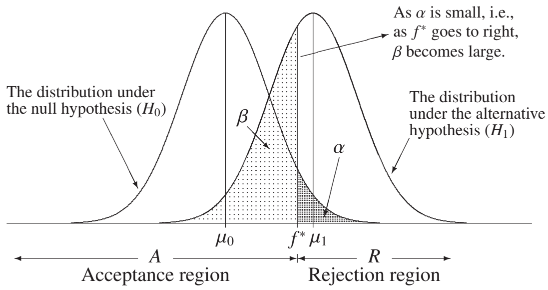
the null hypothesis $H_0: \ \mu = \mu_0$,

the alternative hypothesis $H_1: \ \mu = \mu_1$,

where $\mu_1 > \mu_0$ is taken.



Figure 3: Type I Error ($\alpha$) and Type II Error ($\beta$)






The dark shadow area (probability $\alpha$) corresponds to the probability of a **type I error**, i.e., the **significance level**, while the light shadow area (probability $\beta$) indicates the probability of 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., $1 - \beta$.



The distribution of sample mean $\overline{X}$ is given by:

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

By normalization, we have:

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

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 $\mu$ is replaced by $\mu_0$.

Since the significance level $\alpha$ is the probability which rejects $H_0$ when $H_0$ is true, it is given by:

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

where $z_\alpha$ denotes $100 \times \alpha$ percent point of $N(0,1)$.

Therefore, the rejection region is given by: $\overline{X} > \mu_0 + z_\alpha \dfrac{\sigma}{\sqrt{n}}$.



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

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

where $F(\cdot)$ represents the standard normal cumulative distribution function, which is given by:

$$F(x) = \int_{-\infty}^{x} (2\pi)^{-1/2} \exp(-\frac{1}{2}t^2)\,\mathrm{d}t.$$

The power function is a function of $\mu_1$, given $\mu_0$ and $\alpha$.



## 8.3 Small Sample Test (小標本検定)

### 8.3.1 Testing Hypothesis on Mean

**Known $\sigma^2$:** Let $X_1, X_2, \cdots, X_n$ be mutually independently, identically and normally distributed with $\mu$ and $\sigma^2$.

Consider testing the null hypothesis $H_0: \mu = \mu_0$.

When the null hypothesis $H_0$ is true, the distribution of $\overline{X}$ is:

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



Therefore, the test statistic is given by: $\dfrac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}$.

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

1. **The alternative hypothesis $H_1: \mu < \mu_0$ (one-sided test，片側検定):** 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 $\alpha$.



2. **The alternative hypothesis $H_1: \mu > \mu_0$ (one-sided test，片側検定):** 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 $\alpha$.

3. **The alternative hypothesis $H_1: \mu \neq \mu_0$ (two-sided test，両側検定):** We have: $P\Big(\Big|\dfrac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}\Big| > z_{\alpha/2}\Big) = \alpha$. Therefore, when $\Big|\dfrac{\overline{x}-\mu_0}{\sigma/\sqrt{n}}\Big| > z_{\alpha/2}$, we reject the null hypothesis $H_0: \mu = \mu_0$ at the significance level $\alpha$.



**Unknown $\sigma^2$:** Let $X_1, X_2, \cdots, X_n$ be mutually independently, identically and normally distributed with $\mu$ and $\sigma^2$.

Test the null hypothesis $H_0: \mu = \mu_0$.

When the null hypothesis $H_0$ is true, the distribution of $\overline{X}$ is given by:

$$\frac{\overline{X}-\mu_0}{S/\sqrt{n}} \sim t(n-1).$$

Therefore, the test statistic is given by: $\dfrac{\overline{X}-\mu_0}{S/\sqrt{n}}$.

### 8.3.2 Testing Hypothesis on Variance

**Testing Hypothesis on Variance:** Let $X_1, X_2, \cdots, X_n$ be mutually independently, identically and normally distributed with $\mu$ and $\sigma^2$.

Test the null hypothesis $H_0: \ \sigma^2 = \sigma_0^2$.

When the null hypothesis $H_0$ is true, the distribution of $S^2$ is given by:

$$\frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2(n-1)$$

**Testing Equality of Two Variances:** Let $X_1, X_2, \cdots, X_n$ be mutually independently, identically and normally distributed with $\mu_x$ and $\sigma_x^2$.



Let $Y_1, Y_2, \cdots, Y_m$ be mutually independently, identically and normally distributed with $\mu_y$ and $\sigma_y^2$.

Test the null hypothesis $H_0: \ \sigma_x^2 = \sigma_y^2$.

$$\frac{(n-1)S_x^2}{\sigma_x^2} \sim \chi^2(n-1), \quad \text{where } S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

$$\frac{(m-1)S_y^2}{\sigma_y^2} \sim \chi^2(m-1), \quad \text{where } S_y^2 = \frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \overline{Y})^2$$

Both are independent.



Then, the ratio of two $\chi^2$ random variables divided by degrees of freedom is:

$$\frac{\dfrac{(n-1)S_x^2}{\sigma_x^2}\Big/(n-1)}{\dfrac{(m-1)S_y^2}{\sigma_y^2}\Big/(m-1)} \sim F(n-1, m-1)$$

Therefore, under the null hypothesis $H_0: \sigma_x^2 = \sigma_y^2$,

$$\frac{S_x^2}{S_y^2} \sim F(n-1, m-1)$$



## 8.4 Large Sample Test (大標本検定)

- **Wald Test (ワルド検定)**

- **Likelihood Ratio Test (尤度比検定)**

- **Lagrange Multiplier Test (ラグランジェ乗数検定)**

  $\longrightarrow$ Skipped in this class.



### 8.4.1 Wald Test (ワルド検定)

From (24), under the null hypothesis $H_0: \ \theta = \theta_0$ (scalar case), as $n \longrightarrow \infty$, the maximum likelihood estimator $\hat\theta_n$ is distributed as:

$$\frac{\hat\theta_n - \theta_0}{\sigma(\hat\theta_n)/\sqrt{n}} \longrightarrow N(0,1).$$

Or, equivalently,

$$\Big(\frac{\hat\theta_n - \theta_0}{\sigma(\hat\theta_n)/\sqrt{n}}\Big)^2 \longrightarrow \chi^2(1).$$



For $H_0: \theta = \theta_0$ and $H_1: \theta \neq \theta_0$, replacing $X_1, \cdots, X_n$ in $\hat\theta_n$ by the observed values $x_1, \cdots, x_n$, the testing procedure is as follows.

When we have: $\Big(\dfrac{\hat\theta_n - \theta_0}{\sigma(\hat\theta_n)/\sqrt{n}}\Big)^2 > \chi_\alpha^2(1)$, we reject the null hypothesis $H_0$ at the significance level $\alpha$.

$\chi_\alpha^2(1)$ denotes the $100\times\alpha$ % point of the $\chi^2$ distribution with one degree of freedom.

This testing procedure is called the **Wald test (ワルド検定)**.

**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 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 $\gamma$, $\hat{\gamma}_n$, is asymptotically represented as:

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

or, equivalently

$$\Big(\frac{\hat{\gamma}_n - \gamma}{\sigma(\hat{\gamma}_n)/\sqrt{n}}\Big)^2 \longrightarrow \chi^2(1),$$

where

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



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

$$\Big(\frac{\hat{\gamma}_n - \gamma_0}{\sigma(\hat{\gamma}_n)/\sqrt{n}}\Big)^2 \longrightarrow \chi^2(1).$$

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

$$\Big(\frac{\hat{\gamma}_n - \gamma_0}{\sigma(\hat{\gamma}_n)/\sqrt{n}}\Big)^2 > \chi^2_\alpha(1),$$

we can reject $H_0$ at the significance level $\alpha$.

We need to derive $\sigma^2(\gamma)$ and $\hat{\gamma}_n$ for the testing procedure.



First, $\sigma^2(\gamma)$ is given by:

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

Note that the first- and the second-derivatives of $\log f(X;\gamma)$ with respect to $\gamma$ 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 $\gamma$, 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 $\gamma$ 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 MLE of $\gamma$, $\hat{\gamma}_n$, 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 $1/\overline{X}$.

Or, equivalently,

$$\Big(\frac{\hat{\gamma}_n - \gamma}{\sigma(\hat{\gamma}_n)/\sqrt{n}}\Big)^2 = \Big(\frac{\hat{\gamma}_n - \gamma}{\hat{\gamma}_n/\sqrt{n}}\Big)^2 \longrightarrow \chi^2(1).$$

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

$$\Big(\frac{\hat{\gamma}_n - \gamma_0}{\hat{\gamma}_n / \sqrt{n}}\Big)^2 > \chi^2_\alpha(1),$$

we reject $H_0$ at the significance level $\alpha$.



### 8.4.2 Likelihood Ratio Test (尤度比検定)

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 $(x_1, \cdots, x_n)$ corresponding to the random sample $(X_1, \cdots, X_n)$.

Let $\theta_1$ and $\theta_2$ be $1 \times k_1$ and $1 \times k_2$ vectors, respectively.

$\theta = (\theta_1, \theta_2)$ 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 $\theta_1$.

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 $(\theta_1, \theta_2)$.



That is, $(\widetilde{\theta}_1, \widetilde{\theta}_2)$ indicates the solution of $(\theta_1, \theta_2)$, obtained from the following equations:

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

The solution $(\widetilde{\theta}_1, \widetilde{\theta}_2)$ is called the **unconstrained maximum likelihood estimator (制約なし最尤推定量)**, 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 $\theta_2$ 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 (制約つき最尤推定量)** of $\theta_2$, because the likelihood function is maximized with respect to $\theta_2$ subject to the constraint $\theta_1 = \theta_1^*$.



Define $\lambda$ as follows:

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

which is called the **likelihood ratio (尤度比)**.

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

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

where $k_1$ denotes the number of the constraints.

Let $\chi^2_\alpha(k_1)$ be the $100 \times \alpha$ 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 $\alpha$.

This test is called the **likelihood ratio test (尤度比検定)**

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.



**Example 1.19:** $X_1, X_2, \cdots, X_n$ are mutually independently, identically and exponentially distributed.

Consider the exponential probability density function:

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

for $0 < x < \infty$.

Using the likelihood ratio test, we test the null hypothesis $H_0 : \gamma = \gamma_0$ against the alternative hypothesis $H_1 : \gamma \neq \gamma_0$.



The likelihood ratio is given by:

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

where $\hat\gamma_n$ is derived in Example 1.18, 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 $H_0 : \gamma = \gamma_0$ at the significance level $\alpha$.



**Example 1.20:** Suppose that $X_1, X_2, \cdots, X_n$ are mutually independently, identically and normally distributed with mean $\mu$ and variance $\sigma^2$.

The normal probability density function with mean $\mu$ and variance $\sigma^2$ 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 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:

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

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 $\sigma^2$.

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 estimate $\widetilde{\sigma}^2$ is:

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



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

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



For the denominator of the likelihood ratio, because the unconstrained maximum likelihood estimates 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:

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



Thus, the likelihood ratio is given by:

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

Asymptotically, we have:

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

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



# Exam

July 31, 2012

60–70% from 16 exercises (in my Web) and two homeworks