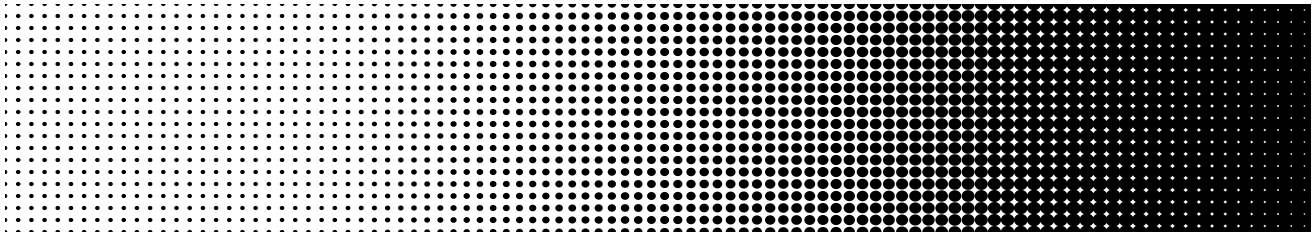
Mathematical Statistics

pehcy (MurphyShark) https://github.com/pehcy



Based on lectures UECM 3363 Mathematical Statistics, UCCM 2263 Applied Statistical Modelling, UECM 3213 Machine Learning in Universiti Tunku Abdul Rahman in 2019
Notes taken by MurphyShark.

Typeset with LaTeX, amsmath

Notations in this notes

Here are a list of symbols you will see a lot throughout this notes:

$\mu_X = \mathbb{E}[X]$ represents the mean or expectation of a random variable X.

$\sigma_X^2 = \text{Var}[X]$ represents the variance of a random variable X.

$\sigma = \sqrt{\text{Var}[X]}$ represents the standard deviation of a random variable X.

$\sigma_{XY} = \text{Cov}(X, Y)$ represents the covariance of two random variables X and Y.

$\rho_{XY} = \text{Corr}(X, Y)$ represents the correlation coefficient of two random variables X and Y.

$\mathbb{P}[X]$ denotes the probability measures of event X.

$\widehat{\beta}$ denotes the estimated value for the unknown parameter β.

This page intentionally left blank.

1

Set Theory and Counting Techniques

Two approaches of the concept of probability will introduced lately in this notes: The classical probability and the experimental probability. The sample space of probability theory is developed using the foundation of set theory.

In set theory, the number of elements in a set has a special name. It is called the **cardinality** of the set. In this notes we write $\#(A)$ to represent the cardinality of the set A.

1.1 Notion of sets

Set is a basic term in mathematics. One can think of a set to be a *class* or *collection*.

Definition 1.1 Sets
A set is a collection of objects called elements or numbers.

Definition 1.2 Empty set
The empty set is the set with no elements, denoted as $\varnothing$ or $\{\}$.

✱ **Example 1.1.1.** Throughout this notes, we are using the following number systems.

- The set of all nonzero positive integers, known as natural numbers

$$\mathbb{N} = \{1, 2, 3, \ldots\}.$$

- The set of all integers, regardless positive or negative,

$$\mathbb{Z} = \{\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots\}.$$

- The set of all rational numbers, the numbers that can be express as ratio of two integers

$$\mathbb{Q} = \left\{ \frac{p}{q} : p, q \in \mathbb{Z} \quad \text{with } q \neq 0 \right\}$$

- The set of all real numbers,

$$\mathbb{R} = \{\mathbb{Q} \text{ along with all irrational numbers such as } \sqrt{2}, \pi, e, \ldots\}.$$

Definition 1.3 Subsets
A set A is a subset of B, $A \subset B$, if $a \in A \implies a \in B$.

✱ Example 1.1.2. What is the cardinality of each of the following sets?

1. $\varnothing$.
2. $\{\varnothing\}$.
3. $A = \{\heartsuit, \{\clubsuit\}, \{\heartsuit, \{\heartsuit\}\}\}$.
4. $B = \{x | x \text{ is an integer such that } 2x^2 + 3x - 2 = 0\}$.

✲ Solution

1. Certainly, $\#(\varnothing) = 0$.
2. This is a set consists of one element $\varnothing$, thus $\#(\{\varnothing\}) = 1$.
3. The set A consists of three elements which are $\boxed{\heartsuit}$, $\boxed{\{\clubsuit\}}$, and $\boxed{\{\heartsuit, \{\heartsuit\}\}}$. So $\#(A) = 3$.
4. Solving the quadratic equation,

$$\begin{aligned} 2x^2 + 3x - 2 = 0 &\implies (2x-1)(x+2) = 0 \\ &\implies x = -2 \quad \text{or} \quad x = \frac{1}{2}. \end{aligned}$$

Only $x = -2$ is integer. Thus $\#(B) = \#(\{-2\}) = 1$.

◀

Definition 1.4 Power Set
If A is a set, then $\mathcal{P}(A) = \{B : B \subset A\}$.

✱ Example 1.1.3. If $A = \varnothing$, then $\mathcal{P}(A) = \{\varnothing\}$. The power set has only one single element, that is the empty set itself.

1.2 Set Operations

Theorem 1.1 Properties of set union and intersect
For any three sets, A, B, and C.

S1 **[Commutativity]**

$$A \cup B = B \cup A,$$
$$A \cap B = B \cap A.$$

S2 **[Associativity]**

$$A \cup (B \cup C) = (A \cup B) \cup C,$$
$$A \cap (B \cap C) = (A \cap B) \cap C,$$

S3 **[Distributive]**

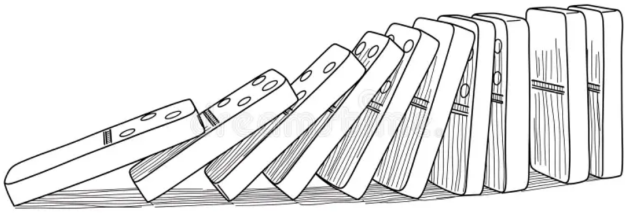
$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C),$$
$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

S4 **[DeMorgan's law]**

$$(A \cup B)^c = A^c \cap B^c,$$
$$(A \cap B)^c = A^c \cup B^c$$

1.2.1 Proof by induction

Mathematical induction is a proof technique for proving that a statement is true for all natural numbers. It works by first proving a base case (the first natural number, often 1) and then proving an inductive step which shows that if the statement holds for some arbitrary integer k.



Theorem 1.2 Mathematical Induction
Let $P(n)$ be a statement relies on $n \in \mathbb{N}$. Assume that

1. (Base case) $P(1)$ is true.

2. (Induction step) If $P(m)$ is true, then $P(m+1)$ is true.

Proof. Let

$$S = \{n \in \mathbb{N} : P(n) \text{ is not true}\}.$$

We want to show S is empty. We will show this by contradiction.

Intuition: By assuming $S \neq \varnothing$ and derive a false statement. The rules of logic $p \to q$ (If p then q) say that $S \neq \varnothing$ is false.

For the sake of contradiction, suppose $S \neq \varnothing$. By Well Ordering Principle of natural numbers, S has a least element $x \in S$. Given that the basis step $P(1)$ is true, then $1 \notin S \implies x \neq 1$. In particular, $x > 1$.

Since x is the least element of S, and $x - 1 < x$, thus $x - 1 \notin S$. From $S \neq \varnothing$, we are now arriving to the conclusion that $\exists x \in \mathbb{N}$ such that $x \in S$ and $x \notin S$ at the same time. This is a contradiction. Therefore S must be empty. □

1.3 Counting Techniques

Theorem 1.3 Multiplication rule of counting
If a choice consists of k steps, of which the first can be made in n_1 ways, for each of these the second can be made in n_2 ways and so on. For each of these k-th can be made in n_k ways, then the whole choice can be made in

$$n_1 \times n_2 \times \cdots \times n_k$$

ways.

Proof. In the notion of set theory, we use S_i to represent the set of outcomes for the i-th step for all $i = 1, 2, \ldots, k$. Then $\#(S_i) = n_i$. The set of outcomes for the entire job is the Cartesian product

$$S_1 \times S_2 \times \cdots \times S_k = \{(s_1, s_2, \ldots, s_k) : s_i \in S_i, \quad 1 \leq i \leq k\}. \tag{1.1}$$

Thus, we just need to show that the number of outcomes is equal to the product of number of choices for each step. That is,

$$\#(S_1 \times S_2 \times \cdots \times S_k) = \#S_1 \cdot \#S_2 \cdot \cdots \cdot \#S_k \tag{1.2}$$

[Basis Step] By Theorem 1.2.5, we have $\#(S_1 \times S_2) = \#(S_1) \times \#(S_2)$. Thus, the property is true for $n = 2$.

[Induction Hypothesis] Suppose

$$\#(S_1 \times S_2 \times \cdots \times S_k) = \#(S_1) \cdot \#(S_2) \cdots \#(S_k)$$

for $k = 2, 3, \cdots, n$.

[Induction Step] We must show

$$\#(S_1 \times S_2 \times \cdots \times S_{n+1}) = \#(S_1) \cdot \#(S_2) \cdots \#(S_{n+1}).$$

To see this, note that there is a one-to-one correspondence between the sets $S_1 \times S_2 \times \cdots \times S_{n+1}$ and $(S_1 \times S_2 \times \cdots \times S_n) \times S_{n+1}$ given by $f(s_1, s_2, \cdots, s_n, s_{n+1}) = ((s_1, s_2, \cdots, s_n), s_{n+1})$. See Problem 1.2.18. Thus,

$$\#(S_1 \times S_2 \times \cdots \times S_{n+1}) = \#((S_1 \times S_2 \times \cdots \times S_n) \times S_{n+1}) = \#(S_1 \times S_2 \times \cdots S_n)\#(S_{n+1})$$

(by Theorem 1.2.5). Now, applying the induction hypothesis gives

$$\#(S_1 \times S_2 \times \cdots S_n \times S_{n+1}) = \#(S_1) \cdot \#(S_2) \cdots \#(S_{n+1}).$$

□

Tutorials

Exercise 1.1 Use mathematical induction to show that for any positive integer n, $6^n - 1$ is divisible by 5.

Exercise 1.2 Show that $n! > 3^n$ for $n \geq 7$.

Exercise 1.3 Consider the Fibonacci sequence $\{x_n\}_{n=1}^{\infty}$, defined the relations $x_1 = 1, x_2 = 1$ and

$$x_n = x_{n-1} + x_{n-2} \quad \text{for } n \geq 3.$$

Use mathematical induction in order to show that for $n \geq 1$.

$$x_n = \frac{1}{\sqrt{5}} \left[\left(\frac{1 + \sqrt{5}}{\sqrt{2}} \right)^n - \left(\frac{1 - \sqrt{5}}{\sqrt{2}} \right)^n \right]$$

This page intentionally left blank.

2

Random Variables

2.1 Density function

By definition, a random variable X is a function with domain the sample space and range a subset of the real numbers. For example, in rolling two dice X might represent the sum of the points on the two dice. Similarly, in taking samples of college students X might represent the number of hours per week a student studies, a students GPA, or a students height. The notation $X(s) = x$ means that x is the value associated with the out come s by the random variable X

There are three types of random variables: discrete random variables, con tinuous random variables, and mixed random variables.

✱ Example 2.1.1. A committee of 4 is selected from a group consisting of 5 men and 5 women. Let X be the random variable that represents the number of women in the committee. Find the probability mass distribution of X.

✲ Solution For $x = 0, 1, 2, 3, 4$ we have

$$p_X(x) = \frac{\binom{5}{x}\binom{5}{4-x}}{\binom{10}{4}} \quad x = 0, 1, 2, 3, 4.$$

The probability mass function can be described by the table

x	0	1	2	3	4
$p(x)$	$\frac{5}{210}$	$\frac{50}{210}$	$\frac{100}{210}$	$\frac{50}{210}$	$\frac{5}{210}$

◀

2.2 Cumulative Distribution

First, we prove that the probability is a continuous set function. In order to do that, we need the following definitions:

Definition 2.1 Increasing and Decreasing sequence of events
A sequence of sets $\{E_n\}_{n=1}^{\infty}$ is said to be increasing if

$$E_1 \subset E_2 \subset \ldots \subset E_n \subset E_{n+1} \subset \ldots$$

whereas it is said to be a decreasing sequence if

$$E_1 \supset E_2 \supset \ldots \supset E_n \supset E_{n+1} \supset \ldots.$$

Lemma 2.1
If $\{E_n\}_{n\geq 1}$ is either an increasing or decreasing sequence of events then

$$\lim_{n\to\infty} \mathbb{P}[E_n] = \mathbb{P}[\lim_{n\to\infty} E_n]. \tag{2.1}$$

that is

$$\mathbb{P}\left[\bigcup_{n=1}^{\infty} E_n\right] = \lim_{n\to\infty} \mathbb{P}[E_n] \quad \text{for increasing sequence,} \tag{2.2}$$

and

$$\mathbb{P}\left[\bigcap_{n=1}^{\infty} E_n\right] = \lim_{n\to\infty} \mathbb{P}[E_n] \quad \text{for decreasing sequence,} \tag{2.3}$$

Proof. Firstly, suppose that $E_n \subset E_{n+1}$ for all $n \geq 1$. Define the events

$$\begin{aligned} F_1 &= E_1 \\ F_n &= E_n \cap E_{n-1}^c, \quad n > 1 \end{aligned}$$

Note that for $n > 1$, F_n consists of those outcomes in E_n that are not in any of the earlier E_n $\forall i < n$. Clearly, for $i \neq j$ we have $F_i \cap F_j = \varnothing$. Also, $\bigcup_{n=1}^{\infty} F_n = \bigcup_{n=1}^{\infty} E_n$ and for $n \geq 1$ we have $\bigcup_{i=1}^{n} F_i = \bigcup_{i=1}^{n} E_i$. From these properties we have

$$\begin{aligned}
\mathbb{P}\left[\lim_{n\to\infty} E_n\right] = \mathbb{P}\left[\bigcup_{n=1}^{\infty} E_n\right] &= \mathbb{P}\left[\bigcup_{n=1}^{\infty} F_n\right] \\
&= \sum_{n=1}^{\infty} \mathbb{P}[F_n] \\
&= \lim_{n\to\infty} \sum_{i=1}^{n} \mathbb{P}[F_i] \\
&= \lim_{n\to\infty} \mathbb{P}\left[\bigcup_{i=1}^{n} F_i\right] \\
&= \lim_{n\to\infty} \mathbb{P}\left[\bigcup_{i=1}^{n} E_i\right] \\
&= \lim_{n\to\infty} \mathbb{P}[E_n].
\end{aligned}$$

On the other hand, now suppose that the sequence $\{E_n\}_{n\geq 1}$ is a decreasing sequence of events. Then $\{E_n^c\}_{n\geq 1}$ is an increasing sequence of events. Hence, from the previous part we have

$$\mathbb{P}\left[\bigcup_{n=1}^{\infty} E_n^c\right] = \lim_{n\to\infty} [E_n^c].$$

By De Morgan's law we have $\bigcup_{n=1}^{\infty} E_n^c = (\bigcap_{n=1}^{\infty} E_n)^c$. And

$$\mathbb{P}\left[\left(\bigcap_{n=1}^{\infty} E_n\right)^c\right] = \lim_{n\to\infty} \mathbb{P}[E_n^c].$$

Equivalently,

$$1 - \mathbb{P}\left[\bigcap_{n=1}^{\infty} E_n\right] = \lim_{n\to\infty} (1 - \mathbb{P}[E_n]) = 1 - \lim_{n\to\infty} \mathbb{P}[E_n]$$

or

$$\mathbb{P}\left[\bigcap_{n=1}^{\infty} E_n\right] = \lim_{n\to\infty} \mathbb{P}[E_n].$$

□

Theorem 2.1 Properties of Cumulative Distribution Function
If $F_X(x)$ is a cumulative distribution function, then

1. $F_X(-\infty) = \lim_{x\downarrow -\infty} F_X(x) = 0$.
2. $F_X(+\infty) = \lim_{x\to +\infty} F_X(x) = 1$.
3. $F_X(x)$ is always *monotonically increasing*. That said, if $x_1 < x_2$, then $F_X(x_1) < F_X(x_2)$.

Proof. 1. Note that $\lim_{x\downarrow-\infty} F(x) = \lim_{n\to\infty} F(x_n)$ where $\{x_n\}$ is a decreasing sequence such that $x_n \downarrow -\infty$. Define

$$E_n = \{s \in \Omega : X(s) \leq x_n\}.$$

Then we have the nested chain $E_1 \supseteq E_2 \supseteq \ldots$. Moreover,

$$\varnothing = \bigcap_{n=1}^{\infty} E_n.$$

By previous proposition, we find

$$\lim_{x\to\infty} F(x) = \lim_{n\downarrow-\infty} F(x_n) = \lim_{n\to\infty} \mathbb{P}[E_n] = \mathbb{P}\left[\bigcap_{n=1}^{\infty} E_n\right] = \mathbb{P}[\varnothing] = 0.$$

2. In the other hand, suppose that $\lim_{x\to\infty} F(x) = \lim_{n\to\infty} F(x_n)$ where $\{x_n\}$ is a increasing sequence such that $x_n \to \infty$. We reuse back the definition of E_n that is

$$E_n = \{s \in \Omega : X(s) \leq x_n\}.$$

Then we have the nested chain in the opposite direction $E_1 \subseteq E_2 \subseteq \ldots$. Moreover,

$$\Omega = \bigcup_{n=1}^{\infty} E_n$$

By previous proposition, we find

$$\lim_{x\to\infty} F(x) = \lim_{n\to\infty} F(x_n) = \lim_{n\to\infty} \mathbb{P}[E_n] = \mathbb{P}\left[\bigcup_{n=1}^{\infty} E_n\right] = \mathbb{P}[\Omega] = 1.$$

3. Consider two real numbers a, b such that $a < b$. Then

$$\{s \in \Omega : X(s) \leq a\} \subset \{s \in \Omega : X(s) \leq b\}.$$

This implies that $\mathbb{P}[X \leq a] < \mathbb{P}[X \leq b]$. Hence, $F(a) < F(b)$.

□

✱ Example 2.2.1. Let X be a random variable with probability density function

$$f_X(x) = \begin{cases} 2 - 4|x| & \text{if } \frac{1}{2} < x < -\frac{1}{2} \\ 0 & \text{otherwise.} \end{cases}$$

1. Find the variance of X.
2. Find the cumulative function $F(x)$ of X.

✲ **Solution** 1. Since the density function $f(x)$ is odd in $\left(-\frac{1}{2}, \frac{1}{2}\right)$, we have $\mathbb{E}[X] = 0$. Therefore

$$\begin{aligned} Var[X] = \mathbb{E}[X^2] - 0 &= \int_{-1/2}^{0} x^2(2+4x)\,\mathrm{d}x + \int_{0}^{1/2} x^2(2-4x)\,\mathrm{d}x \\ &= \frac{1}{24}. \end{aligned}$$

2. The cumulative function is

$$\begin{aligned} F(x) &= \begin{cases} 0 & \text{if } x < -\frac{1}{2} \\ \int_{-1/2}^{x}(2+4t)\,\mathrm{d}t & \text{if } -\frac{1}{2} \le x \le 0 \\ \int_{-1/2}^{0}(2+4t)\,\mathrm{d}t + \int_{0}^{x}(2-4t)\,\mathrm{d}t & \text{if } 0 \le x \le \frac{1}{2} \\ 1 & \text{if } x > \frac{1}{2} \end{cases} \\ &= \begin{cases} 0 & \text{if } x < -\frac{1}{2} \\ 2x^2 + 2x + \frac{1}{2} & \text{if } -\frac{1}{2} \le x \le 0 \\ -2x^2 + 2x + \frac{1}{2} & \text{if } 0 \le x \le \frac{1}{2} \\ 1 & \text{if } x > \frac{1}{2} \end{cases} \end{aligned}$$

◀

2.3 Percentiles and Quantiles

2.3.1 Mode

In the discrete case, the mode is the value that is most likely to be sampled. In the continuous case, the mode is where $f(x)$ is at its peak.

✱ **Example 2.3.1.** The lifetime of a light bulb has density function, f_X, where $f_X(x)$ is proportional to $\frac{x^2}{1+x^3}$, $0 < x < 5$, and 0 otherwise.

Calculate the mode of this distribution.

✲ **Solution** Given the lifetime of a light bulb X has density function

$$f_X(x) = \frac{cx^2}{1+x^3}.$$

Compute the first and second order derivative of f.

$$\frac{\mathrm{d}f}{\mathrm{d}x} = \frac{(1+x^3)\frac{\mathrm{d}}{\mathrm{d}x}(cx^2) - cx^2\frac{\mathrm{d}}{\mathrm{d}x}(1+x^3)}{(1+x^3)^2} = \frac{2cx - cx^4}{(1+x^3)^2}.$$

$$\begin{aligned}
\frac{\mathrm{d}^2 f}{\mathrm{d}x^2} &= \frac{(1+x^3)^2\frac{\mathrm{d}}{\mathrm{d}x}(2cx - cx^4) - (2cx - cx^4)\frac{\mathrm{d}}{\mathrm{d}x}(1+x^3)^2}{(1+x^3)^4} \\
&= \frac{(1+x^3)^2(2c - 4cx^3) - (2cx - cx^4)2(1+x^3)(3x^2)}{(1+x^3)^4} \\
&= \frac{2c(1+x^3)(1 - 2x^3 - 3x^5)}{(1+x^3)^4} \\
&= \frac{2c(1 - 2x^3 - 3x^5)}{(1+x^3)^3}.
\end{aligned}$$

◀

By inspection, $\frac{\mathrm{d}^2 f}{\mathrm{d}x^2} < 0$. And so $\frac{\mathrm{d}f}{\mathrm{d}x} = 0$ is maximum point in $(0, 5)$. Solve for x of the following equation:

$$\frac{2cx - cx^4}{(1+x^3)^2} = 0$$

Since $(1+x^3)^2 > 0$, we can remove it safely from the equation. Then

$$\begin{aligned}
2cx - cx^4 = 0 &\implies x^4 - 2x = 0 \\
&\implies x(x^3 - 2) = 0 \\
&\implies x = 0 \text{ or } x = \sqrt[3]{2}.
\end{aligned}$$

Since x cannot be zero, thus the mode is $\sqrt[3]{2} = 1.26$.

2.4 Expected Value and Moments

For a random variable X, the expected value is denoted $\mathbb{E}[X]$, or μ_X or simply μ. The expected value is called the expectation of X, which is the "average" over the range of values that distribution X can be. You may said the expectation is the "center" of the distribution.

Definition 2.2 Expectation value
Let $(\Omega, \mathbb{P})$ be a probability space, let $E \subseteq \mathbb{R}$ be countable, and let X be a E-valued random

variable on $(\Omega, \mathbb{P})$. The expectation of X, if it exists, is defined by

$$\mathbb{E}[X] := \sum_{e \in E} e\, f_X(e). \tag{2.4}$$

Lemma 2.2
Let $(\Omega, \mathbb{P})$ be a probability space, let $E \subseteq \mathbb{R}$ be countable set and let X be an E-valued random variable on $(\Omega, \mathbb{P})$. Then

$$\mathbb{E}[X] = \sum_{\omega \in \Omega} X(\omega)\mathbb{P}[\{\omega\}].$$

Proof. Recall that

$$\Omega = \bigcup_{e \in E} \{X = e\}$$

and the events $\{X = e\}$ are mutually exclusive. Hence

$$\begin{aligned}
\sum_{\omega \in \Omega} X(\omega)\mathbb{P}[\{\omega\}] &= \sum_{e \in E} \sum_{\omega \in \{X=e\}} X(\omega)\mathbb{P}[\{\omega\}] \\
&= \sum_{e \in E} e\, \mathbb{P}\{X = e\} \\
&= \sum_{e \in E} e\, f_X(e)
\end{aligned}$$

as what we expected. □

✱ Example 2.4.1. Let X be a random variable representing the value shown a fair six-sided die is rolled. Then $X \sim$ discreteU$(\{1, 2, 3, 4, 5, 6\})$, and $f_X(k) = \frac{1}{6}$ for each number.

$$\mathbb{E}[X] = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6} = 3.5$$

So the expected value of a die rolled is 3.5.

✱ Example 2.4.2. If $f(x) = (k+1)x^2$ for $0 < x < 1$, find the moment generating function.

✲ Solution Since f is a density function, thus $\int_0^1 f(x)\mathrm{d}x = 1$, it follows that

$$(k+1) \times \frac{1}{3} = 1$$

so that $k = 2$ and now $f(x) = 3x^2$ for $0 < x < 1$. Then the moment generating function is

$$\begin{aligned} M_X(t) = \int_0^1 e^{tx}(3x^2)\,\mathrm{d}x &= \int_0^1 3x^2 \mathrm{d}\left(\frac{e^{tx}}{t}\right) \\ &= \frac{3x^2e^{tx}}{t}\bigg|_{x=0}^{x=1} - \int_0^1 \frac{6xe^{tx}}{t}\,\mathrm{d}x \\ &= \frac{3e^t}{t} - \left[\frac{6xe^{tx}}{t^2}\bigg|_{x=0}^{x=1} - \int_0^1 \frac{6xe^{tx}}{t^2}\,\mathrm{d}x\right] \\ &= \frac{3e^t}{t} - \frac{6e^t}{t^2} + \frac{6(e^t-1)}{t^3} \\ &= \frac{e^t(6-6t+3t^2)}{t^3} - \frac{6}{t^3}. \end{aligned}$$

◀

2.4.1 Variance

Variance is a measure of the "dispersion" of X about the mean.

Definition 2.3 Variance
The variance of distribution X is sum of squared loss

$$Var[X] := \mathbb{E}_X(X_i - \mu_X)^2 = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \tag{2.5}$$

A large variance indicates significant levels of probability or density from points far away from the mean. The variance must be always ≥ 0 (Since everything is squared). The variance of X is equal to zero only if X is a fixed single point and with probability 1 at that point; In other words, the function of X is a constant function (For example, $x \sim f_X(x) = 6$, then $Var[X] = 0$).

The standard deviation of the random variable X is the square root

Theorem 2.2 Chebyshev's Theorem
Let X be a random variable with mean μ_X and finite variance σ^2. Then,

$$\mathbb{P}[\,|X - \mu_X| < k\sigma] \geq 1 - \frac{1}{k^2} \tag{2.6}$$

or

$$\mathbb{P}[\,|X - \mu_X| \geq k\sigma] \leq \frac{1}{k^2} \tag{2.7}$$

for some constant $k > 0$.

2.4.2 The Coefficient of Variation

Definition 2.4 Coefficient of Variation
The coefficient of variation is

$$CV = \frac{\sigma_X}{\mu_X} = \frac{\sqrt{Var[x]}}{\mathbb{E}[X]}. \tag{2.8}$$

A higher CV implies greater variability, while a lower CV suggests more consistency or reliability of the data. Imagine if we have two datasets:

- Dataset A has a mean of 10 and standard deviation of 2, and $CV_A = 2/10 = 1/5$.
- Dataset B has a mean of 100 and standard deviation of 10, and $CV_B = 10/100 = 1/10$.

While dataset B has higher standard deviation, but it has a lower CV compared to dataset A. This indicating B less reliable variation to the mean.

2.5 Discrete Random Variables

2.5.1 Binomial distribution

A Bernoulli trial is an experiment with only two outcomes: **Success** and **failure**. The probability of a success is denoted by p and that of a failure by $q = 1 - p$. Moreover, p and q are related by the formula

$$p + q = 1.$$

A Bernoulli experiment is a sequence of independent Bernoulli trials. Let X represent the number of successes that occur in n independent Bernoulli trials. Then X is said to be a Binomial random variable (n, p). If $n = 1$, then X is said to be a Bernoulli random variable.

Theorem 2.3
Let $(\Omega, \mathbb{P})$ be a probability space, let $p \in [0, 1]$ and let $X_1, X_2, \ldots, X_n : \Omega \to \{0, 1\}$ be independent random variables such that each $X_i \sim \text{Bernoulli}(p)$. Then

$$X_1 + X_2 + \cdots + X_n \sim \text{Bin}(n, p).$$

2.5.2 Geometric distribution

A geometric random variable with parameter p, $0 < p < 1$ has a probability mass function

$$p_X(n) = \mathbb{P}(X = n) = p(1 - p)^{n-1}, \quad n = 1, 2, \ldots. \tag{2.9}$$

Note that $p_X(n) \geq 0$ and

$$\sum_{n=1}^{\infty} p(1-p)^{n-1} = \frac{p}{1-(1-p)} = 1. \tag{2.10}$$

A geometric random variable models the number of successive independent Bernoulli trials that must be performed to obtain the r-st success . For example, the number of flips of a fair coin until the r-st head appears follows a geometric distribution.

✱ Example 2.5.1. Consider the experiment of rolling a pair of fair dice.

1. What is the probability of getting a sum of 11?

2. If you roll two dice repeatedly, what is the probability that the first sum of 11 occurs on the 8-th roll?

✲ Solution 1. A sum of 11 occurs when the pair of dice show either $(5, 6)$ or $(6, 5)$ so that the required probability is $\frac{2}{36} = \frac{1}{18}$.

2. Let X be the number of rolls on which the first sum of 11 happened. Then X is a geometric random variable with probability $p = \frac{1}{18}$. Thus

$$\mathbb{P}[X = 8] = \frac{1}{18}\left(1 - \frac{1}{18}\right)^7 = 0.0372.$$

◀

2.6 Continuous Probability Distributions

2.6.1 Uniform Probability Distribution

Definition 2.5 Uniform Distribution
If $a < b$, a random variable X is said to have a continuous uniform distribution if

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & \text{if } a \leq x \leq b \\ 0 & \text{elsewhere.} \end{cases} \tag{2.11}$$

Theorem 2.4 Mean and variance of Uniform Distribution

If X is a continuous uniform distribution on the interval $[a, b]$, then the mean is

$$\mu_X = \mathbb{E}[X] = \frac{a+b}{2} \tag{2.12}$$

and

$$\sigma_X^2 = Var[X] = \frac{(b-a)^2}{12} \tag{2.13}$$

Proof. Given X is a continuous uniform distribution on the interval $[a, b]$, with $a < b$. Then the expectation of X is

$$\begin{aligned}
\mu_X = \mathbb{E}[X] = \int_a^b x \, \frac{1}{b-a} dx &= \frac{x^2}{2(b-a)}\Big|_{x=a}^{x=b} \\
&= \frac{b^2 - a^2}{2(b-a)} \\
&= \frac{(b-a)(b+a)}{2(b-a)} \\
&= \frac{a+b}{2}.
\end{aligned}$$

Now we continue to work on the variance for X. But before that, we need to find the expectation of X^2.

$$\begin{aligned}
\mathbb{E}[X^2] = \int_a^b x^2 \, \frac{1}{b-a} dx &= \frac{x^3}{3(b-a)}\Big|_{x=a}^{x=b} \\
&= \frac{b^3 - a^3}{3(b-a)} \\
&= \frac{(b-a)(b^2+ab+a^2)}{3(b-a)} \\
&= \frac{a^2+ab+b^2}{3}
\end{aligned}$$

The variance of X is

$$\begin{aligned}\sigma_X^2 = Var[X] &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \\ &= \frac{a^2+ab+b^2}{3} - \left(\frac{a+b}{2}\right)^2 \\ &= \frac{4(b^2+ab+a^2) - 3(a+b)^2}{12} \\ &= \frac{a^2-2ab+b^2}{12} \\ &= \frac{(b-a)^2}{12}\end{aligned}$$

which is what we expected. □

2.7 Normal Distribution

2.8 Gamma Distribution

Some random variables can yield distributions of data are skewed right and is non-symmetric.

Definition 2.6 Gamma Distribution
Let X be a random variable followed *gamma distribution* with parameters $\alpha > 0$ and β. The density function of X is

$$f_X(x) = \begin{cases} \dfrac{x^{\alpha-1}\, e^{-x/\beta}}{\beta^\alpha\, \Gamma(\alpha)}, & \text{if } x \geq 0 \\ 0 & \text{elsewhere.} \end{cases} \tag{2.14}$$

Theorem 2.5 Mean and variance of Gamma Distribution
If X is a gamma distribution with parameters $\alpha > 0$ and $\beta > 0$, then the mean and variance are

$$\mu_X = \alpha\beta \tag{2.15}$$

$$\sigma^2 = \alpha\beta^2. \tag{2.16}$$

Proof. Using the moment generating function approach to find mean and variance,

$$\begin{aligned}
M_X(t) &= \int_0^\infty e^{tx}\,\frac{x^{\alpha-1}\,e^{-x/\beta}}{\beta^\alpha\,\Gamma(\alpha)}\,dx \\
&= \frac{1}{\beta^\alpha}\int_0^\infty \frac{x^{\alpha-1}\,e^{-x\left(\frac{1}{\beta}-t\right)}}{\Gamma(\alpha)}\,dx \\
&= \frac{1}{\beta^\alpha}\int_0^\infty \frac{x^{\alpha-1}\,e^{-x/\left(\frac{\beta}{1-\beta t}\right)}}{\Gamma(\alpha)}\,dx \\
&= \frac{\left(\frac{\beta}{1-\beta t}\right)^\alpha}{\beta^\alpha}\int_0^\infty \frac{x^{\alpha-1}\,e^{-x/\left(\frac{\beta}{1-\beta t}\right)}}{\Gamma(\alpha)\left(\frac{\beta}{1-\beta t}\right)^\alpha}\,dx
\end{aligned}$$

Now observe that

$$\int_0^\infty \frac{x^{\alpha-1}\,e^{-x/\left(\frac{\beta}{1-\beta t}\right)}}{\Gamma(\alpha)\left(\frac{\beta}{1-\beta t}\right)^\alpha}\,dx = 1$$

as integrating x along the entire curve will give us 1. Thus,

$$M_X(t) = \frac{\left(\frac{\beta}{1-\beta t}\right)^\alpha}{\beta^\alpha} \times 1 = \frac{1}{(1-\beta t)^\alpha}. \tag{2.17}$$

That is to say, recall that the k-th derivative of $M_X(t)$ with respect to t as at $t = 0$, is the $\mathbb{E}[X^k]$. From here we compute

$$\mathbb{E}[X] = \dot{M}_X(0) = \alpha\beta(1-\beta t)^{-\alpha-1}\big|_{t=0} = \alpha\beta.$$
$$\mathbb{E}[X^2] = \ddot{M}_X(0) = -\alpha(1+\alpha)\beta^2(1-\beta t)^{-\alpha-2}\big|_{t=0} = \alpha(1+\alpha)\beta^2.$$

And so the variance is

$$\begin{aligned}
Var[X] &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \\
&= \alpha(\alpha+1)\beta^2 - (\alpha\beta)^2 \\
&= \alpha\beta^2.
\end{aligned}$$

and we are done. □

2.8.1 Chi-square distribution

Definition 2.7
If X is a gamma distribution with parameters $\alpha = \nu/2$ and $\beta = 2$, then X is a χ^2-distribution with ν degree of freedom. (with $\nu > 0$)

This page intentionally left blank.

3

Transformation of Random Variables

We begin with a random variable X with a given distribution. We wish to find the distribution of $Y = g(X)$ where $g : \mathbb{R} \to \mathbb{R}$ is some function.

The **inverse image** of a set $\mathcal{A}$ is defined as

$$g^{-1}(\mathcal{A}) = \{x \in \mathbb{R} \mid g(x) \in \mathcal{A}\} \tag{3.1}$$

In other words, x is in set $g^{-1}(\mathcal{A})$ if and only if $g(x) \in \mathcal{A}$.

✱ **Example 3.0.1.** If $g(x) = x^3$, then $g^{-1}([1, 8]) = [1, 2]$.

3.1 Transform of Discrete Random Variable

Theorem 3.1
For X a discrete random variable with probability mass function $f_X(x)$, the probability mass function of $Y = g(X)$ is

$$f_Y(y) = \sum_{x \in g^{-1}(y)} f_X(x). \tag{3.2}$$

✱ **Example 3.1.1.** Let X be a discrete random variable with probability mass function

x	-2	-1	0	1	2
$f_X(x)$	$\frac{1}{5}$	$\frac{1}{6}$	$\frac{1}{5}$	$\frac{1}{15}$	$\frac{11}{30}$

Find the pdf of $Y = X^2$.

✲ **Solution** The set of mass points of X is $\mathcal{A} = \{-2, -1, 0, 1, 2\}$. The transformation $Y = g(X) = X^2$

maps the set points to $\mathcal{B} = \{0, 1, 4\}$. Hence, the probability mass function of Y is

$$f_Y(y) = \mathbb{P}[Y = y] = \begin{cases} \mathbb{P}[Y = 0] = \mathbb{P}[X = 0] = \dfrac{1}{5}, & \text{if } y = 0, \\ \mathbb{P}[Y = 1] = \mathbb{P}[X = \pm 1] = f_X(1) + f_X(-1) = \dfrac{7}{30}, & \text{if } y = 1, \\ \mathbb{P}[Y = 4] = \mathbb{P}[X = \pm 2] = f_X(2) + f_X(-2) = \dfrac{17}{30}, & \text{if } y = 4, \\ 0, & \text{otherwise.} \end{cases}$$

◀

Theorem 3.2
If X is a random variable defined on $(\Omega, \mathscr{F}, P)$, and if g is a Borel-measurable function, then $g(X)$ is also a random variable defined on $(\Omega, \mathscr{F}, P)$.

Theorem 3.3
Given a random variable X with known distribution mass function, then the distribution mass function of $Y = g(X)$, where g is a Borel-measurable function, can be determined.

✱ Example 3.1.2. Suppose X is a random variable with known distribution function $F_X(x)$, show that the following functions are also random variables.

1. $|X|$.
2. $aX + b$, where a and b are constants.
3. X^k, where k is a positive integer.
4. $X_+ = \max\{X, 0\}$.
5. $X_- = \min\{X, 0\}$.

✲ Solution We can try expressing each transformation in terms of $F_X(x)$, the distribution function of X.

1. Let $g(x) = |x|$, the cumulative function of $Y = g(X)$ is

$$\begin{aligned} G_Y(y) &= \mathbb{P}(|X| \leq y) \\ &= \mathbb{P}(-y \leq X \leq y) \\ &= F_X(y) - F_X(-y). \end{aligned}$$

Since $F_X(x)$ is a distribution function, $G(y)$ is also a distribution function. Thus $|X|$ is a random variable.

2. Let $Y = aX + b$, the cumulative function of Y is

$$H_Y(y) = \mathbb{P}(aX + b \le y) = \begin{cases} \mathbb{P}\left(X \le \dfrac{y-b}{a}\right) & \text{if } a > 0, \\ \mathbb{P}\left(X \ge \dfrac{y-b}{a}\right) & \text{if } a < 0. \\ 0 & \text{if } a = 0. \end{cases}$$
$$= \begin{cases} F_X\left(\dfrac{y-b}{a}\right) & \text{if } a > 0, \\ 1 - F_X\left(\dfrac{y-b}{a}\right) & \text{if } a < 0. \\ 0 & \text{if } a = 0. \end{cases}$$

3. For $Y = X^k$, where k is a positive integer, the cumulative function of Y is

$$\breve{H}_Y(y) = \mathbb{P}(X^k \le y)$$
$$= \begin{cases} \mathbb{P}(X \le y^{1/k}) & \text{if } k \text{ is odd}, \\ \mathbb{P}(-y^{1/k} \le X \le y^{1/k}) & \text{if } k \text{ is even and } y \ge 0, \end{cases}$$
$$= \begin{cases} F_X(y^{1/k}) & \text{if } k \text{ is odd}, \\ F_X(y^{1/k}) - F_X(-y^{1/k}) & \text{if } k \text{ is even and } y \ge 0 \end{cases}.$$

4. The cumulative distribution function of $Y = X_+ = \max\{X, 0\}$ is

$$\mathbb{P}(X_+ \le y) = \mathbb{P}(\max\{X, 0\} \le y)$$
$$= \begin{cases} \mathbb{P}(0 \le y) = 0 & \text{if } y < 0, \\ \mathbb{P}(X_+ = 0) = \mathbb{P}(X \le 0) = F_X(0) & \text{if } y = 0, \\ \mathbb{P}(0 < X \le y) = F_X(y) & \text{if } y > 0. \end{cases}$$

5. The cumulative distribution function of $Y = X_- = \min\{X, 0\}$ is

$$\mathbb{P}(X_- \le y) = \mathbb{P}(\min\{X, 0\} \le y)$$
$$= \begin{cases} \mathbb{P}(X_- \le 0) = 1 & \text{if } y > 0, \\ \mathbb{P}(X \le y) = F_X(y) & \text{if } y \le 0. \end{cases}$$

◀

✱ Example 3.1.3. Let X be a Poisson random variable with probability mass function

$$f_X(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots$$

Find the distribution of $Y = \mathbb{P}(Y = y)$, if $Y = X^2 + 3$.

✲ Solution The transformation $Y = X^2+3$ maps the set points $\mathcal{A} = \{0, 1, 2, \ldots\}$ onto $\mathcal{B} = \{3, 4, 7, 12, \ldots\}$. The inverse mapping is $X = \sqrt{Y - 3}$. Since there is no negative values in $\mathcal{A}$, so we take only the positive square root.

Thus, the probability mass function of Y is

$$\begin{aligned}\mathbb{P}(Y = y) &= \mathbb{P}(X = \sqrt{y-3}) \\ &= \frac{e^{-\lambda}\lambda^{\sqrt{y-3}}}{(\sqrt{y-3})!}, \quad y \in \mathcal{B}.\end{aligned}$$

◀

✱ Example 3.1.4. Given $X \sim BIN(n, p)$ for $x = 0, 1, 2, \ldots, n$. Find the probability mass function of

1. $Y = aX + b$,
2. $Z = X^2$,
3. $W = \sqrt{X}$.

✲ Solution 1. Since $Y = aX + b$ maps the set points $\mathcal{A} = \{0, 1, 2, \ldots, n\}$ onto $\mathcal{B} = \{b, a + b, 2a + b, \ldots, na + b\}$, the inverse mapping is $X = \frac{Y-b}{a}$. And since there is no negative values in $\mathcal{A}$, so we take only the positive value. The pdf of Y is

$$\mathbb{P}(Y = y) = \mathbb{P}\left(X = \frac{y-b}{a}\right) = \binom{n}{\frac{y-b}{a}} p^{\frac{y-b}{a}} (1-p)^{n-\frac{y-b}{a}}, \quad y \in \mathcal{B}.$$

2. The transformation $Z = X^2$ maps the set points $\mathcal{A} = \{0, 1, 2, \ldots, n\}$ onto $\mathcal{B} = \{0, 1, 4, 9, \ldots, n^2\}$. The inverse mapping is $X = \sqrt{Z}$. Also since there is no negative values in $\mathcal{A}$, so we take only the positive square root. The pdf of Z is

$$\begin{aligned}\mathbb{P}(Z = z) &= \mathbb{P}(X^2 = z) \\ &= \mathbb{P}(X = \sqrt{z}) \\ &= \binom{n}{\sqrt{z}} p^{\sqrt{z}}(1-p)^{n-\sqrt{z}}, \quad z \in \mathcal{B}.\end{aligned}$$

3. $W = \sqrt{X}$ maps the set points $\mathcal{A} = \{0, 1, 2, \ldots, n\}$ onto $\mathcal{B} = \{0, 1, \sqrt{2}, \ldots, \sqrt{n}\}$. The inverse mapping is $X = W^2$. Also since there is no negative values in $\mathcal{A}$, so we take only the positive square root. The pdf of W is

$$\mathbb{P}(W = w) = \mathbb{P}(X = w^2) = \binom{n}{w^2} p^{w^2}(1-p)^{n-w^2}, \quad w \in \mathcal{B}.$$

◀

3.2 Continuous random variable

Theorem 3.4
Let X be a random variable of continuous type with probability density function $f_X(x)$, and let $Y = g(X)$ be differentiable for all x and either

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{\mathrm{d}}{\mathrm{d}y} g^{-1}(y) \right|, \quad -\infty < y < \infty. \tag{3.3}$$

Proof. In this case, $g(x)$ is a increasing function. We compute the cumulative distribution of $Y = g(X)$ in terms of $F(x)$, the cumulative distribution function of X. Note that

$$F_Y(y) = \mathbb{P}[Y \leq y] = \mathbb{P}[g(X) \leq y] = \mathbb{P}[X \leq g^{-1}(y)] = F_X(g^{-1}(y)). \tag{♡}$$

Now use the chain rule of differentiation to differentiate (♡) with respect to y.

$$f_Y(y) = F_Y'(y) = \frac{\mathrm{d}}{\mathrm{d}y} F_X(g^{-1}(y)) = f_X(g^{-1}(y)) \frac{\mathrm{d}}{\mathrm{d}y} g^{-1}(y). \tag{★}$$

In the case where $g(x)$ is a decreasing function, we have

$$F_Y(y) = \mathbb{P}[Y \leq y] = \mathbb{P}[g(X) \leq y] = \mathbb{P}[X \geq g^{-1}(y)] = 1 - F_X(g^{-1}(y)).$$

and the density function is

$$f_Y(y) = F_Y'(y) = \frac{\mathrm{d}}{\mathrm{d}y} \left[1 - F_X(g^{-1}(y))\right] = -f_X(g^{-1}(y)) \frac{\mathrm{d}}{\mathrm{d}y} g^{-1}(y). \tag{♠}$$

Now combining (★) and (♠), we obtained

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{\mathrm{d}}{\mathrm{d}y} g^{-1}(y) \right|$$

which is valid for both increasing and decreasing functions $g(x)$. □

Remark. 1. *If the conditions of this theorem are violated, then we should return to the previous method.*

2. *If the pdf f vanishes outside an interval $[a, b]$ of finite length, we need only to assume that $g = f(x)$ is differentiable in (a, b) and either $g'(x) > 0$ or $g'(x) < 0$ throughout the interval. Then we can take $\alpha = \min\{g(a), g(b)\}$ and $\beta = \max\{g(a), g(b)\}$.*

✱ Example 3.2.1. If $X \sim UNIF(0, 1)$ and $Y = X^2$, find the probability density function of Y if

1. $Y = e^X$,
2. $Y = -2 \ln X$.

✲ **Solution** The probability density function of X is

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

1. To find the density function of $Y = e^X$, Let $g(x) = e^x$ and apparently $g'(x) = e^x$. The function g is monotonically increasing. The inverse function is

   $$x = g^{-1}(y) = \ln y, \quad y > 0.$$

   and its derivative is

   $$\frac{\mathrm{d}}{\mathrm{d}y} g^{-1}(y) = \frac{\mathrm{d}}{\mathrm{d}y} \ln y = \frac{1}{y}.$$

   Using the transformation formula, we have

   $$\begin{aligned} f_Y(y) &= f_X(g^{-1}(y)) \left| \frac{\mathrm{d}}{\mathrm{d}y} g^{-1}(y) \right| \\ &= f_X(\ln y) \left| \frac{1}{y} \right| \\ &= \begin{cases} \frac{1}{y} & \text{for } 0 < \ln y < 1 \implies 0 < y < e, \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

   Thus the density function of Y is

   $$f_Y(y) = \begin{cases} \frac{1}{y} & \text{for } 1 < y < e, \\ 0 & \text{otherwise.} \end{cases}$$

2. In another case, let $g(x) = -2 \ln x$. The function g is monotonically decreasing. The inverse function is

   $$x = g^{-1}(y) = e^{-y/2}, \quad y > 0.$$

   Using the transformation formula, we have

   $$\begin{aligned} f_Y(y) &= f_X(g^{-1}(y)) \left| \frac{\mathrm{d}}{\mathrm{d}y} g^{-1}(y) \right| \\ &= f_X(e^{-y/2}) \left| -\frac{1}{2} e^{-y/2} \right| \\ \implies f_Y(y) &= \begin{cases} \dfrac{1}{2} e^{-y/2} & \text{for } 0 < e^{-y/2} < 1 \implies 0 < y < \infty, \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

◀

Theorem 3.5
If the transformation $Y = g(X)$ is not one-to-one transformation from $\mathfrak{X}$ onto $\mathcal{D}$. For instance, for a point in $\mathfrak{D}$, there exists more than one point in $\mathfrak{X}$. Then $\mathfrak{X}$ can be partitioned into n mutually exclusive subsets $\mathfrak{X}_1, \mathfrak{X}_2, \ldots, \mathfrak{X}_n$, say, such that $y = g(x)$ is one-to-one from each $\mathfrak{X}_i$ onto $\mathfrak{D}$. Let $g_i^{-1}(y)$ denote the inverse function of $y = g(x)$ on $\mathfrak{X}_i$. Then the density function of Y is given by

$$f_Y(y) = \sum_{i=1}^{n} f_X(g_i^{-1}(y)) \left| \frac{\mathrm{d}}{\mathrm{d}y} g_i^{-1}(y) \right|, \quad \text{if } y \in \mathfrak{D}. \tag{3.4}$$

✱ Example 3.2.2. Let X be a normal random variable with mean 0 and variance 1. Find the density function of $Y = X^2$.

✲ Solution Here, let $X \sim N(0, 1)$ and the density function of X is

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad x \in \mathfrak{X} = \mathbb{R}.$$

Let $y = g(x) = x^2$. The transformation $y = x^2$ is not one-to-one from $\mathfrak{X} = \mathbb{R}$ onto $\mathfrak{D} = \{y \mid y \geq 0\}$. Note that

$$y = x^2 \implies x = \pm\sqrt{y}.$$

From here we can decompose $\mathfrak{X}$ into two disjoint subsets: $\mathfrak{X}_1 = \{x \in \mathbb{R} | x < 0\}$ and $\mathfrak{X}_2 = \{x \in \mathbb{R} | x > 0\}$. And now $y = x^2$ can be one-to-one from each $\mathfrak{X}_i$ onto $\mathfrak{D}$ for $i = 1, 2$. The inverse functions for each $\mathfrak{X}_i$ are

$$x = \begin{cases} g_1^{-1}(y) = -\sqrt{y}, & \text{for } x \in \mathfrak{X}_1, \\ g_2^{-1}(y) = \sqrt{y}, & \text{for } x \in \mathfrak{X}_2. \end{cases}$$

The pdf for $Y = X^2$ is

$$\begin{aligned} f_Y(y) &= f_X(g_1^{-1}(y)) \left| \frac{\mathrm{d}}{\mathrm{d}y} g_1^{-1}(y) \right| + f_X(g_2^{-1}(y)) \left| \frac{\mathrm{d}}{\mathrm{d}y} g_2^{-1}(y) \right|, \quad \text{if } y \in \mathfrak{D} \\ &= \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \left| -\frac{1}{2\sqrt{y}} \right| + \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \left| \frac{1}{2\sqrt{y}} \right| \\ &= \frac{2}{\sqrt{2\pi}} \frac{e^{-y/2}}{\sqrt{y}} \\ &= \frac{e^{-y/2}\, y^{-1/2}}{\sqrt{\pi} 2^{1/2}}, \quad \text{if } y > 0 \end{aligned}$$

Recall that $\sqrt{\pi} = \Gamma\left(\frac{1}{2}\right)$, we can rewrite the density function of Y as

$$f_Y(y) = \frac{y^{\frac{1}{2}-1} e^{-y/2}}{\Gamma\left(\frac{1}{2}\right) 2^{1/2}}, \quad \text{if } y > 0.$$

This is actually the pdf of a Gamma distribution with parameters $\alpha = \frac{1}{2}$ and $\beta = \frac{1}{2}$. So $Y = X^2 \sim Gamma\left(\frac{1}{2}, \frac{1}{2}\right)$. ◀

✱ Example 3.2.3. Let X be a random variable with probability density function

$$f_X(x) = \begin{cases} \dfrac{2x}{\pi^2} & \text{for } 0 < x < \pi, \\ 0 & \text{otherwise.} \end{cases}$$

Find the pdf of $Y = \sin X$.

✲ Solution Here we have $\mathfrak{X} = \{x \mid 0 < x < \pi\}$ and $\mathfrak{D} = \{y \mid 0 < y < 1\}$. However, $y = \sin x$ is not one-to-one from $\mathfrak{X}$ onto $\mathfrak{D}$. We can decompose $\mathfrak{X}$ into two disjoint subsets:

$$\mathfrak{X}_1 = \left\{x \,\middle|\, 0 < x < \frac{\pi}{2}\right\}$$

and

$$\mathfrak{X}_2 = \left\{x \,\middle|\, \frac{\pi}{2} \le x < \pi\right\}.$$

Then the inverse functions for each $\mathfrak{X}_i$ are

$$x = \begin{cases} g_1^{-1}(y) = \arcsin y, & \text{for } x \in \mathfrak{X}_1, \\ g_2^{-1}(y) = \pi - \arcsin y, & \text{for } x \in \mathfrak{X}_2. \end{cases}$$

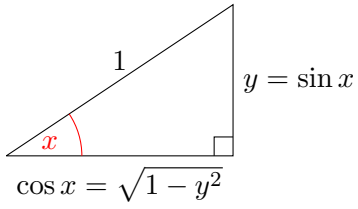
The derivative of each inverse function is

$$\frac{\mathrm{d}x}{\mathrm{d}y} = \begin{cases} \frac{\mathrm{d}}{\mathrm{d}y} g_1^{-1}(y) = \frac{1}{\sqrt{1-y^2}}, & \text{for } x \in \mathfrak{X}_1, \\ \frac{\mathrm{d}}{\mathrm{d}y} g_2^{-1}(y) = -\frac{1}{\sqrt{1-y^2}}, & \text{for } x \in \mathfrak{X}_2. \end{cases}$$

In case you want to find the derivative of $x = \arcsin y$ from scratch, we rewrite $y = \sin x$ and use implicit differentiation.

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \cos x \implies \frac{\mathrm{d}x}{\mathrm{d}y} = \frac{1}{\cos x}$$

Here we can draw a triangle to illustrate the relationship.



From the triangle, we have $\cos x = \sqrt{1-y^2}$. Thus, we get $\dfrac{\mathrm{d}x}{\mathrm{d}y} = \dfrac{1}{\cos x} = \dfrac{1}{\sqrt{1-y^2}}$.

Continue to work on finding the pdf of $Y = \sin x$. Now using the transformation formula, we

have

$$\begin{aligned} f_Y(y) &= f_X(g_1^{-1}(y)) \left| \frac{\mathrm{d}}{\mathrm{d}y} g_1^{-1}(y) \right| + f_X(g_2^{-1}(y)) \left| \frac{\mathrm{d}}{\mathrm{d}y} g_2^{-1}(y) \right|, \quad \text{if } y \in \mathfrak{D} \\ &= f_X(\arcsin y) \left| \frac{1}{\sqrt{1-y^2}} \right| + f_X(\pi - \arcsin y) \left| -\frac{1}{\sqrt{1-y^2}} \right| \\ &= \frac{2 \arcsin y}{\pi^2} \cdot \frac{1}{\sqrt{1-y^2}} + \frac{2(\pi - \arcsin y)}{\pi^2} \cdot \frac{1}{\sqrt{1-y^2}} \\ &= \frac{2}{\pi\sqrt{1-y^2}}, \quad 0 < y < 1. \end{aligned}$$

and now we are done. ◀

3.3 Transformation of Joint distribution

Theorem 3.6
Let X_1, X_2 be two random variables with joint density function $f_{X_1,X_2}(x_1, x_2)$. Let the sample space of (X_1, X_2) be

$$\mathfrak{X} = \{(x_1, x_2) \mid x_1 \in \mathbb{R}, x_2 \in \mathbb{R}\}.$$

The following properties hold:

JT1 The transformation $Y_1 = g_1(X_1, X_2)$ and $Y_2 = g_2(X_1, X_2)$ is one-to-one from $\mathfrak{X}$ onto $\mathfrak{D}$.

JT2 The 1st order partial derivatives of $x_1 = u(y_1, y_2)$ and $x_2 = v(y_1, y_2)$ with respect to y_1 and y_2 are continuous.

JT3 The jacobian

$$J = \begin{vmatrix} \dfrac{\partial u(x_1,x_2)}{\partial y_1} & \dfrac{\partial u(x_1,x_2)}{\partial y_2} \\ \dfrac{\partial v(x_1,x_2)}{\partial y_1} & \dfrac{\partial v(x_1,x_2)}{\partial y_2} \end{vmatrix} \tag{3.5}$$

is non-zero for all $(y_1, y_2) \in \mathfrak{D}$.

Then the joint density function of Y_1 and Y_2 is

$$f_{Y_1,Y_2}(y_1, y_2) = f_{X_1,X_2}(h(y_1, y_2), g(y_1, y_2)) \, |J|, \quad (y_1, y_2) \in \mathfrak{D}. \tag{3.6}$$

Remark. *The Jacobian J is also known as the local magnification factor.*

✱ **Example 3.3.1.** Let X and Y be two independent random variables with common probability

density function,

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

Find the distribution of $U = X - Y$ and $V = X + Y$.

✲ Solution The joint density function of X and Y is

$$f_{X,Y}(x,y) = f_X(x) f_Y(y) = \begin{cases} e^{-(x+y)} & x > 0, y > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Notice that X and Y are independent variables, the joint density function is the product of the marginal density functions of X and Y. Let $u = x - y$ and $v = x + y$. The inverse transformation is

$$x = \frac{u+v}{2}, \quad \text{and} \quad y = \frac{v-u}{2}. \tag{3.7}$$

such that,

$$\begin{aligned} & 0 < x < \infty, && \text{and } 0 < y < \infty \\ \implies & 0 < u + v < \infty, && 0 < v - u < \infty \\ \implies & -u < v < \infty, && u < v < \infty. \\ \implies & 0 < |u| < v < \infty. \end{aligned}$$

The Jacobian is

$$J = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{vmatrix} = \frac{1}{2}.$$

Applying formula, the joint density function of U and V is

$$f_{U,V}(u,v) = f_{X,Y}\left(\frac{u+v}{2}, \frac{v-u}{2}\right)|J| = \begin{cases} \frac{1}{2}e^{-v} & 0 < |u| < v < \infty, \\ 0 & \text{otherwise.} \end{cases} \tag{3.8}$$

◀

✱ Example 3.3.2. Let X_1, X_2 be two iid normal random variables with mean 0 and variance 1.

1. Find the joint density function of $Y_1 = \dfrac{X_1 + X_2}{\sqrt{2}}$ and $Y_2 = \dfrac{X_1 - X_2}{\sqrt{2}}$.

2. Argue that $2X_1X_2$ and $X_1^2 - X_2^2$ are independent random variables that have the same distribution.

✲ **Solution** The joint density function of X_1 and X_2 is

$$f_{X_1,X_2}(x_1,x_2)=\frac{1}{2\pi}e^{-(x_1^2+x_2^2)/2},\quad (x_1,x_2)\in\mathbb{R}^2.$$

1. Let

$$y_1=g_1(x_1,x_2)=\frac{x_1+x_2}{\sqrt{2}},\quad y_2=g_2(x_1,x_2)=\frac{x_1-x_2}{\sqrt{2}}.$$

The inverse transformation is

$$x_1=u(y_1,y_2)=\frac{y_1+y_2}{\sqrt{2}},\quad x_2=v(y_1,y_2)=\frac{y_1-y_2}{\sqrt{2}}.$$

Clearly, $(y_1,y_2)\in\mathbb{R}^2$. The Jacobian is

$$J=\begin{vmatrix}\dfrac{\partial u(y_1,y_2)}{\partial y_1} & \dfrac{\partial u(y_1,y_2)}{\partial y_2}\\ \dfrac{\partial v(y_1,y_2)}{\partial y_1} & \dfrac{\partial v(y_1,y_2)}{\partial y_2}\end{vmatrix}=\begin{vmatrix}\dfrac{1}{\sqrt{2}} & \dfrac{1}{\sqrt{2}}\\ \dfrac{1}{\sqrt{2}} & -\dfrac{1}{\sqrt{2}}\end{vmatrix}=-1.$$

Thus the joint density function of Y_1 and Y_2 is

$$\begin{aligned}f_{Y_1,Y_2}(y_1,y_2)&=f_{X_1,X_2}(u(y_1,y_2),v(y_1,y_2))\,|J|\\&=f_{X_1,X_2}\left(\frac{y_1+y_2}{\sqrt{2}},\frac{y_1-y_2}{\sqrt{2}}\right)\cdot 1\\&=\frac{1}{2\pi}e^{-\frac{1}{2}\left(\left(\frac{y_1+y_2}{\sqrt{2}}\right)^2+\left(\frac{y_1-y_2}{\sqrt{2}}\right)^2\right)}\\&=\frac{1}{2\pi}e^{-(y_1^2+y_2^2)/2}\\&=\underbrace{\frac{1}{\sqrt{2\pi}}e^{-y_1^2/2}}_{f_{Y_1}(y_1)}\cdot\underbrace{\frac{1}{\sqrt{2\pi}}e^{-y_2^2/2}}_{f_{Y_2}(y_2)}.\end{aligned}$$

◀

✱ **Example 3.3.3.** Let $X_1,X_2\sim UNIF(0,1)$. Find the cumulative distribution function and the joint density function of X_1+X_2. How should the above result be modified if $X_1,X_2\sim \texttt{Rect}(a,b)$?

✲ **Solution** Let $U=X_1+X_2$. The cumulative distribution function of U is

$$\begin{aligned}F_U(u)&=\mathbb{P}[U\le u]=\mathbb{P}[X_1+X_2\le u]\\&=\iint\limits_{x_1+x_2\le u} f_{X_1,X_2}(x_1,x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2\end{aligned}$$

Since $X_1,X_2\sim UNIF(0,1)$, the range of $U=X_1+X_2$ should takes between 0 and 2. For $0<u\le 1$,

we have

$$F_U(u) = \mathbb{P}[X_1 + X_2 \leq u] = \frac{\text{Area of the shaded region } A}{\text{Area of the sample space } \Omega}.$$

Using the concept of geometric area, as (X_1, X_2) is uniformly distributed over the unit square, we have

$$\Omega = \{(x_1, x_2) \mid 0 < x_1, x_2 < 1\}.$$

and

$$\mathcal{A} = \{(x_1, x_2) \mid 0 < x_1, x_2 < 1 \text{ and } x_1 + x_2 \leq u\} \subseteq \Omega.$$

There are two possible cases to consider: $0 < u < 1$ and $1 < u < 2$. First we consider if $0 < u < 1$, we have

$$F_U(u) = \mathbb{P}[X_1 + X_2 \leq u] = \frac{\int_0^u \int_0^{u-x_1} 1 \, \mathrm{d}x_2 \mathrm{d}x_1}{1^2} = \frac{1}{2}u^2, \quad \text{for } 0 < u \leq 1. \tag{3.9}$$

For $1 < u < 2$, we have

$$\begin{aligned} F_U(u) = \mathbb{P}[X_1 + X_2 \leq u] &= \frac{\text{Area of the shaded region } A}{\text{Area of the sample space } \Omega} \\ &= \frac{\int_{u-1}^1 \int_{u-x_1}^1 1 \, \mathrm{d}x_2 \mathrm{d}x_1}{1^2} \\ &= 1 - \frac{1}{2}(2-u)^2, \quad \text{for } 0 < u \leq 1. \end{aligned}$$

Hence the cumulative distribution function of U is

$$F_U(u) = \begin{cases} 0, & u \leq 0 \\ \frac{1}{2}u^2, & 0 < u \leq 1 \\ 1 - \frac{1}{2}(2-u)^2, & 1 < u < 2 \\ 1, & u \geq 2 \end{cases}. \tag{♠}$$

Differentiating (♠) with respect to u, we have the density function of U as

$$f_U(u) = \begin{cases} u, & 0 < u \leq 1 \\ 2 - u, & 1 < u < 2 \\ 0, & \text{otherwise} \end{cases}.$$

and we are done. ◄

✱ Example 3.3.4. If $X_1, X_2, X_3 \sim N(0,1)$, find the joint density function of

$$Y_1 = \frac{X_1 + X_2 + X_3}{\sqrt{3}}, \quad Y_2 = \frac{X_1 - X_2}{\sqrt{2}}, \quad Y_3 = \frac{X_1 + X_2 - 2X_3}{\sqrt{6}}.$$

✲ Solution The joint density function of X_1, X_2 and X_3 is given by

$$f_{X_1,X_2,X_3}(x_1, x_2, x_3) = \left(\frac{1}{2}\right)^{3/2} \exp\left\{-\frac{1}{2}(x_1^2 + x_2^2 + x_3^2)\right\}.$$

Note that

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & -\frac{2}{\sqrt{6}} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = A\mathbf{x}. \tag{3.10}$$

where A is orthogonal matrix such that $A^T A = I_{3\times 3}$. Thus $A^T = A$ and $\mathbf{x} = A^T\mathbf{y}$. Hence

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & 0 & -\frac{2}{\sqrt{6}} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}.$$

compute the matrix multiplication and we have

$$\begin{cases} x_1 = \frac{y_1}{\sqrt{3}} + \frac{y_2}{\sqrt{2}} + \frac{y_3}{\sqrt{6}} \\ x_2 = \frac{y_1}{\sqrt{3}} - \frac{y_2}{\sqrt{2}} + \frac{y_3}{\sqrt{6}} \\ x_3 = \frac{y_1}{\sqrt{3}} - \frac{2y_3}{\sqrt{6}} \end{cases}$$

The Jacobian is the derivative of the old variables with respect to the new variables. That is

$$J = \left| \frac{\partial(\text{Old variable})}{\partial(\text{New variable})} \right| \quad \text{on} \quad \left| \frac{\partial(x_1, x_2, x_3)}{\partial(y_1, y_2, y_3)} \right|$$

$$= \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \frac{\partial x_1}{\partial y_3} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} & \frac{\partial x_2}{\partial y_3} \\ \frac{\partial x_3}{\partial y_1} & \frac{\partial x_3}{\partial y_2} & \frac{\partial x_3}{\partial y_3} \end{vmatrix} = \begin{vmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & 0 & -\frac{2}{\sqrt{6}} \end{vmatrix} = |A^T| = \pm 1.$$

And since A is an orthogonal matrix, so

$$\mathbf{y}'\mathbf{y} = \mathbf{x}'A'A\mathbf{x} = \mathbf{x}'\mathbf{x} \implies \sum_{i=1}^{3} y_i^2 = \sum_{i=1}^{3} x_i^2.$$

Clearly, $y_i \in \mathbb{R}^3$ for $i = 1, 2, 3$. Thus the joint density function of Y_1, Y_2 and Y_3 is

$$\begin{aligned} f_{Y_1,Y_2,Y_3}(y_1, y_2, y_3) &= \frac{1}{(2\pi)^{3/2}} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{3} y_i^2 \right\} \cdot |\pm 1|, \quad \forall y_i \in \mathbb{R} \\ &= \prod_{i=1}^{3} \left\{ \frac{1}{\sqrt{2\pi}} e^{-y_i^2/2} \right\} \\ &= \prod_{i=1}^{3} f_{Y_i}(y_i). \end{aligned}$$

◀

3.4 Moment Generating Methods

3.5 Order Statistics

We can ordering the observed random variables based on their magnitudes or ranking. These ordered variables are known as **order statistics**.

Consider $X_1, X_2, \ldots, X_n$ are independent countinuous random variables with cdf $F_X(y)$ and mass function $f_X(y)$. We ordered them into order statistics $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ such that

$$X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}.$$

In this notion, the maximum random variable is

$$X_{(n)} = \max\{X_1, X_2, \ldots, X_n\}$$

and the minimum random variable is

$$X_{(1)} = \min\{X_1, X_2, \ldots, X_n\}$$

$X_{(n)}$ is the largest among $X_1, X_2, \ldots, X_n$, the event $X_{(n)} \leq y$ will happened only if each $X_i \leq y$. Then the joint probability is

$$G_{(n)}(y) = \mathbb{P}[X_{(n)} \leq y] = \mathbb{P}[X_1 \leq y,\ X_2 \leq y,\ \ldots,\ X_n \leq y] = \prod_{i=1}^{n} \mathbb{P}[X_i \leq y]. \qquad (\heartsuit)$$

Because $\mathbb{P}[X_i \leq y] = F_X(y)$ for all $i = 1, 2, \ldots, n$. It follows that

$$(\heartsuit) \ \Rightarrow\ \mathbb{P}[X_1 \leq y]\,\mathbb{P}[X_2 \leq y] \cdots \mathbb{P}[X_n \leq y] = [F_X(y)]^n.$$

Now letting $g_{(n)}$ denote the density function of $Y_{(n)}$, we see that, on taking derivative on $G_{(n)}$ with respect to y.

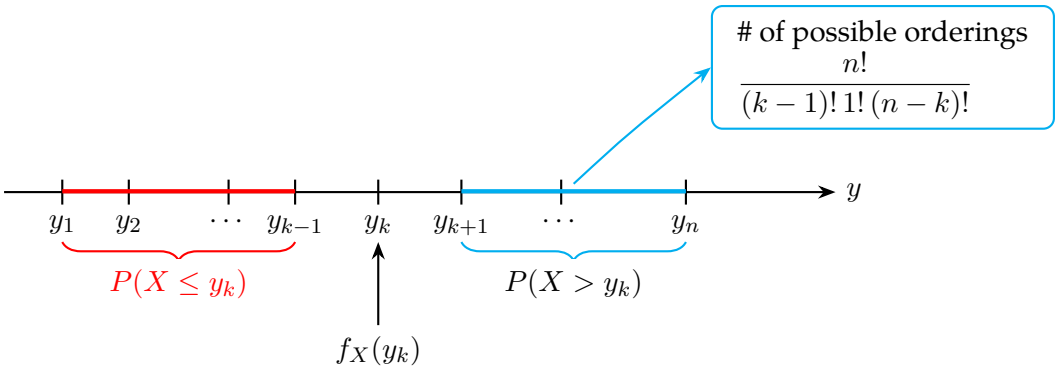
$$\begin{aligned} g_{(n)}(y) &= \frac{\mathrm{d}}{\mathrm{d}y}[F_X(y)]^n \\ &= n[F(y)]^{n-1}\,\frac{\mathrm{d}}{\mathrm{d}y}F_X(y) && \text{By Chain rule of derivative} \\ &= n[F(y)]^{n-1}\,f_X(y) \end{aligned}$$

Now we get the maximum variable. For the minimum variable $X_{(1)}$ can be found using the similar way. The cdf of $X_{(1)}$ is

$$F_{(1)}(y) = \mathbb{P}[X_{(1)} \leq y] = 1 - \mathbb{P}[X_{(1)} > y].$$

Since $X_{(1)}$ is the minimum of $X_1, X_2, \ldots, X_n$, and the event $Y_i > y$ can be occurs for $i = 1, 2, 3, \ldots, n$. It other words, any X_i in $X_1, X_2, \ldots, X_n$ can be the minimum variable. Hence

$$\begin{aligned} F_{(1)}(y) &= \mathbb{P}[X_{(1)} \leq y] = 1 - 1 - \mathbb{P}[X_{(1)} > y] \\ &= 1 - \mathbb{P}[X_1 > y,\ X_2 > y,\ \ldots\ X_n > y] \\ &= 1 - \mathbb{P}[X_1 > y]\,\mathbb{P}[X_2 > y]\ \ldots\ \mathbb{P}[X_n > y] \\ &= 1 - [1 - F_X(y)]^n. \end{aligned}$$



Theorem 3.7 k-th order statistics
Let $X_1, X_2, \ldots, X_n$ be i.i.d continuous random variable with common cdf $F_X(y)$ and common density function $f_X(y)$. Let $X_{(k)}$ denote the k-th order Statistics, then the density function of $X_{(k)}$ is

$$g_{(n)}(y) = \frac{n!}{(k-1)!(n-k)!}[F_X(y)]^{k-1}[1-F_X(y)]^{n-k} f_X(y), \quad -\infty < y < \infty. \tag{3.11}$$

✱ Example 3.5.1. Let $Y \sim \text{Uniform}(0, \theta)$ be the waiting time of bus arrival. A random samples of size $n = 5$ is taken. Then,

1. Find the distribution of minimum variable.
2. Find the probability that $Y_{(3)}$ is less than $\frac{2}{3}\theta$.
3. Suppose that the waiting time for bus arrival is uniformly distributed on 0 to 15 minutes, find $\mathbb{P}[Y_{(5)} < 10]$.

✲ Solution 1. The density of $X_{(1)}$ is

$$\begin{aligned} Y_{(1)} \sim g_{(1)}(y) &= \frac{5!}{(1-1)!\,(5-1)!}[F_Y(y)]^{1-1}[1-F_Y(y)]^{5-1} f_Y(y) \\ &= \frac{5!}{0!\,4!}[1-F_Y(y)]^4 f_Y(y) \\ &= 5\left(1-\frac{y}{\theta}\right)^4\left(\frac{1}{\theta}\right) \\ &= \frac{5(\theta-y)^4}{\theta^5}. \end{aligned}$$

Hence compute the mean of $X_{(1)}$,

$$\mathbb{E}[Y_{(1)}] = \int_0^\theta y\left[\frac{5(\theta-y)^4}{\theta^5}\right]\mathrm{d}y = \int_0^\theta \frac{5y(\theta-y)^4}{\theta^5}\mathrm{d}y \tag{♣}$$

using the substitution method and letting $u = \theta - y$, and for that

$$y = \theta - u \Longrightarrow -du = dy$$

substitute back into (♣) and we have

$$\begin{aligned}
(\clubsuit) = \int_0^\theta \frac{5(\theta-u)u^4}{\theta^5}\,(-\mathrm{d}u) &= -\frac{1}{\theta^5}\int_0^\theta (5\theta u^4 - u^5)\,\mathrm{d}u \\
&= -\frac{1}{\theta^5}\left[\theta u^5 - \frac{1}{6}\theta^6\right]_{u=0}^{u=\theta} \\
&= -\frac{1}{\theta^5}\left[0 - \frac{1}{6}\theta^6\right] \\
&= \frac{\theta}{6} = \mathbb{E}[Y_{(1)}].
\end{aligned}$$

2. First we need to find the probability density function of $Y_{(3)}$, that is,

$$\begin{aligned}
Y_{(3)} \sim g_{(3)}(y) &= \frac{5!}{(3-1)!\,(5-3)!}[F_Y(y)]^{3-1}[1-F_Y(y)]^{5-3}\,f_Y(y) \\
&= 30\left(\frac{y}{\theta}\right)^2\left(1-\frac{y}{\theta}\right)^2\frac{1}{\theta}, \quad 0 < y < \theta.
\end{aligned}$$

Compute the probability on which that $Y_{(3)}$ is smaller that $\frac{2\theta}{3}$.

$$\begin{aligned}
\mathbb{P}[Y_{(3)} < \tfrac{2}{3}\theta] &= \int_0^{\frac{2}{3}\theta} 30\left(\frac{y}{\theta}\right)^2\left(1-\frac{y}{\theta}\right)^2\frac{1}{\theta}\,\mathrm{d}y \\
&= \frac{30}{\theta^5}\int_0^{\frac{2}{3}\theta} y^2(\theta^2 - 2\theta y + y^2)\,\mathrm{d}y \\
&= \frac{30}{\theta^5}\left[\frac{1}{3}\theta^2 y^3 - \frac{1}{2}\theta y^4 + \frac{1}{5}y^5\right]_{y=0}^{y=\frac{2}{3}\theta} \\
&= 30\left(\frac{1}{3}\right)\left(\frac{2}{3}\right)^3 - 15\left(\frac{2}{3}\right)^4 + 6\left(\frac{2}{3}\right)^5 \\
&= \frac{64}{81}.
\end{aligned}$$

3. The probability that $Y_{(5)}$ less than 10 minutes is equivalent to taking the bus five times. That is

$$\begin{aligned}\mathbb{P}[Y_{(5)} < 10] &= \mathbb{P}[Y_{(1)} < 10,\ Y_{(2)} < 10,\ \ldots, Y_{(5)} < 10] \\ &= \mathbb{P}[Y_{(1)} < 10] \times \mathbb{P}[Y_{(2)} < 10] \times \cdots \mathbb{P}[Y_{(5)} < 10] \\ &= \left(\frac{10}{15}\right)^5 = \frac{32}{243}.\end{aligned}$$

◀

This page intentionally left blank.

4

Limiting Distribution

4.1 Probability inequality

We first start off with some useful probability inequalities that will be used later in this chapter – the inequalities which contain the probability measure in either left side or right side or in both sides. This is known as **probability inequalities**.

Theorem 4.1 Markov's Inequality
Suppose X is a random variable that having finite expectation, said $\mathbb{E}[X]$ converges. Then for any nonzero quantity a are having the inequality

$$\mathbb{P}[X \geq a] \leq \frac{\mathbb{E}[X]}{a}. \tag{4.1}$$

4.2 Covergence in distribution

Convergence in distribution state that the sequence of random variables $X_1, X_2, \ldots, X_n$ converges to some distribution X when n goes to infinity. It does not require any dependence between the X_n and X. Convergence in distribution is consider as the weakest type of convergence. The sequence $\{X_n\}$ is said to weakly converges to X if the cdf $F_n(x) \to F(x)$ whenever $n \to \infty$ at all continuity points of $F(x)$. Symbolically, we denote this as $X_n \xrightarrow{\mathcal{D}} X$.

Definition 4.1 Covergence in distribution
Let $\{X_n\}_{n\geq 1}$ be a sequence of random variable and let X be a random variable. Let F_{X_n} and F_X be the cdf of X_n and X respectively. And let $C(F)$ be the set of all continuous points of F. We say that X_n converges in distribution to X if

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x) \tag{4.2}$$

for all $x \in C(F)$. We denote $X_n \xrightarrow{\mathcal{D}} X$.

✱ Example 4.2.1. Let $X_1, X_2, X_3, \ldots$ be a sequence of random variable such that

$$X_n \sim \text{Geom}(\lambda/n), \quad \forall n = 1, 2, 3, \ldots$$

where $\lambda > 0$ is a constant. Define a new sequence Y_n as

$$Y_n = \frac{1}{n} X_n, \quad \forall n = 1, 2, 3, \ldots$$

Show that Y_n converges in distribution to $\text{Exp}(\lambda)$.

✲ Solution The cdf of Y_n is

$$\begin{aligned} F_{Y_n}(y) &= \mathbb{P}\left[\frac{1}{n} X_n \leq y\right] \\ &= \mathbb{P}\left[X_n \leq ny\right] \\ &= 1 - \left(1 - \frac{\lambda}{n}\right)^{\text{floor}(ny)}. \end{aligned}$$

and taking limit for $n \to +\infty$ we have

$$\begin{aligned} \lim_{n\to\infty} \left(1 - \left(1 - \frac{\lambda}{n}\right)^{\lfloor ny \rfloor}\right) &= 1 - \lim_{n\to\infty} \left(1 + \frac{1}{\left(-\frac{n}{\lambda}\right)}\right)^{-n(-y)} \\ &= 1 - \lim_{n\to\infty} \left(1 + \frac{1}{\left(-\frac{n}{\lambda}\right)}\right)^{-\frac{n}{\lambda}(-\lambda y)} \\ &= 1 - e^{-\lambda y} \end{aligned}$$

which is the cdf of exponential distribution with parameter λ. ◀

✱ Example 4.2.2. Let $\{X_n\}_{n\geq 1}$ be a sequence of random variables drawn from a uniform distribution $(0, \theta)$ distribution. We define the order statistics

$$X_{(n)} = \max\{X_1, X_2, \ldots, X_n\}.$$

Show that $\lim_{n\to\infty} X_{(n)} = \theta$.

✲ **Solution** From the cdf of $X_{(n)}$ we have

$$\begin{aligned}
\mathbb{P}[X_{(n)} \le x] &= F_n(x) \\
&= \mathbb{P}[X_1, X_2, \ldots, X_n \le x] \\
&= \prod_{i=1}^{n} \mathbb{P}[X_i \le x] && \text{Since all } X_i \text{ are identically independent} \\
&= (\mathbb{P}[X_1 \le x])^n \\
&= \begin{cases} 0 & \text{if } x < 0 \\ \left(\frac{x}{\theta}\right)^n & \text{if } 0 \le x < \theta \\ 1 & \text{if } x > \theta \end{cases}
\end{aligned}$$

which is the joint density function of n uniform distribution. Note that

$$F_n(x) \xrightarrow{p} F(x) = \begin{cases} 0 & \text{if } x < \theta \\ 1 & \text{if } x \ge \theta \end{cases}$$

and now we can see that the limiting distribution become degenerate at θ. ◀

Definition 4.2 Converges in probability
Let $\{X_n\}_{n\ge1}$ be a sequence of random variable defined on some probability space $(\Omega, \mathscr{F}, P)$, and let X be a random variable. We say that the sequence X_n **converges in probability** to X if for all $\epsilon > 0$, we have

$$\mathbb{P}\{|X_n - X| > \epsilon\} \to 0 \quad \text{as } n \to \infty. \tag{4.3}$$

Or, equivalently,

$$\mathbb{P}\{|X_n - X| < \epsilon\} \to 1 \quad \text{as } n \to \infty. \tag{4.4}$$

We denote this as $X_n \xrightarrow{p} X$.

Remark. *We emphasize that the definition of convergence in probability says nothing about the convergence of the random variables X_n in the sense in which it is understood in real analysis.*

In particular, $X_n \xrightarrow{p} X$ does not imply that given $\epsilon > 0$, we can find an integer N such that $|X_n - X| < \epsilon$ for all $n \ge N$. The definition above speaks about the convergence of the sequence of probabilities $\mathbb{P}\{|X_n - X| > \epsilon\}$ to 0 as n goes to infinity.

✱ **Example 4.2.3.** Let $\{X_n\}_{n\ge1}$ be a sequence of random variable with probability mass function

$$\mathbb{P}[X_n = 1] = \frac{1}{n}, \quad \mathbb{P}[X_n = 0] = 1 - \frac{1}{n}.$$

Show that X_n converges in probability to 0.

✲ Solution From the definition of convergence in probability, we need to show that for all $\epsilon > 0$, that is,

$$\mathbb{P}\left(|X_n| > \epsilon\right) = \begin{cases} \mathbb{P}(X_n = 1) = \frac{1}{n} & \text{if } 0 < \epsilon < 1 \\ 0 & \text{if } \epsilon \geq 1. \end{cases}$$

It follows that $\mathbb{P}\left(|X_n| > \epsilon\right) \to 0$ as $n \to \infty$, which proves that $X_n \xrightarrow{p} 0$. ◀

Theorem 4.2 Slutsky's Theorem
If X_n *converges in distribution* to a random variable X, and Y_n *converges in probability* to a constant c, then

- $Y_n X_n \xrightarrow{\mathcal{D}} cX$
- $X_n + Y_n \xrightarrow{\mathcal{D}} X + c$.

Proof. We will prove both parts of Slutsky's theorem.

Part 1: We want to show that $Y_n X_n \xrightarrow{d} cX$.

Since $Y_n \xrightarrow{p} c$, for any $\epsilon > 0$, we have $\mathbb{P}[|Y_n - c| > \epsilon] \to 0$ as $n \to \infty$. This implies that Y_n is bounded in probability, and we can write $Y_n = c + (Y_n - c)$ where $(Y_n - c) \xrightarrow{p} 0$.

Now, $Y_n X_n = cX_n + (Y_n - c)X_n$. Since $X_n \xrightarrow{d} X$ and multiplication by the constant c is a continuous operation, we have $cX_n \xrightarrow{d} cX$.

For the second term, we need to show that $(Y_n - c)X_n \xrightarrow{p} 0$. Since $Y_n - c \xrightarrow{p} 0$ and X_n is bounded in probability (as it converges in distribution), their product converges to 0 in probability.

By Slutsky's theorem for sums (which we prove next), we get $Y_n X_n = cX_n + (Y_n - c)X_n \xrightarrow{d} cX + 0 = cX$.

Part 2: We want to show that $X_n + Y_n \xrightarrow{d} X + c$.

We use characteristic functions. Let $\phi_{X_n}(t)$, $\phi_X(t)$, and $\phi_{Y_n}(t)$ denote the characteristic functions of X_n, X, and Y_n, respectively.

The characteristic function of $X_n + Y_n$ is given by:

$$\phi_{X_n+Y_n}(t) = \mathbb{E}[e^{it(X_n+Y_n)}] = \mathbb{E}[e^{itX_n} e^{itY_n}]$$

Since $Y_n \xrightarrow{p} c$, we have $e^{itY_n} \xrightarrow{p} e^{itc}$ by the continuous mapping theorem.

Using the fact that $X_n \xrightarrow{d} X$ implies $\phi_{X_n}(t) \to \phi_X(t)$, and that convergence in probability pre-

serves the limit of expectations for bounded random variables, we get:

$$\begin{aligned}\lim_{n\to\infty}\phi_{X_n+Y_n}(t) &= \lim_{n\to\infty}\mathbb{E}[e^{itX_n}e^{itY_n}]\\ &= \mathbb{E}[e^{itX}]\cdot e^{itc}\\ &= \phi_X(t)\cdot e^{itc}\\ &= \phi_{X+c}(t)\end{aligned}$$

Since the characteristic function of $X_n + Y_n$ converges pointwise to the characteristic function of $X + c$, we conclude that $X_n + Y_n \xrightarrow{d} X + c$. □

✱ Example 4.2.4. If the random variable $X \sim \text{Gamma}(\mu, 1)$, show that

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sqrt{\overline{X}_n}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$$

✲ Solution Slutsky's theorem stated that If X_n converges in distribution to a random variable X and if Y_n converges in probability to a constant c. Then X_n/Y_n converges in distribution to X/c. By the central limit theorem we have

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{\mathcal{D}} \mathcal{N}(0, Var[X_i]).$$

and in this case $\mathbb{E}X_i = Var[X_i] = \mu$, thus we obtained

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mu).$$

Replacing the theorem denominator Y_n with $\overline{X}_n$, which $\overline{X}_n$ converges to constant μ in probability. Hence

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sqrt{\overline{X}_n}} \xrightarrow{\mathcal{D}} \mathcal{N}\left(\frac{0}{\mu}, \frac{\mu}{\mu}\right) \Longrightarrow \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sqrt{\overline{X}_n}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$$

and we are done with the proof. ◀

Theorem 4.3
If the random variable X_n converges to constant c in probability, then

$$\sqrt{X_n} \xrightarrow{p} \sqrt{c}, \quad c > 0. \tag{4.5}$$

Theorem 4.4
If the random variable X_n converges to constant c in probability, and Y_n converges to constant d in probability, then

- $aX_n + bY_n \xrightarrow{p} ac + bd.$

- $X_n Y_n \xrightarrow{p} cd$.
- $\frac{1}{X_n} \xrightarrow{p} \frac{1}{c}$ for all $c \neq 0$.

Proof. We will prove each part of the theorem.

Part 1: We want to show that $aX_n + bY_n \xrightarrow{p} ac + bd$.

For any $\epsilon > 0$, we have:

$$\begin{aligned} |aX_n + bY_n - (ac + bd)| &= |a(X_n - c) + b(Y_n - d)| \\ &\leq |a||X_n - c| + |b||Y_n - d| \end{aligned}$$

By the triangle inequality, for any $\delta > 0$:

$$\begin{aligned} &\mathbb{P}[|aX_n + bY_n - (ac + bd)| > \epsilon] \\ &\leq \mathbb{P}[|a||X_n - c| + |b||Y_n - d| > \epsilon] \\ &\leq \mathbb{P}[|a||X_n - c| > \epsilon/2] + \mathbb{P}[|b||Y_n - d| > \epsilon/2] \\ &= \mathbb{P}[|X_n - c| > \frac{\epsilon}{2a}] + \mathbb{P}[|Y_n - d| > \frac{\epsilon}{2b}], \quad \forall a, b > 0 \end{aligned}$$

Since $X_n \xrightarrow{p} c$ and $Y_n \xrightarrow{p} d$, both terms on the right approach 0 as $n \to \infty$.

Part 2: We want to show that $X_n Y_n \xrightarrow{p} cd$.

We can write:

$$\begin{aligned} X_n Y_n - cd &= X_n Y_n - cY_n + cY_n - cd \\ &= Y_n(X_n - c) + c(Y_n - d) \end{aligned}$$

Since $Y_n \xrightarrow{p} d$, the sequence $\{Y_n\}$ is bounded in probability. That is, for any $\delta > 0$, there exists $M > 0$ such that $\mathbb{P}[|Y_n| > M] < \delta$ for all n sufficiently large.

For any $\epsilon > 0$:

$$\begin{aligned} |X_n Y_n - cd| &= |Y_n(X_n - c) + c(Y_n - d)| \\ &\leq |Y_n||X_n - c| + |c||Y_n - d| \end{aligned}$$

Given $\epsilon > 0$, choose $\delta > 0$ such that:

$$\begin{aligned} &\mathbb{P}[|X_n Y_n - cd| > \epsilon] \\ &\leq \mathbb{P}[|Y_n||X_n - c| + |c||Y_n - d| > \epsilon] \\ &\leq \mathbb{P}[|Y_n||X_n - c| > \epsilon/2] + \mathbb{P}[|c||Y_n - d| > \epsilon/2] \end{aligned}$$

For the first term, using the boundedness of Y_n and convergence of X_n, and for the second term using convergence of Y_n, both approach 0 as $n \to \infty$.

Part 3: We want to show that $\frac{1}{X_n} \xrightarrow{p} \frac{1}{c}$ for $c \neq 0$.

Since $c \neq 0$, there exists $\delta > 0$ such that $|c| > \delta > 0$. Because $X_n \xrightarrow{p} c$, for any $\epsilon > 0$, we have $\mathbb{P}[|X_n - c| > \epsilon] \to 0$.

In particular, $\mathbb{P}[|X_n - c| > \delta/2] \to 0$, which implies $\mathbb{P}[|X_n| > \delta/2] \to 1$. This means X_n is bounded away from 0 in probability.

Now, for any $\epsilon > 0$:

$$\left|\frac{1}{X_n} - \frac{1}{c}\right| = \left|\frac{c - X_n}{X_n c}\right| = \frac{|X_n - c|}{|X_n||c|}$$

On the event $\{|X_n| > \delta/2\}$, we have:

$$\left|\frac{1}{X_n} - \frac{1}{c}\right| \leq \frac{2|X_n - c|}{\delta|c|}$$

Therefore:

$$\begin{aligned}\mathbb{P}\left[\left|\frac{1}{X_n} - \frac{1}{c}\right| > \epsilon\right] &\leq \mathbb{P}[|X_n| \leq \delta/2] + \mathbb{P}\left[\frac{2|X_n - c|}{\delta|c|} > \epsilon, |X_n| > \delta/2\right] \\ &\leq \mathbb{P}[|X_n| \leq \delta/2] + \mathbb{P}\left[|X_n - c| > \frac{\epsilon\delta|c|}{2}\right]\end{aligned}$$

As $n \to \infty$, both terms vanishing to zero, this completes the proof. □

Definition 4.3 Almost sure converges
We said that X_n converges to X **almost surely** if the probability that the sequence $X_n(s)$ converges to $X(s)$ is equal to 1.

✱ Example 4.2.5. Consider the sample space $S = [0, 1]$ with uniform probability distribution, for instance,

$$\mathbb{P}([a, b]) = b - a \quad \forall 0 \leq a \leq b \leq 1.$$

Define the sequence $\{X_n, n = 1, 2, 3, \ldots\}$ as

$$X_n(s) = \frac{n}{n+1}s + (1 - s)^n.$$

Also, define the random variable X on the sample space as $X(s) = s$. Show that X_n *almost sure* converges to X.

✲ Solution For any $s \in [0,1]$, taking the limit when $n >> \infty$ we have

$$\begin{aligned}\lim_{n\to\infty} X_n(s) &= \lim_{n\to\infty} \left[\frac{n}{n+1}s + (1-s)^n\right] \\ &= \lim_{n\to\infty} \frac{n}{n+1}s + \lim_{n\to\infty} (1-s)^n \\ &= 1 \cdot s + 0 \\ &= s = X(s).\end{aligned}$$

However, if $s = 0$ then

$$\lim_{n\to\infty} X_n(0) = \lim_{n\to\infty} \left[\frac{n}{n+1}(0) + (1-0)^n\right] = 1.$$

Thus, we conclude that $\lim_{n\to\infty} X_n(s) = X(s)$ for all s lies between 0 and 1. And because $\mathbb{P}([0,1]) = 1$, we conclude that

$$X_n \xrightarrow{a.s.} X$$

and we are done. ◄

4.3 Law of Large Numbers

In this section we will discuss the Weak and Strong Law of Large Numbers. The Law of Large Numbers are consider as a form of convergence in probability.

In practice estimates are made of an unknown quantity (or parameter) by taking the average $\overline{X}_n$ of a number of repeated measurements of the quantity each of which may have been affected by random errors. It is, therefore, of interest to study the properties of such an estimate. An initial enquiry is made concerning its behavior as the number of measurements increases without bound (as $n \to \infty$). The Law of Large Numbers gives an answer to this question. Does the estimate $\overline{X}_n$ converge to the true value ξ of the parameter under study?

The probability can be formulated as follows: Let $X_1, X_2, \ldots, X_n$ be a sequence of i.i.d observations and $\overline{X}_n$ be the mean of the first n observations. Under what condition can we say that

$$\overline{X}_n \to \xi \quad \text{as } n \to \infty?$$

In one or other sense we shall generalise this problem further and ask for the conditions under which

$$\overline{X}_n - \overline{\xi}_n \to 0 \quad \text{as } n \to \infty,$$

where $\{\xi_n\}$ is a sequence of constants that is measured by the sequence of observations $\{X_n\}$. We shall say that the Weak Law of Large Numbers (WLLN) holds if the convergence such that

$$\overline{X}_n \to \xi \quad \text{or} \quad (\overline{X}_n - \overline{\xi}_n) \to 0 \tag{4.6}$$

takes place.

When the convergence is in probability we shall said that WLLN holds. Thus the theorem on

WLLN states the conditions under which the WLLN holds for a given sequence of observations $\{X_n\}$.

In other words, our problem is to answer the question in the affirmative sense that whether there exists a sequence of constants $\{A_n\}$ and $\{B_n\}$ such that $B_n \to \infty$ as $n \to \infty$ and such that

$$\frac{\sum_{k=1}^{n} X_k - A_n}{B_n} \xrightarrow{p} 0 \quad \text{as } n \to \infty. \tag{4.7}$$

Theorem 4.5 Weak Law of Large Numbers (WLLN)
Let $\{X_n\}_{n\geq 1}$ be a sequence of i.i.d random variables, each with mean $\mathbb{E}[X_i] = \mu$ and standard deviation σ, we define

$$\overline{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

The Weak Law of Large Numbers (WLLN) states that for all $\epsilon > 0$, we have

$$\lim_{n\to\infty} \mathbb{P}[\,|\overline{X}_n - \mu| > \epsilon] = 0. \tag{4.8}$$

Proof. Suppose that $Var[X_i] = \sigma^2 > 0$ for finite i. Since $X_1, X_2, \ldots, X_n$ are identically independent, there is no correlation between them, thus

$$\begin{aligned} Var[\overline{X}_n] &= Var\left[\frac{X_1 + X_2 + \cdots + X_n}{n}\right] \\ &= \frac{1}{n^2} Var[X_1 + X_2 + \cdots + X_n] \\ &= \frac{1}{n^2}[VarX_1 + VarX_2 + \cdots + VarX_n] \\ &= \frac{1}{n^2}(\underbrace{\sigma^2 + \sigma^2 + \cdots + \sigma^2}_{n \text{ times}}) \\ &= \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}. \end{aligned}$$

Notice that the mean of each X_i in the sequence is also equal to the mean of the sample average, said $\mathbb{E}[X_i] = \mu$. We can now apply Chebyshev's inequality on $\overline{X}_n$ to get, for all $\epsilon > 0$,

$$\mathbb{P}[\,|\overline{X}_n - \mu| > \epsilon] \leq \frac{\sigma^2}{n\epsilon^2}.$$

So that □

Lemma 4.1
Denote Z the "standard normal random variable" with density $\frac{1}{\sqrt{2\pi}} e^{-x^2/2}$. Then

$$\mathop{\mathbb{E}}_{x}\left[\, e^{itZ} \,\right] = e^{-t^2/2}. \tag{4.9}$$

Proof. We may use the same calculation as for the moment generating function:

$$\int_{-\infty}^{\infty} \exp\left(itx - \frac{1}{2}x^2\right)\,\mathrm{d}x - e^{-t^2/2}\int_{-\infty}^{\infty}\exp\left\{-\frac{1}{2}(x-it)^2\right\}\,\mathrm{d}x - \sqrt{2\pi}.$$

Note that $e^{-z^2/2}$ is an analytic function and $\oint_\gamma e^{-z^2/2}\,\mathrm{d}z = 0$ on closed path. □

Theorem 4.6 Strong Law of Large Numbers
Let $\{X_n\}_{n\geq 1}$ be a sequence of i.i.d random variables, each with mean $\mathbb{E}[X_i] = \mu$ and standard deviation σ, then

$$\mathbb{P}[\lim_{n\to\infty} \overline{X}_n = \mu] = 1. \tag{4.10}$$

Proof. By Markov's inequality that

$$\mathbb{P}\left[\frac{1}{n}|\overline{X}_n - n\mu| \geq n^{-\gamma}\right] \leq \frac{\mathbb{E}[(\frac{\overline{X}_n}{n} - \mu)^4]}{n^{-4\gamma}} = Kn^{-2+4\gamma}.$$

Define for all $\gamma \in \left(0, \frac{1}{4}\right)$, and let

$$A_n = \left\{\frac{1}{n}|\overline{X}_n - n\mu| \geq n^{-\gamma}\right\} \Rightarrow \sum_{n\geq 1}\mathbb{P}[A_n] < \infty \Rightarrow \mathbb{P}[A] = 0$$

by the first Borel-Cantelli lemma, where $A = \bigcap_{n\geq 1}\bigcup_{m\geq n} A_n$. But now, the event A^c happened if and only if

$$\exists N\,\forall n \geq N\ \left|\frac{\overline{X}_n}{n} - \mu\right| < n^{-\gamma} \Rightarrow \frac{\overline{X}_n}{n} \xrightarrow{p} \mu.$$

□

Lemma 4.2
We can write $f(x) = \mathcal{O}(x)$ if $\frac{f(x)}{x} \to 0$ as $x \to 0$. We have

$$\lim_{x\to 0}\left[1 + \frac{a}{n} + \mathcal{O}\left(\frac{1}{n}\right)\right]^n = e^a \quad \forall a \in \mathbb{R}. \tag{4.11}$$

Proof. Using Taylor's expansion we have

$$\begin{aligned} &f(x) = f(0) + xf'(\theta x), \quad \text{for } 0 < \theta < 1 \\ &\Rightarrow f(x) = f(0) + xf'(0) + x[f'(\theta x) - f'(0)]. \end{aligned}$$

If $f'(x)$ is continuous at $x = 0$, then $f'(\theta x) - f'(0) = \mathcal{O}(x)$ as $x \to 0$. Now let $f(x) = \ln(1 + x)$. Taking derivative we have $f'(x) = \frac{1}{1+x}$, which is continuous at $x = 0$. From here we have

$$\ln(1 + x) = \ln 1 + x + \mathcal{O}(x) \implies \ln(1 + x) = x + \mathcal{O}(x).$$

Then for sufficiently large n, we have

$$\begin{aligned} n \ln \left[1 + \frac{a}{n} + \mathcal{O}\left(\frac{1}{n}\right)\right] &= n\left[\frac{a}{n} + \mathcal{O}\left(\frac{1}{n}\right) + \mathcal{O}\left(\frac{a}{n} + \mathcal{O}\left(\frac{1}{n}\right)\right)\right] \\ &= a + n\mathcal{O}\left(\frac{1}{n}\right) + n\mathcal{O}\left(\frac{a}{n} + \mathcal{O}\left(\frac{1}{n}\right)\right) \\ &= a + n\mathcal{O}\left(\frac{1}{n}\right) + n\mathcal{O}\left(\frac{1}{n}\right) \to a \quad \text{as } n \to \infty. \end{aligned}$$

Taking exponential on both sides we have

$$\lim_{x \to 0} \left[1 + \frac{a}{n} + \mathcal{O}\left(\frac{1}{n}\right)\right]^n = e^a$$

for every real a, and now we are done. □

Remark. *Notice that*

$$\mathcal{O}\left(\frac{1}{n}\right) = \frac{k_1}{n^2} + \frac{k_2}{n^3} + \cdots.$$

and

$$\begin{aligned} \mathcal{O}\left(\frac{a}{n} + \mathcal{O}\left(\frac{1}{n}\right)\right) &= k_1\left(\frac{a}{n} + \mathcal{O}\left(\frac{1}{n}\right)\right)^2 + k_2\mathcal{O}\left(\frac{a}{n} + \mathcal{O}\left(\frac{1}{n}\right)\right)^3 + \cdots \\ &= \frac{c_1}{n^2} + \frac{c_2}{n^3} + \cdots. \end{aligned}$$

for arbitrary constants $c_1, c_2, \ldots$, this implies that $\mathcal{O}\left(\frac{a}{n} + \mathcal{O}\left(\frac{1}{n}\right)\right)$ is also $\mathcal{O}\left(\frac{1}{n}\right)$.

4.4 Central Limit Theorem

Theorem 4.7 Central Limit Theorem
Let $\{X_n\}_{n \geq 1}$ be a sequence of i.i.d random variable whose moment generating function exist in a neighborhood of 0. Let $\mathbb{E}[X_i] = \mu$ and $Var[X_i] = \sigma^2 > 0$. Define $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$. Then

$$Z = \frac{\sqrt{n}(\overline{X}_i - \mu)}{\sigma} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1) \tag{4.12}$$

or

$$Z = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1) \tag{4.13}$$

✱ Example 4.4.1. Let $\overline{X}_n$ be the sample mean from a random sampling of size $n = 100$ from χ^2_{50}. Compute approximate value of $\mathbb{P}(49 < \overline{X} < 51)$.

✲ Solution Because $\overline{X}$ followed Chi-squared distribution with degree of freedom 50, then the mean and variance are $\mathbb{E}[X_i] = 50$ and $Var[X_i] = 2(50) = 100$. By Central Limit Theorem,

$$\begin{aligned}
\mathbb{P}(49 < \overline{X} < 51) &\simeq \mathbb{P}\left[\frac{\sqrt{100}(49 - \mathbb{E}[X_i])}{\sqrt{Var[X_i]}} < Z < \frac{\sqrt{100}(51 - \mathbb{E}[X_i])}{\sqrt{Var[X_i]}}\right] \\
&\simeq \mathbb{P}\left[\frac{\sqrt{100}(49 - 50)}{\sqrt{100}} < Z < \frac{\sqrt{100}(51 - 50)}{\sqrt{100}}\right] \\
&\simeq \mathbb{P}\left[-1 < Z < 1\right] \\
&\simeq \Phi(1) - \Phi(-1) \\
&\simeq 0.84134 - 0.15866 = 0.68268.
\end{aligned}$$

◀

Theorem 4.8 De Moivre-Laplace Limit Theorem
Let $\{X_n\}_{n\geq 1}$ be a sequence of identically independent Bernoulli random variable with parameter p. Such as

$$\mathbb{P}[X_n = 1] = p, \quad \mathbb{P}[X_n = 0] = 1 - p \quad \text{where } 0 < p < 1.$$

Then

$$\lim_{n\to\infty} \mathbb{P}\left[\frac{\sum_{k=1}^{n} X_k - np}{\sqrt{npq}} \leq x\right] = \Phi(x). \tag{4.14}$$

That is, the sum $\sum_{k=1}^{n} X_k \xrightarrow{\mathcal{D}} X \sim N(np, npq)$. Also, $\frac{\sum_{k=1}^{n} X_k - np}{\sqrt{npq}}$ is *asymptotically normal* with mean zero and unit variance.

Proof. Let $S_n = \sum_{k=1}^{n} X_k$ and then the mean and variance are $\mathbb{E}[S_n] = np$ and $Var[S_n] = \sqrt{npq}$. The

characteristic function is

$$\begin{aligned}\phi_n(t) &= \mathbb{E}\left[\exp\left\{\left(\frac{S_n - np}{\sqrt{npq}}\right) it\right\}\right] \\ &= \prod_{k=1}^{n} \mathbb{E}\left[\exp\left\{\frac{it}{\sqrt{n}}\left(\frac{X_k - p}{\sqrt{p(1-p)}}\right)\right\}\right] \\ &= \left[\phi\left(\frac{t}{\sqrt{n}}\right)\right]^n\end{aligned}$$

□

Alternative proof

Proof. Since each X_k's are identically independent Bernoulli random variable with mean $\mathbb{E}X_k = p$ and $Var(X_k) = pq \leq \frac{1}{4} < \infty$. By Lindeberg-Lévy CLT theorem,

$$\frac{\sqrt{n}(\overline{X}_n - p)}{\sqrt{pq}} \xrightarrow{\mathcal{D}} X \sim N(0,1). \tag{4.15}$$

Let the sum be $S_n = \sum_{k=1}^{n} X_k$, the probability mass function of S_n is given by

$$\begin{aligned}p(x) &= \frac{n!}{x!(n-x)!} p^x q^{n-x} \\ &\approx \frac{\sqrt{2\pi} e^{-n} n^{n+\frac{1}{2}} p^x q^{n-x}}{\sqrt{2\pi} e^{-x} x^{x+\frac{1}{2}} \sqrt{2\pi} e^{-(n-x)} (n-x)^{(n-x)+\frac{1}{2}}} \\ &\approx \frac{1}{\sqrt{2\pi}\sqrt{npq}} \left(\frac{np}{x}\right)^{x+\frac{1}{2}} \left(\frac{nq}{n-x}\right)^{n-x+\frac{1}{2}}\end{aligned}$$

Now let

$$\delta = \frac{x - np}{\sqrt{npq}}$$

where $\delta > 0$.

$$\begin{aligned}&\Rightarrow x = np + \delta\sqrt{npq} \\ &\Rightarrow n - x = nq - \delta\sqrt{npq}\end{aligned}$$

Hence, the Binomial density function can be written as

$$f(x) = \frac{1}{\sqrt{2\pi}\sqrt{npq}} \left(1 + \delta\sqrt{\frac{q}{np}}\right)^{-x-\frac{1}{2}} \left(1 - \delta\sqrt{\frac{p}{nq}}\right)^{-(n-x)-\frac{1}{2}}$$

Taking logarithm on both sides,

$$\ln f(x) = \ln \frac{1}{\sqrt{2\pi}\sqrt{npq}} - \left(x + \frac{1}{2}\right) \ln\left(1 + \delta\sqrt{\frac{q}{np}}\right) - \left(n - x + \frac{1}{2}\right) \ln\left(1 - \delta\sqrt{\frac{p}{nq}}\right)$$

To make it easier, observe that $\ln \frac{1}{\sqrt{2\pi}\sqrt{npq}}$ is a constant. We let $c = \ln \frac{1}{\sqrt{2\pi}\sqrt{npq}}$ and continue evaluate this algebraic expression.

$$\begin{aligned}
\ln f(x) &= c - \left(np + \delta\sqrt{npq} + \frac{1}{2}\right)\left[\delta\sqrt{\frac{q}{np}} - \delta^2\frac{q}{2np} + \delta^3\frac{q^{3/2}}{3(np)^{3/2}} + \cdots\right] \\
&\quad + \left(nq - \delta\sqrt{npq} + \frac{1}{2}\right)\left[\delta\sqrt{\frac{q}{np}} - \delta^2\frac{q}{2np} + \delta^3\frac{q^{3/2}}{3(np)^{3/2}} + \cdots\right] \\
&= c + [-\delta\sqrt{npq} + \delta\sqrt{npq}] + \left[-(\delta^2 q + \delta^2 p) + \left(\delta^2\frac{q}{2} + \delta^2\frac{p}{2}\right)\right] \\
&\quad + \left[\left(-\frac{\delta}{2}\sqrt{\frac{q}{np}} + \frac{\delta}{2}\sqrt{\frac{p}{nq}}\right) + \left(\delta^3\frac{q^{3/2}}{2(np)^{3/2}} - \delta^3\frac{p^{3/2}}{2(nq)^{3/2}}\right) + \left(-\delta^3\frac{q^{3/2}}{3(np)^{3/2}} + \delta^3\frac{p^{3/2}}{3(nq)^{3/2}}\right)\right] \\
&\quad + \left[\left(\frac{\delta^2 q}{4np} + \frac{\delta^2 p}{4nq}\right) - \left(\frac{\delta^2 q^2}{3np} + \frac{\delta^2 p^2}{3nq}\right) + \left(\frac{\delta^4 q^2}{4np} + \frac{\delta^4 p^2}{4nq}\right)\right] + \cdots
\end{aligned}$$

Assuming that $\frac{|\delta|}{n^{1/\delta}} \to 0$ for large n. Then

$$\frac{\delta}{\sqrt{n}} \to 0, \quad \frac{\delta^2}{n} \to 0, \quad \frac{\delta^3}{\sqrt{n}} \to 0, \quad \frac{\delta^4}{n^{2/3}} \to 0 \quad \text{as } n \to \infty.$$

So that $\ln f(x)$ is approximate to $c - \frac{1}{2}\delta^2$. Hence

$$f(x) = e^{c - \frac{1}{2}\delta^2} = \frac{1}{\sqrt{2\pi}\sqrt{npq}} \exp\left\{-\frac{1}{2}\left(\frac{x - np}{\sqrt{npq}}\right)^2\right\} \sim N(np, npq).$$

□

Theorem 4.9 Lyapunov Limit Theorem
Let $\{X_n\}_{n\geq 1}$ be a sequence of identically independent random variables with mean $\mathbb{E}[X_n] = \mu_n$ and variance $Var(X_n) = \sigma_n^2$, and

$$\mathbb{E}|X_n - \mu_n|^{2+\delta} < \infty \quad \text{for some } \delta > 0.$$

Define the sum of sequence as

$$S_n = \sum_{k=1}^{n} X_k.$$

Then

$$\mathbb{P}\left[\frac{S_n - \sum_{k=1}^n \mu_k}{\sqrt{\sum_{k=1}^n \sigma_k^2}}\right] \to \Phi(x) \sim N(0,1) \quad \text{as } n \to \infty. \tag{4.16}$$

Provided that the Lyapunov condition is satisfied, namely

$$\lim_{n\to\infty} \frac{\sum_{k=1}^n \mathbb{E}|X_k - \mu_k|^{2+\delta}}{\left(\sum_{k=1}^n \sigma_k^2\right)^{1+\delta/2}} = 0. \tag{4.17}$$

The cases when $\delta > 1$ can be reduced to the case $\delta = 1$, thus it is enough to consider that $0 < \delta < 1$.

Proof. We use the following bound to verify Lyapunov's condition:

$$\frac{1}{\widehat{\sigma_n^2}} \sum_{k=1}^{r_n} \int_{|X_{nk}|>\epsilon\widehat{\sigma}_n} X_{nk}^2 \, \mathrm{d}P \le \frac{1}{\epsilon^\delta \widehat{\sigma}_n^{2+\delta}} \sum_{k=1}^{r_n} \int_{|X_{nk}|>\epsilon\widehat{\sigma}_n} |X_{nk}|^{2+\delta} \, \mathrm{d}P \le \frac{1}{\epsilon^\delta \widehat{\sigma}_n^{2+\delta}} \sum_{k=1}^{n} \mathbb{E}|X_{nk}|^{2+\delta}.$$

□

Corollary 4.1
Suppose X_k are independent with mean zero, variance σ^2 and that $\sup_k \mathbb{E}|X_k|^{2+\delta} < \infty$. Then

$$\frac{S_n}{\sigma\sqrt{n}} \xrightarrow{\mathcal{D}} \sigma N(0,1).$$

Proof. Let $C = \sup_k \mathbb{E}|X_k|^{2+\delta}$. Then $s_n = \sqrt{n}$ and

$$\frac{1}{s_n^{2+\delta}} \sum_{k=1}^n \mathbb{E}|X_k|^{2+\delta} \le \frac{nC}{n^{1+\delta/2}} = \frac{C}{n^{\delta/2}} \to 0.$$

as $n \to \infty$. So Lyapunov's condition is satisfied, and we are done. □

Tutorials

Exercise 4.1 Let $X_1, X_2, \ldots, X_{2020}$ be a random sample of size 2020 from a Poisson distribution with density function

$$f_{X_i}(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots, \infty.$$

What is the distribution of $2020\overline{X}$?

This page intentionally left blank.

5

Decision theory

For the given observation $\mathcal{X}$, we decide to take an action $a \in \mathcal{A}$. An action is a map $a : \mathcal{X} \to \mathcal{A}$ with $a(X)$ being the decision taken.

$L(\theta, a)$ denoted as the "loss function", it is the loss incurred when state is θ and an action a is taken.

$$L : \Theta \times \mathcal{A} \to \mathbb{R}. \tag{5.1}$$

5.1 Conditional Distributions

Recall the definition of conditional probabilities: For two sets A and B, with $P(A) \neq 0$, the conditional probability of B given that A is true is defined as

$$\mathbb{P}(B \mid A) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}. \tag{5.2}$$

✱ Example 5.1.1. Let X and Y be two jointly continuous random variable with joint density function

$$f_{XY}(x, y) = \begin{cases} x^2 + \frac{1}{3}y, & -1 \leq x \leq 1,\ 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}.$$

For $0 \leq y \leq 1$, find the conditional pdf of X given $Y = y$.

✲ Solution First we find the marginal distribution of Y, which we can obtain by integrating along with x.

$$\begin{aligned} f_Y(y) = \int_{-1}^{1} f_{XY}(x, y)\, \mathrm{d}x &= \int_{-1}^{1} \left(x^2 + \frac{1}{3}y \right) \mathrm{d}x \\ &= \frac{1}{3}x^3 + \frac{1}{3}xy \Big|_{-1}^{1} \\ &= \frac{2}{3}(1 + y). \end{aligned}$$

The conditional distribution of X given $Y = y$ is

$$f_{X|Y=y}(x) = \frac{f_{XY}(x,y)}{f_Y(y)} = \frac{x^2 + \frac{1}{3}y}{\frac{2}{3}(1+y)} = \frac{3x^2+y}{2(1+y)}, \quad -1 \le x \le 1,\ 0 \le y \le 1$$

◀

✱ Example 5.1.2 (Two-sample mean problems). Consider the observations $X_1, X_2, \ldots, X_m \sim \mathcal{N}(\mu, \sigma^2)$ response under control treatment. And $Y_1, Y_2, \ldots, Y_n \sim \mathcal{N}(\mu + \Delta, \sigma^2)$ are explanatory data response under test treatment where $\mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+$. σ^2 is unknown variance and $\Delta \in \mathbb{R}$ is unknown treatment effect.

We define two testing hypotheses:

$$H_0 : P \in \{P : \Delta = 0\} = \{P_\theta : \theta \in \Theta_0\}$$

$$H_1 : P \in \{P : \Delta \neq 0\} = \{P_\theta : \theta \notin \Theta_0\}$$

By construct decision rule accepting null hypothesis H_0 if estimate of Δ is significantly far away from zero. For instance, $\hat{\Delta} = \bar{Y} - \bar{X}$ to be the estimate difference in sample means. Since σ is unknown, we use $\hat{\sigma}$ to estimate true σ. The decision procedure is

$$\delta(X,Y) = \begin{cases} 1 & \text{if } |\frac{\hat{\Delta}}{\hat{\sigma}}| < c \\ 0 & \text{if } |\frac{\hat{\Delta}}{\hat{\sigma}}| \ge c \end{cases}$$

We again define a zero-one loss function to make decision

$$L(\theta, a) = \begin{cases} 0 & \text{if } \theta \in \Theta_a \quad \text{(correct action)} \\ 1 & \text{if } \theta \notin \Theta_a \quad \text{(wrong action)} \end{cases}.$$

The risk function is linear combination of the loss of correct and wrong actions,

$$\begin{aligned} R(\theta, \delta) &= L(\theta,0) P_\theta\left(\delta(X,Y) = 0\right) + L(\theta,1) P_\theta\left(\delta(X,Y) = 1\right) \\ &= \begin{cases} P_\theta\left(\delta(X,Y) = 1\right) & \text{if } \theta \in \Theta_0 \\ P_\theta\left(\delta(X,Y) = 0\right) & \text{if } \theta \notin \Theta_0 \end{cases} \end{aligned}$$

✱ Example 5.1.3 (Statistical testing). We are going to use the random variable $X \sim P_\theta$ with sample space $\mathcal{X}$ and parameter space Θ, we want to test the testing hypothesis

$$H_0 : \theta \in \Theta_0$$

against

$$H_1 : \theta \notin \Theta_0.$$

We construct the critical region of a test δ as

$$C = \{x : \delta(x) = 1\}.$$

with zero-one loss. Note that

- Type I error: the test $\delta(X)$ rejects H_0 when H_0 is true.
- Type II error: the test $\delta(X)$ accepts H_0 when H_0 is false.

The risk under zero-one loss as

$$\begin{aligned} R(\theta, \delta) &= P_\theta(\delta(X) = 1 \mid \theta) \quad \text{if } \theta \in \Theta_0 \\ &= \text{Probability of Type I error.} \end{aligned}$$

$$\begin{aligned} R(\theta, \delta) &= P_\theta(\delta(X) = 0 \mid \theta) \quad \text{if } \theta \notin \Theta_0 \\ &= \text{Probability of Type II error.} \end{aligned}$$

✱ Example 5.1.4 (Statistical testing with two different hypothesis subspace)**.** We are going to use the random variable $X \sim P_\theta$ with sample space $\mathcal{X}$ and parameter space Θ, we want to test the testing hypothesis

$$H_0 : \theta \in \Theta_0$$

against

$$H_1 : \theta \notin \Theta_0.$$

We construct the critical region of a test δ as

$$C = \{x : \delta(x) = 1\}.$$

with zero-one loss. Note that

- Type I error: the test $\delta(X)$ rejects H_0 when H_0 is true.
- Type II error: the test $\delta(X)$ accepts H_0 when H_0 is false.

The risk under zero-one loss as

$$\begin{aligned} R(\theta, \delta) &= P_\theta(\delta(X) = 1 \mid \theta) \quad \text{if } \theta \in \Theta_0 \\ &= \text{Probability of Type I error.} \end{aligned}$$

$$\begin{aligned} R(\theta, \delta) &= P_\theta(\delta(X) = 0 \mid \theta) \quad \text{if } \theta \notin \Theta_0 \\ &= \text{Probability of Type II error.} \end{aligned}$$

5.2 Value-at-risk

✱ Example 5.2.1 (Confidence Interval). We altering the previous decision framework setup:

- X is a random variable with probability P_θ.
- The parameter of interest is $\mu(\theta)$.
- Define $\mathfrak{U} = \{\mu = \mu(\theta) : \theta \in \Theta\}$.
- Objective: we want to construct an interval estimation of $\mu(\theta)$.
- Action space: $\mathcal{A} = \{\mathbf{a} = [\underline{a}, \overline{a}] : \underline{a} < \overline{a} \in \mathfrak{U}\}$.
- Interval Estimator: define a map $\hat{\mu}(X) : \mathcal{X} \to \mathcal{A}$, that is $\hat{\mu}(X) = [\hat{\mu}_{\text{Lower}}(X), \hat{\mu}_{\text{Upper}}(X)]$

Note that θ is not random, the interval is random given a fixed θ. We have to use Bayesian models to compute

$$\mathbb{P}\left[\mu(\theta) \in [\hat{\mu}_{\text{Lower}}(X), \hat{\mu}_{\text{Upper}}(X)] \mid X = x\right].$$

We define the zero-one loss function

$$L(\theta, (\underline{a}, \overline{a})) = \begin{cases} 1 & \text{if } \underline{a} > \mu(\theta) \text{ or } \overline{a} < \mu(\theta) \\ 0 & \text{otherwise.} \end{cases}$$

The risk function under zero-one loss is

$$\begin{aligned} R(\theta, \hat{\mu}(X)) &= \mathbb{E}_X[L(\theta, \hat{\mu}(X)) \mid \theta] \\ &= P_\theta\left(\hat{\mu}_{\text{Lower}}(X) > \mu(\theta) \text{ or } \hat{\mu}_{\text{Upper}}(X) < \mu(\theta)\right) \\ &= 1 - P_\theta\left(\hat{\mu}_{\text{Lower}}(X) \leq \mu(\theta) \leq \hat{\mu}_{\text{Upper}}(X) \mid \theta\right). \end{aligned}$$

It is said that the interval estimator $\hat{\mu}(\theta)$ has confidence level $1 - \alpha$ if

$$P_\theta\left(\hat{\mu}_{\text{Lower}}(X) \leq \mu(\theta) \leq \hat{\mu}_{\text{Upper}}(X) \mid \theta\right) \geq (1 - \alpha) \quad \forall \theta \in \Theta.$$

Equivalently, we can said $R(\theta, \hat{\mu}(X)) \leq \alpha$ for all $\theta \in \Theta$.

5.3 Admissible

On basis of performance measure by the risk function $R(\theta, \delta)$, some rules are obviously bad. We said that a decision procedure $\delta(\cdot)$ is inadmissible if $\exists \delta'$ such that

$$R(\theta, \delta') \leq R(\theta, \delta) \quad \forall \theta \in \Theta \tag{5.3}$$

with strict inequality for some θ.

✱ Example 5.3.1. Suppose, for $n \geq 2$, the observations $X_1, X_2, \ldots, X_n$ be i.i.d with mean $g(\theta) := \mathbb{E}_\theta[X_i] = \mu$, and $Var[X_i] = 1$ for all i. We take quadratic loss

$$L(\theta, a) := |\mu_X - a|^2.$$

Consider the decision

$$\delta'(X_1, X_2, \ldots, X_n) := \bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n},$$

and $\delta(X_1, X_2, \ldots, X_n) := X_1$. Then for all θ, we have

$$R(\theta, \delta') = \frac{1}{n}, \quad R(\theta, \delta) = 1.$$

Therefore δ is inadmissible.

This page intentionally left blank.

6

Sampling Distributions

6.1 Snedecor's F-distribution

TheF-distribution was named in honor of Sir Ronald Fisher by George Snedecor. F-distribution arises as the distribution of a ratio of variances. Like, the other two distributions this distribution also tends to normal distribution as ν_1 and ν_2 become very large. The following figure illustrates the shape of the graph of this distribution for various degrees of freedom.

Theorem 6.1
If the random variable X is F-distributed with degrees of freedom ν_1 and ν_2, then its mean is

$$\mathbb{E}[X] = \begin{cases} \dfrac{\nu_2}{\nu_2 - 2} & \text{if } \nu_2 \geq 3 \\ DNE & \text{if } \nu_2 = 1, 2 \end{cases} \tag{6.1}$$

and the variance is

$$Var[X] = \begin{cases} \dfrac{2\nu_2^2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 2)^2(\nu_2 - 4)} & \text{if } \nu_2 \geq 5 \\ DNE & \text{if } \nu_2 = 1, 2, 3, 4. \end{cases} \tag{6.2}$$

Theorem 6.2
If a random variable $X \sim F(\nu_1, \nu_2)$, then its reciprocal $\frac{1}{X} \sim F(\nu_1, \nu_2)$.

Theorem 6.3
If the random variables $U \sim \chi^2(\nu_1)$ and $V \sim \chi^2(\nu_2)$, and U and V are independent, then

$$\frac{U/\nu_1}{V/\nu_2} \sim F(\nu_1, \nu_2). \tag{6.3}$$

✱ **Example 6.1.1.** Let $X_1, X_2, \ldots, X_4$ and $Y_1, Y_2, \ldots, Y_5$ be two random samples of size 4 and 5,

respectively, from a standard normal population. What is the variance of the statistic

$$T = \left(\frac{5}{4}\right) \frac{X_1^2 + X_2^2 + X_3^2 + X_4^2}{Y_1^2 + Y_2^2 + Y_3^2 + Y_4^2 + Y_5^2}.$$

✲ Solution Since the population is standard normal, we have

$$X_1^2 + X_2^2 + X_3^2 + X_4^2 \sim \chi^2(4).$$

Similarly,

$$Y_1^2 + Y_2^2 + Y_3^2 + Y_4^2 + Y_5^2 \sim \chi^2(5).$$

Therefore,

$$\begin{aligned} T &= \left(\frac{5}{4}\right) \frac{X_1^2 + X_2^2 + X_3^2 + X_4^2}{Y_1^2 + Y_2^2 + Y_3^2 + Y_4^2 + Y_5^2} \\ &= \frac{\dfrac{X_1^2 + X_2^2 + X_3^2 + X_4^2}{4}}{\dfrac{Y_1^2 + Y_2^2 + Y_3^2 + Y_4^2 + Y_5^2}{5}} \\ &\sim F(4,5). \end{aligned}$$

Applying theorem, the variance of this statistic is

$$\begin{aligned} Var[T] &= Var[F(4,5)] \\ &= \frac{2(5)^2(4+5-2)}{4(5-2)^2(5-4)} \\ &= \frac{350}{36}. \end{aligned}$$

◀

This page intentionally left blank.

7

Descriptive Statistics

7.1 What is Statistics?

"Statistics" is a science of decision making on the basis of sample observations drawn from a population under uncertainty. That is, it is a mathematical discipline concerned with the collection of data, summarization of data, analysis of data and interpretation of data toward a valid decision.

Given a sample (a set of outcomes), we are to say (infer) about the population or the model. Statistics primarily deals with situations in which the occurrence of some event can't be predicted with certainty. Here are the major objectives of statistics:

- Make inference about a population from an analysis of information contained in the sample data.
- To make assessments of the extent of uncertainty involved in these inferences.
- A third objective, no less important, is to design the process and the extent of sampling so that the observations from a basis for drawing valid and accurate inferences.

7.2 Collecting Data

Statistical data are frequently obtained by a process in which the desired information is obtained from the source, either by having an enumerator visit to the informant, ask the necessary questions and enter the replies on a schedule, or by sending to the informant a list of questions or some questionnaire which he may answer at his convenience. The term "questionnaire" means a list of certain systematically arranged questions relating to the subject of enquiry. It is necessary that questionnaire is designed with due care so that necessary data may be easily collected.

In the schedule one finds a list of items, on which information will be collected, the exact forms of the questions to be put to the informants are not given and task of questioning, explaining the desired information is left to the investigator.

7.3 Presentation of Data

7.3.1 Bar chart

A Bar diagram which consists of a number of rectangles (usually called bars) is used for one-dimensional comparison. It is used to show absolute changes in magnitudes overtime (chronological) or space (geographical/regional). Changes in time or space, as the case may be, are shown along the x-axis with equally spaced magnitudes. Rectangles of equal width are drawn with lengths varying with the magnitude represented. While a line graph is not suitable for representation of data classified geographically or qualitatively, a bar diagram is suitable for representation of such data. Vertical bars should also be used for data classified quantitatively. When making comparisons of data classified qualitatively or geographically, on the other hand, horizontal bars are generally used.

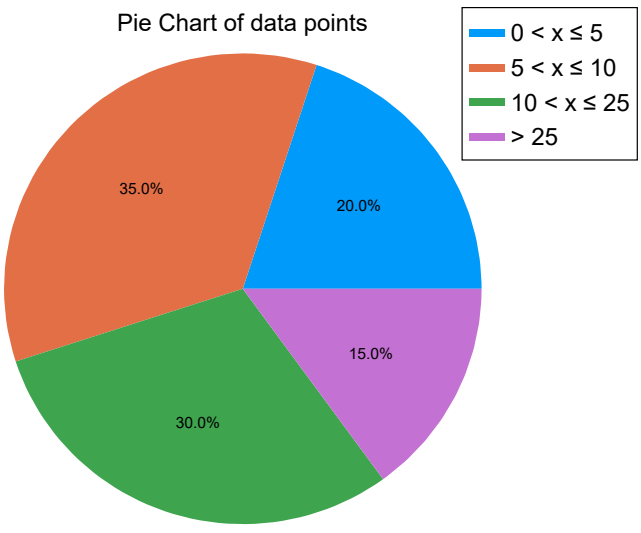
Pie chart

When an aggregate is divided into different components, we may be interested in the relative importance of the different components, rather than their absolute contribution to the aggregate. For representing breakdown of an aggregate into components a pie diagram is used. For pie diagram, one circle is used and the area enclosed by it being taken as 100. It is then divided into a number of sectors by drawing angles at the centre, the area of each sector representing the corresponding percentage. Since the full angle at the center is 360°, it is clear that for any particular category the angle (in degrees) should be 3.6 times the corresponding percentage.

When observations on discrete or continuous variables are available on a single characteristic of a large number of members often it becomes necessary to condense the data as far as possible without losing any information of interest. If the data is non-frequency type, then the first step of condensation is to classify different values or is to divide the observed range of the variable into a suitable member of groups or classes, according to their increasing order in terms of magnitude and to record the number of observations corresponding to each distinct value or falling in each class.

7.4 Frequency distribution

When observations on discrete or continuous variables are available on a single characteristic of a large number of members often it becomes necessary to condense the data as far as possible without losing any information of interest. If the data is non-frequency type, then the first step of condensation is to classify different values or is to divide the observed range of the variable into a suitable member of groups or classes, according to their increasing order in terms of magnitude and to record the number of observations corresponding to each distinct value or falling in each class.



This page intentionally left blank.

8

Estimation

Definition 8.1 Estimator
An **estimator** is a formula, that tells how to calculate the value of an estimate based on the observations contained in a sample.

For example, the sample mean

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

is a rule that tells us how to calculate the estimate of the population mean μ based on the observations in a sample.

Theorem 8.1 Mean Squared Error
For an estimator $\hat{\mu}(X)$ of $\mu = \mu(\theta)$, the mean-squared error is

$$MSE(\hat{\mu}) = Var[\hat{\mu}(X) \mid \theta] + \text{Bias}(\hat{\mu} \mid \theta)^2 \tag{8.1}$$

where $\text{Bias}(\hat{\mu} \mid \theta) = \mathbb{E}_\theta[\hat{\mu}(X) \mid \theta] - \mu$.

Proof. Consider the following decision framework:

- $X \sim P_\theta,\ \theta \in \Theta$.
- The parameter of interest, $\mu(\theta)$ is a certain function.
- Action space, $\mathcal{A} = \{\mu = \mu(\theta),\ \theta \in \Theta\}$.
- Decision procedure (or estimator), $\hat{\mu}(X) : \mathcal{X} \to \mathcal{A}$.
- Squared error loss as loss function: $L(\theta, a) = [a - \mu(\theta)]^2$.

with the setup above, the MSE is equal to the risk of decision,

$$\begin{aligned} R(\theta, \hat{\mu}(X)) &= \mathbb{E}[L((\theta, \hat{\mu}(X)) \mid \theta] \\ &= \mathbb{E}[(\hat{\mu}(X) - \mu(\theta))^2 \mid \theta] \\ &= \mathbb{E}[(\hat{\mu}(X) - \mu)^2 \mid \theta] \\ &= Var[\hat{\mu}(X) \mid \theta] + \big(\underbrace{\mathbb{E}[\hat{\mu}(X) \mid \theta] - \mu}_{\text{Bias}(\hat{\mu}|\theta)}\big)^2 \end{aligned}$$

□

8.1 Point Estimators

A **point estimator** is a function of the sample data that provides a single value as an estimate of an unknown population parameter. Since the estimator is calculated from a random sample, it is itself a random variable and has a probability distribution, called the **sampling distribution**.

The sampling distribution of a point estimator describes how the estimator varies from sample to sample. Key properties of the sampling distribution include its mean (which relates to bias) and its variance (which relates to the precision of the estimator). Understanding the sampling distribution is fundamental for assessing the reliability of an estimator, constructing confidence intervals, and performing hypothesis tests.

	Target Parameter	Sample size	Point Estimator	$\mathbb{E}[\theta]$	Standard Error
Population Mean	μ	n	$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$	μ	$\frac{\sigma}{\sqrt{n}}$
Proportion	p	n	$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$	p	$\sqrt{\frac{p(1-p)}{n}}$
Difference in Means	$\mu_1 - \mu_2$	m, n	$\overline{X} - \overline{Y}$	$\mu_1 - \mu_2$	$\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$
Difference in Proportions	$p_1 - p_2$	m, n	$\hat{p}_1 - \hat{p}_2$	$p_1 - p_2$	$\sqrt{\frac{p_1(1-p_1)}{m} + \frac{p_2(1-p_2)}{n}}$

✱ Example 8.1.1. In a random sample of 80 components of a certain type, 12 are found to be defective.

1. Find a point estimate of the proportion of non-defective components.
2. Find the standard error of the point estimate.

✲ Solution 1. With p as the proportion of non-defective components, the point estimate for proportion is

$$\hat{p} = \frac{80 - 12}{80} = 0.85.$$

2. The standard error of the point estimate of non-defective proportion is

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.85 \times 0.15}{80}} \approx 0.0399.$$

◀

✱ Example 8.1.2. Let X and Y denote the strengths of concrete beam and cylinder specimens, respectively. The following data were obtained:

X	5.9	7.2	7.3	6.3	8.1	6.8	7.0
	7.6	6.8	6.5	7.0	6.3	7.9	9.0
	8.2	8.7	7.8	9.7	7.4	7.7	9.7
	7.8	7.7	11.6	11.3	11.8	10.7	
Y	6.1	5.8	7.8	7.1	7.2	9.2	6.6
	8.3	7.0	8.3	7.8	8.1	7.4	8.5
	8.9	9.8	9.7	14.1	12.6	11.2	

Suppose $\mathbb{E}[X] = \mu_1$, $Var[X] = \sigma_1^2$, $\mathbb{E}[Y] = \mu_2$, and $Var[Y] = \sigma_2^2$.

1. Show that $\overline{X} - \overline{Y}$ is an unbiased estimator of $\mu_1 - \mu_2$.
2. Find the mean and standard error of the point estimate of $\mu_1 - \mu_2$.

✲ Solution 1. Since X and Y are independent, we have

$$\mathbb{E}[\overline{X} - \overline{Y}] = \mathbb{E}[\overline{X}] - \mathbb{E}[\overline{Y}] = \mu_1 - \mu_2.$$

Thus, $\overline{X} - \overline{Y}$ is an unbiased estimator of $\mu_1 - \mu_2$.

2. The mean of the point estimate is

$$\mathbb{E}[\overline{X} - \overline{Y}] = \bar{x} - \bar{y} = 8.141 - 8.575 = 0.434.$$

The variance of the difference in means is

$$Var[\overline{X} - \overline{Y}] = Var[\overline{X}] + Var[\overline{Y}] = \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}.$$

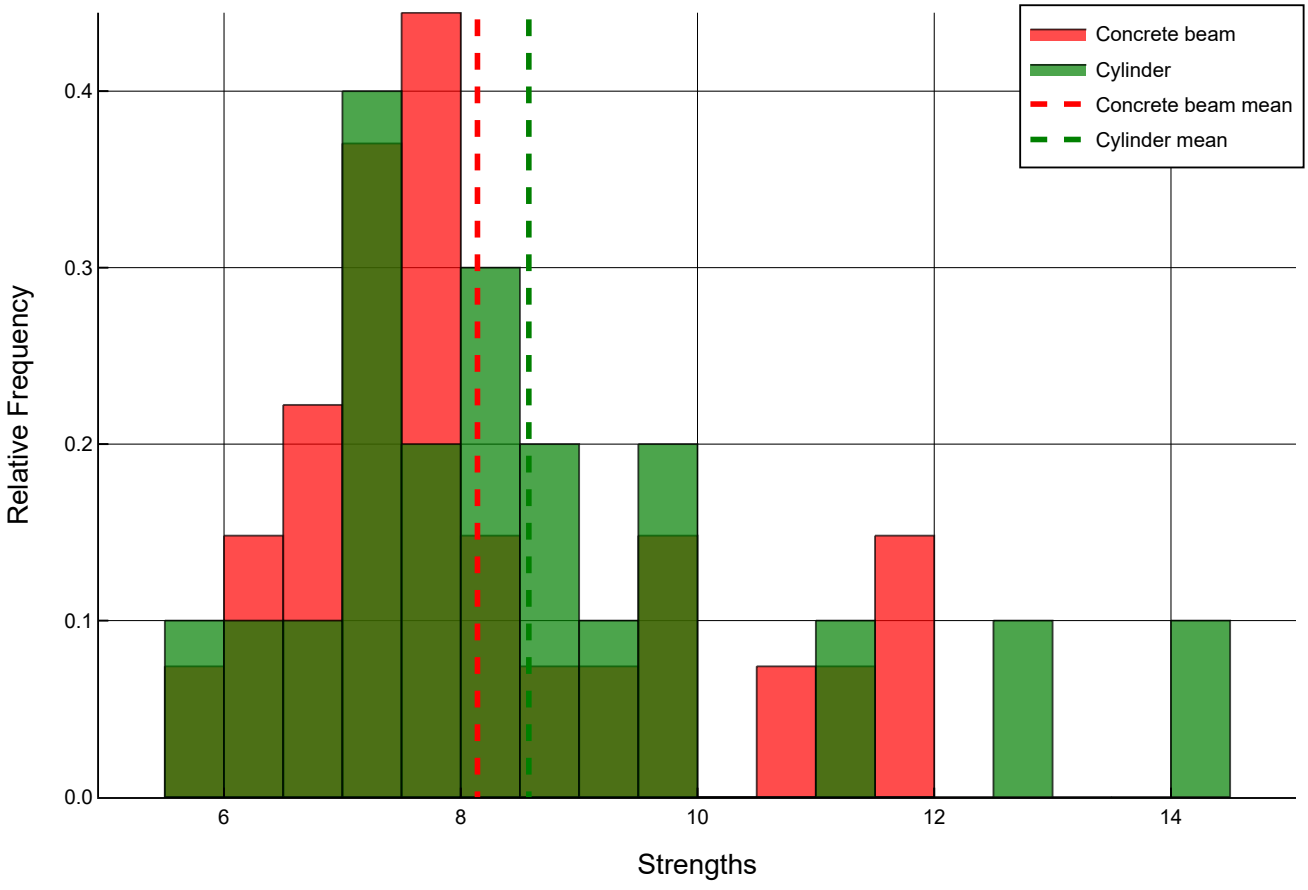
And

$$\sigma_{\overline{X}-\overline{Y}} = \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}.$$

Since σ_1^2 and σ_2^2 are unknown, we use s_X^2 and s_Y^2 to estimate σ_1^2 and σ_2^2 respectively. Thus, The standard error of the point estimate is

$$S_{\overline{X}-\overline{Y}} = \sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}} = \sqrt{\frac{1.666^2}{27} + \frac{2.104^2}{20}} = 0.5687.$$

◀



Remark. *Note that S_1 is not an unbiased estimator of σ_1. Similarly, S_1/S_2 is not an unbiased estimator of σ_1/σ_2.*

8.2 Evaluating the Estimators

Suppose $\hat{\theta}_1$ and $\hat{\theta}_2$ are two estimators of θ that are both unbiased. Then, although the distribution of each estimator is centered at the true value of θ, the spreads of the distributions about the true value may be different.

Among all estimators of θ that are unbiased, we will always choose the one that has minimum variance. WHY?

The resulting $\hat{\theta}$ is called the **minimum variance unbiased estimator (MVUE)** of θ.

Definition 8.2 Unbiased estimator
The estimator $\hat{\mu}$ is unbiased if $\text{Bias}(\hat{\mu} \mid \theta) = 0$

✱ Example 8.2.1. Let X_1, X_2, X_3 be a random sample of size 3 from a population with pmf

$$f(x|\lambda) = \begin{cases} \frac{\lambda^x e^{-\lambda}}{x!} & x = 0, 1, 2, \ldots \\ 0 & \text{otherwise} \end{cases}$$

where $\lambda > 0$ is a parameter. Are the following estimators of λ unbiased?

$$\hat{\lambda}_1 = \frac{1}{4}(X_1 + 2X_2 + X_3), \quad \hat{\lambda}_2 = \frac{1}{9}(4X_1 + 3X_2 + 2X_3)$$

Given, $\hat{\lambda}_1$ and $\hat{\lambda}_2$ which one is more efficient?

Hence, find an unbiased estimator of λ that is more efficient than both $\hat{\lambda}_1$ and $\hat{\lambda}_2$.

✲ Solution Given the observations X_1, X_2, X_3 are i.i.d with $X_i \sim Poisson(\lambda)$, we have

$$\mathbb{E}[X_i] = Var[X_i] = \lambda \quad \forall i = 1, 2, 3.$$

It is easy to see that

$$\mathbb{E}[\hat{\lambda}_1] = \frac{1}{4}(\mathbb{E}[X_1] + 2\mathbb{E}[X_2] + \mathbb{E}[X_3]) = \frac{1}{4}(\lambda + 2\lambda + \lambda) = \lambda,$$
$$\mathbb{E}[\hat{\lambda}_2] = \frac{1}{9}(4\mathbb{E}[X_1] + 3\mathbb{E}[X_2] + 2\mathbb{E}[X_3]) = \frac{1}{9}(4\lambda + 3\lambda + 2\lambda) = \lambda.$$

Thus, both $\hat{\lambda}_1$ and $\hat{\lambda}_2$ are unbiased estimators of λ. Next, we compute the variances of both estimators,

$$Var[\hat{\lambda}_1] = \frac{1}{16}(Var[X_1] + 4Var[X_2] + Var[X_3]) = \frac{1}{16}(\lambda + 4\lambda + \lambda) = \frac{3\lambda}{8},$$
$$Var[\hat{\lambda}_2] = \frac{1}{81}(16Var[X_1] + 9Var[X_2] + 4Var[X_3]) = \frac{1}{81}(16\lambda + 9\lambda + 4\lambda) = \frac{29\lambda}{81}.$$

By inspection, since $\frac{3}{8} = 0.375 > \frac{29}{81} \approx 0.358$, the estimator $\hat{\lambda}_2$ is more efficient than $\hat{\lambda}_1$. We have seen in previous section that the sample mean is always an unbiased estimator of the population mean irrespective of the population distribution. The variance of the sample mean is always equal to $\frac{\sigma^2}{n}$, where σ^2 is the population variance and n is the sample size. Thus

$$Var[\overline{X}] = \frac{Var[X_i]}{3} = \frac{1}{3}\lambda.$$

The sample mean has even smaller variance than both $\hat{\lambda}_1$ and $\hat{\lambda}_2$. Thus, $\overline{X} = \frac{1}{3}\lambda$ is an unbiased estimator of λ that is more efficient than both $\hat{\lambda}_1$ and $\hat{\lambda}_2$. ◀

✱ Example 8.2.2. Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be unbiased estimators of θ. Suppose $Var\left(\hat{\theta}_1\right) = 1$, $Var\left(\hat{\theta}_2\right) = 2$ and $\text{Cov}\left(\hat{\theta}_1, \hat{\theta}_2\right) = \frac{1}{2}$. What are the values of c_1 and c_2 for which $c_1\hat{\theta}_1 + c_2\hat{\theta}_2$ is an unbiased estimator of θ with minimum variance among unbiased estimators of this type?

✲ **Solution** We want to find c_1 and c_2 such that $c_1\hat{\theta}_1 + c_2\hat{\theta}_2$ to be a minimum variance unbiased estimator of θ. Then

$$\begin{aligned}\mathbb{E}[c_1\hat{\theta}_1 + c_2\hat{\theta}_2] = \theta &\implies c_1\mathbb{E}[\hat{\theta}_1] + c_2\mathbb{E}[\hat{\theta}_2] = \theta \\ &\implies c_1\theta + c_2\theta = \theta \\ &\implies c_1 + c_2 = 1 \\ &\implies c_2 = 1 - c_1.\end{aligned}$$

Therefore,

$$\begin{aligned}Var[c_1\hat{\theta}_1 + c_2\hat{\theta}_2] &= c_1^2 Var[\hat{\theta}_1] + c_2^2 Var[\hat{\theta}_2] + 2c_1c_2\,\mathrm{Cov}[\hat{\theta}_1, \hat{\theta}_2] \\ &= c_1^2(1) + 2(1-c_1)^2 + 2c_1(1-c_1)\left(\frac{1}{2}\right) \\ &= 3c_1^2 - 3c_1 + 2.\end{aligned}$$

To find the minimum variance, we differentiate $Var[c_1\hat{\theta}_1 + c_2\hat{\theta}_2]$ with respect to c_1 and set it to zero, that is

$$\frac{\mathrm{d}}{\mathrm{d}c_1} Var[c_1\hat{\theta}_1 + c_2\hat{\theta}_2] = 6c_1 - 3 = 0 \quad \implies c_1 = \frac{1}{2}.$$

Thus, $c_2 = 1 - c_1 = \frac{1}{2}$. Therefore, the minimum variance unbiased estimator of θ is

$$\hat{\theta} = \frac{1}{2}\hat{\theta}_1 + \frac{1}{2}\hat{\theta}_2.$$

In fact, if θ_1 and θ_2 are both unbiased estimators of θ, then the linear combination $c_1\theta_1 + c_2\theta_2$ is also an unbiased estimator of θ for any c_1, c_2 such that $c_1 + c_2 = 1$. Hence

$$\mathcal{C} = \{\hat{\theta} = c\hat{\theta}_1 + (1-c)\hat{\theta}_2 \mid c \in \mathbb{R}\}$$

◀

Rule of thumb choosing a good estimator:

- Unbiasedness: $\mathbb{E}[\hat{\theta}] = \theta$.
- Minimum variance: A good estimator should has smaller $Var[\hat{\theta}]$, the smaller the better.

8.2.1 Method of Moments (MoM) Estimator

The method of moments is a technique for estimating population parameters by equating sample moments to theoretical moments. Moments are quantitative measures related to the shape of a distribution, such as the mean (first moment), variance (second moment), skewness (third moment), and kurtosis (fourth moment). The method of moments involves the following steps:

1. Calculate the theoretical (population) moments as functions of the unknown parameters.

2. Calculate the corresponding sample moments from the observed data.

3. Set the population moments equal to the sample moments to create a system of equations.

4. Solve the system of equations for the unknown parameters to obtain the MoM estimators.

✱ Example 8.2.3. Let $X_1, X_2, \ldots, X_n$ be a random sample from a population with pdf

$$f(x|\theta) = \begin{cases} \theta x^{\theta-1} & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

where $\theta > 0$ is an unknown parameter. Find the method of moments estimator of θ.

✲ Solution To find the method of moments estimator, we shall equate the first population moment to the sample moment. The first population moment $\mathbb{E}[X]$ is given by

$$\begin{aligned} \mathbb{E}[X] &= \int_0^1 x f(x|\theta)\, \mathrm{d}x \\ &= \int_0^1 x(\theta x^{\theta-1})\, \mathrm{d}x \\ &= \theta \int_0^1 x^{\theta}\, \mathrm{d}x \\ &= \theta \left[\frac{x^{\theta+1}}{\theta+1}\right]_{x=0}^{x=1} = \frac{\theta}{\theta+1} = M_X(x). \end{aligned}$$

We know that the first moment $M_X(x) = \overline{X}$. Now setting $M_X(x) = \mathbb{E}X$ and solving for θ, we have

$$\overline{X} = \frac{\theta}{\theta+1}$$

that is

$$\theta = \frac{\overline{X}}{1-\overline{X}}.$$

where $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the sample mean. Thus, the statistic $\dfrac{\overline{X}}{1-\overline{X}}$ is an estimator of parameter θ. We write

$$\hat{\theta} = \frac{\overline{X}}{1-\overline{X}}.$$

Now let say we have the following sample data:

$$0.44, \quad 0.55, \quad 0.60, \quad 0.30$$

we have $\overline{X} = \frac{0.44 + 0.55 + 0.60 + 0.30}{4} = 0.4725$, and the estimate of θ is

$$\hat{\theta} = \frac{0.4725}{1 - 0.4725} = 0.8957.$$

◀

✱ Example 8.2.4. Let $X \sim \mathcal{N}(\mu, \sigma^2)$, and $X_1, X_2, \ldots X_n$ be a random sample of size n from X. Find the method of moments estimators of μ and σ^2.

✲ Solution The first population moment is

$$\mathbb{E}[X] = \mu.$$

The second population moment is

$$\mathbb{E}[X^2] = Var[X] + (\mathbb{E}[X])^2 = \sigma^2 + \mu^2.$$

The first sample moment is

$$M_1 = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

The estimator of the parameter μ is $\hat{\mu} = \overline{X}$, that is

$$\hat{\mu} = \overline{X} = \frac{1}{n}\sum_{i=0}^{n} X_i.$$

Next, we equate the second population moment to the second sample moment. Note that the variance of the population is

$$\begin{aligned}
\sigma^2 &= \mathbb{E}[X^2] - \mu^2 \\
&= M_2 - \mu^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\overline{X})^2. \\
&= \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X})^2.
\end{aligned}$$

The last line follows from the fact that

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

Thus, the estimator of the parameter σ^2 is

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

◀

Theorem 8.2
Let $X_1, X_2, \ldots, X_n$ be a random sample with $\mathbb{E}[X_i]=\mu$ and $Var[X_i]=\sigma^2$. Then

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

is a *biased estimator* of σ^2 but that

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

is an *unbiased estimator* of σ^2.

Proof. From previous example, we can see that

$$\sum_{i=1}^{n}(X_i-\overline{X})^2=\sum_{i=1}^{n}X_i^2-n(\overline{X})^2 \qquad (\bigstar)$$

Hence, we use this and the fact that

$$\mathbb{E}[X_i^2]=Var[X_i]+(\mathbb{E}[X_i])^2=\sigma^2+\mu^2,$$

and

$$\mathbb{E}[\overline{X}^2]=Var[\overline{X}]+(\mathbb{E}[\overline{X}])^2=\frac{\sigma^2}{n}+\mu^2,$$

and take expectation on both sides of (★), we have

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^{n}(X_i-\overline{X})^2\right] &= \mathbb{E}\left[\sum_{i=1}^{n}X_i^2 - n(\overline{X})^2\right] \\ &= \sum_{i=1}^{n}\mathbb{E}[X_i^2] - n\mathbb{E}[(\overline{X})^2] \\ &= n(\sigma^2+\mu^2) - n\left(\frac{\sigma^2}{n}+\mu^2\right) \\ &= (n-1)\sigma^2.\end{aligned}$$

It follows that

$$\mathbb{E}[\tilde{S}^2] = \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n}(X_i-\overline{X})^2\right] = \frac{n-1}{n}\sigma^2 \neq \sigma^2,$$

and that $\tilde{S}^2$ is biased since $\mathbb{E}[\tilde{S}^2] \neq \sigma^2$. However,

$$\mathbb{E}[S^2] = \frac{1}{n-1}\mathbb{E}\left[\sum_{i=1}^{n}(X_i-\overline{X})^2\right] = \sigma^2,$$

thus we can see that S^2 is an unbiased estimator of σ^2. □

8.3 Maximum Likelihood Estimator (MLE)

The maximum likelihood estimation (MLE) is a method used to estimate the parameters of a statistical model. The MLE is the parameter value that maximizes the likelihood function, which measures how likely it is to observe the given sample data for different parameter values. Next, we describe this method in detail.

Definition 8.3 Likelihood function and MLE
Let $X_1, X_2, \ldots, X_n$ be a random sample from a population with pdf/pmf $f(x|\theta)$, where $\theta \in \Theta$ is an unknown parameter. The likelihood function is defined as

$$L(\theta|x) = \prod_{i=1}^{n} f(x_i|\theta).$$

The maximum likelihood estimator (MLE) of θ is the value of θ that maximizes the likelihood function $L(\theta)$.

This definition state that the likelihood function $L(\theta|x)$ is the product of the individual pdf evaluated at each observation in the sample, given the parameter θ. The likelihood function represents the joint density of a random sample $X_1, X_2, \ldots, X_n$ given the parameter θ. The MLE is the value of θ that makes the observed data most probable.

The θ that maximizes $L(\theta|x)$ is called the maximum likelihood estimate and is denoted by $\hat{\theta}$.

$$\hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta|x).$$

In practice, it is often more convenient to work with the natural logarithm of the likelihood function, known as the log-likelihood function:

$$\ell(\theta) = \ln L(\theta) = \sum_{i=1}^{n} \ln f(x_i|\theta).$$

Maximizing the log-likelihood function is equivalent to maximizing the likelihood function itself, as the logarithm is a monotonically increasing function.

✱ Example 8.3.1. Let $X \sim B(1, p)$, a Bernoulli random variable with parameter p, with pmf

$$f(x|p) = \mathbb{P}[X = x|p] = \begin{cases} p^x(1-p)^{1-x} & x = 0, 1 \\ 0 & \text{otherwise} \end{cases}$$

where $0 < p < 1$. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from X. Find the maximum likelihood estimator of p.

✲ Solution Our goal is to find the value of p that maximizes the likelihood function based on the observed sample data $X = (X_1, X_2, \ldots, X_n)$. Note that $X_1, X_2, \ldots, X_n$ are i.i.d. Thus, the likelihood function is given by

$$\begin{aligned} L(p|x) &= \prod_{i=1}^{n} f(x_i|p) \\ &= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \\ &= p^{\sum_{i=1}^n x_i}(1-p)^{\sum_{i=1}^n (1-x_i)} \\ &= p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}. \end{aligned}$$

We can simplify the notation by letting $S_n = \sum_{i=1}^n x_i$. We want to choose p such that $L(p|x)$ is maximized. Take the logarithm of the likelihood function, we have

$$\ell(p|x) = \ln L(p|x) = S_n \ln p + (n - S_n) \ln(1-p).$$

To find the maximum, we take the derivative of $\ell(p|x)$ with respect to p and set it to zero:

$$\begin{aligned}
\frac{\partial \ell(p|x)}{\partial p} &= \frac{S_n}{p} - \frac{n - S_n}{1-p} = 0 \\
\Rightarrow S_n(1-p) &= (n - S_n)p \\
\Rightarrow S_n - S_n p &= np - S_n p \\
\Rightarrow S_n &= np \\
\Rightarrow \hat{p} = \frac{S_n}{n} &= \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{X}.
\end{aligned}$$

The sample mean (proportion) $\overline{X}$ is the maximum likelihood estimator of p. ◀

8.3.1 MLE based on grouped data

In some cases, data may be grouped into intervals or categories, and we may not have access to the individual data points. In such situations, we can still use the maximum likelihood estimation (MLE) method to estimate parameters based on the grouped data.

In the complete data case, the likelihood is measured by the density (or probability) $f(x_i)$ at the known data point x_i. The likelihood function is the product of those densities for the points in the sample. In the interval grouped data case, we measure likelihood of a point as the probability of the interval in which that point occurs. For a data point in the interval $(c_{j-1}, c_j]$, the probability of that interval is $[F(c_j;\theta) - F(c_{j-1};\theta)]$. The likelihood function is the product of those interval probabilities for all of the sample points. Since there are n_j sample points in the interval $(c_{j-1}, c_j]$, the likelihood function will include a factor of $[F(c_j;\theta) - F(c_{j-1};\theta)]^{n_j}$ for those n_j points. The overall likelihood function is the product of all of those factors:

Definition 8.4 MLE based on grouped data
If the data is grouped into k intervals with counts $n_1, n_2, \ldots, n_k$ in each interval, the likelihood function for the grouped data is given by

$$L(\theta|x) = \prod_{j=1}^{k} [F(c_j|\theta) - F(c_{j-1}|\theta)]^{n_j},$$

where $F(x|\theta)$ is the cumulative distribution function (CDF) of the underlying distribution, and $[c_{j-1}, c_j)$ is the j-th interval.

✱ Example 8.3.2. For a group of insurance policies, you are given:

1. The losses follow the distribution function

$$F(x|\theta) = 1 - \frac{\theta}{x}, \quad \theta < x < \infty.$$

2. A sample of 20 losses is grouped as follows:

Interval	Number of loss
$x \leq 10$	9
$10 < x \leq 25$	6
$x > 25$	5

Calculate the maximum likelihood estimate of θ.

✲ Solution The likelihood function is the product of the probabilities of observing the data in each interval, The probability for the interval $x \leq 10$ is given by

$$F(10|\theta) = 1 - \frac{\theta}{10},$$

the probability for the interval $10 < x \leq 25$ is given by

$$F(25|\theta) - F(10|\theta) = \left(1 - \frac{\theta}{25}\right) - \left(1 - \frac{\theta}{10}\right) = \frac{\theta}{10} - \frac{\theta}{25} = \frac{3}{50}\theta,$$

and the probability for the interval $x > 25$ is given by

$$1 - F(25|\theta) = 1 - \left(1 - \frac{\theta}{25}\right) = \frac{\theta}{25}.$$

Then, the likelihood function is given by

$$\begin{aligned} L(\theta) &= [F(10|\theta)]^{n_1}[F(25|\theta) - F(10|\theta)]^{n_2}[1 - F(25|\theta)]^{n_3} \\ &= \left(1 - \frac{\theta}{10}\right)^9 \left(\frac{3\theta}{50}\right)^6 \left(\frac{\theta}{25}\right)^5 \\ &= c(10-\theta)^9\theta^{11}, \end{aligned}$$

where $c = \dfrac{3^6}{50^6 \times 25^5}$ is a constant. To find the value of θ that maximizes $L(\theta)$, we take the logarithm of the likelihood function:

$$\ell(\theta) = \ln L(\theta) = \ln c + 9\ln(10-\theta) + 11\ln\theta.$$

Next, we differentiate $\ell(\theta)$ with respect to θ and set it to zero:

$$\begin{aligned} \frac{\mathrm{d}\ell(\theta)}{\mathrm{d}\theta} = \frac{9}{10-\theta}(-1) + \frac{11}{\theta} &= 0 \\ \Rightarrow -\frac{9}{10-\theta} + \frac{11}{\theta} &= 0 \\ \Rightarrow 11(10-\theta) &= 9\theta \\ \Rightarrow 110 - 11\theta &= 9\theta \\ \Rightarrow 110 &= 20\theta \\ \Rightarrow \hat{\theta} &= 5.5 \end{aligned}$$

Thus, the maximum likelihood estimate of θ is $\hat{\theta} = 5.5$. ◀

✱ Example 8.3.3. A grouped data set has 20 data points grouped into the following intervals:

Interval	$0 < x \leq 5$	$5 < x \leq 10$	$10 < x \leq 25$	$x > 25$
Number of data points	4	7	6	3

Apply the maximum likelihood method to estimate the parameter θ of the following two cases:

1. The data follow the exponential distribution with parameter θ,
2. The data follow the uniform distribution on the interval $(0, \theta)$.

✲ Solution 1. If we assume that the data follow the exponential distribution with parameter θ, then the cdf is given by

$$F(x|\theta) = 1 - e^{-\theta x}, \quad x > 0, \theta > 0.$$

The likelihood function is given by

$$\begin{aligned} L(\theta) &= [F(5|\theta)]^{n_1}[F(10|\theta) - F(5|\theta)]^{n_2}[F(25|\theta) - F(10|\theta)]^{n_3}[1 - F(25|\theta)]^{n_4} \\ &= (1 - e^{-5\theta})^4(e^{-5\theta} - e^{-10\theta})^7(e^{-10\theta} - e^{-25\theta})^6(e^{-25\theta})^3 \\ &= c(1 - e^{-5\theta})^4(e^{-5\theta})^7(1 - e^{-5\theta})^6(e^{-25\theta})^3 \\ &= c(1 - e^{-5\theta})^{10}(e^{-5\theta})^{10}(e^{-25\theta})^3 \\ &= c(1 - e^{-5\theta})^{10}e^{-50\theta}, \end{aligned}$$

where $c = 1$ is a constant. To find the value of θ that maximizes $L(\theta)$, we take the logarithm of the likelihood function:

2. In another case, if we assume that the data follow the uniform distribution on the interval $(0, \theta)$, then the cdf is given by

$$F(x|\theta) = \begin{cases} \frac{x}{\theta} & 0 < x < \theta \\ 0 & \text{otherwise} \end{cases}$$

The likelihood function is given by

$$\begin{aligned} L(\theta) &= [F(5|\theta)]^{n_1}[F(10|\theta) - F(5|\theta)]^{n_2}[F(25|\theta) - F(10|\theta)]^{n_3}[1 - F(25|\theta)]^{n_4} \\ &= \left(\frac{5}{\theta}\right)^4 \left(\frac{10}{\theta} - \frac{5}{\theta}\right)^7 \left(\frac{25}{\theta} - \frac{10}{\theta}\right)^6 \left(1 - \frac{25}{\theta}\right)^3 \\ &= c\theta^{-20}(\theta - 25)^3, \end{aligned}$$

where $c = 5^4 \times 5^7 \times 15^6$ is a constant. To find the value of θ that maximizes $L(\theta)$, we take the logarithm of the likelihood function:

$$\ell(\theta) = \ln L(\theta) = \ln c - 20 \ln \theta + 3 \ln(\theta - 25).$$

Next, we differentiate $\ell(\theta)$ with respect to θ and set it to zero:

$$\begin{aligned} \frac{\mathrm{d}\ell(\theta)}{\mathrm{d}\theta} &= -\frac{20}{\theta} + \frac{3}{\theta - 25} = 0 \\ \Rightarrow -20(\theta - 25) + 3\theta &= 0 \\ \Rightarrow -20\theta + 500 + 3\theta &= 0 \\ \Rightarrow -17\theta + 500 &= 0 \\ \Rightarrow \hat{\theta} &= \frac{500}{17} \approx 29.41. \end{aligned}$$

Thus, the maximum likelihood estimate of θ is $\hat{\theta} = \frac{500}{17} \approx 29.41$.

◀

Tutorials

Exercise 8.1 Given 20 observations on breakdown voltage for some materials

24.46	25.61	26.25	26.42	26.66	27.15	27.31	27.54	27.74	27.94
27.98	28.04	28.28	28.49	28.50	28.87	29.11	29.13	29.50	30.88

Assume that after looking at the histogram, we think that the distribution of breakdown voltage is normal with mean value μ. What are some point estimators for μ?

Exercise 8.2 The probability density function of the random variable X is defined by

$$f(x|\lambda) = 1 - \frac{2}{3}\lambda + \lambda\sqrt{x}, \quad 0 \leq x \leq 1,$$

and zero otherwise. What is the maximum likelihood estimate of the parameter λ based on the two independent observations $x_1 = \frac{1}{4}$ and $x_2 = \frac{9}{16}$?

Exercise 8.3 Consider that $X_1, X_2, \ldots, X_n$ is a random sample from a distribution with density function

$$f(x|\beta) = \begin{cases} \frac{x^6 e^{-x/\beta}}{\Gamma(7)\,\beta^7} & \text{if } x > 0 \\ 0 & \text{otherwise.} \end{cases}$$

What is the maximum likelihood estimate of β?

Exercise 8.4 Let $X_1, X_2, \ldots, X_n$ be a random sample from the uniform density function

$$f(x|\theta) = \begin{cases} \frac{1}{\theta} & \text{if } 2\theta \leq x \leq 3\theta \\ 0 & \text{otherwise} \end{cases}$$

where $\theta > 0$. Show that the maximum likelihood estimate of θ is

$$\frac{1}{3} \max\{X_1, X_2, \ldots, X_n\}.$$

Exercise 8.5 What is the maximum likelihood estimate of β if five values $\frac{4}{5}, \frac{2}{3}, 1, \frac{3}{2}, \frac{5}{4}$ were drawn from the population for which $f(x|\beta) = \frac{1}{2}(1+\beta)^5 \left(\frac{x}{2}\right)^\beta$?

Exercise 8.6 Eight independent trials are conducted of a given system with the following results:

$$S, F, S, F, S, S, S, S$$

where S denotes the success and F denotes the failure. What is the maximum likelihood estimate of the probability of successful operation p?

Exercise 8.7 Suppose fertilizer-1 has a mean yield per acre of μ_1 with variance σ^2, whereas the expected yield for fertilizer-2 is μ_2 with the same variance σ^2. Let S_i^2 denote the sample variances of yields based on sample sizes n_1 and n_2 respectively, of the two fertilizers.

1. Show that the pooled (combined) estimator

$$\hat{\sigma}^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

   is an unbiased estimator of σ^2.

2. The measurements for the two types of fertilizers were obtained independently, with testing samples of $n_1 = n_2 = 100$, the following sample means and variances were computed:

$$\bar{x}_1 = 179.3 \text{ yield per acre}, \quad \bar{x}_2 = 190.0 \text{ yield per acre},$$

$$s_1^2 = 1440.80, \quad s_2^2 = 1960,$$

   Estimate the difference mean yied per acre, and the pooled variance of two fertilizers.

Exercise 8.8 A random sample $X_1, X_2, \ldots, X_n$ of size n is selected from a normal distribution with variance σ^2. Let S^2 be the unbiased estimator of σ^2, and T be the maximum likelihood

estimator of σ^2. If $20T - 19S^2 = 0$, then what is the sample size?

Exercise 8.9 A box contains 50 red and blue balls out of which are red. A sample of 30 balls is to be selected without replacement. If X denotes the number of red balls in the sample, then find an estimator for using the *moment method*.

Exercise 8.10 What is the difference between the parameters and statistics?

Exercise 8.11 You are given:

1. Losses follow an exponential distribution with mean θ.
2. A random sample of 20 losses is observed as follows:

Loss Range	Frequency
$[0, 1000]$	7
$(1000, 2000]$	6
$(2000, \infty)$	7

Calculate the maximum likelihood estimate of θ.

This page intentionally left blank.

9

Evaluating the goodness of estimators

9.1 Relative efficiency

It is usually possible to retrieve more than one unbiased estimator for the same target parameter θ. But we only prefer to use the estimator with the **smaller variance**. That is, if $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are unbiased estimators for the same parameter θ, we said that $\widehat{\theta}_1$ is *relatively more efficient* than $\widehat{\theta}_2$ if

$$Var(\widehat{\theta}_1) < Var(\widehat{\theta}_2).$$

In fact, this can be expressed as the ratio $Var(\widehat{\theta}_1)/Var(\widehat{\theta}_2)$ to measure the relative efficiency of these two unbiased estimators.

Definition 9.1 Relative efficiency
Given two unbiased estimators $\widehat{\theta}_1$ and $\widehat{\theta}_2$ for the same parameter θ, with variances $Var(\widehat{\theta}_1)$ and $Var(\widehat{\theta}_2)$, respectively. Then the efficiency of $\widehat{\theta}_1$ relative to $\widehat{\theta}_2$, wrote as $\mathsf{Eff}(\widehat{\theta}_1, \widehat{\theta}_2)$, is defined to be the ratio

$$\mathsf{Eff}(\widehat{\theta}_1, \widehat{\theta}_2) := \frac{Var(\widehat{\theta}_2)}{Var(\widehat{\theta}_1)}. \tag{9.1}$$

✱ **Example 9.1.1.** If $Y_1, Y_2, \ldots, Y_n$ denote a random sample from the uniform distribution on the interval $(0, \theta)$. The two unbiased estimators for θ are

$$\widehat{\theta}_1 = 2\overline{Y}, \quad \widehat{\theta}_2 = \left(\frac{n+1}{n}\right) Y_{(n)},$$

where $Y_{(n)} = \max\{Y_1, Y_2, \ldots, Y_n\}$. Find the efficiency of $\widehat{\theta}_1$ relative to $\widehat{\theta}_2$.

✲ **Solution** Because each $Y_i \sim U(0, \theta)$, we have

$$\mu = \mathbb{E}[Y_i] = \frac{\theta}{2}, \quad \sigma^2 = Var(Y_i) = \frac{\theta^2}{12}.$$

Therefore,

$$\mathbb{E}[\widehat{\theta}_1] = \mathbb{E}(2\overline{Y}) = 2\mathbb{E}(\overline{Y}) = 2\mu = \theta,$$

and that $\widehat{\theta}_1$ is unbiased, as claimed. Further we check that

$$Var(\widehat{\theta}_1) = Var(2\overline{Y}) = 4Var(\overline{Y}) = 4 \cdot \frac{\sigma^2}{n} = \frac{\theta^2}{3n},$$

The mean of this order statistic is

$$\begin{aligned}\mathbb{E}[Y_{(n)}] = \int_0^\theta y \cdot n \left(\frac{y}{\theta}\right)^{n-1} \left(\frac{1}{\theta}\right) \mathrm{d}y &= \frac{n}{\theta^n} \int_0^\theta y^n \, \mathrm{d}y \\ &= \frac{n}{\theta^n} \left[\frac{y^{n+1}}{n+1}\right]_{y=0}^{y=\theta} \\ &= \frac{n\theta}{n+1}.\end{aligned}$$

and it follows that

$$\mathbb{E}\left[\frac{n+1}{n} Y_{(n)}\right] = \theta;$$

that is, $\widehat{\theta}_2$ is also an unbiased estimator of θ. Since

$$\begin{aligned}\mathbb{E}[Y_{(n)}^2] = \int_0^\theta y^2 \cdot n \left(\frac{y}{\theta}\right)^{n-1} \left(\frac{1}{\theta}\right) \mathrm{d}y &= \frac{n}{\theta^n} \int_0^\theta y^{n+1} \, \mathrm{d}y \\ &= \frac{n}{\theta^n} \left[\frac{y^{n+2}}{n+2}\right]_{y=0}^{y=\theta} \\ &= \frac{n\theta^2}{n+2}.\end{aligned}$$

we obtain

$$\begin{aligned}Var[Y_{(n)}] = \mathbb{E}[Y_{(n)}^2] - (\mathbb{E}[Y_{(n)}])^2 &= \frac{n\theta^2}{n+2} - \left(\frac{n\theta}{n+1}\right)^2 \\ &= \theta^2 \left[\frac{n}{n+2} - \frac{n^2}{(n+1)^2}\right] \\ &= \theta^2 \left[\frac{n(n+1)^2 - n^2(n+2)}{(n+2)(n+1)^2}\right] \\ &= \theta^2 \left[\frac{n^3 + 2n^2 + n - n^3 - 2n^2}{(n+2)(n+1)^2}\right] \\ &= \frac{n\theta^2}{(n+2)(n+1)^2}.\end{aligned}$$

and hence

$$\begin{aligned} Var[\widehat{\theta}_2] &= Var\left[\frac{n+1}{n}Y_{(n)}\right] \\ &= \left(\frac{n+1}{n}\right)^2 Var[Y_{(n)}] \\ &= \left(\frac{n+1}{n}\right)^2 \cdot \frac{n\theta^2}{(n+2)(n+1)^2} \\ &= \frac{\theta^2}{n(n+2)}. \end{aligned}$$

Therefore, the efficiency of $\widehat{\theta}_1$ relative to $\widehat{\theta}_2$ is

$$\text{Eff}(\widehat{\theta}_1, \widehat{\theta}_2) = \frac{Var(\widehat{\theta}_2)}{Var(\widehat{\theta}_1)} = \frac{\theta^2/[n(n+2)]}{\theta^2/(3n)} = \frac{3}{n+2}.$$

This efficiency is less than 1 if $n > 1$. That is, if $n > 1$, then $\widehat{\theta}_2$ must have a smaller variance than $\widehat{\theta}_1$. And in result, θ_2 is generally more preferable to θ_1 when estimating θ. ◀

9.2 Consistency

Definition 9.2 Consistency
An estimator $\widehat{\theta}_n$ is consistent for parameter θ if for every $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|\widehat{\theta}_n - \theta| \geq \epsilon) = 0. \tag{9.2}$$

In other words, $\widehat{\theta}_n$ converges in probability to θ as n goes to infinity, denoted by

$$\widehat{\theta}_n \xrightarrow{p} \theta, \quad \text{as } n \to \infty.$$

Theorem 9.1
An unbiased estimator $\widehat{\theta}_n$ is consistent for parameter θ if

$$\lim_{n\to\infty} Var(\widehat{\theta}_n) = 0. \tag{9.3}$$

✱ **Example 9.2.1.** Let $X_1, X_2, \ldots, X_n$ be a random sample such that $\mathbb{E}[X_i] = \mu$, $\mathbb{E}[X_i^2] = \mu_2'$ and $\mathbb{E}[X_i^4] = \mu_4'$ are all finite. Show that

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

is a consistent estimator of $\sigma^2 = Var(X_i)$.

✲ Solution We are going to use subscript n on both $\widehat{\sigma}_n^2$ and $\overline{X}$ to explicitly convey their dependence on the value of the sample size n.

From previous example, we had derived that

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}_n^2\right) = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}_n^2\right).$$

The statistic $\frac{1}{n}\sum_{i=1}^{n} X_i^2$ is the average of n independent and identically distributed random variables $X_1^2, X_2^2, \ldots, X_n^2$ with $\mathbb{E}[X_i^2] = \mu_2'$ and $V(X_i^2) = \mu_4' - (\mu_2')^2$.

By the weak law of large numbers, we have

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{p} \mu_2', \quad \text{as } n \to \infty,$$

and certainly $\overline{X}_n \xrightarrow{p} \mu$. And because the function $g(x) = x^2$ is continuous for all $x \in \mathbb{R}$. This implies that $\overline{X}_n^2 \xrightarrow{p} \mu^2$. It follows that

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}_n^2 \xrightarrow{p} \sigma^2.$$

Note that $\frac{n}{n-1} \to 1$ when n goes to infinity. We can conclude that

$$\widehat{\sigma}_n^2 \xrightarrow{p} (1)\sigma^2 = \sigma^2, \quad \text{as } n \to \infty.$$

Equivalently, $\widehat{\sigma}_n^2$, the sample variance, is a consistent estimator of σ^2, the population variance. ◀

✱ Example 9.2.2. Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal population X with mean μ and variance $\sigma^2 > 0$. Is the likelihood estimator

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

of σ^2 a consistent estimator of σ^2?

✲ Solution Since $\widehat{\sigma}^2$ depends on the sample size n, we denote $\widehat{\sigma}^2$ as $\widehat{\sigma}_n^2$. Hence

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

The variance of $\widehat{\sigma}_n^2$ is given by

$$\begin{aligned} Var\left(\widehat{\sigma}_n^2\right) = Var\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2\right) &= \frac{1}{n^2}Var\left(\sigma^2\frac{(n-1)S^2}{\sigma^2}\right) \\ &= \frac{\sigma^4}{n^2}Var\left(\frac{(n-1)S^2}{\sigma^2}\right) \\ &= \frac{\sigma^4}{n^2}Var\left(\chi^2(n-1)\right) \\ &= \frac{2(n-1)\sigma^4}{n^2} \\ &= \left[\frac{1}{n}-\frac{1}{n^2}\right]2\sigma^4. \end{aligned}$$

Hence

$$\lim_{n\to\infty} Var\left(\widehat{\theta}_n\right) = \lim_{n\to\infty}\left[\frac{1}{n}-\frac{1}{n^2}\right]2\sigma^4 = 0.$$

The biased $B\left(\widehat{\theta}_n,\theta\right)$ is given by

$$\begin{aligned} B\left(\widehat{\theta}_n,\theta\right) = E\left(\widehat{\theta}_n\right)-\sigma^2 = E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2\right)-\sigma^2 &= \frac{1}{n}E\left(\sigma^2\frac{(n-1)S^2}{\sigma^2}\right)-\sigma^2 \\ &= \frac{\sigma^2}{n}E\left(\chi^2(n-1)\right)-\sigma^2 \\ &= \frac{(n-1)\sigma^2}{n}-\sigma^2 \\ &= -\frac{\sigma^2}{n}. \end{aligned}$$

Thus

$$\lim_{n\to\infty} B\left(\widehat{\theta}_n,\theta\right) = -\lim_{n\to\infty}\frac{\sigma^2}{n} = 0.$$

Hence $\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2$ is a consistent estimator of σ^2.

In the last example we saw that the likelihood estimator of variance is a consistent estimator. In general, if the density function $f(x|\theta)$ of a population satisfies some mild conditions, then the maximum likelihood estimator of θ is consistent. Similarly, if the density function $f(x;\theta)$ of a population satisfies some mild conditions, then the estimator obtained by moment method is also consistent.

Since consistency is a large sample property of an estimator, some statisticians suggest that consistency should not be used alone for judging the goodness of an estimator; rather it should be used along with other criteria. ◀

9.3 Sufficiency

Definition 9.3 Sufficiency
A statistic $T(X)$ is sufficient for parameter $\theta \in \Theta$ if the conditional distribution of the sample X given the statistic $T(X)$ does not depend on the parameter θ. In other words, once we know the value of the sufficient statistic, the sample provides no additional information about the parameter.

✱ Example 9.3.1. If $X_1, X_2, \ldots, X_n$ are i.i.d. random samples from the Bernoulli distribution with parameter p, then the sum of the samples with denisty function

$$f_X(x|\theta) = \begin{cases} \theta^x(1-\theta)^{1-x}, & x = 0, 1 \\ 0 & \text{otherwise.} \end{cases}$$

where $0 < \theta < 1$. Show that the statistic $T(X) = \sum_{i=1}^n X_i$ is a sufficient statistic for θ.

✲ Solution First, we find the joint density function of the sample:

$$f_X(x|\theta) = \prod_{i=1}^n f_{X_i}(x_i|\theta) = \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}.$$

Since each X_i is either 0 or 1, the sum $\sum_{i=1}^n x_i$ counts the number of successes (1s) in the sample. Let $T(X) = \sum_{i=1}^n X_i$. Then we can rewrite the joint density function as:

$$Y = \sum_{i=1}^n X_i \sim Bin(n, \theta),$$

Thus, the joint density function can be expressed as:

$$Y \sim g(y) = \binom{n}{y}\theta^y(1-\theta)^{n-y}, \quad y = 0, 1, \ldots, n.$$

◀

✱ Example 9.3.2. Let $X_1, \ldots, X_n$ be iid $N(\theta, \sigma_0^2)$ r.v.'s where σ_0^2 is known. Evaluate whether $T(X) = (\sum_{i=1}^n X_i)$ is sufficient for θ.

✲ Solution We consider the transformation of $X = (X_1, X_2, \ldots, X_n)$ to $Y = (T, Y_2, Y_3, \ldots, Y_n)$ where $T = \sum X_i$ and $Y_2 = X_2 - X_1, Y_3 = X_3 - X_1, \ldots, Y_n = X_n - X_1$. The transformation is 1-1, and the Jacobian of the transformation is 1.

The joint distribution of $X|\theta$ is $N_n(\mu \times 1, \sigma_0^2 I_n)$, where μ represents the mean parameter θ and 1 is the vector of ones. The joint distribution of $Y|\theta$ is $N_n(\mu_Y, \Sigma_{YY})$ where $\mu_Y = (n\theta, 0, 0, \ldots, 0)^T$ and the

covariance matrix is

$$\Sigma_{YY} = \left[\begin{array}{c|cccccc} n\sigma_0^2 & 0 & 0 & 0 & \cdots & 0 \\ \hline 0 & 2\sigma_0^2 & \sigma_0^2 & \sigma_0^2 & \cdots & \sigma_0^2 \\ 0 & \sigma_0^2 & 2\sigma_0^2 & \sigma_0^2 & \cdots & \sigma_0^2 \\ 0 & \sigma_0^2 & \sigma_0^2 & 2\sigma_0^2 & \cdots & \sigma_0^2 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \sigma_0^2 & \sigma_0^2 & \sigma_0^2 & \cdots & 2\sigma_0^2 \end{array}\right].$$

Since T and $(Y_2, \ldots, Y_n)$ are independent, it follows that $(Y_2, \ldots, Y_n)$ given $T = t$ has the unconditional distribution, which means T is a sufficient statistic for θ.

We note that all functions of $(Y_2, \ldots, Y_n)$ are independent of θ and T, which yields independence of $\bar{X}$ and s^2 where

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n-1}\left[\sum_{j=1}^{n}(X_j - X_j)^2\right].$$

◀

However, the sufficient statistics are not unique. For example, if $T(X)$ is a sufficient statistic, then any one-to-one function of $T(X)$ is also a sufficient statistic. the observation X itself is always a sufficient for θ, but it is not very useful since it does not reduce the data; Said, if we take $T(X) = X$, then $g(t, \theta) = f_X(t|\theta)$ and $h(x) = 1$.

But this is not much useful since it does not reduce the data.

For example, if the sample space $\mathcal{X}^n$ is partitioned into subsets $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_k$ such that the conditional distribution of X given $X \in \mathcal{A}_i$ does not depend on θ, then the statistic $T(X)$ that indicates which subset X belongs to is a sufficient statistic for θ.

Definition 9.4 Minimal sufficient statistic
A sufficient statistic $T(X)$ is minimal sufficient if it is a function of every other sufficient statistic. For example, if $S(X)$ is another sufficient statistic, then

$$S(X) = S(Y) \implies T(X) = T(Y).$$

Theorem 9.2
Consider a statistical decision problem with sample space

- random variable X with measure $\mathbb{P}_\theta$, the parameter $\theta \in \Theta$

Theorem 9.3 Factorization theorem
A statistic $T(X)$ with range $\mathcal{T}$ sufficient for parameter $\theta \in \Theta$ if and only if there exists functions

$$g(t, \theta) : \mathcal{T} \times \Theta \to [0, \infty) \quad \text{and} \quad h(x) : \mathcal{X}^n \to [0, \infty) \tag{9.4}$$

such that the joint density function of the sample can be factored as

$$p(x|\theta) = g(T(x), \theta)h(x), \quad \forall x \in \mathcal{X}^n, \theta \in \Theta. \tag{9.5}$$

Proof. ($\Rightarrow$) Consider the *discrete case* where

$$p(x|\theta) = P(X = x|\theta).$$

First of all, suppose T is sufficient for θ. Then, by definition, the conditional distribution of X given $T(X) = t$ is independent of θ and we can write

$$\begin{aligned} \mathbb{P}_\theta(x) &= \mathbb{P}_\theta(X = x, T = t(x)) \\ &= \mathbb{P}_\theta(X = x|T = t(x))\mathbb{P}_\theta(T = t(x)) \\ &= g(t(x), \theta)h(x) \end{aligned}$$

where $g(t, \theta) = \mathbb{P}_\theta(T = t)$ and

$$h(x) = \begin{cases} \mathbb{P}_\theta(X = x|T = t(x)), & \text{if } \mathbb{P}_\theta(T = t(x)) > 0, \\ 0, & \text{otherwise.} \end{cases}$$

($\Leftarrow$) Next, suppose that $\mathbb{P}_\theta(x)$ satisfies the factorization theorem, i.e.,

$$\mathbb{P}_\theta(x) = g(T(x), \theta)h(x).$$

Then, fix a statistic t_0 on $\mathbb{P}_\theta(T = t_0) > 0$ for some $\theta \in \Theta$. Then □

✱ Example 9.3.3. Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with density function

$$f_X(x_i|\theta) = \begin{cases} \frac{1}{\theta} e^{-x_i/\theta}, & 0 \le x < \infty, \\ 0, & \text{elsewhere.} \end{cases}$$

where parameter $\theta > 0$ for $i = 1, 2, \ldots, n$. Show that $\overline{X}$ is a sufficient statistic for the parameter θ.

✲ **Solution** The likelihood $L(\theta)$ of the sample is the joint density of each X_i, that is

$$\begin{aligned} L(x_1, x_2, \ldots, x_n|\theta) &= f(x_1, x_2, \ldots, x_n|\theta) \\ &= f(x_1|\theta) \times f(x_2|\theta) \times \cdots \times f(x_n|\theta) \\ &= \frac{1}{\theta^n} e^{-(x_1+x_2+\cdots+x_n)/\theta} \\ &= \frac{1}{\theta^n} e^{-n\overline{x}/\theta} \end{aligned}$$

Notice that $L(\theta)$ is a function only of two parameter: θ and $\overline{x}$, and that if

$$g(\overline{x}, \theta) = \frac{1}{\theta^n} e^{-n\overline{x}/\theta} \quad \text{and} \quad h(x_1, x_2, \ldots, x_n) = 1,$$

then the likelihood can be factored as

$$L(x_1, x_2, \ldots, x_n|\theta) = g(\overline{x}, \theta) h(x_1, x_2, \ldots, x_n).$$

By the factorization theorem, $\overline{X}$ is a sufficient statistic for the parameter θ. ◀

Lemma 9.1 Rao Blackwell Theorem
If $T(X)$ is a sufficient statistic for parameter θ, and $\widehat{\theta}$ is an unbiased estimator of θ with $\mathbb{E}[\widehat{\theta}] < \infty$ for all $\theta \in \Theta$. Let $\widehat{\theta}^* = \mathbb{E}[\widehat{\theta}|T(X)]$, then $\widehat{\theta}^* = \mathbb{E}[\widehat{\theta}|T]$, then

$$\mathbb{E}[(\widehat{\theta}^* - \theta)^2] \leq \mathbb{E}[(\widehat{\theta} - \theta)^2]. \tag{9.6}$$

The inequality is strict unless $\widehat{\theta}$ is a function of $T(X)$.

Proof. By the law of conditional expectation, we have

$$\mathbb{E}[\widehat{\theta}^*] = \mathbb{E}[\mathbb{E}[\widehat{\theta}|T]] = \mathbb{E}[\widehat{\theta}] = \theta,$$

so $\widehat{\theta}$ and $\widehat{\theta}^*$ are having the same bias. By the conditional variance formula, we have

$$\mathbb{E}[\widehat{\theta}] = \mathbb{E}[Var(\widehat{\theta}|T)] + Var(\mathbb{E}[\widehat{\theta}|T]) = \mathbb{E}[Var(\widehat{\theta}|T)] + Var(\widehat{\theta}^*).$$

Hence $Var[\widehat{\theta}^*] \geq Var[\widehat{\theta}]$, and so $MSE[\widehat{\theta}^*] \geq MSE[\widehat{\theta}]$. □

9.4 Variance of estimators based on sufficient statistics

All estimators can be regarded random variables, and we can compare their variances. therefore the maximum likelihood estimator can also be a random variable.

Definition 9.5 Fisher information
The Fisher information of a random variable X with density function $f(x|\theta)$ is defined as

$$I_X(\theta) = \mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\ln f(X|\theta)\right] = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\ell(\theta)\right)^2\right].$$

If $T(X)$ is a sufficient statistic for θ, then the Fisher information contained in $T(X)$ is equal to the Fisher information contained in the sample X, i.e.,

$$I_T(\theta) = I_X(\theta).$$

One way to find a minimum variance unbiased estimator for a parameter is to use the Cramér-Rao lower bound or the Fisher information inequality.

Remark (Multivariate Information Inequality)**.** *A multivariate version of the Information Inequality exists as well. If $\Theta \subset \mathbb{R}^k$ for some $k \in \mathbb{N}$, and if $T : \mathcal{X} \to \mathbb{R}^n$ is an n-dimensional statistic for some $n \in \mathbb{N}$ for data $X \sim f(x \mid \theta)$ taking values in a space $\mathcal{X}$ of arbitrary dimension, define the mean function $m : \mathbb{R}^k \to \mathbb{R}^n$ by $m(\theta) := \mathbb{E}_\theta T(X)$ and its $n \times k$ Jacobian matrix by*

$$J_{ij} := \frac{\partial m_i(\theta)}{\partial\theta_j}.$$

Then the multivariate Information Inequality asserts that

$$Cov_\theta[T(X)] \geq J\, I(\theta)^{-1} J^\top$$

where $I(\theta) := Cov_\theta[\nabla_\theta \log f(X \mid \theta)]$ is the Fisher information matrix, where the notation "$A \geq B$" for $n \times n$ matrices A, B means that $[A - B]$ is positive semi-definite, and where $C^\top$ denotes the $k \times n$ transpose of an $n \times k$ matrix C. This gives lower bounds on the variance of $z^\top T(X)$ for all vectors $z \in \mathbb{R}^n$ and, in particular, lower bounds for the variance of components $T_i(X)$.

Theorem 9.4 Cramér-Rao lower bound
Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with density function $f(x|\theta)$ where $\theta \in \Theta \subseteq \mathbb{R}$. Suppose the following regularity conditions hold:

1. The support of $f(x|\theta)$ does not depend on θ.
2. $\frac{\partial}{\partial\theta}\ln f(x|\theta)$ exists for all x and θ.
3. $\mathbb{E}\left[\frac{\partial}{\partial\theta}\ln f(X|\theta)\right] = 0$ for all θ.
4. $0 < I_X(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\ln f(X|\theta)\right)^2\right] < \infty$ for all θ.

If $T = T(X_1, \ldots, X_n)$ is any unbiased estimator of θ with finite variance, then

$$\text{Var}(T) \geq \frac{1}{nI_X(\theta)},$$

where $I_X(\theta)$ is the Fisher information of a single observation. Equality holds if and only if there exists a function $g(\theta)$ such that

$$T - \theta = g(\theta) \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \ln f(X_i|\theta).$$

Proof. Let $S(\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \ln f(X_i|\theta)$ be the score function for the sample. By the regularity conditions, we have $\mathbb{E}[S(\theta)] = 0$ and $\text{Var}(S(\theta)) = nI_X(\theta)$.

Since T is an unbiased estimator of θ, we have $\mathbb{E}[T] = \theta$. Taking the derivative with respect to θ and using the regularity conditions to interchange the order of differentiation and integration:

$$1 = \frac{d}{d\theta}\mathbb{E}[T] = \mathbb{E}\left[\frac{\partial T}{\partial \theta}\right] = \mathbb{E}\left[T \cdot \frac{\partial}{\partial \theta} \ln f(X_1, \ldots, X_n|\theta)\right] = \mathbb{E}[T \cdot S(\theta)].$$

Now, since $\mathbb{E}[T] = \theta$ and $\mathbb{E}[S(\theta)] = 0$, we have:

$$\text{Cov}(T, S(\theta)) = \mathbb{E}[T \cdot S(\theta)] - \mathbb{E}[T]\mathbb{E}[S(\theta)] = 1 - \theta \cdot 0 = 1.$$

By the Cauchy-Schwarz inequality:

$$(\text{Cov}(T, S(\theta)))^2 \leq \text{Var}(T) \cdot \text{Var}(S(\theta)),$$

which gives us:

$$1 \leq \text{Var}(T) \cdot nI_X(\theta).$$

Therefore:

$$\text{Var}(T) \geq \frac{1}{nI_X(\theta)}.$$

Equality holds in the Cauchy-Schwarz inequality if and only if $T - \mathbb{E}[T]$ and $S(\theta) - \mathbb{E}[S(\theta)]$ are linearly dependent, i.e., there exists a constant $g(\theta)$ such that:

$$T - \theta = g(\theta)(S(\theta) - 0) = g(\theta)S(\theta).$$

□

Remark. *An unbiased estimator T that achieves the Cramér-Rao lower bound is called an* efficient estimator *or* minimum variance unbiased estimator (MVUE)*. When such an estimator exists, it is unique and coincides with the maximum likelihood estimator under regularity conditions. The Fisher information $I_X(\theta)$ measures the amount of information about θ contained in a single observation, and the Cramér-Rao bound shows that no unbiased estimator can have variance smaller than the reciprocal of the total Fisher information.*

Lemma 9.2 Cramér-Rao lower bound - 1st theorem
Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with density function $f(x|\theta)$, where $\theta \in \Theta$ is a scalar parameter. Suppose $\widehat{\theta}$ be any unbiased estimator of θ with finite variance. Suppose the likelihood function $L(\theta)$ is differentiable with respect to θ and satisfies

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(x_1, \ldots, x_n) L(\theta)\, dx_1 \cdots dx_n = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(x_1, \ldots, x_n) \frac{\mathrm{d}L(\theta)}{\mathrm{d}\theta}\, dx_1 \cdots dx_n. \quad (9.7)$$

for any function $h(x_1, \ldots, x_n)$ with $\mathbb{E}[h(x_1, \ldots, x_n)] < \infty$. Then

$$\mathrm{Var}(\widehat{\theta}) \geq \frac{1}{\mathbb{E}\left[\left(\frac{\partial \ln L(\theta)}{\partial \theta}\right)^2\right]} = \frac{1}{nI_X(\theta)}, \quad (9.8)$$

Proof. Since $L(\theta)$ is the joint density function of the sample, we have

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} L(\theta)\, dx_1 \cdots dx_n = 1. \quad (\clubsuit)$$

Differentiating ($\clubsuit$) with respect to θ, we get

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{\mathrm{d}L(\theta)}{\mathrm{d}\theta}\, dx_1 \cdots dx_n = 0. \quad (\spadesuit)$$

Rewriting ($\spadesuit$) as

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{1}{L(\theta)} \frac{\mathrm{d}L(\theta)}{\mathrm{d}\theta} L(\theta)\, dx_1 \cdots dx_n = 0,$$

so that

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \theta \frac{\mathrm{d} \ln L(\theta)}{\mathrm{d}\theta} L(\theta)\, dx_1 \cdots dx_n = 0. \quad (\blacklozenge)$$

Since $\widehat{\theta}$ is an unbiased estimator of θ, we can see that

$$\mathbb{E}[\widehat{\theta}] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \widehat{\theta} L(\theta)\, dx_1 \cdots dx_n = \theta. \quad (\bigstar)$$

Differentiating (★) with respect to θ, we get

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \widehat{\theta} L(\theta)\, dx_1 \cdots dx_n = 1.$$

Again, using the fact (♣) with $h(X_1, X_2, \ldots, X_n) = \widehat{\theta}$, we have

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \widehat{\theta}\, \frac{\mathrm{d}L(\theta)}{\mathrm{d}\theta}\, dx_1 \cdots dx_n = 1.$$

Rewriting the above equation as

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \widehat{\theta} \frac{1}{L(\theta)} \frac{\mathrm{d}L(\theta)}{\mathrm{d}\theta} L(\theta)\, dx_1 \cdots dx_n = 1,$$

so that

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \widehat{\theta}\, \frac{\mathrm{d} \ln L(\theta)}{\mathrm{d}\theta} L(\theta)\, dx_1 \cdots dx_n = 1. \qquad (♡)$$

From (♦) and (♡), we have

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} (\widehat{\theta} - \theta)\, \frac{\mathrm{d} \ln L(\theta)}{\mathrm{d}\theta} L(\theta)\, dx_1 \cdots dx_n = 1.$$

By the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
1^2 &= \left(\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} (\widehat{\theta} - \theta)\, \frac{\mathrm{d} \ln L(\theta)}{\mathrm{d}\theta} L(\theta)\, dx_1 \cdots dx_n \right)^2 \\
&\leq \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} (\widehat{\theta} - \theta)^2 L(\theta)\, dx_1 \cdots dx_n \\
&\quad \times \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left(\frac{\mathrm{d} \ln L(\theta)}{\mathrm{d}\theta} \right)^2 L(\theta)\, dx_1 \cdots dx_n \\
1 &= \mathrm{Var}(\widehat{\theta})\, \mathbb{E}\left[\left(\frac{\mathrm{d} \ln L(\theta)}{\mathrm{d}\theta} \right)^2 \right].
\end{aligned}$$

Therefore,

$$\mathrm{Var}(\widehat{\theta}) \geq \frac{1}{\mathbb{E}\left[\left(\frac{\partial \ln L(\theta)}{\partial \theta} \right)^2 \right]}.$$

and the proof is complete. □

What this saying is, for any unbiased estimator $\widehat{\theta}$ of θ, its variance (MSE) is at least $\frac{1}{I(\theta)}$. If we achieve this lower bound, meaning that our variance is exactly equal to $\frac{1}{I(\theta)}$, then we have found the best possible unbiased estimator for θ. That is, we have found the **minimum variance unbiased estimator (MVUE)** for θ.

✱ Example 9.4.1. Suppose that $Y_1, Y_2, \ldots, Y_n$ denote a random sample from the Weibull distribution with pdf

$$f_Y(y|\theta) = \begin{cases} \frac{2y}{\theta} e^{-y^2/\theta} & \text{for } y > 0 \\ 0 & \text{otherwise} \end{cases}.$$

Find an MVUE for θ.

✲ Solution We begin using the factorization criterion to find the sufficient statistic that best summarizes the information about θ.

$$\begin{aligned} L(y_1, y_2, \ldots, y_n|\theta) &= f_Y(y_1, y_2, \ldots, y_n|\theta) \\ &= \left(\frac{2y_1}{\theta}\right) e^{-y_1^2/\theta} \times \left(\frac{2y_2}{\theta}\right) e^{-y_2^2/\theta} \times \cdots \times \left(\frac{2y_n}{\theta}\right) e^{-y_n^2/\theta} \\ &= \underbrace{\left(\frac{2}{\theta}\right)^n \exp\left\{-\frac{1}{\theta}\sum_{i=1}^n y_i^2\right\}}_{g'(\sum y_i|\theta)} \underbrace{(y_1 \times y_2 \times \cdots \times y_n)}_{h(y_1, y_2, \ldots, y_n)} \end{aligned}$$

Thus, $U = \sum_{i=1}^n Y_i^2$ is the minimal sufficient statistic for θ (by Factorization theorem).

We now need to find a function of this statistic that is unbiased for θ. Now let $W = Y_i^2$. Using the method of transformation,

$$f_W(w) = f_Y(h^{-1}(w))\frac{\mathrm{d}h^{-1}(w)}{\mathrm{d}w} = f_Y(\sqrt{w})\frac{\mathrm{d}}{\mathrm{d}w}(\sqrt{w}),$$

continue simplify the expression gives

$$f_W(w) = \frac{2}{\theta}\left(\sqrt{w}e^{-w/\theta}\right)\left(\frac{1}{2\sqrt{w}}\right) = \frac{1}{\theta}e^{-w/\theta}, \quad w > 0.$$

That is, $Y_i^2 \sim Exp(\theta)$. As

$$\mathbb{E}[Y_i^2] = \theta \tag{9.9}$$

and

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^{n} Y_i^2\right] &= \mathbb{E}[Y_1^2 + Y_2^2 + \cdots + Y_n^2] \\ &= \mathbb{E}[Y_1^2] + \mathbb{E}[Y_2^2] + \cdots + \mathbb{E}[Y_n^2] && \text{linearity or expectation} \\ &= \underbrace{\theta + \cdots + \theta}_{n \text{ times}} \\ &= n\theta.\end{aligned}$$

◀

Corollary 9.1 Cramér-Rao lower bound - 2nd theorem
If $L(\theta)$ is twice differentiable with respect to θ, then the inequality can be stated equivalently as

$$\text{Var}(\widehat{\theta}) \geq \frac{-1}{\mathbb{E}\left[\frac{\partial^2 \ln L(\theta)}{\partial \theta^2}\right]}. \tag{9.10}$$

Proof. If $L(\theta)$ is twice differentiable, then $\frac{\partial^2 \ln L(\theta)}{\partial \theta^2}$ exists. And since that $L(\theta)$ is maximum likelihood so we have

$$\frac{\partial^2 \ln L(\theta)}{\partial \theta^2} < 0.$$

Hence the sign of the right-hand side of the Cramér-Rao inequality must be negative. □

✱ Example 9.4.2. Let $X_1, X_2, \ldots, X_n$ be a random samples from a distribution with density function

$$f(x|\theta) = \begin{cases} 3\theta x^2 e^{-\theta x^3} & \text{if } 0 < x < \infty \\ 0 & \text{otherwise.} \end{cases}$$

What is the Cramér-Rao lower bound for the variance of unbiased estimator of the parameter θ?

✲ Solution Let $\widehat{\theta}$ be an unbiased estimator of θ. Cramér-Rao lower bound for the variance of $\widehat{\theta}$ is given by

$$\text{Var}(\widehat{\theta}) \geq \frac{-1}{\mathbb{E}\left[\left(\frac{\mathrm{d} \ln L(\theta)}{\mathrm{d}\theta}\right)^2\right]}.$$

First, compute the likelihood function for the sample:

$$L(\theta) = \prod_{i=1}^{n} f(x_i|\theta) = \prod_{i=1}^{n} 3\theta x_i^2 e^{-\theta x_i^3} = 3^n \theta^n \left(\prod_{i=1}^{n} x_i^2\right) \exp\left(-\theta \sum_{i=1}^{n} x_i^3\right).$$

Take the log-likelihood:

$$\ell(\theta) = \ln L(\theta) = n \ln 3 + n \ln \theta + 2 \sum_{i=1}^{n} \ln x_i - \theta \sum_{i=1}^{n} x_i^3.$$

Compute the first derivative with respect to θ:

$$\frac{\partial}{\partial \theta} \ell(\theta) = \frac{n}{\theta} - \sum_{i=1}^{n} x_i^3.$$

Compute the second derivative:

$$\frac{\partial^2}{\partial \theta^2} \ell(\theta) = -\frac{n}{\theta^2}.$$

The Fisher information for one observation is

$$I_X(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial \theta^2} \ln f(X|\theta)\right] = \frac{1}{\theta^2}.$$

For n independent observations, the total Fisher information is $nI_X(\theta) = \frac{n}{\theta^2}$.

Therefore, the Cramér-Rao lower bound is

$$\text{Var}(\widehat{\theta}) \geq \frac{1}{nI_X(\theta)} = \frac{\theta^2}{n}.$$

where $L(\theta)$ denotes the likelihood function of the given random sample. ◀

✱ Example 9.4.3. Let $X_1, X_2, \ldots, X_n$ denote a random sample from $Bin(1, p)$. We knew that $\overline{X}$ is an unbiased estimator of p and that

$$Var(\overline{X}) = \frac{p(1-p)}{n}.$$

Find the Cramér-Rao lower bound for the variance of every unbiased estimator of p.

✲ Solution This is a Bernoulli distribution with parameter p. The denisty function for each X_i is

$$f(x|p) = p^x (1-p)^{1-x}, \quad x = 0, 1.$$

Taking logarithm on f,

$$\ln f(x|p) = \ln p^x + \ln(1-p)^{1-x} = \boxed{x \ln p + (1-x) \ln(1-p)} \qquad (\blacklozenge)$$

Compute the first and second order derivative of $(\blacklozenge)$ with respect to p. The second order derivative will be the Fisher information. In this case we have only one parameter which is p, so the information is just a simple algebraic expression rather than a matrix.

$$\frac{\partial \ln f(x|p)}{\partial p} = \frac{x}{p} + \frac{x-1}{1-p}$$

$$\begin{aligned}\frac{\partial^2 \ln f(x|p)}{\partial p^2} &= -\frac{x}{p^2} + (-1)^2(x-1)(1-p)^{-2} \\ &= -\frac{x}{p^2} + \frac{x-1}{(1-p)^2}. \end{aligned} \tag{▲}$$

Hence we find the expectation of x in (▲), that is

$$\begin{aligned}\mathbb{E}_X\left[\frac{\partial^2 \ln f(x|p)}{\partial p^2}\right] &= \mathbb{E}_X\left[-\frac{x}{p^2} + \frac{x-1}{(1-p)^2}\right] \\ &= -\frac{1}{p^2}\mathbb{E}_X[X] + \frac{1}{(1-p)^2}\mathbb{E}_X[X-1] \\ &= -\frac{p}{p^2} + \frac{p-1}{(1-p)^2} \\ &= -\frac{1}{p(1-p)}.\end{aligned}$$

Therefore, the Cramér-Rao lower bound for the variance of unbiased estimator of p is

$$Var(\widehat{p}) \geq -\frac{1}{-nI_X(p)} = \frac{1}{\frac{-n}{-p(1-p)}} = \frac{p(1-p)}{n}.$$

◀

9.4.1 Delta method – Variance of functions of estimators

The Delta Method (DM) states that we can approximate the asymptotic behaviour of functions over a random variable, if the random variable is itself asymptotically normal. In practice, this theorem tells us that even if we do not know the expected value and variance of the function $g(X)$ we can still approximate it reasonably. Note that by Central Limit Theorem we know that several important random variables and estimators are asymptotically normal, including the sample mean. We can therefore approximate the mean and variance of some transformation of the sample mean using its variance.

More specifically, suppose that we have some sequence of random variables $\{X_n\}$, as $n \to \infty$,

Given this, if g is some smooth function (i.e. there are no discontinuous jumps in values) then the Delta Method states that:

$$\frac{\sqrt{n}(g(X_n) - g(\mu))}{|\dot{g}(\mu)|\sigma} \approx \mathcal{N}(0,1) \tag{9.11}$$

DM also generalizes to multidimensional functions, where instead of converging on the standard normal the random variable must converge in distribution to a multivariate normal, and the derivatives of g are replaced with the gradient of g (a vector of all partial derivatives).

Theorem 9.5 Delta method
Suppose that $g(\theta)$ is a function of estimator. The delta method provides a way to approximate

the variance of a function of an estimator. This approximate of the variance is given by

$$Var(g(\widehat{\theta})) \approx [\dot{g}(\theta)]^2 \, Var[\widehat{\theta}]. \tag{9.12}$$

✱ Example 9.4.4. Given $g(s,t) = \frac{s}{t}$, $h(s,t) = \ln s$ and $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are unbiased estimators of θ_1 and θ_2. Based on a particular sample, the maximum likelihood estimates of θ_1 and θ_2 are $\widehat{\theta}_1 = 3.2$ and $\widehat{\theta}_2 = 11.8$, and the log-likelihood is $\ell(\theta_1, \theta_2) = -2\theta_1^2\theta_2 - \theta_2^3$.

✲ Solution We first compute the Fisher information matrix:

$$\frac{\partial^2}{\partial\theta_1^2}\ell(\theta_1,\theta_2) = -4\theta_2, \quad \frac{\partial^2}{\partial\theta_2^2}\ell(\theta_1,\theta_2) = -6\theta_2, \quad \frac{\partial^2}{\partial\theta_1\partial\theta_2}\ell(\theta_1,\theta_2) = -4\theta_1.$$

The information matrix is

$$I_X(\theta) = -\mathbb{E}\begin{bmatrix} \frac{\partial^2}{\partial\theta_1^2}\ell(\theta_1,\theta_2) & \frac{\partial^2}{\partial\theta_1\partial\theta_2}\ell(\theta_1,\theta_2) \\ \frac{\partial^2}{\partial\theta_1\partial\theta_2}\ell(\theta_1,\theta_2) & \frac{\partial^2}{\partial\theta_2^2}\ell(\theta_1,\theta_2) \end{bmatrix} = \begin{bmatrix} 4\theta_2 & 4\theta_1 \\ 4\theta_1 & 6\theta_2 \end{bmatrix}.$$

The covariance matrix of the MLEs is given by the inverse of the Fisher information matrix evaluated at the MLEs:

$$\Sigma = I_X(\theta)^{-1} = \begin{bmatrix} 4\theta_2 & 4\theta_1 \\ 4\theta_1 & 6\theta_2 \end{bmatrix}^{-1} = \frac{1}{12\theta_2^2 - 8\theta_1^2}\begin{bmatrix} 3\theta_2 & -2\theta_1 \\ -2\theta_1 & 2\theta_2 \end{bmatrix}.$$

The estimated covariance matrix is obtained by substituting the MLEs:

$$\begin{aligned} \widehat{\Sigma} = \frac{1}{12\widehat{\theta}_2^2 - 8\widehat{\theta}_1^2}\begin{bmatrix} 3\widehat{\theta}_2 & -2\widehat{\theta}_1 \\ -2\widehat{\theta}_1 & 2\widehat{\theta}_2 \end{bmatrix} &= \frac{1}{12(11.8)^2 - 8(3.2)^2}\begin{bmatrix} 3(11.8) & -2(3.2) \\ -2(3.2) & 2(11.8) \end{bmatrix} \\ &= \begin{bmatrix} 0.0222787 & -0.00402779 \\ -0.00402779 & 0.0148525 \end{bmatrix} \\ &= \begin{bmatrix} Var(\widehat{\theta}_1) & \text{Cov}(\widehat{\theta}_1, \widehat{\theta}_2) \\ \text{Cov}(\widehat{\theta}_1, \widehat{\theta}_2) & Var(\widehat{\theta}_2) \end{bmatrix}. \end{aligned}$$

1. Now take partial derivatives of $g(s,t)$ with respect to s and t:

$$g_s(s,t) = \frac{\partial g}{\partial s} = \frac{1}{t} \Longrightarrow g_s(\widehat{\theta}_1, \widehat{\theta}_2) = \frac{1}{11.8},$$

$$g_t(s,t) = \frac{\partial g}{\partial t} = -\frac{s}{t^2} \Longrightarrow g_t(\widehat{\theta}_1, \widehat{\theta}_2) = -\frac{3.2}{(11.8)^2}.$$

Hence, let $\mathbf{w} = \begin{bmatrix} g_s(\widehat{\theta}_1, \widehat{\theta}_2) & g_t(\widehat{\theta}_1, \widehat{\theta}_2) \end{bmatrix} = \begin{bmatrix} \frac{1}{11.8} & -\frac{3.2}{11.8^2} \end{bmatrix}$. the approximate variance of $g(\widehat{\theta}_1, \widehat{\theta}_2)$ is

$$\begin{aligned} Var(g(\widehat{\theta}_1, \widehat{\theta}_2)) &\approx \mathbf{w}\widehat{\Sigma}\mathbf{w}^T \\ &= \begin{bmatrix} \frac{1}{11.8} & -\frac{3.2}{11.8^2} \end{bmatrix} \begin{bmatrix} 0.0222787 & -0.00402779 \\ -0.00402779 & 0.0148525 \end{bmatrix} \begin{bmatrix} \frac{1}{11.8} \\ -\frac{3.2}{11.8^2} \end{bmatrix} \\ &= 0.000208785. \end{aligned}$$

2. Continue with $h(s,t)$: take partial derivatives of $h(s,t)$ with respect to s and t:

$$h_s(s,t) = \frac{\partial h}{\partial s} = \frac{1}{s} \Longrightarrow h_s(\widehat{\theta}_1, \widehat{\theta}_2) = \frac{1}{3.2},$$

$$h_t(s,t) = \frac{\partial h}{\partial t} = 0 \Longrightarrow h_t(\widehat{\theta}_1, \widehat{\theta}_2) = 0.$$

The estimated covariance between $h(\widehat{\theta}_1, \widehat{\theta}_2)$ and $g(\widehat{\theta}_1, \widehat{\theta}_2)$ is

$$\begin{aligned} \text{Cov}(h(\widehat{\theta}_1, \widehat{\theta}_2), g(\widehat{\theta}_1, \widehat{\theta}_2)) &= \begin{bmatrix} \frac{1}{\widehat{\theta}_1} & -\frac{\widehat{\theta}_1}{\widehat{\theta}_2^2} \end{bmatrix} \begin{bmatrix} 0.0222787 & -0.00402779 \\ -0.00402779 & 0.0148525 \end{bmatrix} \begin{bmatrix} \frac{1}{\widehat{\theta}_1} \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{11.8} & -\frac{3.2}{11.8^2} \end{bmatrix} \begin{bmatrix} 0.0222787 & -0.00402779 \\ -0.00402779 & 0.0148525 \end{bmatrix} \begin{bmatrix} \frac{1}{3.2} \\ 0 \end{bmatrix} \\ &= 6.2 \times 10^{-3} \end{aligned}$$

◀

Tutorials

Exercise 9.1 If X is uniformly distributed on the interval $(2\theta, 3\theta)$ where $\theta > 0$. And that $X_1, X_2, \ldots, X_n$ is a random sample from the distribution of X. Find the bias in the maximum likelihood estimator of θ.

Exercise 9.2 Let $X_1, X_2, \ldots, X_n$ be a random sample from a population $X \sim Poisson(\lambda)$, where $\lambda > 0$ is a parameter. Is the estimator $\overline{X}$ of λ a consistent estimator of λ?

Exercise 9.3 Let $X_1, X_2, \ldots, X_n$ be a random sample from a population X with density function

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

where $\theta > 1$ is a parameter. Show that

$$-\frac{1}{n}\sum_{i=1}^{n} \ln(X_i)$$

is a uniform minimum variance unbiased estimator of $\frac{1}{\theta}$.

Exercise 9.4 Let $Y_1, Y_2, \ldots, Y_n$ denote a random sample from a population with mean μ and variance σ^2. Consider the following three estimators for μ:

$$\widehat{\mu}_1 = \frac{1}{2}(Y_1 + Y_2), \quad \widehat{\mu}_2 = \frac{1}{4}Y_1 + \frac{Y_2 + \ldots + Y_{n-1}}{2(n-2)} + \frac{1}{4}Y_n, \quad \widehat{\mu}_3 = \overline{Y}.$$

1. Show that each of the three estimators is unbiased.
2. Find the efficiency of $\widehat{\mu}_3$ relative to $\widehat{\mu}_2$ and $\widehat{\mu}_1$, respectively.

Exercise 9.5

1. State the definition of unbiased estimator.
2. Show that if $\widehat{\Theta}_1$ is an unbiased estimator for θ, and W is a zero mean random variable, then

   $$\widehat{\Theta}_2 = \widehat{\Theta}_1 + W$$

   is also an unbiased estimator for θ.
3. Show that if $\widehat{\Theta}_1$ is an unbiased estimator for θ such that $\mathbb{E}[\widehat{\Theta}_1] = a\theta + b$, where $a \neq 0$. Then

   $$\widehat{\Theta}_2 = \frac{\widehat{\Theta}_1 - b}{a}$$

   is an unbiased estimator for θ.

Exercise 9.6 Given $\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}$ is a vector of parameters being estimated by maximum likelihood. You are given that the current value of the vector of estimates is $\widehat{\theta} = \begin{bmatrix} 30 \\ 2 \end{bmatrix}$ as well as the estimated

information matrix

$$I(\widehat{\theta}) = \begin{bmatrix} .075 & -.620 \\ -.620 & 10.0 \end{bmatrix}.$$

Determine the approximate variance of $\widehat{\theta}_1$.

Exercise 9.7 Let X be a random variable, and $X_n = X + Y_n$, where

$$\mathbb{E}[Y_n] = \frac{1}{n} \quad \text{and} \quad Var(Y_n) = \frac{\sigma^2}{n}.$$

where $\sigma > 0$ is constant. Show that $X_n \xrightarrow{p} X$ as $n \to \infty$. Hint: $|Y_n| \leq |Y_n - \mathbb{E}[Y_n]| + \frac{1}{n}$.

This page intentionally left blank.

10

Confidence Intervals

The reason of using an interval estimator $[L(\mathbf{X}), U(\mathbf{X})]$ instead of a point estimator $\widehat{\theta}$ is that the interval estimator can have some level of confidence that the unknown parameter θ lies within the interval. The certainty of this guarantee is qualified by the following definitions.

Definition 10.1 Interval Estimator
Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be a random sample of size n from a population with density function $f(x|\theta)$, where $\theta \in \Theta$ is an unknown parameter. The **interval estimator** is a pair of statistics $[L(\mathbf{X}), U(\mathbf{X})]$ such that $L(\mathbf{X}) < U(\mathbf{X})$ for all possible samples $\mathbf{X}$.

Recall that a sample is a portion of the population usually chosen by some method of random sampling and as such is a set of random variables $X_1, X_2, \ldots, X_n$ with the same probability density function $f(x|\theta)$ as the population. Once the sampling is done, we will get the sample data

$$X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n.$$

Another well known method for constructing confidence sets is that using the pivotal quantities.

Definition 10.2 Confidence Coefficient
Consider $X_1, X_2, \ldots, X_n$ be a random sample of size n from a population with density $f(x|\theta)$, where θ is unknown parameter. An interval estimator $[L(\mathbf{X}), U(\mathbf{X})]$ of θ is called a $100(1-\alpha)\%$ **confidence interval** for θ if

$$\mathbb{P}_{\theta}[\, L \leq \theta \leq U \,] = 1 - \alpha, \quad \forall \theta \in \Theta, \tag{10.1}$$

The random variable L is called the **lower confidence limit** and the random variable U is called the **upper confidence limit**. The value $1 - \alpha$ is called the **confidence coefficient** or **coverage probability** of the interval estimator.

One can find inifitely many pairs of L, U, such that

$$1 - \alpha = \mathbb{P}_{\theta}[\, L \leq \theta \leq U \,]$$

for a given confidence coefficient $1-\alpha$. Thus, there are infinitely many confidence intervals for a given confidence coefficient.

However, we only need the confidence interval that is the shortest in length among all possible

confidence intervals. If a confidence interval is constructed by omitting equal tail areas then we obtain what we known as the central confidence interval. In a symmetric distribution, the central confidence interval is also the shortest confidence interval.

10.1 Pivotal Quantities

Definition 10.3 Pivotal Quantity
A function $Q(X, \theta)$ is called a pivotal quantity (or pivot) if and only if the distribution of $Q(X, \theta)$ does not depend on any unknown parameter θ.

Remark. *A pivot is not a statistic, although its distribution is known.*

With a pivot $Q(\mathbf{X}, \theta)$, a confidence set on level $1 - \alpha$ for any $\alpha \in (0, 1)$, can be obtained by finding a Borel set $\mathcal{A} = [c_1, c_2]$ such that $\mathbb{P}[Q(\mathbf{X}, \theta) \in \mathcal{A}] \geq 1 - \alpha$. Then the set

$$C(\mathbf{X}) = \{\theta \in \Theta \mid Q(\mathbf{X}, \theta) \in \mathcal{A}\} \tag{10.2}$$

is a confidence set on level $1 - \alpha$ since

$$\inf_{\theta \in \Theta} \mathbb{P}_\theta(Q(\mathbf{X}, \theta) \in \mathcal{A}) = \mathbb{P}[Q(\mathbf{X}, \theta) \in \mathcal{A}] \geq 1 - \alpha. \tag{10.3}$$

If $Q(\mathbf{X}, \theta)$ has a continuous cdf, then we can choose c_1 and c_2 such that $C(x)$ has exact coverage probability $1 - \alpha$.

Definition 10.4 Location-Scale Family
Let $g : \mathbb{R} \to \mathbb{R}$ be a probability density function. Then for any μ and any $\sigma > 0$, the family of functions

$$\mathcal{F} = \left\{ f(x; \mu, \sigma) = \frac{1}{\sigma} g\left(\frac{x - \mu}{\sigma}\right) \mid \mu \in (-\infty, \infty), \sigma \in (0, \infty) \right\} \tag{10.4}$$

is called the *location-scale family* with standard probability density $f(x; \theta)$. The parameter μ is called the *location parameter* and the parameter σ is called the *scale parameter*. If $\sigma = 1$, then $\mathcal{F}$ is called the *location family*. If $\mu = 0$, then $\mathcal{F}$ is called the *scale family*.

It should be noted that each member $f(x; \mu, \sigma)$ of the location-scale family is a probability density function. If we take $g(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$, then the normal density function

$$f(x|\mu, \sigma) = \frac{1}{\sigma} g\left(\frac{x - \mu}{\sigma}\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty < x < \infty$$

belongs to the location-scale family. The density function

$$f(x|\theta) = \begin{cases} \frac{1}{\theta} e^{-\frac{x}{\theta}} & \text{if } 0 < x < \infty \\ 0 & \text{otherwise,} \end{cases} \tag{10.5}$$

belongs to the scale family. However, the density function

$$f(x|\theta) = \begin{cases} \theta x^{\theta-1} & \text{if } 0 < x < 1 \\ 0 & \text{otherwise,} \end{cases} \tag{10.6}$$

does not belong to the location-scale family.

Form of pdf	Type of pdf	Pivots
$f(x-\mu)$	Location	$\overline{X} - \mu$
$\frac{1}{\sigma} f(\frac{1}{\sigma})$	Scale	$\frac{\overline{X}}{\sigma}, \frac{S^2}{\sigma^2}, \frac{X_{(n)}}{\sigma}$
$\frac{1}{\sigma} f(\frac{x-\mu}{\sigma})$	Location-scale	$\frac{\overline{X}-\mu}{S}, \frac{S^2}{\sigma^2}$

Table 10.1: The location and scale families and some common pivots. Here, $\overline{X}$ is the sample mean, S^2 is the sample variance, and $X_{(n)}$ is the maximum order statistic.

✱ Example 10.1.1 (Scale uniform interval estimation)**.** Let $X_1, X_2, \ldots, X_n$ be a random sample from a uniform distribution $U(0, \theta)$, and let $Y = \max\{X_1, X_2, \ldots, X_n\}$. We are interested in finding an interval estimator of θ.

We consider two candidate interval estimators:

1. $[aY, bY]$, $\quad 1 \leq a < b$.

2. $[Y + c, Y + d]$, $\quad 0 \leq c < d$.

where a, b, c, d are specified constants. Note that the parameter θ must always larger than y.

[First candidate] For the first candidate, we have

$$\begin{aligned} \mathbb{P}_\theta[\, \theta \in [aY, bY]\,] &= \mathbb{P}_\theta[\, aY \leq \theta \leq bY\,] \\ &= \mathbb{P}_\theta\left[\, \frac{1}{b} \leq \frac{Y}{\theta} \leq \frac{1}{a}\,\right] \end{aligned}$$

From previuous example we know that

$$f_Y(y) = \begin{cases} \dfrac{n y^{n-1}}{\theta^n} & \text{if } 0 < y < \theta \\ 0 & \text{otherwise.} \end{cases}.$$

so the pdf of T is $f_T(t) = nt^{n-1}$ for $0 < t < 1$. Thus,

$$\mathbb{P}_\theta\left[\frac{1}{b} \leq \frac{Y}{\theta} \leq \frac{1}{a}\right] = \int_{1/b}^{1/a} nt^{n-1}\,\mathrm{d}t = \frac{1}{a^n} - \frac{1}{b^n}.$$

The covergae probability of the first interval estimator is independent of the parameter θ, and thus $\frac{1}{a^n} - \frac{1}{b^n}$ is the confidence coefficient of the interval estimator $[aY, bY]$.

[Second candidate] On the other hand, for the second candidate, for $\theta \geq d$ and a similar calculation we obtained

$$\begin{aligned}
\mathbb{P}_\theta[\,\theta \in [Y+c, Y+d]\,] &= \mathbb{P}_\theta[\,Y + c \leq \theta \leq Y + d\,] \\
&= \mathbb{P}_\theta\left[1 - \frac{d}{\theta} \leq T \leq 1 - \frac{c}{\theta}\right] \\
&= \int_{1-d/\theta}^{1-c/\theta} nt^{n-1}\,\mathrm{d}t \\
&= (1 - c/\theta)^n - (1 - d/\theta)^n
\end{aligned}$$

In this case, the covergae probability of the second interval estimator depends on the unknown parameter θ. As $\theta \to \infty$, for any fixed c and d, the coverage probability is

$$\lim_{\theta\to\infty} (1 - c/\theta)^n - (1 - d/\theta)^n = 0.$$

Showing that the second interval estimator has zero confidence coefficient.

10.2 Confidence Interval for Population Mean

At the outset, we use the pivotal quantity method to construct a confidence interval for the mean of a normal population. First we assume that the population is normal and the population variance is known, but the variance is unknown. Next, we construct the confidence interval for the mean of a population with continuous, symmetric and unimodal probability distribution by applying the central limit theorem.

We know that $\widehat{\mu} = \overline{X}$. Because each X_i is identically distributed as $N(\mu, \sigma^2)$, the distribution of the sample mean $\overline{X}$ is given by

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

It is easy to verify that the distribution of the estimator $\widehat{\mu}$ is not independent of the parameter μ. If we standardize $\widehat{\mu}$, we have

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

However, the distribution of the standardized variable Z is independent of the parameter μ. Thus, Z is a pivotal quantity since it is a function of the sample $X_1, X_2, \ldots, X_n$ and parameter μ. Using this standardized variable as the pivotal quantity, we can construct a confidence interval for the population mean μ as follows:

$$\begin{aligned}
1 - \alpha &= \mathop{\mathbb{P}}_{\mu \sim \overline{X}}\big[-z_{\alpha/2} \leq Z \leq z_{\alpha/2} \big] \\
&= \mathop{\mathbb{P}}_{\mu \sim \overline{X}}\left[-z_{\alpha/2} \leq \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2} \right] \\
&= \mathop{\mathbb{P}}_{\mu \sim \overline{X}}\left[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \right].
\end{aligned}$$

Hence, the $(1-\alpha)100\%$ confidence interval for μ when the population X is normal and known variance σ^2 is given by

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{10.7}$$

To interpret the confidence interval of μ, we say that if you repeating the same experiment process many times, and generate a confidence intervals using the same method. Then approximately $(1-\alpha)100\%$ of the intervals will contain the true value of μ.

10.2.1 Confidence interval for small sample mean

Suppose $X_1, X_2, \ldots, X_n$ is random sample from a normal population X with mean μ and variance $\sigma^2 > 0$. Let the sample mean and sample variances be $\overline{X}$ and S^2 respectively. Then

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

and

$$\frac{\overline{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1).$$

Therefore, the random variable defined by the ratio of $\frac{(n-1)S^2}{\sigma^2}$ to $\frac{\overline{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}}$ has a t-distribution with

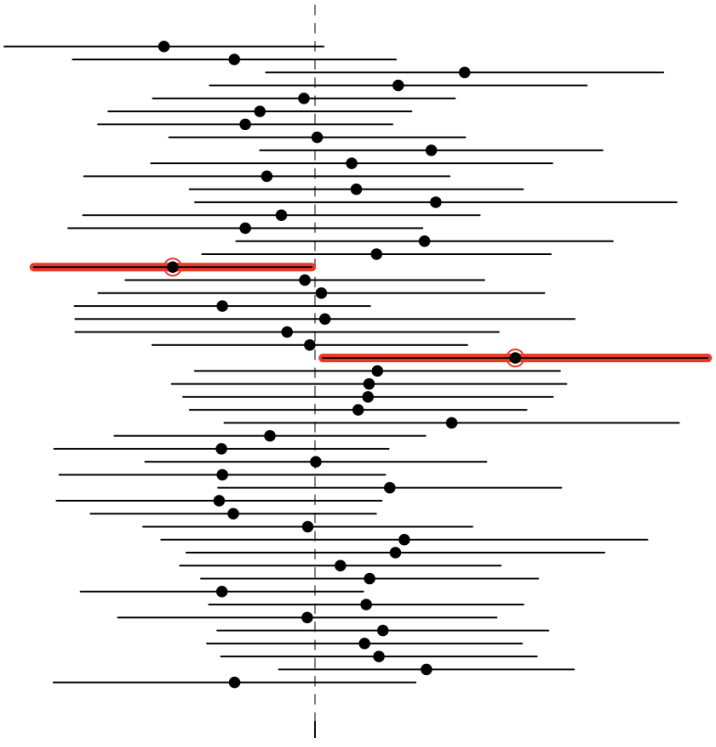


Figure 10.1: This is a simulation of the confidence interval for the population mean. We generate 50 samples of size 30 from a normal population with mean 50 and standard deviation 15. The red horizontal line indicates the true mean of the population do not include the true population mean. Observe that 48 out of 50, of the intervals contain the true mean. Thus, the coverage probability is approximately 96%.

$(n-1)$ degrees of freedom, that is

$$Q(X_1, X_2, \ldots, X_n, \mu) = \frac{\frac{\overline{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}}}{\sqrt{\frac{(n-1)S^2}{(n-1)\sigma^2}}} = \frac{\overline{X}-\mu}{\sqrt{\frac{S^2}{n}}} \sim t(n-1),$$

where Q is the pivotal quantity to be used for the construction of the confidence interval for μ. Using this pivotal quantity, we construct the confidence interval as follows:

$$\begin{aligned} 1-\alpha &= P\left(-t_{\frac{\alpha}{2}}(n-1) \leq \frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}} \leq t_{\frac{\alpha}{2}}(n-1)\right) \\ &= P\left(\overline{X} - \left(\frac{S}{\sqrt{n}}\right) t_{\frac{\alpha}{2}}(n-1) \leq \mu \leq \overline{X} + \left(\frac{S}{\sqrt{n}}\right) t_{\frac{\alpha}{2}}(n-1)\right) \end{aligned}$$

Hence, the $100(1-\alpha)\%$ confidence interval for μ when the population X is normal with the unknown variance σ^2 is given by

$$\left[\overline{X} - \left(\frac{S}{\sqrt{n}}\right) t_{\frac{\alpha}{2}}(n-1), \quad \overline{X} + \left(\frac{S}{\sqrt{n}}\right) t_{\frac{\alpha}{2}}(n-1)\right]. \tag{10.8}$$

Remark. *For small sample size $n < 30$, we should use the t-distribution to construct the confidence interval for the population mean μ as the population variance σ^2 is unknown.*

✱ Example 10.2.1. Let $X_1, X_2, \ldots, X_{11}$ be a random sample from a normal population with unknown mean μ and variance $\sigma^2 = 9.9$. Given that $\sum_{i=1}^{11} x_i = 132$. Find a 95% confidence interval for μ.

✲ Solution From the information above, the sample mean is

$$\bar{x} = \frac{\sum_{i=1}^{11} x_i}{11} = \frac{132}{11} = 12.$$

Furthermore, since μ is unknown, we use $\widehat{\mu} = \bar{x} = 12$. Since each $X_i \sim N(\mu, \sigma^2 = 9.9)$, the confidence interval for μ at 95% confidence level is

$$\begin{aligned} \bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} &= 12 \pm z_{\frac{0.05}{2}}\sqrt{\frac{9.9}{11}} \\ &= 12 \pm 1.96\sqrt{0.9}. \end{aligned}$$

That is

$$[10.141, 13.859].$$

◀

Definition 10.5 Minimum sample size for estimating population mean
Let $X_1, X_2, \ldots, X_n$ be a random sample from a population with unknown mean μ and known variance σ^2. For a specified margin of error $E > 0$ and confidence level $100(1-\alpha)\%$, the minimum sample size n required to estimate the population mean μ is given by

$$n = \left\lceil \left(\frac{z_{\alpha/2}\,\sigma}{E}\right)^2 \right\rceil. \tag{10.9}$$

The value of n determined from the above formula is the **minimum** sample size required to ensure that the desired level of confidence is achieved. So we always round up to the next larger integer.

✱ Example 10.2.2. Suppose that we wanted to estimate the true average number of eggs a queen

honeybee lays with 95% confidence. The margin of error we are willing to accept is 2 eggs. Suppose we all know that the sample variance is $s^2 = 100$. What is the minimum sample size required?

✲ Solution For a 95% confidence interval, we have $\alpha = 0.05$ and $z_{\alpha/2} = z_{0.025} = 1.96$. The minimum sample size required is

$$n = \left\lceil \left(\frac{1.96 \times \sqrt{100}}{2} \right)^2 \right\rceil = \lceil 96.04 \rceil = 97.$$

Thus, we need a sample size of at least 97 queen honeybees. ◀

10.3 Confidence interval for population proportion

Let $X_1, X_2, \ldots, X_n$ be a random sample from a population with unknown proportion p. The sample proportion is given by

$$f(x|p) = p^x \, (1-p)^{1-x}, \quad x = 0, 1.$$

We want to construct a $100(1-\alpha)\%$ approximate confidence interval for the parameter p.

To do this, we note that the likelihood function of the sample is given by

$$L(p|X) = \prod_{i=1}^{n} f(X_i|p) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i} = p^{\sum_{i=1}^n X_i}(1-p)^{n-\sum_{i=1}^n X_i}.$$

Taking the logarithm of the likelihood function, we get

$$\ln L(p) = \sum_{i=1}^{n} [x_i \ln p + (1-x_i) \ln(1-p)].$$

Differentiating, the above expression, we get

$$\frac{d \ln L(p)}{dp} = \frac{1}{p} \sum_{i=1}^{n} x_i - \frac{1}{1-p} \sum_{i=1}^{n} (1-x_i).$$

Setting this equals to zero and solving for p, we get

$$\frac{n\overline{x}}{p} - \frac{n - n\overline{x}}{1-p} = 0,$$

that is

$$(1-p)\, n\, \overline{x} = p\, (n - n\, \overline{x}),$$

which is

$$n\,\overline{x} - p\,n\,\overline{x} = p\,n - p\,n\,\overline{x}.$$

Hence

$$p = \overline{x}.$$

Therefore, the maximum likelihood estimator of p is given by

$$\widehat{p} = \overline{X}.$$

The variance of $\overline{X}$ is

$$Var\left(\overline{X}\right) = \frac{\sigma^2}{n}.$$

Since $X \sim Ber(p)$, the variance $\sigma^2 = p(1-p)$, and

$$Var\left(\widehat{p}\right) = Var\left(\overline{X}\right) = \frac{p(1-p)}{n}.$$

Since $Var\left(\widehat{p}\right)$ is not free of the parameter p, we replace p by $\widehat{p}$ in the expression of $Var\left(\widehat{p}\right)$ to get

$$Var\left(\widehat{p}\right) \simeq \frac{\widehat{p}(1-\widehat{p})}{n}.$$

The $100(1-\alpha)\%$ approximate confidence interval for the parameter p is given by

$$\left[\widehat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}, \quad \widehat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right].$$

10.4 Confidence interval for unknown variance

Consider a random sample $X_1, X_2, \ldots, X_n$ from a normal population with mean μ and variance σ^2. When both μ and σ^2 are unknown, we can use the sample variance S^2 to estimate the population variance σ^2. We know that

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1} \implies \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2} \sim \chi^2_{n-1}.$$

We take $Q(X_1, \ldots, X_n, \sigma^2) = \dfrac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2}$ as a pivotal quantity to construct the confidence interval for σ^2. Hence, we have

$$\begin{aligned}
1 - \alpha &= \mathop{\mathbb{P}}_{\sigma^2}\left[\frac{1}{\chi^2_{n-1,\alpha/2}} \leq Q \leq \frac{1}{\chi^2_{n-1,\,1-\frac{\alpha}{2}}} \right] \\
&= \mathop{\mathbb{P}}_{\sigma^2}\left[\frac{1}{\chi^2_{n-1,\alpha/2}} \leq \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2} \leq \frac{1}{\chi^2_{n-1,\,1-\frac{\alpha}{2}}} \right] \\
&= \mathop{\mathbb{P}}_{\sigma^2}\left[\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\chi^2_{n-1,\,1-\frac{\alpha}{2}}} \leq \sigma^2 \leq \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\chi^2_{n-1,\,\frac{\alpha}{2}}} \right]
\end{aligned}$$

Hence, the $100(1-\alpha)\%$ confidence interval for σ^2 when the population mean is unknown is given by

$$\left[\frac{(n-1)S^2}{\chi^2_{n-1,\,1-\frac{\alpha}{2}}}, \quad \frac{(n-1)S^2}{\chi^2_{n-1,\,\frac{\alpha}{2}}} \right].$$

10.4.1 Confidence Interval for standard deviation

Let S^2 be the sample variance. We know that

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1).$$

Using this random variable as a pivot, we can construct a $100(1-\alpha)\%$ confidence interval for σ from

$$1 - \alpha = P\left(a \leq \frac{(n-1)S^2}{\sigma^2} \leq b \right)$$

by suitably choosing the constants a and b. Hence, the confidence interval for σ is given by

$$\left[\sqrt{\frac{(n-1)S^2}{b}}, \quad \sqrt{\frac{(n-1)S^2}{a}} \right].$$

The length of this confidence interval is given by

$$L(a,b) = S\sqrt{n-1}\left[\frac{1}{\sqrt{a}} - \frac{1}{\sqrt{b}} \right]$$

In order to find the shortest confidence interval, we should find a pair of constants a and b such

that $L(a,b)$ is minimum. Thus, we have a constraint minimization problem. That is

$$\begin{aligned} &\text{Minimize } L(a,b) \\ &\text{Subject to the condition} \\ &\int_a^b f(u)\mathrm{d}u = 1-\alpha, \end{aligned} \qquad \text{(MP)}$$

where

$$f(x) = \frac{1}{\Gamma\left(\frac{n-1}{2}\right)} 2^{\frac{1-n}{2}} x^{\frac{n-3}{2}} e^{-\frac{x}{2}}.$$

Differentiating L with respect to a, we get

$$\frac{dL}{da} = S\sqrt{n-1}\left(-\frac{1}{2}a^{-\frac{3}{2}} + \frac{1}{2}b^{-\frac{3}{2}}\frac{db}{da}\right).$$

From

$$\int_a^b f(u)\,\mathrm{d}u = 1-\alpha,$$

we find the derivative of b with respect to a as follows:

$$\frac{d}{da}\int_a^b f(u)\,\mathrm{d}u = \frac{d}{da}(1-\alpha)$$

that is

$$f(b)\frac{db}{da} - f(a) = 0.$$

Thus, we have

$$\frac{db}{da} = \frac{f(a)}{f(b)}.$$

Letting this into the expression for the derivative of L, we get

$$\frac{dL}{da} = S\sqrt{n-1}\left(-\frac{1}{2}a^{-\frac{3}{2}} + \frac{1}{2}b^{-\frac{3}{2}}\frac{f(a)}{f(b)}\right).$$

Setting this derivative to zero, we get

$$S\sqrt{n-1}\left(-\frac{1}{2}a^{-\frac{3}{2}} + \frac{1}{2}b^{-\frac{3}{2}}\frac{f(a)}{f(b)}\right) = 0$$

which yields

$$a^{\frac{3}{2}} f(a) = b^{\frac{3}{2}} f(b).$$

Using the form of f, we get from the above expression

$$a^{\frac{3}{2}} a^{\frac{n-3}{2}} e^{-\frac{a}{2}} = b^{\frac{3}{2}} b^{\frac{n-3}{2}} e^{-\frac{b}{2}}$$

that is

$$a^{\frac{n}{2}} e^{-\frac{a}{2}} = b^{\frac{n}{2}} e^{-\frac{b}{2}}.$$

From this we get

$$\ln\left(\frac{a}{b}\right) = \left(\frac{a-b}{n}\right).$$

Hence to obtain the pair of constants a and b that will produce the shortest confidence interval for σ, we have to solve the following system of nonlinear equations

$$\begin{cases} \displaystyle\int_a^b f(u)\,\mathrm{d}u = 1-\alpha \\ \ln\left(\dfrac{a}{b}\right) = \dfrac{a-b}{n}. \end{cases} \qquad (\star)$$

If a_o and b_o are solutions of $(\star)$, then the shortest confidence interval for σ is given by

$$\left[\sqrt{\frac{(n-1)S^2}{b_o}}, \quad \sqrt{\frac{(n-1)S^2}{a_o}}\right].$$

Since this system of nonlinear equations is hard to solve analytically, numerical solutions are given in statistical literature in the form of a table for finding the shortest interval for the variance.

10.5 Approximate Confidence Interval with MLE

When the sample size n is large, we can use the asymptotic normality of the maximum likelihood estimator to construct an approximate confidence interval for the parameter θ.

$$\frac{\widehat{\theta} - \mathbb{E}[\widehat{\theta}]}{\sqrt{Var[\widehat{\theta}]}} \sim N(0,1) \quad \text{as } n \to \infty. \qquad (10.10)$$

Since that for large n, the maximum likelihood estimator $\widehat{\theta}$ is unbiased, so $\mathbb{E}[\widehat{\theta}] = \theta$. Thus

$$\frac{\widehat{\theta} - \theta}{\sqrt{Var[\widehat{\theta}]}} \sim N(0,1) \quad \text{as } n \to \infty. \tag{10.11}$$

Now using

$$Q = \frac{\widehat{\theta} - \theta}{\sqrt{Var[\widehat{\theta}]}}$$

as the pivotal quantity, we construct an *approximate* $100(1-\alpha)\%$ confidence interval for θ as follows:

$$\begin{aligned}
1 - \alpha &= \mathop{\mathbb{P}}_{\theta}\big[\, -z_{\alpha/2} \le Q \le z_{\alpha/2} \,\big] \\
&= \mathop{\mathbb{P}}_{\theta}\left[-z_{\alpha/2} \le \frac{\widehat{\theta} - \theta}{\sqrt{Var[\widehat{\theta}]}} \le z_{\alpha/2} \right] \\
&= \mathop{\mathbb{P}}_{\theta}\left[\widehat{\theta} - z_{\alpha/2}\sqrt{Var[\widehat{\theta}]} \le \theta \le \widehat{\theta} + z_{\alpha/2}\sqrt{Var[\widehat{\theta}]} \right].
\end{aligned}$$

✱ Example 10.5.1. If $X_1, X_2, \ldots, X_n$ is a random sample from an exponential distribution with pdf

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

where $\theta > 0$ is an unknown parameter. Find an approximate $100(1-\alpha)\%$ confidence interval for θ when the sample size n is large.

✲ Solution The log-likelihood function from the previous example is

$$\ell(\theta) = n \ln \theta + (\theta - 1) \sum_{i=1}^{n} \ln x_i.$$

The likelihood function $L(\theta)$ of the sample is

$$L(\theta) = \prod_{i=1}^{n} \theta x_i^{\theta-1}.$$

Hence

$$\ln L(\theta) = n \ln \theta + (\theta - 1) \sum_{i=1}^{n} \ln x_i.$$

The first derivative of the logarithm of the likelihood function is

$$\frac{d}{d\theta}\ln L(\theta) = \frac{n}{\theta} + \sum_{i=1}^{n}\ln x_i.$$

Setting this derivative to zero and solving for θ, we obtain

$$\theta = -\frac{n}{\sum_{i=1}^{n}\ln x_i}.$$

Hence, the maximum likelihood estimator of θ is given by

$$\widehat{\theta} = -\frac{n}{\sum_{i=1}^{n}\ln X_i}.$$

Finding the variance of this estimator is difficult. We compute its variance by computing the Cramér-Rao bound for this estimator. The second derivative of the logarithm of the likelihood function is given by

$$\frac{d^2}{d\theta^2}\ln L(\theta) = \frac{d}{d\theta}\left(\frac{n}{\theta} + \sum_{i=1}^{n}\ln x_i\right) = -\frac{n}{\theta^2}.$$

Hence

$$E\left(\frac{d^2}{d\theta^2}\ln L(\theta)\right) = -\frac{n}{\theta^2}.$$

Therefore

$$Var\left(\widehat{\theta}\right) \geq \frac{\theta}{n}.$$

Thus we take

$$Var\left(\widehat{\theta}\right) \simeq \frac{\theta}{n}.$$

Since $Var\left(\widehat{\theta}\right)$ has θ in its expression, we replace the unknown θ by its estimate $\widehat{\theta}$ so that

$$Var\left(\widehat{\theta}\right) \simeq \frac{\widehat{\theta}^2}{n}.$$

The $100(1-\alpha)\%$ approximate confidence interval for θ is given by

$$\left[\widehat{\theta} - z_{\frac{\alpha}{2}}\frac{\widehat{\theta}}{\sqrt{n}}, \widehat{\theta} + z_{\frac{\alpha}{2}}\frac{\widehat{\theta}}{\sqrt{n}}\right],$$

which is

$$\left[-\frac{n}{\sum_{i=1}^{n}\ln X_i} + z_{\frac{\alpha}{2}}\left(\frac{\sqrt{n}}{\sum_{i=1}^{n}\ln X_i}\right), -\frac{n}{\sum_{i=1}^{n}\ln X_i} - z_{\frac{\alpha}{2}}\left(\frac{\sqrt{n}}{\sum_{i=1}^{n}\ln X_i}\right)\right].$$

✲ **Remark** (Remark 17.7). *In the next section 17.2, we derived the exact confidence interval for θ when the population distribution in exponential. The exact $100(1-\alpha)\%$ confidence interval for θ was given by*

$$\left[-\frac{\chi^2_{\frac{\alpha}{2}}(2n)}{2\sum_{i=1}^{n}\ln X_i}, -\frac{\chi^2_{1-\frac{\alpha}{2}}(2n)}{2\sum_{i=1}^{n}\ln X_i}\right].$$

Note that this exact confidence interval is not the shortest confidence interval for the parameter θ.

◀

✱ **Example 10.5.2.** If $X_1, X_2, \ldots, X_{49}$ is a random sample from a population with density

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

where $\theta > 0$ is an unknown parameter, what are 90% approximate and exact confidence intervals for θ if $\sum_{i=1}^{49}\ln X_i = -0.7567$?

✲ **Solution** We are given the followings:

$$n = 49$$
$$\sum_{i=1}^{49}\ln X_i = -0.7576$$
$$1-\alpha = 0.90.$$

Hence, we get

$$z_{0.05} = 1.64,$$
$$\frac{n}{\sum_{i=1}^{n}\ln X_i} = \frac{49}{-0.7567} = -64.75$$

and

$$\frac{\sqrt{n}}{\sum_{i=1}^{n}\ln X_i} = \frac{7}{-0.7567} = -9.25.$$

Hence, the approximate confidence interval is given by

$$[64.75 - (1.64)(9.25), 64.75 + (1.64)(9.25)]$$

that is $[49.58, 79.92]$.

Next, we compute the exact 90% confidence interval for θ using the formula

$$\left[-\frac{\chi^2_{\frac{\alpha}{2}}(2n)}{2\sum_{i=1}^{n}\ln X_i}, -\frac{\chi^2_{1-\frac{\alpha}{2}}(2n)}{2\sum_{i=1}^{n}\ln X_i}\right].$$

From chi-square table, we get

$$\chi^2_{0.05}(98) = 77.93 \quad \text{and} \quad \chi^2_{0.95}(98) = 124.34.$$

Hence, the exact 90% confidence interval is

$$\left[\frac{77.93}{(2)(0.7567)}, \frac{124.34}{(2)(0.7567)}\right]$$

that is $[51.49, 82.16]$. ◀

✱ Example 10.5.3. If $X_1, X_2, \ldots, X_n$ is a random sample from a population with density

$$f(x;\theta) = \begin{cases} (1-\theta)\theta^x & \text{if } x = 0, 1, 2, \ldots, \infty \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < \theta < 1$ is an unknown parameter, what is a $100(1-\alpha)\%$ approximate confidence interval for θ if the sample size is large?

✲ Solution The logarithm of the likelihood function of the sample is

$$\ln L(\theta) = \ln\theta \sum_{i=1}^{n} x_i + n\ln(1-\theta).$$

Differentiating we see obtain

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\ln L(\theta) = \frac{\sum_{i=1}^{n} x_i}{\theta} - \frac{n}{1-\theta}.$$

Equating this derivative to zero and solving for θ, we get $\theta = \frac{\overline{x}}{1+\overline{x}}$. Thus, the maximum likelihood estimator of θ is given by

$$\widehat{\theta} = \frac{\overline{X}}{1+\overline{X}}.$$

Next, we find the variance of this estimator using the Cramér-Rao lower bound. For this, we need the second derivative of $\ln L(\theta)$. Hence

$$\frac{d^2}{d\theta^2}\ln L(\theta) = -\frac{n\overline{x}}{\theta^2} - \frac{n}{(1-\theta)^2}.$$

Therefore

$$\begin{aligned}
\mathbb{E}\left(\frac{d^2}{d\theta^2}\ln L(\theta)\right) &= \mathbb{E}\left(-\frac{n\overline{X}}{\theta^2} - \frac{n}{(1-\theta)^2}\right) \\
&= -\frac{n}{\theta^2}\mathbb{E}\left(\overline{X}\right) - \frac{n}{(1-\theta)^2} \\
&= -\frac{n}{\theta^2}\cdot\frac{1}{(1-\theta)} - \frac{n}{(1-\theta)^2} \quad \text{(since each } X_i \sim \text{GEO}(1-\theta)\text{)} \\
&= -\frac{n}{\theta(1-\theta)}\left[\frac{1}{\theta} + \frac{\theta}{1-\theta}\right] \\
&= -\frac{n(1-\theta+\theta^2)}{\theta^2(1-\theta)^2}.
\end{aligned}$$

Therefore

$$Var\left(\widehat{\theta}\right) \simeq \frac{\widehat{\theta}^2\left(1-\widehat{\theta}\right)^2}{n\left(1-\widehat{\theta}+\widehat{\theta}^2\right)}.$$

The $100(1-\alpha)\%$ approximate confidence interval for θ is given by

$$\left[\widehat{\theta} - z_{\frac{\alpha}{2}}\frac{\widehat{\theta}\left(1-\widehat{\theta}\right)}{\sqrt{n\left(1-\widehat{\theta}+\widehat{\theta}^2\right)}}, \widehat{\theta} + z_{\frac{\alpha}{2}}\frac{\widehat{\theta}\left(1-\widehat{\theta}\right)}{\sqrt{n\left(1-\widehat{\theta}+\widehat{\theta}^2\right)}}\right],$$

where

$$\widehat{\theta} = \frac{\overline{X}}{1+\overline{X}}.$$

◀

Tutorials

Exercise 10.1 A machine produces steel rods with lengths that are normally distributed with unknown mean μ and standard deviation σ^2.

A quality control inspector uses a gauge to measure the length, x centimeters, of each rod in a random sample of 100 rods from the machine's production. The summarised data are as follows:

$$\sum_{i=1}^{100} x_i = 1040.0, \quad \sum_{i=1}^{100} x_i^2 = 11102.11.$$

Construct a 95% confidence interval for the mean length of rods produced by the machine.

Exercise 10.2 In laboratory work, it is desirable to run careful checks on the variability of readings produced on standard samples. In a study of the amount of calcium in drinking water undertaken as part of a water quality assessment, the same standard sample was run through the laboratory six times, yielding the following results (in mg/L):

$$40.1,\ 39.8,\ 40.0,\ 40.2,\ 39.9,\ 40.1$$

Construct a 95% confidence interval for the mean calcium content in the water.

Exercise 10.3 The executives at Aperture Inc. having recently solved their widget crises, have another major problem with one of their products. Many cities are sending complaints that their manhole covers are defective and people are falling into the sewers. Aperture Inc. is pretty sure that only 4% of their manhole covers are defective, but they would like to do a study to confirm this number. They are hoping to construct a 95% confidence interval to get within 0.01 of the true proportion of defective manhole covers. How many manhole covers need to be tested?

Exercise 10.4 Let $X_1, X_2, \ldots, X_n$ be a random sample of with gamma density function

$$f_{X_i}(x|\theta,\beta) = \frac{1}{\Gamma(\beta)\,\theta^{\beta}}\,x^{\beta-1}e^{-x/\theta}, \quad x > 0.$$

Where θ is an unknown parameter and β is a known constant. Show that

$$\left[\frac{2\sum_{i=1}^{n} X_i}{\chi^2_{1-\frac{\alpha}{2}(2n\beta)}},\ \frac{2\sum_{i=1}^{n} X_i}{\chi^2_{\frac{\alpha}{2}(2n\beta)}}\right]$$

is a $100(1-\alpha)\%$ confidence interval for θ.

Exercise 10.5 Let $X_1, X_2, \ldots, X_n$ be a random sample from population with Pareto density function

$$f_{X_i}(x|\theta,\beta) = \begin{cases} \theta\,\beta^{\theta}\,x^{-(\theta+1)} & x \geq \beta \\ 0 & \text{otherwise,} \end{cases}$$

Where θ is an unknown parameter and β is a known constant. Show that

$$\left[\frac{2\sum_{i=1}^{n} \ln\left(\frac{X_i}{\beta}\right)}{\chi^2_{1-\frac{\alpha}{2}(2n\beta)}},\ \frac{2\sum_{i=1}^{n} \ln\left(\frac{X_i}{\beta}\right)}{\chi^2_{\frac{\alpha}{2}(2n\beta)}}\right]$$

is a $100(1-\alpha)\%$ confidence interval for $\frac{1}{\theta}$.

Exercise 10.6 Let $X_1, X_2, \ldots, X_4$ be a random sample of sample size $n = 4$ from population

with density function

$$f_{X_i}(x|\theta) = \begin{cases} (1+\theta)x^{\theta} & 0 < x < 1 \\ 0 & \text{otherwise,} \end{cases}$$

Where $\theta > 0$ is an unknown parameter. Construct a 90% confidence interval for θ if the observed sample data is

$$x_1 = 0.92,\ x_2 = 0.75,\ x_3 = 0.85,\ x_4 = 0.80.$$

This page intentionally left blank.

11

Hypothesis testing

Definition 11.1 Testing hypotheses
The hypothesis to be tested is called the null hypothesis. The negation of the null hypothesis is called the alternative hypothesis. The null and alternative hypotheses are denoted by H_0 and H_1, respectively.
If θ denotes a population parameter, then the general format of the null hypothesis and alternative hypothesis is

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta_1 \tag{★}$$

where Θ_0 and Θ_1 are disjoint subsets of the parameter space Θ such that

$$\Theta_0 \cap \Theta_1 = \varnothing \quad \text{and} \quad \Theta_0 \sqcup \Theta_1 \subseteq \Theta. \tag{11.1}$$

Remark. *If $\Theta_0 \cup \Theta_1 = \Theta$, then the test (♣) becomes*

$$H_0 : \theta \in \Theta_0 \quad \textit{versus} \quad H_1 : \theta \notin \Theta_0 \tag{11.2}$$

Definition 11.2 Errors in hypothesis
A **type I error** is when the null hypothesis is rejected, but it is true.
A **type II error** is not rejecting null hypothesis H_0, but in fact H_0 is false.

11.1 What is p-value?

One way to think of it is the courtroom example. One day you get arrested as a suspect, you are innocent until proven guilty. The null hypothesis H_0 is that you are innocent. But now all evidence is compiled against you. The question is: given that we are in the world where you are in fact innocent, how likely are we to see this much evidence compiled against you? As opposed to asking "what is the probability that you are innocent?", that is what p-value means.

In statistical terminology, the p-value measures the "extremeness" of the sample.

Definition 11.3 p-value
The p-value is the probability we would get the sample we have or something more extreme if the *null hypothesis* were true.

So, the smaller the p-value, the more evidence there is in the sample data against the null hypothesis and for the alternative hypothesis.

So what constitutes "sufficiently small" and "extreme enough" to make a decision about the null hypothesis?

11.2 Two sample testing

If sample X_1 drawn from $N(\mu_1, \sigma_1^2)$ and is independent of $X_2 \sim N(\mu_2, \sigma_2^2)$, then

$$X_1 - X_2 \sim N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$$

By Central Limit theorem we obtained the result

$$\overline{X_1} - \overline{X_2} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right).$$

where n_1 and n_2 are the sample size for X_1 and X_2 correspondingly. Hence the test statistic is

$$z = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}. \tag{11.3}$$

In general, the test statistic for comparing two normal samples with known variances is as follow:

Definition 11.4 Two sample means test – Large sample size with different variances
Assume that $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$ are two independent samples with $\sigma_1^2 \neq \sigma_2^2$. The test statistic value is

$$z = \frac{\bar{x}_1 - \bar{x}_2 - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}. \tag{11.4}$$

for null hypothesis

$$H_0 : \mu_1 - \mu_2 = \Delta_0$$

with all possible alternative hypotheses:

Alternative Hypothesis	Rejection Region at level α
$H_1 : \mu_1 - \mu_2 > \Delta_0$	$z \geq z_\alpha$ (Upper-tailed)
$H_1 : \mu_1 - \mu_2 < \Delta_0$	$z \leq -z_\alpha$ (Lower-tailed)
$H_1 : \mu_1 - \mu_2 \neq \Delta_0$	either $z \geq z_{\alpha/2}$ or $z \leq -z_{\alpha/2}$ (Two-tailed)

11.2.1 Two samples with equal unknown population variances

Back to previous section, the test statistic value is

$$z = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Since $\sigma_1^2 = \sigma_2^2 = \sigma^2$, then

$$z = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Now both sample variances, S_1^2 and S_2^2, are estimates of σ^2. so this information can be combine to form a *pooled* (or *weighted*) estimate of variance, that is,

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}.$$

Replace σ^2 with pooled variance estimator S_p^2, hence

$$t = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{\nu=n_1+n_2-2}$$

is a t-statistic with degrees of freedom given by $n_1 + n_2 - 2$.

Definition 11.5 Two sample means test – Small sample size with equal variances
Assume that X_1 and X_2 are two independent samples with $\sigma_1^2 = \sigma_2^2 = \sigma^2$. The test statistic value is

$$t = \frac{\bar{x}_1 - \bar{x}_2 - \Delta_0}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{11.5}$$

where the pooled variance id defined as

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}. \tag{11.6}$$

The null hypothesis is

$$H_0 : \mu_1 - \mu_2 = \Delta_0$$

with all possible alternative hypotheses:

Alternative Hypothesis	Rejection Region at level α
$H_1 : \mu_1 - \mu_2 > \Delta_0$	$t \geq t_\alpha$ (Upper-tailed)
$H_1 : \mu_1 - \mu_2 < \Delta_0$	$t \leq -t_\alpha$ (Lower-tailed)
$H_1 : \mu_1 - \mu_2 \neq \Delta_0$	either $t \geq t_{\alpha/2}$ or $t \leq -t_{\alpha/2}$ (Two-tailed)

11.2.2 Paired samples t-test

Assume that we have two dependent samples $X_{\blacksquare}$ and $X_{\clubsuit}$, each of size n. The paired samples are normally distributed with means μ_1 and μ_2, respectively. From earlier work we define the differences between the paired observations as

$$D_i = X_{\blacksquare\, i} - X_{\clubsuit\, i} \quad \text{for } i = 1, 2, \ldots, n. \tag{11.7}$$

In fact, a test of $H_0 : \mu_0 = \mu_1$ is equivalent to a test of $H_0 : \mu_D = 0$. Even though the original samples $X_{\blacksquare}$ and $X_{\clubsuit}$ are may well be normally distributed. The key assumption is that the differences D_i are approximately normally distributed with mean μ_D and variance σ_D^2.

Let $\overline{D}$ and $\widehat{\sigma}_D^2$ be the sample mean and sample standard deviation of the differences D_i, respectively. Where

$$\mu_D = \overline{D} = \frac{\sum_{i=1}^{n} D_i}{n}$$

and

$$\widehat{\sigma}_D^2 = \frac{\sum_{i=1}^{n} (D_i - \overline{D})^2}{n-1}.$$

Then by Central Limit Theorem, the sampling distribution of $\overline{D}$ is approximately normal with mean μ_D and variance σ_D^2/n. That is,

$$\overline{D} \sim N\left(\mu_D, \frac{\sigma_D^2}{n}\right).$$

The test statistic is given by

$$t = \frac{\overline{D} - \mu_D}{S_D/\sqrt{n}}$$

which is a t-statistic with $\nu = n - 1$ degrees of freedom.

The null hypothesis is

$$H_0 : \mu_1 - \mu_2 = 0$$

with all possible alternative hypotheses:

Alternative Hypothesis	Rejection Region at level α
$H_1 : \mu_1 - \mu_2 > 0$	$t \geq t_\alpha$ (Upper-tailed)
$H_1 : \mu_1 - \mu_2 < 0$	$t \leq -t_\alpha$ (Lower-tailed)
$H_1 : \mu_1 - \mu_2 \neq 0$	either $t \geq t_{\alpha/2}$ or $t \leq -t_{\alpha/2}$ (Two-tailed)

11.2.3 Large-Sample Test

Assume that we have a single pooled sample of size $n_1 + n_2$ rather than having two separate samples of size n_1 and n_2. For example, we have two different populations $X \sim BIN(n_1, p_1)$ and $Y \sim BIN(n_2, p_2)$ with $p_1 = p_2$. By combining them together we will have a single sample of size $n_1 + n_2$ from one population with proportion p, that is,

$$X + Y \sim BIN(n_1 + n_2, p_1 = p_2 = p)$$

Definition 11.6 Large samples test – difference in proportion
Assume that X_1 and X_2 are two populations. The large-sample test statistic is

$$z = \frac{\widehat{p}_1 - \widehat{p}_2}{\widehat{p}\,\widehat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}. \tag{11.8}$$

The null hypothesis is

$$H_0 : p_1 - p_2 = 0$$

with all possible alternative hypotheses:

Alternative Hypothesis	Rejection Region at level α
$H_1 : p_1 - p_2 > 0$	$z \geq z_\alpha$ (Upper-tailed)
$H_1 : p_1 - p_2 < 0$	$z \leq -z_\alpha$ (Lower-tailed)
$H_1 : p_1 - p_2 \neq 0$	either $z \geq z_{\alpha/2}$ or $z \leq -z_{\alpha/2}$ (Two-tailed)

11.3 Power of test

Definition 11.7 Power of test
For testing null hypothesis $H_0 : \theta \in \Theta_0$, the power function of a test ϕ with rejection region RR is the function of θ defined as

$$\beta_\phi(\theta) = \mathbb{P}_\theta[X \in RR]. \tag{11.9}$$

Note that

$$\beta(\theta) = \begin{cases} \text{Type I error probability} & \theta \in \Theta_0 \\ \text{One minus the Type II error probability} & \theta \in \Theta_0^c \end{cases}$$

A good test should has power value near 0 for most $\theta \in \Theta_0$, and near 1 for most $\theta \in \Theta_0^c$. The power function is similar to the MSE or risk function in estimation in that typically a test is better than another for some θ's but worse for other θ's.

When testing $H_0 : \theta \leq \theta_0$ with a univariate parameter θ, a reasonable test should have the following properties for its power function $\beta(\theta)$:

- $\beta(\theta)$ is a increasing function of θ.
- $\lim_{n\to\theta_-}\beta(\theta)=0$ and $\lim_{n\to\theta_+}\beta(\theta)=1$, where θ_- is the smallest θ (might be $-\infty$) and θ_+ is the largest θ (might be $+\infty$).

11.3.1 Neyman-Pearson Lemma and powerful test

Lemma 11.1 Neyman-Pearson Lemma
Consider we want to test $H_0:\theta=\theta_0$ versus $H_0:\theta=\theta_1$, where θ_0,θ_1 are two distinct parameter values.

Existence

Any test with the rejection region $\mathcal{R}$ satisfying two conditions:

$$x\in\mathcal{R} \quad \text{if } f_{\theta_1}(x)>cf_{\theta_0}(x)$$

$$x\notin\mathcal{R} \quad \text{if } f_{\theta_1}(x)<cf_{\theta_0}(x)$$

for some $c\geq 0$ is a UMP test of size α_c with

$$\alpha_c=\mathbb{P}_{\theta_0}[X\in\mathcal{R}]. \tag{11.10}$$

And note that nothing will be specified when $f_{\theta_1}(x)=cf_{\theta_0}(x)$.

Uniqueness

If the previously specified test has a positive c, then every level α_c the UMP test has size α_c and has the same form previously stated except perhaps on a set $\mathcal{A}$ satisfying

$$\mathbb{P}_{\theta_0}[\mathcal{A}]=\mathbb{P}_{\theta_1}[\mathcal{A}]=0. \tag{11.11}$$

Proof. We consider the continuous case with pdfs, and that every test can be represented by the indicator function of its rejection region.

[Existence]

Let $\phi(x)$ be the indicator function of the rejection region of the test in the theorem and ψ be the indicator function of the rejection region of any other level α test.

From the construction of ϕ we have

$$[\phi(x)-\psi(x)][f_{\theta_1}(x)-cf_{\theta_0}(x)]\geq 0 \quad \forall x\in\mathfrak{X}$$

Thus,

$$\begin{aligned} 0 &\leq \int_{\mathfrak{X}} [\phi(x) - \psi(x)][f_{\theta_1}(x) - c f_{\theta_0}(x)] \, \mathrm{d}x \\ &= \beta_\phi(\theta_1) - \beta_\psi(\theta_1) - c[\beta_\phi(\theta_0) - \beta_\psi(\theta_0)] \\ &= \beta_\phi(\theta_1) - \beta_\psi(\theta_1) - c[\alpha_c - \beta_\psi(\theta_0)] \\ &\leq \beta_\phi(\theta_1) - \beta_\psi(\theta_1). \end{aligned}$$

This proves $\beta_\phi(\theta_1) \geq \beta_\psi(\theta_1)$.

Since θ_1 is the only point in Θ_0^c, thus ϕ is a UMP test of size $\alpha_c = \mathbb{P}_{\theta_0}[X \in \mathcal{R}]$.

We now further continue to proof the uniqueness part.

[Uniqueness]

Let ψ be the indicator function of another UMP test of level α_c. From the previous proof, we know ϕ is a UMP test of size α_c and hence

$$\beta_\phi(\theta_1) = \beta_\psi(\theta_1)$$

and

$$0 \leq \int_{\mathfrak{X}} [\phi(x) - \psi(x)][f_{\theta_1}(x) - c f_{\theta_0}(x)] \, \mathrm{d}x = -c[\alpha_c - \beta_\psi(\theta_0)].$$

Because $c > 0$, we have $\alpha_c - \beta_\psi(\theta_0) \leq 0$. Since ψ is a level α_c test, $\beta_\psi(\theta_0) \leq \alpha_c$ and hence $\beta_\psi(\theta_0) = \alpha_c$. i.e. ψ has size α_c, which implies

$$\int_{\mathfrak{X}} [\phi(x) - \psi(x)][f_{\theta_1}(x) - c f_{\theta_0}(x)] \, \mathrm{d}x = 0$$

Now let

$$\mathcal{A} := \{x : [\phi(x) - \psi(x)][f_{\theta_1}(x) - c f_{\theta_0}(x)] > 0\},$$

then

$$\int_{\mathcal{A}} [\phi(x) - \psi(x)][f_{\theta_1}(x) - c f_{\theta_0}(x)] \, \mathrm{d}x = 0.$$

We pick $h(x) = [f_{\theta_0}(x) + f_{\theta_0}(x)]/2$, $h(x)$ is a pdf and $h(x) > 0$ on $\mathcal{A}$, which is now

$$\int_{\mathcal{A}} \frac{[\phi(x) - \psi(x)][f_{\theta_1}(x) - c f_{\theta_0}(x)]}{h(x)} \, h(x) \, \mathrm{d}x = 0.$$

From the result we established, this become

$$\mathbb{P}_h \left(\mathbb{1}_{x \in \mathcal{A}} [\phi(x) - \psi(x)][f_{\theta_1}(x) - c f_{\theta_0}(x)] = 0 \right) = 1.$$

□

✱ Example 11.3.1. Suppose that X represents a single observation from a population with density function given by

$$f_X(x|\theta) = \begin{cases} \theta x^{\theta-1} & \text{for } 0 < x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Find the most powerful test with significance level $\alpha = 0.05$ to test

$$H_0 : \theta = 2$$

against the alternative

$$H_1 : \theta = 1.$$

✲ Solution Neyman-Pearson lemma can be applied to derive this test. In this case we comparing the likelihood ratio between H_0 and H_1,

$$\frac{L(\theta_0)}{L(\theta_1)} = \frac{f_X(x|\theta_0 = 2)}{f_X(x|\theta_1 = 1)} = \frac{2x}{1x^0} = 2x, \quad 0 < x < 1$$

and the form of the rejection region for the most powerful test is

$$RR = \{x \mid 2x < k\} = \left\{x \mid x < \tfrac{1}{2}k\right\}$$

for some constant k. Equivalently, $k/2$ is a constant. By letting $k' = k/2$. The rejection region can be simplify to

$$RR = \{x \mid x < k'\}.$$

At significance level $\alpha = 0.05$, the value of k' is determined by

$$\begin{aligned} 0.05 &= \mathbb{P}[X \in RR \mid \theta = 2] \\ &= \mathbb{P}[X < k' \mid \theta = 2] \\ &= \int_0^{k'} 2x^{2-1} \, \mathrm{d}x \\ &= \boxed{(k')^2}. \end{aligned}$$

Therefore $(k')^2 = 0.05$. On solving, the rejection region of the most powerful test is actually

$$RR = \{x \mid x < \sqrt{0.05} = 0.2236\}.$$

What is the actual value for $\mathsf{power}(\theta)$ when $\theta = 1$?

$$\begin{aligned}\mathsf{power}(1) &= \mathbb{P}[X \in RR \mid \theta = 1] \\ &= \mathbb{P}[X < 0.2236 \mid \theta = 1] \\ &= \int_0^{0.2236} 1 \, \mathrm{d}x \\ &= \boxed{0.2236}.\end{aligned}$$

We can see that even though the rejection region gives the maximum value for $\mathsf{power}(1)$ among all tests with $\alpha = 0.05$. But $\beta(1) = 1 - 0.2236 = 0.7764$ is still very large. ◀

✱ Example 11.3.2. Let X_1, X_2, X_3 denote three independent observations from a population with pdf

$$f_X(x|\theta) = \begin{cases} (1+\theta)x^\theta & \text{for } 0 \le x \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

What is the form of the best critical region of size 0.034 for testing

$$H_0 : \theta = 1 \quad \text{versus} \quad H_1 : \theta = 2?$$

✲ Solution By Neyman-Pearson lemma, the rejection region of the most powerful test is of the form

$$\begin{aligned} RR &= \{(x_1, x_2, x_3) \in \mathfrak{X} \mid \frac{L(\theta_0|x_1, x_2, x_3)}{L(\theta_1|x_1, x_2, x_3)} \le k\} \\ &= \{(x_1, x_2, x_3) \in \mathcal{I}^3 \mid \frac{\prod_{i=1}^3 f_X(x_i|\theta_0 = 1)}{\prod_{i=1}^3 f_X(x_i|\theta_1 = 2)} \le k\} \\ &= \{(x_1, x_2, x_3) \in \mathcal{I}^3 \mid \frac{8x_1x_2x_3}{27x_1^2x_2^2x_3^2} \le k\} \\ &= \{(x_1, x_2, x_3) \in \mathcal{I}^3 \mid \frac{1}{x_1x_2x_3} \le \frac{27}{8}k\} \\ &= \{(x_1, x_2, x_3) \in \mathcal{I}^3 \mid x_1x_2x_3 \ge k'\} \end{aligned}$$

where k' is some constant. Hence the most powerful test rejects H_0 is of the form "Reject H_0 if $x_1x_2x_3 \ge k'$."

The value of k' is determined by the size of the test, that is, $\alpha = 0.034$. To evaluate the constant k', we have to find the probability distribution of $X_1X_2X_3$. The distribution of $X_1X_2X_3$ is quite challenging to get. But we have shown that

$$-2(1+\theta)\sum_{i=1}^3 \ln X_i \sim \chi^2_6.$$

Now we proceed to find the constant k'. Since

$$
\begin{aligned}
0.034 &= \alpha \\
&= \mathbb{P}[\text{Reject } H_0 \mid H_0 \text{ is true}] \\
&= \mathbb{P}[X_1X_2X_3 \geq k' \mid \theta = 1] \\
&= \mathbb{P}[\ln(X_1X_2X_3) \geq \ln k' \mid \theta = 1] && \text{Taking logarithm.} \\
&= \mathbb{P}\left[\sum_{i=1}^{3} \ln X_i \geq \ln k' \,\middle|\, \theta = 1\right] \\
&= \mathbb{P}\left[-2(1+\theta)\sum_{i=1}^{3} \ln X_i \geq -2(1+\theta)\ln k' \,\middle|\, \theta = 1\right] && \text{Multiply } -2(1+\theta) \text{ on both sides.} \\
&= \mathbb{P}\left[-4\sum_{i=1}^{3} \ln X_i \geq -4\ln k'\right] \\
&= \mathbb{P}\left[\chi^2_6 \geq -4\ln k'\right]
\end{aligned}
$$

From the χ^2 table, we have

$$-4\ln k' = 1.4.$$

Therefore

$$k' = e^{-0.35} = \boxed{0.7047}.$$

Thus, the most powerful test is given by "Reject H_0 if $x_1x_2x_3 \geq 0.7047$."

The critical region is the region above the surface $x_1x_2x_3 = 0.7047$ in the unit cube $\mathcal{I}^3 = [0,1]^3$. The following figure illustrates the rejection region.

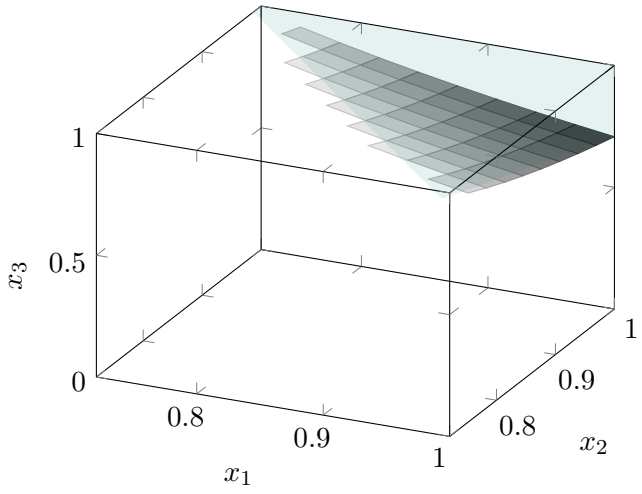


Figure 11.1: The rejection region is to the right of the surface $x_1x_2x_3 = 0.7074$. (The shaded volume)

◀

11.3.2 Likelihood Ratio Test

Definition 11.8 Likelihood Ratio Test
The likelihood ratio test statistic for testing $H_0 : \theta \in \Theta_0$ versus $H_1 :\in \Theta_0^c$ is

$$\Lambda(X) = \frac{\sup\limits_{\theta\in\Theta_0} L(\theta|X)}{\sup\limits_{\theta\in\Theta} L(\theta|X)} \tag{11.12}$$

where $L(\theta|x)$ is the likelihood function based on $X = x$. A **likelihood ratio test (LRT)** is any test that has a rejection region of the form

$$\{x \mid \Lambda(x) \leq c\}$$

where c is a constant satisfying $0 \leq c \leq 1$.

The logic behind LRTs is that the likelihood ratio $\Lambda(x)$ is likely to be small if there are parameter points in Θ_0^c for which x is much more likely than for any parameter in Θ_0. Note that in the denominator of the likelihood ratio, the supremum is taken over Θ, not Θ_0^c.

✱ Example 11.3.3. Suppose that $X_1, X_2, \ldots, X_n$ be a random sample from a $N(\mu, \sigma^2)$ population, where both μ and σ^2 are unknown. What is the likelihood ratio test of significance of size α for testing the null hypothesis

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu \neq \mu_0\,?$$

✲ Solution We illustrate the following sets:

$$\Omega = \{(\mu, \sigma^2) \in \mathbb{R}^2 \mid -\infty < \mu < \infty, \sigma^2 > 0\},$$

$$\Omega_0 = \{(\mu, \sigma^2) \in \mathbb{R}^2 \mid \mu = \mu_0, \sigma^2 > 0\},$$

where $\mu_0 \in \mathbb{R}$ is a constant. The alternative parameter space is

$$\begin{aligned}\Omega_1 &= \{(\mu, \sigma^2) \in \mathbb{R}^2 \mid \mu \neq \mu_0, \sigma^2 > 0\} \\ &= \ \setminus\ \end{aligned}$$

so that $\Omega_0 \cup \Omega_1 = \Omega$.

The likelihood function is given by

$$L(\mu, \sigma^2|x) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\}$$

Next we find the maximum of $L(\mu, \sigma^2)$ on the set Ω_0. Since the set Ω_0 is equal to the set $\{(\mu_0, \sigma^2) \in$

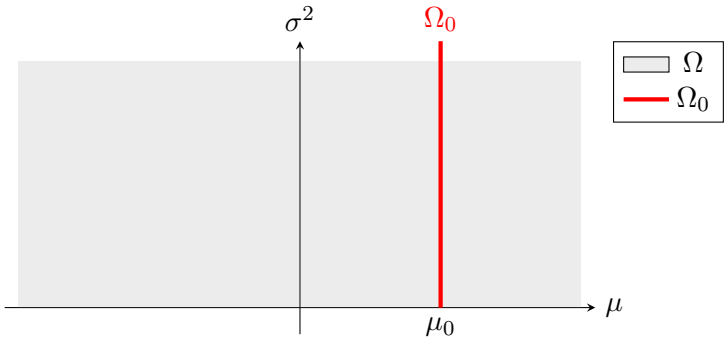


Figure 11.2: The illustrated graphs of the parameter spaces Ω and Ω_0.

$\mathbb{R}^2 \mid \sigma^2 > 0\}$, we have

$$\max_{(\mu,\sigma^2)\in\Omega_0} L(\mu, \sigma^2|x) = \max_{\sigma^2>0} L(\mu_0, \sigma^2|x).$$

Since both $L(\mu_0, \sigma^2|x)$ and $\ln L(\mu, \sigma^2|x)$ are maximized at the same point σ, we determine the value of σ where $\ln L(\mu_0, \sigma^2|x)$ is maximized. Taking the natural logarithm of the likelihood function, we have

$$\ln L(\mu_0, \sigma^2|x) = -\frac{n}{2}\ln(\sigma^2) - \frac{n}{2}\ln(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu_0)^2. \qquad (\clubsuit)$$

differentiating with respect to σ^2 and setting the result equal to zero, we have

$$\frac{\mathrm{d}}{\mathrm{d}\sigma^2}\ln L(\mu_0, \sigma^2|x) = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i - \mu_0)^2 = 0.$$

Solving for σ, we obtain

$$\widehat{\sigma}^2 = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_0)^2}. \qquad (\spadesuit)$$

Thus $\ln L(\mu, \sigma^2)$ attains maximum at $(\spadesuit)$. Since this value of σ is also yield maximum value of $L(\mu, \sigma^2)$, we have

$$\max_{\substack{(\mu,\sigma^2)\in\Omega_0,\\ \sigma>0}} L(\mu, \sigma^2|x) = \left(\frac{2\pi}{n}\sum_{i=1}^{n}(x_i - \mu_0)^2\right)^{-n/2} e^{-\frac{n}{2}}. \qquad (\heartsuit)$$

Next, we need to determine the maximum of $L(\mu, \sigma^2|x)$ on the set Ω. As before, we consider the natural logarithm of the likelihood function is already achieves its maximum at the same point as the likelihood function itself. Taking the partial derivatives of $(\clubsuit)$ with respect to μ and σ^2 respectively, we have

$$\frac{\partial}{\partial\mu}\ln L(\mu, \sigma^2|x) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = 0,$$

and

$$\frac{\partial}{\partial \sigma^2} \ln L(\mu, \sigma^2 | x) = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^{n} (x_i - \mu)^2 = 0.$$

Solving for μ and σ^2, we obtain

$$\mu = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad \text{and} \quad \sigma^2 = \widehat{\sigma}^2 = \frac{n-1}{n} s^2.$$

where $s^2 := \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$ is the sample variance.

Plugging these optimal values of μ and σ into the likelihood function $L(\mu, \sigma)$, we have

$$\max_{\substack{(\mu,\sigma^2)\in\Omega, \\ \sigma>0}} L(\mu, \sigma^2 | x) = \left(\frac{2\pi}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right)^{-n/2} e^{-\frac{n}{2}}. \tag{◆}$$

The likelihood ratio test statistic is given by

$$\Lambda = \frac{(♡)}{(◆)} = \frac{\max\limits_{\substack{(\mu,\sigma^2)\in\Omega_0, \\ \sigma>0}} L(\mu, \sigma^2 | x)}{\max\limits_{\substack{(\mu,\sigma^2)\in\Omega, \\ \sigma>0}} L(\mu, \sigma^2 | x)} = \frac{\left(\frac{2\pi}{n} \sum_{i=1}^{n} (x_i - \mu_0)^2 \right)^{-n/2} e^{-\frac{n}{2}}}{\left(\frac{2\pi}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right)^{-n/2} e^{-\frac{n}{2}}} = \left[\frac{\sum_{i=1}^{n} (x_i - \mu_0)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right]^{-n/2} \tag{11.13}$$

Since

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = (n-1)s^2$$

and

$$\sum_{i=1}^{n} (x_i - \mu)^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 + n(\bar{x} - \mu_0)^2.$$

We have

$$\Lambda(x_1, x_2, \ldots, x_n) = \left[\frac{\sum_{i=1}^{n} (x_i - \mu_0)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right]^{-n/2} = \left(1 + \frac{n}{n-1} \frac{(\bar{x} - \mu_0)^2}{s^2} \right)^{-n/2}.$$

Now the inequality $\Lambda(x_1, x_2, \ldots, x_n) \leq k$ becomes

$$\left(1 + \frac{n}{n-1} \frac{(\bar{x} - \mu_0)^2}{s^2} \right)^{-n/2} \leq k \tag{11.14}$$

and which can be rewritten in the form of

$$\left(\frac{\bar{x}-\mu_0}{s}\right)^2 \geq \frac{n-1}{n}\left(k^{-\frac{2}{n}}-1\right)$$

or

$$\left|\frac{\bar{x}-\mu_0}{s/\sqrt{n}}\right| \geq K$$

where $K := \sqrt{(n-1)\left(k^{-\frac{2}{n}}-1\right)}$. In view of the above inequality, the critical region can be described as

$$RR = \left\{(x_1, x_2, \ldots, x_n) : \left|\frac{\bar{x}-\mu_0}{s/\sqrt{n}}\right| \geq K\right\}.$$

That is, the best likelihood ratio test is described as: " Reject H_0 if $\left|\frac{\bar{x}-\mu_0}{s/\sqrt{n}}\right| \geq K$". Since we are given α, the size of the critical region. We can determine the constant K.

In order to find K, we need the density function of the statistic $\frac{\bar{x}-\mu_0}{s/\sqrt{n}}$ when the population X is normally distributed and the null hypothesis $H_0 : \mu = \mu_0$ is true.

Since the population is normal with mean μ and variance σ^2, so the statistic

$$\frac{\overline{X}-\mu_0}{S/\sqrt{n}} \sim t_{\nu=n-1}$$

where S is the sample standard deviation that equals to $\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2}$. Hence

$$K = t_{\frac{\alpha}{2};\,n-1}\,\frac{s}{\sqrt{n}},$$

where $t_{\frac{\alpha}{2};\,n-1}$ is a real number such that the integral of the t-distribution with $n-1$ degrees of freedom from $t_{\frac{\alpha}{2};\,n-1}$ to $+\infty$ with area under curve equals $\alpha/2$.

Therefore, the likelihood ratio test is given by "Reject H_0 if

$$|\overline{X}-\mu_0| \geq t_{\frac{\alpha}{2};\,n-1}\,\frac{s}{\sqrt{n}}."$$

if we denote

$$T = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$$

then the above inequality becomes

$$|T| \geq t_{\frac{\alpha}{2};\,n-1}.$$

Thus the critical region is now simplified to

$$RR = \{(x_1, x_2, \ldots, x_n) \mid |t| \geq t_{\frac{\alpha}{2};\,n-1}\}. \tag{11.15}$$

This tells us that the null hypothesis must be rejected when the absolute value of t takes on a value

greater than or equal to $t_{\frac{\alpha}{2};\, n-1}$. ◀

Remark. *In the example above, if we had right-sided alternative*

$$H_0 : \mu \neq \mu_0 \quad \textit{versus} \quad H_1 : \mu > \mu_0.$$

Then the critical region becomes

$$RR = \{(x_1, x_2, \ldots, x_n) \mid t \geq t_{\alpha;\, n-1}\}.$$

Similarly, if the alternative would have been left-sided, that is,

$$H_0 : \mu \neq \mu_0 \quad \textit{versus} \quad H_1 : \mu < \mu_0.$$

then the critical region would have been

$$RR = \{(x_1, x_2, \ldots, x_n) \mid t \leq -t_{\alpha;\, n-1}\}.$$

We summarize the three cases of hypotheses test of the mean of the normal population (with unknown variance) in the following table.

H_0	H_1	*Critical region*
$\mu = \mu_0$	$\mu > \mu_0$	$t \geq t_{\alpha;\, n-1}$
$\mu = \mu_0$	$\mu < \mu_0$	$t \leq -t_{\alpha;\, n-1}$
$\mu = \mu_0$	$\mu \neq \mu_0$	$\lvert t \rvert \geq t_{\frac{\alpha}{2};\, n-1}$

11.4 Hypothesis testing for variance

Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal population with mean μ and variance σ^2. What is the likelihood ratio test of significance of size α for testing the null hypothesis

$$H_0 : \sigma^2 = \sigma_0^2 \quad \text{versus} \quad H_1 : \sigma^2 \neq \sigma_0^2 \,?$$

We illustrate the following example.

$$\Theta = \{(\mu, \sigma^2) \in \mathbb{R}^2 \mid -\infty < \mu < \infty, \sigma^2 > 0\}$$

$$\Theta_0 = \{(\mu, \sigma^2) \in \mathbb{R}^2 \mid -\infty < \mu < \infty, \sigma^2 = \sigma_0^2\}$$

where $\sigma_0^2 > 0$ is a constant. The alternative parameter space is

$$\begin{aligned} \Theta_1 &= \{(\mu, \sigma^2) \in \mathbb{R}^2 \mid -\infty < \mu < \infty, \sigma^2 \neq \sigma_0^2\} \\ &= \quad \setminus \end{aligned}$$

and that $\Theta_0 \cup \Theta_1 = \Theta$.

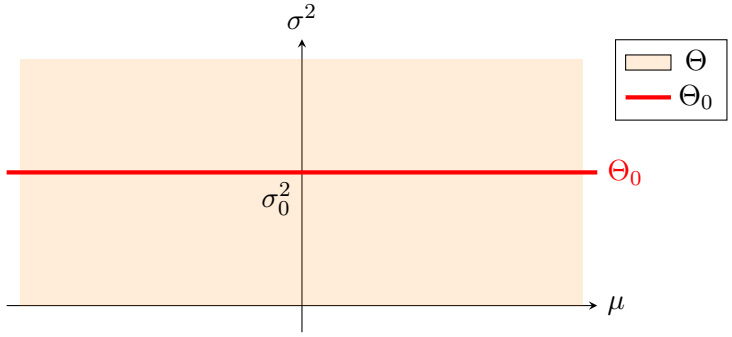


Figure 11.3: The illustrated graphs of the parameter spaces Θ and Θ_0.

The likelihood function is given by

$$L(\mu, \sigma^2|x) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right).$$

Next we find the maximum of $L(\mu, \sigma^2)$ on the set Θ_0. Since the set Θ_0 is equal to $\{(\mu, \sigma_0^2) \in \mathbb{R}^2 \,|\, \mu \in \mathbb{R}\}$, we have

$$\max_{(\mu,\sigma^2)\in\Theta_0} L(\mu, \sigma^2|x) = \max_{\mu\in\mathbb{R}} L(\mu, \sigma_0^2|x).$$

From here we can see that both $L(\mu, \sigma^2|x)$ and $\ln L(\mu, \sigma^2|x)$ are achieving at the same value of μ. We further determine the value of μ that maximizes $\ln L(\mu, \sigma_0^2|x)$. Taking the natural logarithm of the likelihood function and differentiating with respect to μ, we have

$$\begin{aligned}\frac{\mathrm{d}\ln L(\mu, \sigma_0^2|x)}{\mathrm{d}\mu} &= \frac{\mathrm{d}}{\mathrm{d}\mu}\left[-\frac{n}{2}\ln(\sigma_0^2) - \frac{n}{2}\ln(2\pi) - \frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(x_i-\mu)^2\right] \\ &= \frac{1}{\sigma_0^2}\sum_{i=1}^{n}(x_i-\mu) \qquad (\heartsuit)\end{aligned}$$

Setting $(\heartsuit)$ to zero and solving for μ, we have

$$\frac{1}{\sigma_0^2}\sum_{i=1}^{n}(x_i-\mu) = 0 \implies \widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}.$$

Hence, we continue to maximize $L(\bar{x}, \sigma_0^2|x)$ with respect to σ^2. Let $\sigma^2 = \varsigma$, we have

$$\begin{aligned}\frac{\mathrm{d}\ln L(\bar{x}, \varsigma|x)}{\mathrm{d}\varsigma} &= \frac{\mathrm{d}}{\mathrm{d}\varsigma}\left[-\frac{n}{2}\ln(\varsigma) - \frac{n}{2}\ln(2\pi) - \frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(x_i-\bar{x})^2\right] \\ &= -\frac{n}{2\varsigma} + \frac{1}{2(\varsigma)^2}\sum_{i=1}^{n}(x_i-\bar{x})^2 \qquad (\clubsuit)\end{aligned}$$

Setting the above equation (♣) to zero and solving for ς, we have

$$\begin{aligned}
&-\frac{n}{2\varsigma}+\frac{1}{2(\varsigma)^2}\sum_{i=1}^{n}(x_i-\bar{x})^2=0\\
&\implies \frac{1}{2(\varsigma)^2}\sum_{i=1}^{n}(x_i-\bar{x})^2=\frac{n}{2\varsigma}\\
&\implies \widehat{\varsigma}=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\\
&\implies \widehat{\varsigma}=\frac{n-1}{n}\cdot\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2=\frac{n-1}{n}s^2.
\end{aligned}$$

where $s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$ is the sample variance. Therefore, we obtain

$$\sup_{(\mu,\sigma^2)\in\Theta}L(\mu,\sigma^2|x)=L(\bar{x},\widehat{\varsigma}|x)=\left(\frac{n}{2\pi(n-1)s^2}\right)^{n/2}\exp\left[-\frac{n}{2(n-1)s^2}\sum_{i=1}^{n}(x_i-\bar{x})^2\right].$$

Thus using the optimal values we found, the likelihood ratio test statistic is

$$\begin{aligned}
\Lambda(x_1,x_2,\ldots,x_n)&=\frac{\sup\limits_{(\mu,\sigma^2)\in\Theta_0}L(\mu,\sigma^2|x)}{\sup\limits_{(\mu,\sigma^2)\in\Theta}L(\mu,\sigma^2|x)}\\
&=\frac{\left(\frac{n}{2\pi\sigma_0^2}\right)^{n/2}\exp\left[-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(x_i-\bar{x})^2\right]}{\left(\frac{n}{2\pi(n-1)s^2}\right)^{n/2}\exp\left[-\frac{n}{2(n-1)s^2}\sum_{i=1}^{n}(x_i-\bar{x})^2\right]}\\
&=n^{-n/2}e^{n/2}\left(\frac{(n-1)s^2}{\sigma_0^2}\right)^{n/2}\exp\left[-\frac{1}{2}\left(\frac{1}{\sigma_0^2}-\frac{n}{(n-1)s^2}\right)\sum_{i=1}^{n}(x_i-\bar{x})^2\right]\\
&=n^{-n/2}e^{n/2}\left(\frac{(n-1)s^2}{\sigma_0^2}\right)^{n/2}\exp\left[-\frac{(n-1)s^2}{2\sigma_0^2}\right]\leq k.
\end{aligned}$$

Now this inequality can be rearranged to

$$\left(\frac{(n-1)s^2}{\sigma_0^2}\right)^{n}\exp\left[-\frac{(n-1)s^2}{\sigma_0^2}\right]\leq\left[k\left(\frac{n^{n/2}}{e}\right)\right]^2:=K_0.$$

where K_0 is some constant. Now let H be a function defined by

$$H(w):=w^ne^{-w}\quad\text{ for }w>0.$$

With this notation, we see that the above inequality is equivalent to

$$H\left(\frac{(n-1)s^2}{\sigma_0^2}\right) \leq K_0.$$

From this it follows that

$$\frac{(n-1)s^2}{\sigma_0^2} \leq K_1 \quad \text{or} \quad \frac{(n-1)s^2}{\sigma_0^2} \geq K_2.$$

In view of these inequalities, the rejection region is of the form

$$RR = \left\{(x_1, x_2, \ldots, x_n) \,\middle|\, \frac{(n-1)s^2}{\sigma_0^2} \leq K_1 \quad \text{or} \quad \frac{(n-1)s^2}{\sigma_0^2} \geq K_2\right\} \tag{11.16}$$

and the best likelihood ratio test can be described as follows: "Reject H_0 if

$$\frac{(n-1)s^2}{\sigma_0^2} \leq K_1 \quad \text{or} \quad \frac{(n-1)s^2}{\sigma_0^2} \geq K_2." \tag{11.17}$$

Since we are given the size of the test α, the values of K_1 and K_2 can be determined. As the sample $X_1, X_2, \ldots, X_n$ is drawn from a normal population with mean μ and variance σ^2, so

$$\frac{(n-1)s^2}{\sigma_0^2} \sim \chi^2_{n-1}. \tag{11.18}$$

when the null hypothesis $H_0 : \sigma^2 = \sigma_0^2$ is true. Therefore, the likelihood ratio test of size α rejects H_0 if

$$RR = \left\{(x_1, x_2, \ldots, x_n) \,\middle|\, \frac{(n-1)s^2}{\sigma_0^2} \leq \chi^2_{\frac{\alpha}{2},n-1} \quad \text{or} \quad \frac{(n-1)s^2}{\sigma_0^2} \geq \chi^2_{1-\frac{\alpha}{2},n-1}\right\} \tag{11.19}$$

where $\chi^2_{\frac{\alpha}{2},n-1}$ and $\chi^2_{1-\frac{\alpha}{2},n-1}$ are the lower and upper $\frac{\alpha}{2}$ points of the chi-square distribution with $n-1$ degrees of freedom, respectively.

✱ Example 11.4.1. A random sample of 16 recorded deaths in the city of Urbana was compiled. The sample average is 71.8 years old and the sample standard deviation is 9 years. Assuming that life expectancy is normally distributed but with no known standard deviation, can we claim that the standard deviation is equal to 7 years? Or is it different than that? Use a 5% level of significance.

✲ Solution We want to test

$$H_0 : \sigma^2 = 49 \quad \text{versus} \quad H_1 : \sigma^2 \neq 49,$$

where σ^2 is the variance of life expectancy in Urbana. The sample size is $n = 16$, and the test statistic is

$$\chi_0^2 = \frac{(16-1) \times 9^2}{49} = \boxed{24.796}.$$

The corresponding critical region is

$$RR = \left\{\chi_0^2 \leq \chi^2_{0.025,15} = 27.488 \quad \text{or} \quad \chi_0^2 \geq \chi^2_{0.975,15} = 6.262\right\}.$$

Hence, we do not reject H_0 at the 5% level of significance, as

$$\chi^2_{0.975,15} \leq \chi_0^2 \leq \chi^2_{0.025,15}.$$

◀

Tutorials

Exercise 11.1 A firm obtains its supply of steel wire of a particular gauge from each of two manufacturers A and B. The firm suspects that the mean breaking strength, in newtons (N), of wire from manufacturer A differs from that supplied to manufacturer B.

The table below shows the breaking strengths of random samples of wire

A	80.5	83.1	73.6	70.4	68.9	71.6	82.3	78.6	73.4
B	71.4	86.2	81.4	72.3	78.9	80.3	81.4	78.0	

Assuming all such breaking strengths to be normally distributed with a standard deviation of 5N. Test, at the 5% significance level, the firm's suspicion.

Exercise 11.2 A microbiologist wishes to determine whether there is any difference in the time it takes to make yoghurt from two different starters; lactobacillus acidophilus (A) and bulgarius (B). Seven batches of yoghurt were made with each of the starters. The table below shows the time taken, in hours, to make each batch.

Starter A	6.8	6.3	7.4	6.1	8.2	7.3	6.9
Starter B	6.1	6.4	5.7	5.5	6.9	6.3	6.7

Assuming that both sets of times may be considered to be random samples from normal populations with the same variance, test the hypothesis that the mean time taken to make yoghurt is the same for both starters.

Exercise 11.3 A new chemical process is developed for the manufacture of nickel-cadmium batteries. The company believes that this new process will increase the mean lifetime of a battery by 5 hours as compared to that of batteries produced by the old process. Sixteen batteries produced by the old process were randomly selected and the mean and the standard deviation of the lifetimes of these batteries were 105.2 hours and 9.1 hours, respectively. Fifteen batteries produced by the

new process were also randomly selected and calculations gave corresponding values of 112.4 and 8.3 hours.

Assuming all battery lifetimes to be normally distributed, test at the 5% significance level whether there is

1. a difference in the variability of the two processes,
2. an increase of 5 hours in the mean lifetime of batteries produced by the new process as compared to that of batteries produced by the old process.

Exercise 11.4 What is the difference between simple and composite hypothesis?

Exercise 11.5 If an observation X is drawn from a population with probability mass function

$$f_X(x|\theta) = \begin{cases} \frac{2x}{\theta^2} & \text{for } 0 \leq x < \theta \\ 0 & \text{otherwise.} \end{cases}$$

Use Neyman-Pearson lemma to find the most powerful test for testing

$$H_0 : \theta = 4 \quad \text{versus} \quad H_1 : \theta = 5.$$

Hence, find the power of the test.

Exercise 11.6 Let X be a random variable whose pdf under H_0 and H_1 are given by

x	1	2	3	4	5	6	7
$f(x\|H_0)$	.01	.01	.01	.01	.01	.01	.94
$f(x\|H_1)$	.06	.05	.04	.03	.02	.01	.79

Use the Neyman-Pearson lemma to find the most powerful test for testing H_0 against H_1 at level of significance $\alpha = 0.04$. Hence, compute the probability of type II error for this test.

Exercise 11.7 Let X be a random sample from Bernoulli distribution with probability of success θ. It is proposed to test

$$H_0 : \theta = 0.5 \quad \text{against} \quad H_1 : \theta = 0.3$$

based on sample of size 5.

1. Show that the rejection region for the test is

$$RR = \left\{ \sum_{i=1}^{5} X_i > 3 \right\}.$$

2. Find the probabilities of type I and type II errors, as well as the power of test.

Exercise 11.8 Construct the UMP test for testing

$$H_0 : \lambda = 2 \quad \text{against} \quad H_1 : \lambda > 3$$

when observations X of size n are drawn from population with density function

$$f_X(x|\lambda) = \begin{cases} \dfrac{\lambda^x e^{-\lambda}}{x!} & \text{for } x = 0, 1, 2 \dots \\ 0 & \text{otherwise.} \end{cases}$$

Considering level of significance $\alpha = 0.02$.

Exercise 11.9 Find the likelihood ratio test for testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ when a random sample is drawn from $N(\mu, 625)$.

Exercise 11.10 The daily output of 9 randomly selected operators was recorded before and after a two-week training programme:

Operator	A	B	C	D	E	F	G	H	I
Before	52	72	58	55	67	63	56	69	57
After	50	90	62	56	80	72	58	84	60

Assuming the change in the daily output follows a normal distribution, examine the hypothesis that the two-week training programme results in a significant increase in the mean daily output of the operators at the 5% significance level.

This page intentionally left blank.

12

Analysis of Variance

The main idea of an ANOVA test is that if researchers ask if a set of sample means gives evidence of differences in the population means, what matters is not how far apart the sample means are, but how far apart they are *relative to the variability of individual observations*.

12.1 One-way ANOVA

Lemma 12.1
The total sum of squares is equal to the *sum of within* and *between sum of squares*, that is

$$SS_T = SS_W + SS_B. \tag{12.1}$$

Proof. Rewriting we have

$$\begin{aligned} SS_T &= \sum_{i=1}^{m}\sum_{j=1}^{n}(Y_{ij} - \overline{Y}_{\bullet\bullet}) \\ &= \sum_{i=1}^{m}\sum_{j=1}^{n}\left[(Y_{ij} - \overline{Y}_{i\bullet}) + (Y_{i\bullet} - \overline{Y}_{\bullet\bullet})\right]^2 \\ &= \sum_{i=1}^{m}\sum_{j=1}^{n}(Y_{ij} - \overline{Y}_{i\bullet})^2 + \sum_{i=1}^{m}\sum_{j=1}^{n}(Y_{i\bullet} - \overline{Y}_{\bullet\bullet})^2 + 2\sum_{i=1}^{m}\sum_{j=1}^{n}(Y_{ij} - \overline{Y}_{i\bullet})(Y_{i\bullet} - \overline{Y}_{\bullet\bullet}) \end{aligned}$$

Hence we obtain the asserted result

$$SS_T = SS_W + SS_B$$

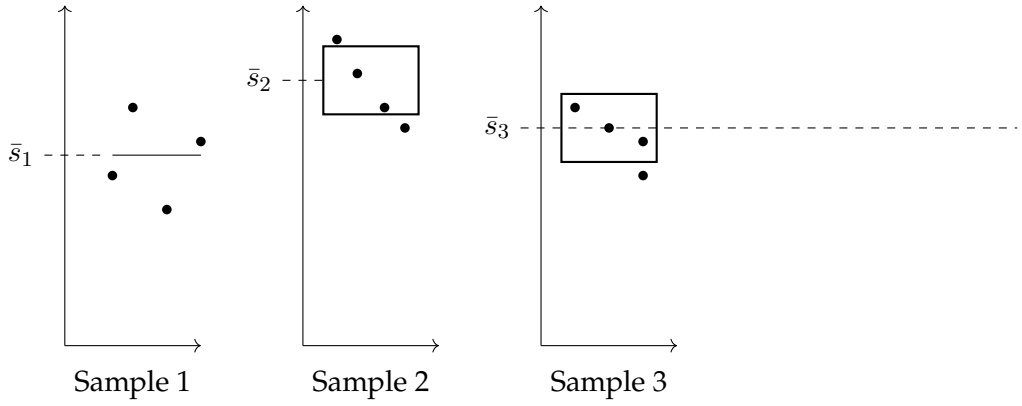
and the proof of the lemma is complete. □

Theorem 12.1
Suppose the one-way ANOVA model is given by the equation where the ε_{ij}'s are independent and normally distributed random variables with mean zero and variance σ^2 for $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$.

Case 1

In this case the variation within samples is roughly on a par with that occurring between samples.

Figure 1



Case 2

In this case the variation within samples is considerably less than that occurring between samples.

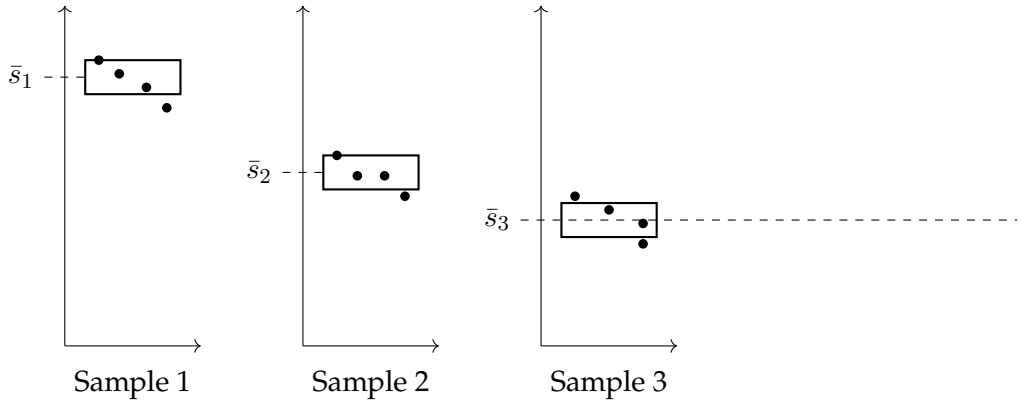


Figure 12.1: Comparison of within-sample and between-sample variation in ANOVA

The null hypothesis

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_m = \mu$$

is rejected whenever the statistics $\mathcal{F}$ satisfies

$$\mathcal{F} = \frac{SS_B/(m-1)}{SS_W/[m(n-1)]} > F_\alpha(m-1, m(n-1)). \tag{12.2}$$

where α is the significance level of the hypothesis test and $F_\alpha(m-1, m(n-1))$ denotes the $100(1-\alpha)$-th percentile of the F-distribution with $m-1$ numerator and $m(n-1)$ denominator degrees of freedom.

Proof. Under the null hypothesis $H_0 : \mu_1 = \mu_2 = \cdots = \mu_m = \mu$, the likelihood function takes the form

$$\begin{aligned} L(\mu, \sigma^2|Y) &= \prod_{i=1}^{m}\prod_{j=1}^{n}\left\{\frac{1}{\sqrt{2\sigma^2}}\exp\left[-\frac{(Y_{ij}-\mu)^2}{2\sigma^2}\right]\right\} \\ &= \left(\frac{1}{\sqrt{2\sigma^2}}\right)^{nm}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{m}\sum_{j=1}^{n}(Y_{ij}-\mu)^2\right] \end{aligned} \tag{♡}$$

Maximizing the natural logarithm of the likelihood function (♡), we obtain

$$\widehat{\mu} = \overline{Y}_{\bullet\bullet} \quad \text{and} \quad \widehat{\sigma_{H_0}} = \frac{1}{mn}SS_T$$

as the maximum likelihood estimators of μ and σ^2, respectively. Plugging these estimators back into (♡), we have the maximum likelihood function, that is,

$$\max L(\mu, \sigma^2|Y) = \left(\frac{1}{\sqrt{2\widehat{\sigma^2_{H_0}}}}\right)^{nm}\exp\left[-\frac{1}{2\widehat{\sigma^2_{H_0}}}\sum_{i=1}^{m}\sum_{j=1}^{n}(Y_{ij}-\overline{Y}_{\bullet\bullet})^2\right].$$

Simplifying the above expression, we see that

$$\begin{aligned} \max L(\mu, \sigma^2|Y) &= \left(\frac{1}{\sqrt{2\widehat{\sigma^2_{H_0}}}}\right)^{nm}\exp\left[-\left(\frac{2}{nm}SS_T\right)^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n}(Y_{ij}-\overline{Y}_{\bullet\bullet})^2\right] \\ &= \left(\frac{1}{\sqrt{2\widehat{\sigma^2_{H_0}}}}\right)^{nm}\exp\left[-\frac{nm}{2\,SS_T}SS_T\right] \\ &= \left(\frac{1}{\sqrt{2\widehat{\sigma^2_{H_0}}}}\right)^{nm}e^{-\frac{nm}{2}} \end{aligned} \tag{♣}$$

When no restrictions imposed, we obtain the maximum of the likelihood function from lemma 12.1 as

$$\max L(\mu_1, \mu_2, \ldots, \mu_m, \sigma^2|Y) = \left(\frac{1}{\sqrt{2\widehat{\sigma^2_{H_0}}}}\right)^{nm}\exp\left[-\frac{1}{2\widehat{\sigma^2_{H_0}}}\sum_{i=1}^{m}\sum_{j=1}^{n}(Y_{ij}-\overline{Y}_{i\bullet})^2\right] \tag{★}$$

notice that the grand mean $\overline{Y}_{\bullet\bullet}$ now replace by $\overline{Y}_{i\bullet}$, and $\widehat{\sigma^2_{H_0}}$ being replace by $\widehat{\sigma^2}$. Again simplying the

expression above and

$$\begin{aligned}\max L(\mu_1, \mu_2, \ldots, \mu_m, \sigma^2|Y) &= \left(\frac{1}{\sqrt{2\widehat{\sigma^2}}}\right)^{nm} \exp\left[-\left(\frac{2}{nm}SS_W\right)^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n}(Y_{ij}-\overline{Y}_{i\bullet})^2\right] \\ &= \left(\frac{1}{\sqrt{2\widehat{\sigma^2}}}\right)^{nm} \exp\left[-\frac{nm}{2\,SS_W}SS_W\right] \\ &= \left(\frac{1}{\sqrt{2\widehat{\sigma^2}}}\right)^{nm} e^{-\frac{nm}{2}}.\end{aligned}$$

Next, we are going to find the likelihood ratio statistic Λ for testing the null hypothesis H_0. Recall that the likelihood ratio statistic Λ can be found by evaluating

$$\Lambda = \frac{\max L(\mu, \sigma^2)}{\max L(\mu_1, \mu_2, \ldots, \mu_m, \sigma^2)}.$$

Using (♣) divide (★), we have

$$\Lambda = \left(\frac{\widehat{\sigma^2}}{\widehat{\sigma^2_{H_0}}}\right)^{\frac{nm}{2}}.$$

Recall that the likelihood ratio test to reject the null hypothesis is

$$\Lambda < k_0 \implies \left(\frac{\widehat{\sigma^2}}{\widehat{\sigma^2_{H_0}}}\right)^{\frac{nm}{2}} < k_0 \implies \frac{\widehat{\sigma^2_{H_0}}}{\widehat{\sigma^2}} > \left(\frac{1}{k_0}\right)^{\frac{2}{nm}}$$

Applying lemma 12.1,

$$\frac{SS_W + SS_B}{SS_W} > \left(\frac{1}{k_0}\right)^{\frac{2}{nm}} \implies \frac{SS_B}{SS_W} > k' \qquad (♠)$$

where $k' := \left(\frac{1}{k_0}\right)^{\frac{2}{nm}} - 1$. In order to find the cutoff point k' in (♠). We apply lemma 12.1. Thus

$$\mathcal{F} = \frac{SS_B/(m-1)}{SS_W/(m(n-1))} > \frac{m(n-1)}{m-1}k'.$$

□

Source of Variation	Sums of squares	Degree of freedom	Mean squares	$\mathcal{F}$-statistic
Between	SS_B	$m-1$	$MS_B = \frac{SS_B}{m-1}$	$\mathcal{F} = \frac{MS_B}{MS_W}$
Within	SS_W	$N-m$	$MS_W = \frac{SS_W}{N-m}$	
Total	SS_T	$N-1$		

Table 12.1: One-way ANOVA table with unequal sample size

12.2 Test for the Homogeneity of Variances

One of the assumptions behind the ANOVA test is that the variances of each samples under consideration should be the same for all population.

The test statistic B_c is given by

$$B_c = \frac{(N-m)\ln S_p^2 - \sum_{i=1}^{m}(n_i-1)\ln s_i^2}{1+\frac{1}{3(m-1)}\left[\sum_{i=1}^{m}\frac{1}{n_i-1} - \frac{1}{N-m}\right]} \tag{12.3}$$

In the formula above,

- s_i^2 is the sample variance of the i-th group.
- n_i is the sample size of i-th group.
- $N = \sum n_i$ is the total sample size.
- m is the number of groups.

and the pooled variance S_p^2 is given by

$$S_p^2 = \frac{\sum_{i=1}^{m}(n_i-1)s_i^2}{N-m} = MS_W. \tag{12.4}$$

The sampling distribution of B_c is approximately chi-square with $m-1$ degrees of freedom, that is,

$$B_c \sim \chi^2(m-1)$$

when $(n_i - 1) \geq 3$. Therefore the Barlett test rejects the null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_m^2$ at a significance level α if

$$B_c \sim \chi^2_{1-\alpha}(m-1)$$

where $\chi^2_{1-\alpha}(m-1)$ denotes the upper $(1-\alpha)\times 100$ percentile of the chi-square with $m-1$ degrees of freedom.

Definition 12.1 Barlett's test
The test statistic for Barlett's test is

$$B_c = \frac{(N-m)\ln S_p^2 - \sum_{i=1}^{m}(n_i-1)\ln S_i^2}{1+\frac{1}{3(m-1)}\left[\sum_{i=1}^{m}\frac{1}{n_i-1}-\frac{1}{N-m}\right]}. \tag{12.5}$$

Reject $H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_m^2$ if $B_c > \chi_\alpha^2(m-1)$.

Barlett's test is the uniformly most powerful (UMP) test for the homogeneity of variances problem under the assumption that each treatment population is normally distributed. Bartlett's test is useful whenever the assumption of equal variances is made. In particular, this assumption is made for the frequently used one-way analysis of variance.

However, Barlett's test has crucial weaknesses if the normality assumption is not met:

- The tests reliability is sensitive (not robust) to non-normality.
- If the treatment populations are not approximately normal, the true significance level can be very different from the nominal significance level (say, $\alpha = 0.05$). This difference depends on the kurtosis (4th moment) of the distribution.

In this case, Bartlett's or Levene's test should be applied to verify the assumption.

12.2.1 Levene's test

The Bartlett's test assumes that the grouped samples should be taken from normal populations. Thus Bartlett test is sensitive to departures from normality. The Levene's test is an alternative to the Bartlett's test that is less sensitive to departures from normality. Levene (1960) proposed a test for the homogeneity of population variances that considers the random variables

Definition 12.2 Levene's Test
To perform Levene's Test:

1. Calculate each $Z_{ij} = |Y_{ij} - \overline{Y}_{i\bullet}|$.
2. Run an ANOVA on the set of Z_{ij} values.
3. If $\mathcal{F} > F_\alpha(m-1, N-m)$, reject null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_m^2$ and conclude that the variances are not all equal.

Brown and Forsythe (1974) proposed using the transformed variables based on the absolute deviations from the median,that is

$$Z_{ij} = |Y_{ij} - \widetilde{Y}_{i\bullet}|, \tag{12.6}$$

where $\widetilde{Y}_{i\bullet}$ denotes the median of i-th group. Again if the F-test is significant, the homogeneity of variances is rejected.

Definition 12.3 Brown-Forsythe Test
To perform Brown-Forsythe's Test:

1. Calculate each $Z_{ij} = |Y_{ij} - \widetilde{Y}_{i\bullet}|$, where $\widetilde{Y}_{i\bullet}$ is the i-th median of the treatment.
2. Run an ANOVA on the set of Z_{ij} values.
3. If $\mathcal{F} > F_\alpha(m-1, N-m)$, reject null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_m^2$ and conclude that the variances are not all equal.

✱ Example 12.2.1. For the following data set contains 10 measurements of gear diameter (in centimeter) for five different batches for a total of 50 measurements.

Batches				
1	2	3	4	5
1.006	0.998	0.991	1.005	0.998
0.996	1.006	0.987	1.002	0.998
0.998	1.000	0.997	0.994	0.982
1.000	1.002	0.999	1.000	0.990
0.992	0.997	0.995	0.995	1.002
0.993	0.998	0.994	0.994	0.984
1.002	0.996	1.000	0.998	0.996
0.999	1.000	0.999	0.996	0.993
0.994	1.006	0.996	1.002	0.980
1.000	0.988	0.996	0.996	1.018

The Cochran's C test statistic is popular for test equivalence of variances among multiple sample groups. It is a one-sided upper limit variance outlier test, and it is simple to apply with the following assumptions:

- The data in each group are normally distributed.
- The data set is balanced design, i.e., each group has the same sample size.

The Cochran's C-test has been used as an alternative to Bartlett, Levene and Brown Forsythe's tests in the evaluation of homoscedasticity (same variance) such as in a linear regression model. Note that this Cochran's C test is totally different with the Cochran's Q test, which the last one is being

used in the analysis of two-way randomized block designs with different treatments in a design of experiments.

$$C = \frac{\max\limits_{1 \le i \le m} s_i^2}{\sum\limits_{i=1}^{m} s_i^2}. \tag{12.7}$$

The test statistic is the ratio of the maximum variance among the data set, and the sum of all the variances.

Tutorials

Exercise 12.1 Given 20 observations on breakdown voltage for some materials

24.46	25.61	26.25	26.42	26.66	27.15	27.31	27.54	27.74	27.94
27.98	28.04	28.28	28.49	28.50	28.87	29.11	29.13	29.50	30.88

This page intentionally left blank.

13

Simple Linear Regression

The regression model is to explain variability in independent variable by means of one or more of independent pr control variables.

A **simple regression model** can be expressed as

$$\text{Value of dependent variable} = y\text{-intercept} + (\text{Slope} \times \text{Value of Indep. variable}) + \text{Error}$$
$$\mathbf{y} = \beta_0 + \beta\mathbf{x} + \varepsilon.$$

where $\epsilon \sim N(0, \sigma^2)$ is the error term. The sufficient statistics for the parameters β_0, β_1, and σ^2 can be derived from the likelihood function.

Now consider that we observe five pairs of x and y observations as follow: $(-2, 0)$, $(-1, 0)$, $(0, 1)$, $(1, 1)$ and $(2, 3)$.

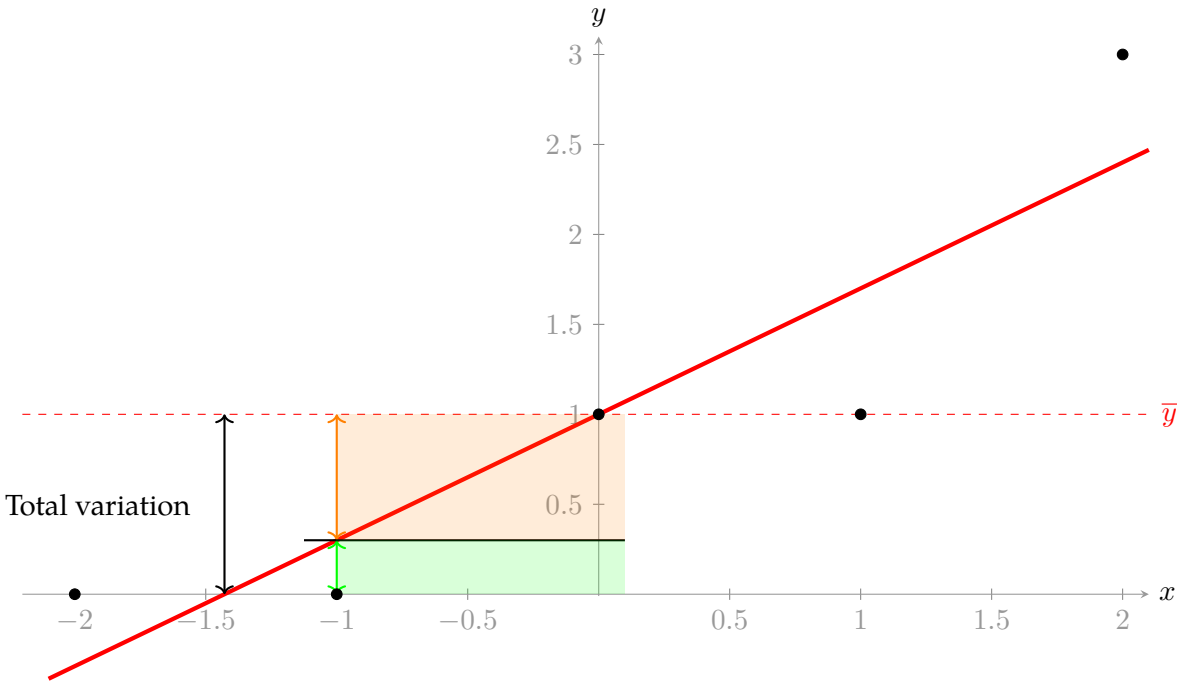


Figure 13.1: The illustrated graphs of the parameter spaces Θ and Θ_0.

13.1 Least squares Method

Let $\mathcal{X} \times \mathcal{Y} = \{(x_i, y_i) \mid i = 1, 2, \ldots, n\}$ be a set of data with sample size n. Consider we have a linear regression

$$\mathop{\mathbb{E}}_{x}[\, Y_i | x_i \,] = \beta_0 + \beta_1 x_i, \tag{13.1}$$

that is a simple straight line

$$y_i = \beta_0 + \beta_1 x_i, \quad i = 1, 2, \ldots, n. \tag{13.2}$$

Take the distance of each data points to the straight line and summing up. The sum of the squares of the error is given by

$$\mathcal{E}(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2. \tag{13.3}$$

The least squares estimates of the fitting parameters β_0 and β_1 are said to be those values which minimize the sum of squares error. That is,

$$\underline{\beta} = (\widehat{\beta}_0, \widehat{\beta}_1) = \mathop{\arg\min}_{(\beta_0, \beta_1) \in \mathbb{R}^2} \mathcal{E}(\beta_0, \beta_1). \tag{13.4}$$

✱ Example 13.1.1. The article "Relating the Cetane Number of Biodiesel Fuels to Their Fatty Acid Composition: A Critical Study" (J. of Automobile Engr., 2009: 565-583) included the following data on x = iodine value (g) and y = cetane number for a sample of 14 biofuels (see next slide). The iodine value (x) is the amount of iodine necessary to saturate a sample of 100 g of oil.

x	132.0	129.0	120.0	113.2	105.0	92.0	84.0	83.2	88.4	59.0	80.0	81.5	71.0	69.2
y	46.0	48.0	51.0	52.1	54.0	52.0	59.0	58.7	61.6	64.0	61.4	54.6	58.8	58.0

Under the additional assumption that ε_i's are iid $N(0, \sigma^2)$, the likelihood function of the sample is

$$\begin{aligned} L(\alpha, \beta, \sigma^2 | \mathbf{y}, \mathbf{x}) = \prod_{i=1}^{n} f_Y(y_i | x_i, \alpha, \beta, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \alpha - \beta x_i)^2}{2\sigma^2} \right) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 \right). \end{aligned} \tag{♠}$$

Considering $(y_1, x_1), \ldots, (y_n, x_n)$ as n pairs of data points plotted in the xy-plane as the scatterplot in the previous figure.

Think of drawing through this cloud of points a straight line that comes "as close as possible" to all the points, measured by the vertical distances from the points to the straight line. For any line $y = \alpha + \beta x$, the vertical distance from the point (x_i, y_i) to the line is $y_i - (\alpha + \beta x_i)$. Letting

$$\psi(\alpha, \beta) := \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2,$$

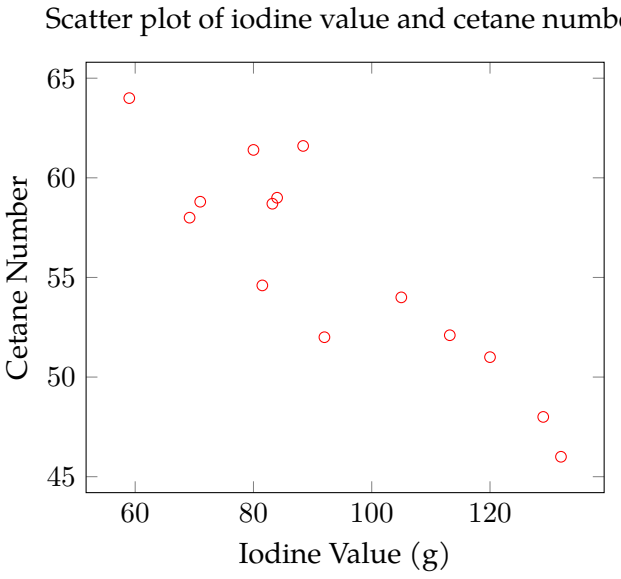


Figure 13.2: The scatter plot of iodine value (x) and cetane number (y).

Maximizing the likelihood is equivalent to

$$\min_{\alpha,\beta} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2,$$

Taking the partial derivatives of $\psi(\alpha, \beta)$ with respect to α and β, we have

$$\frac{\partial \psi(\alpha,\beta)}{\partial \alpha} = -2 \sum_{i=1}^{n} (y_i - \alpha - \beta x_i) = 0,$$

$$\frac{\partial \psi(\alpha,\beta)}{\partial \beta} = -2 \sum_{i=1}^{n} x_i (y_i - \alpha - \beta x_i) = 0. \qquad (\clubsuit)$$

The first equation gives

$$\overline{y} - a - b\overline{x} \implies a = \overline{y} - b\overline{x}.$$

Substituting a into the second equation ($\clubsuit$) by $\overline{y} - b\overline{x}$ results in

$$\sum_{i=1}^{n} x_i (y_i - \overline{y}) + b \sum_{i=1}^{n} x_i (\overline{x} - x_i) = 0.$$

This equation is the same as $S_{xy} = bS_{xx}$, that is

$$S_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}), \quad S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2.$$

Therefore, replacing y_i by the random variable Y_i for all $i = 1, 2, \ldots, n$ (and we still use S_{xy} when y_i is replaced by Y_i), we obtain the MLE or LSE as

$$\widehat{\beta} = \frac{S_{xy}}{S_{xx}}, \quad \widehat{\alpha} = \overline{Y} - \widehat{\beta}\overline{x} = \overline{Y} - \frac{S_{xy}}{S_{xx}}\overline{x}. \tag{13.5}$$

We can always assume that $S_{xx} > 0$, since $S_{xx} = 0$ is the trivial case of identical x_i's.

We now proceed to show that $\widehat{\alpha}$ and $\widehat{\beta}$ are UMVUE of α and β, respectively. First of all, we had already shown that they are unbiased estimators of α and β, respectively.

$$\mathbb{E}[S_{xy}] = \sum_{i=1}^{n}(x_i - \overline{x})\mathbb{E}_y(y_i - \overline{y}) = \sum_{i=1}^{n}(x_i - \overline{x})\beta(x_i - \overline{x}) = \beta S_{xx}.$$

Since $\widehat{\beta}$ is unbiased for β and

$$\mathbb{E}[\widehat{\alpha}] = \mathbb{E}[\overline{Y}] - \overline{x}\mathbb{E}[\widehat{\beta}] = \alpha + \beta\overline{x} - \overline{x}\beta = \alpha.$$

Continue with (♠), the likelihood function of the sample is

$$\begin{aligned} L(\alpha, \beta, \sigma^2|\mathbf{y}, \mathbf{x}) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2\right) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(y_i - \overline{y})^2 + n(\overline{y} - \alpha - \beta\overline{x})^2 + (\beta - \widehat{\beta})^2 S_{xx}\right]\right). \end{aligned}$$

where

$$S_{yy} = \sum_{i=1}^{n}(y_i - \overline{y})^2.$$

We still using the notation S_{xy} and S_{xx} when y_i is replaced by Y_i. From the properties of the exponential family, a complete and sufficient statistic for $\underline{\theta} = (\alpha, \beta, \sigma^2)$ is given by $(\widehat{\alpha}, \widehat{\beta}, S_{yy})$.

Since $\widehat{\alpha}$ and $\widehat{\beta}$ are both unbiased estimators and functions of the sufficient and complete statistic, thus they are UMVUE of α and β, respectively.

✱ Example 13.1.2 ((cont.)). **Question: What if we remove the normality assumption?** $\widehat{\alpha}$ and $\widehat{\beta}$ are still least squares estimators, but they are no longer MLEs. A statistical property that holds for LSE is that it is the best linear unbiased estimator (BLUE) in the sense that $\widehat{\alpha}$ or $\widehat{\beta}$ has the smallest variance among all linear unbiased estimators of the form

$$\sum_{i=1}^{n} d_i Y_i \tag{13.6}$$

If the estimator of this form is unbiased for β, then

$$\begin{aligned}\beta = \mathbb{E}_y\left(\sum_{i=1}^n d_i Y_i\right) &= \sum_{i=1}^n d_i \mathbb{E}_y[Y_i] \\ &= \sum_{i=1}^n d_i(\alpha + \beta x_i) && \text{from the regression line, } \widehat{Y} = \alpha + \beta x \\ &= \alpha \sum_{i=1}^n d_i + \beta \sum_{i=1}^n d_i x_i && \text{linearity of summation.}\end{aligned}$$

holds for all α and β, which imples that

$$\sum_{i=1}^n d_i = 0, \quad \sum_{i=1}^n d_i x_i = 1.$$

A geometric description of the BLUE of β is given in the filgure below.

We want to show the LSE $\widehat{\beta}$ is BLUE. Since

$$Var\left(\sum_{i=1}^n d_i Y_i\right) = \sigma^2 \sum_{i=1}^n d_i^2,$$

the BLUE of β is the solution of the optimization problem

$$\begin{aligned}\min_{d_i} \quad & \sum_{i=1}^n d_i^2 && (13.7a)\\ \text{subject to} \quad & \sum_{i=1}^n d_i = 0, && (13.7b)\\ & \sum_{i=1}^n d_i x_i = 1. && (13.7c)\end{aligned}$$

Consider the Lagrange multiplier method by minimizing

$$g(d_1, \ldots, d_n, \lambda_1, \lambda_2) = \sum_{i=1}^n d_i^2 + \lambda_1\left(\sum_{i=1}^n d_i\right) + \lambda_2\left(\sum_{i=1}^n d_i x_i - 1\right). \tag{13.8}$$

Taking derivatives with respect to d_i and setting them to zero, we have

$$\frac{\partial g}{\partial d_i} = 2d_i + \lambda_1 + \lambda_2 x_i = 0.$$

Then

$$0 = \sum_{i=1}^{n}(2d_i + \lambda_1 + \lambda_2 x_i) = 2\sum_{i=1}^{n} d_i + n\lambda_1 + \lambda_2 \sum_{i=1}^{n} x_i \qquad \text{linearity of summation}$$

$$= \lambda_1 n + \lambda_2 \sum_{i=1}^{n} x_i \qquad \text{since } \sum d_i = 0$$

which yields $\lambda_1 = -\lambda_2 \overline{x}$ and, hence

$$0 = 2d_i + \lambda_2(x_i - \overline{x}).$$

Then

$$0 = \sum_{i=1}^{n}(x_i - \overline{x})[2d_i + \lambda_2(x - \overline{x})] = 2 + \lambda_2 S_{xx}$$

which gives us $\lambda_2 = -2/S_{xx}$. Then

$$d_i = \frac{-(\lambda_1 + \lambda_2 x_i)}{2} = \frac{-\lambda_2(x_i - \overline{x})}{2} = \frac{(x_i - \overline{x})}{S_{xx}}$$

and the BLUE of β is

$$\sum_{i=1}^{n} d_i Y_i = \sum_{i=1}^{n} \frac{(x_i - \overline{x})}{S_{xx}} Y_i = \frac{S_{xy}}{S_{xx}} = \widehat{\beta}.$$

And because

$$\widehat{\beta} = \sum_{i=1}^{n} \frac{(x_i - \overline{x})(\beta x_i + \varepsilon_i)}{S_{xx}} = \beta + \sum_{i=1}^{n} d_i \varepsilon_i$$

where $d_i = (x_i - \overline{x})/S_{xx}$, and the random error $\varepsilon \sim N(0, 1)$. We obtain that

$$Var(\widehat{\beta}) = \sum_{i=1}^{n} d_i^2 Var(\varepsilon_i) = \frac{\sigma^2}{S_{xx}}.$$

And we are done.

✱ Example 13.1.3. Assume monthly data of sales and advertising costs for a company below and independence between monthly sales. What is the relationship between sales and advertising costs for a company?

Advertising costs (in $\$100,000$)	1	2	3	4	5
Sales (in $10,000$ unit)	1	1	2	2	4

1. Find the estimated parameters $\widehat{\beta}_0$ and $\widehat{\beta}_1$. What do these estimated parameters mean?
2. What are the estimated sales when the advertising cost is $\$100,000$ and $\$250,000$, respectively?

✲ Solution 1. Using LSE to find unbiased estimators of the parameters:

	x	y	x^2	y^2	xy
	1	1	1	1	1
	2	1	4	1	2
	3	2	9	4	6
	4	2	16	4	8
	5	4	25	16	20
$\sum =$	15	10	55	26	37

Now we can applying the formulas,

$$\widehat{\beta}_1 = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2} = \frac{5(37) - 15(10)}{5(55) - 15^2} = 0.70.$$

and

$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X} = \frac{\sum y}{n} - \widehat{\beta}_0 \left(\frac{\sum x}{n}\right) = \frac{10}{5} - 0.70 \times \frac{15}{5} = -0.10.$$

The fitted regression line is $\widehat{y} = -0.1 + 0.7x$. We take a look on each estimated parameters:

- $\widehat{\beta}_0 = -0.1$, since there is no advertising cost, the sales is -1000 units, which is meaningless.
- $\widehat{\beta}_1 = 0.7$, as the advertising cost increased by $\$100,000$, the sales is increased by 7000 units.

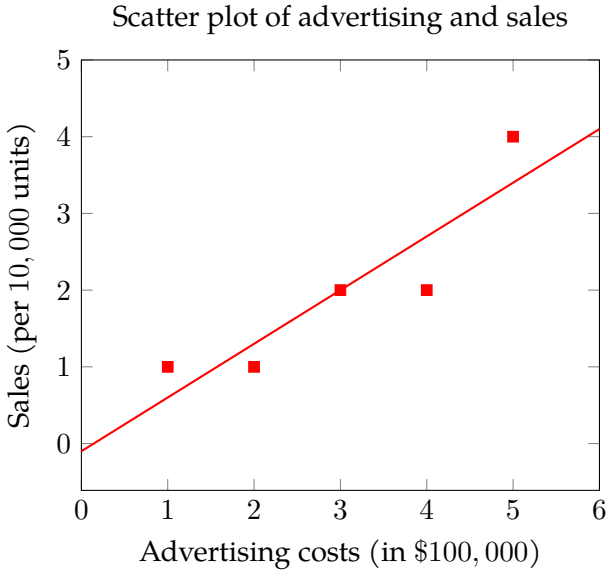


Figure 13.3: The scatter plot of advertising costs (x) and sales (y). The scatterplot shows a positive linear relationship.

2. If advertising cost is $\$100,000$, that means $x = 1$. The estimated sales will be $\widehat{y}_1 = -0.1 + 0.7(1) = 0.6$. Which is 6,000 units.

If advertising cost is $\$250,000$, that means $x = 2.5$. The estimated sales will be $\widehat{y}_{2.5} = -0.1 + 0.7(2.5) = 1.65$. Which is 16,500 units.

◀

Definition 13.1 Extrapolation
Extrapolation is using the regression line to predict the value of a response corresponding to a x values that is **outside the range** of the data used to determine the regression line. Extrapolation may lead to **unreliable predictions**.

13.2 Making Inference

Lemma 13.1
Given the simple linear regression $Y_i = \beta_0 + \beta_1 x_i$, and each x_i to be a set of arbitrary but fixed real numbers, then

1. $SS_E = SS_T - \widehat{\beta}_1 S_{xY}$.
2. $\text{Cov}(X, Y) = S_{xx}\widehat{\beta}_1$.
3. $\text{Cov}(\widehat{\beta}_0, \widehat{\beta}_1) = -\dfrac{\overline{x}\sigma^2}{S_{xx}}$.

Proof. 1.

$$\begin{aligned} SS_T = SS_E + SS_R &\implies SS_E = SS_T - SS_R \\ &\implies SS_E = SS_T - SS_R \end{aligned}$$

□

13.2.1 Estimate σ^2

The error sum of squares (equivalently, *residual sum of squares*), denoted by SS_E, is

$$SS_E = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n} \left[y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i) \right]^2 \tag{13.9}$$

and the estimate of σ^2 is

$$\widehat{\sigma}^2 = s^2 = \frac{SS_E}{n-2} = \frac{1}{n-2} \sum_{i=1}^{n} (y - \widehat{y}_i)^2 = \frac{1}{n-2} \sum_{i=1}^{n} \widehat{\epsilon}_i^2. \tag{13.10}$$

The divisor $n-2$ in S^2 is the number of degrees of freedom associated with SS_E and the estimate s^2. This is because to obtain the sample variance s^2, the two parameters β_0 and β_1 must first be estimated, which results in a loss of 2 degree of freedom, just as μ had to be estimated in one sample problems, resulting in an estimated variance based on $n-1$ degree of freedom in our previous t-tests assumption. Indeed, S^2 is an unbiased estimator for σ^2.

Theorem 13.1 Unbiased Estimator for σ^2
Given the simple linear regression $Y_i = \beta_0 + \beta_1 x_i$, then

$$MS_E = S^2 = \frac{1}{n-2} SS_E = \frac{1}{n-2} \sum_{i=1}^{n} \left(Y_i - \widehat{Y}_i \right)^2 \tag{13.11}$$

is an unbiased estimator for population variance σ^2.

Proof. To show that S^2 is an unbiased estimator for σ^2, we want verify that

$$\mathbb{E}[S^2] = \mathbb{E}\left[\frac{SS_E}{n-2} \right] = \frac{1}{n-2} \mathbb{E}[SS_E] \stackrel{?}{=} \sigma^2,$$

it is important to find $\mathbb{E}[SS_E]$ in order to verify the identity above.

Notice that

$$\begin{aligned}
\mathbb{E}[SS_E] &= \mathbb{E}\left[\sum_{i=1}^{n} \left(Y_i - \widehat{Y}_i \right)^2 \right] \\
&= \mathbb{E}\left[\sum_{i=1}^{n} (Y_i - \overline{Y} + \widehat{\beta}_1 \overline{x} - \widehat{\beta}_1 x_i)^2 \right] \\
&= \mathbb{E}\left[\sum_{i=1}^{n} \left((Y_i - \overline{Y}) + \widehat{\beta}_1 (\overline{x} - x_i) \right)^2 \right]
\end{aligned}$$

continue to expand the squares and we have

$$\mathbb{E}[SS_E] = \mathbb{E}\left[\sum (Y_i - \overline{Y})^2 + \widehat{\beta}_1^2 \sum (x_i - \overline{x})^2 - 2\widehat{\beta}_1 \sum (x_i - \overline{x})(Y_i - \overline{Y}) \right]. \tag{♣}$$

Since $\sum (x_i - \overline{x})(Y_i - \overline{Y}) = \widehat{\beta}_1 S_{xx}$, and also $\sum (Y_i - \overline{Y})^2 = \sum Y_i^2 - n\overline{Y}$. Combining these two results and replace the first and last term of (♣). We have

$$\begin{aligned}
\mathbb{E}[SS_E] &= \mathbb{E}\left[\sum Y_i^2 - n\overline{Y} + \widehat{\beta}_1^2 S_{xx} - 2\widehat{\beta}_1 \left(\widehat{\beta}_1 S_{xx} \right) \right] \\
&= \mathbb{E}\left[\sum Y_i^2 - n\overline{Y} - \widehat{\beta}_1^2 S_{xx} \right] \\
&= \sum \mathbb{E}[Y_i^2] - n\mathbb{E}[\overline{Y}^2] - S_{xx}\, \mathbb{E}[\widehat{\beta}_1^2]. \qquad (♡)
\end{aligned}$$

Applying the theorem that for any random variable U, $\mathbb{E}[U^2] = Var[U] - \mathbb{E}[U]^2$. We can see that

$$\begin{aligned}
(\heartsuit) = \mathbb{E}[SS_E] &= \sum_{i=1}^{n} \left[V(Y_i) - \mathbb{E}[Y_i]^2\right] - n\mathbb{E}[V(\overline{Y}) - \mathbb{E}[\overline{Y}]^2] \\
&\quad - S_{xx}\left(\frac{\sigma^2}{S_{xx}} + \beta_1^2\right) \\
&= n\sigma^2 + \sum_{i=1}^{n} (\beta_0 + \beta_1 x_i)^2 - n\left[\frac{\sigma^2}{n} + (\beta_0 + \beta_1 x_i)^2\right] \\
&\quad - S_{xx}\left(\frac{\sigma^2}{S_{xx}} + \beta_1^2\right) \\
&= n\sigma^2 + n\left(\beta_0 + \beta_1 x_i\right)^2 - \sigma^2 - n\left(\beta_0 + \beta_1 x_i\right)^2 - \sigma^2 - S_{xx}\beta_1 \\
&= \sigma^2(n-2).
\end{aligned}$$

□

13.2.2 Inference about the Slope parameter β_1

Lemma 13.2
In normal regression, the distributions of the slope parameters $\widehat{\beta}_1$ and $\widehat{\beta}_0$ are given by

$$\widehat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right) \quad \text{and} \quad \widehat{\beta}_0 \sim N\left(\beta_0, \frac{\sigma^2}{n} + \frac{\overline{x}^2\sigma^2}{S_{xx}}\right). \tag{13.12}$$

Proof. On the other hand, now we need to determine the distribution of intercept $\widehat{\beta}_0$. Since $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$, the distribution of $\overline{Y}$ is given by

$$\overline{Y} \sim N\left(\beta_0 + \beta_1 \bar{x}, \frac{\sigma^2}{n}\right).$$

We proved that

$$\widehat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right).$$

Multiply it by $\overline{x}$ and the distribution of $\overline{x}\widehat{\beta}_1$ is given by

$$\overline{x}\widehat{\beta}_1 \sim N\left(\overline{x}\beta_1, \overline{x}^2\frac{\sigma^2}{S_{xx}}\right).$$

Notice that $\widehat{\beta}_0 = \overline{Y} - \overline{x}\widehat{\beta}_1$, and $\overline{Y}$ and $\overline{x}\widehat{\beta}_1$ both are normal random variables. $\widehat{\beta}_0$ is also normal with

mean equal to $\beta_0 + \beta_1\overline{x} - \beta_1\overline{x} = \beta_0$ and variance is equal to

$$Var[\widehat{\beta}_0] = \frac{\sigma^2}{n} + \frac{\overline{x}^2\sigma^2}{S_{xx}}.$$

That is,

$$\widehat{\beta}_0 \sim N\left(\beta_0, \frac{\sigma^2}{n} + \frac{\overline{x}^2\sigma^2}{S_{xx}}\right)$$

provided the distribution of $Y_i|x_i$ is normal. The proof of the theorem is now complete. □

Theorem 13.2
The test statistic to test the hypotheses

$$H_0 : \beta_1 = \beta_1^* \quad \text{against} \quad H_0 : \beta_1 \neq \beta_1^*$$

is

$$T = \frac{\widehat{\beta}_1 - \beta_1^*}{S/\sqrt{S_{xx}}} = \frac{\widehat{\beta}_1 - \beta_1^*}{\widehat{\sigma}}\sqrt{\frac{(n-2)S_{xx}}{n}}, \tag{13.13}$$

which has a t-distribution with $n-2$ degree of freedom. Reject H_0 if

$$|T| < t_{n-2,\,\frac{\alpha}{2}}.$$

The term $Se(\widehat{\beta}_1) = \sqrt{\frac{MS_E}{S_{xx}}}$ is the estimated standard error of $\widehat{\beta}_1$. The p-value based on $n-2$ df can be computed as was done previously for t-tests.

✱ Example 13.2.1. The frequency of chirping of a cricket is thought to be related to temperature. This study suggests the possibility that the temperature can be estimated through the chirp frequency. The following data shown the number of chirping per second, x by the striped ground cricket and the temperature, y in Fahrenheit:

x	20	16	20	18	17	16	15	17	15	16
y	89	72	93	84	81	75	70	82	69	83

Assuming the regression of temperature on the number of chirps per second is normal random variable. Test the hypotheses

$$H_0 : \beta_1 = 4 \quad \text{against} \quad H_1 : \beta_1 \neq 4$$

at the significance level 0.1.

✱ Example 13.2.2. Table below shows the cost of installing a sample of communications nodes for a large manufacturing firm. The number of access ports at each of the sampled nodes is also shown. The data scientist wants to develop a linear model that is helpful in pricing the installation of new communications nodes on the network.

Cost	Number of ports	Cost	Number of ports
52,388	68	33,969	28
51,761	52	31,309	24
50,221	44	23,444	24
36,095	32	24,269	12
27,500	16	53,479	52
57,088	56	33,543	20

The R codes and outputs are shown below. Using the outputs to answer the following questions:

1. What is the regression line that relate the number of ports to cost.
2. Is there sufficient evidence to conclude that a linear relationship exists between the cost and number of ports under $\alpha = 0.05$?
3. Test whether there is a direct (positive) relationship between cost and number of ports. Use $\alpha = 0.05$.
4. A claim is made that each new access port adds at least $\$1,000$ to the installation cost of a communications node. Examine this claim at $\alpha = 0.05$.
5. The following test has no practical significance in this problem and is performed merely to illustrate the test for the intercept provided on computer output. Test whether the intercept is zero.

✲ Solution From the outputs provided above:

1. The regression line is $\widehat{y}_i = 16596.84 + 644.64x_i$.
2. The test hypotheses are $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$.

   The T statistic value is $T = 8.261 \sim t_{\nu=10}$, and $T > t_{0.025,10} = 2.228$.

   The p-value is $2[1 - \mathbb{P}(T_{10} > 8.261)] = 8.88 \times 10^{-6}$ and is smaller than $\alpha = 0.05$.

   We reject H_0, we have sufficient evidence to conclude that a linear relationship exists between the cost and the number of ports.
3. We want to test $H_0 : \beta_1 \leq 0$ against $H_1 : \beta_1 > 0$.

$$p - \text{value} = \frac{8.88 \times 10^{-6}}{2} = 4.44 \times 10^{-6} < 0.05.$$

   Reject H_0, we have sufficient evidence to conclude that there is a direct (positive) relationship between the cost and the number of ports.
4. The test hypotheses are $H_0 : \beta_1 \geq 1000$ versus $H_1 : \beta_1 < 1000$.

$$T = \frac{\widehat{\beta}_1 - \beta_1^*}{\sqrt{\frac{MS_E}{S_{xx}}}}$$

◀

13.3 Prediction

13.3.1 Prediction vs Estimation

What is the main reason we fitting a model to data? It is often to accomplish one of two goals. We can either use a model to estimate the relationship between response and the predictors, or to predict the response based on the predictors. Often, a good model can do both, but here we'll discuss both goals separately since the process of finding models for explaining and predicting are slightly different. In most cases, prediction is less precise than estimation with a bigger standard error.

Rather than calculate an interval estimate for μ_{Y,x^*}, an investigator may wish to obtain a range or an interval of possible values of Y associated with some future observation when the independent variable has value x^*. Consider, for example, relating vocabulary size y to age of a child x. The CI with $x^* = 6$ would provide a range that covers with 95% confidence the true average vocabulary size for all 6-year-old children.

Alternatively, we might wish an interval of plausible values for the vocabulary size of a particular 6-year-old child. How can you tell that a child is "off the chart" for example?

A **confidence interval** refers to a *parameter*, or population characteristic, whose value is *fixed but unknown* to us.

In contrast, a **future value** of Y is not a parameter but instead a *random variable*; for this reason we refer to an *interval of plausible values for a future* Y as a **prediction interval** rather than a confidence interval. Determining a prediction interval for Y requires that we model the error involved in the prediction of the Y variable.

One important application in linear regression is the prediction of a future value Y_0 at the covariate value x_0,

$$Y_0 = \beta_0 + \beta x_0 + \varepsilon_0$$

Since the mean of Y_0 is $\beta_0 + \beta x_0$, the commonly used predict value for Y is

$$\widehat{Y_0} = \widehat{\beta_0} + \widehat{\beta} x_0$$

which has mean $\beta_0 + \beta x_0$ and variance

$$\begin{aligned} Var(\widehat{Y_0}) &= Var(\widehat{\beta_0}) + x_0^2 Var(\widehat{\beta}) + 2x_0 \operatorname{Cov}(\widehat{\beta_0}, \widehat{\beta}) + Var(\varepsilon_0) \\ &= \frac{\sigma^2}{S_{xx}} \left(\frac{1}{n} S_{xx} + \overline{x}^2 + x_0^2 - 2x_0 \overline{x} \right) \\ &= \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}} \right]. \end{aligned}$$

The prediction error $Y_0 - \widehat{Y_0}$ must have zero mean, which means it is an unbiased prediction and with

variance

$$\sigma^2 + Var(\widehat{Y}_0) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right]. \tag{13.14}$$

We would like to construct an interval C based on samples data such that

$$\mathbb{P}[Y_0 \in C] = 1 - \alpha$$

This is called a prediction interval, where $\mathbb{P}$ is the probability measure with respect to both Y_0 and $Y_1, \ldots, Y_n$, then we can apply the fact that, under the normality assumption on ε's,

$$\frac{Y_0 - \widehat{Y}_0}{\widehat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}} \sim t_{\nu = n-2} \tag{13.15}$$

The resulting interval is $C = [\widehat{Y}_-, \widehat{Y}_+]$, where

$$\widehat{Y}_\pm = \widehat{Y}_0 \pm t_{n-2,\,\alpha/2}\,\widehat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}} \tag{13.16}$$

Theorem 13.3 Prediction interval for Future value Y
The prediction interval of Y is

$$\widehat{Y}_\pm = \widehat{Y}_0 \pm t_{n-2,\,\alpha/2}\,\widehat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}} \tag{13.17}$$

Suppose that we want to construct a confidence interval for $\theta = \alpha + \beta x_0$, the mean of Y-values at x_0, instead of a prediction interval for the future value Y_0.

We can use $\widehat{Y}_0 = \widehat{\alpha} + \widehat{\beta} x_0$ as an estimator of θ and we still have

$$Var(\widehat{Y}_0) = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right)$$

Under the normality assumption on ε's,

$$\frac{\widehat{Y}_0 - \theta}{\widehat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}} \sim t\text{-distribution with degrees of freedom } n - 2$$

Hence, the resulting confidence interval is $C_0 = [\widehat{\theta}_-, \widehat{\theta}_+]$, where

$$\widehat{\theta}_\pm = \widehat{Y}_0 \pm t_{n-2,\,\alpha/2}\,\widehat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}$$

This interval is **shorter** than the prediction interval for a future Y_0, because of the additional variability

from Y_0 in the prediction interval. Without the normality assumption, C_0 is still asymptotically valid but C is not asymptotically valid.

The interpretation of the prediction level $100(1-\alpha)\%$ is similar to that of previous confidence levels—if is used repeatedly, in the long run the resulting interval will actually contain **the observed y values** $100(1-\alpha)\%$ of the time.

Notice that the 1 underneath the initial square root symbol makes the PI wider than the CI, though the intervals are both centered at $\widehat{\beta}_0 + \widehat{\beta}_1 x^*$.

Also, as $n \to \infty$, the width of the CI approaches 0, whereas the width of the PI does not (because even with perfect knowledge of β_0 and β_1, there will still be randomness in prediction).

13.4 Correlation: How strong is the linear relationship?

In regression analysis, the linear correlation coefficient (which also called **Pearson correlation coefficient**), ρ, is the measure of the degree of association between the response x and explanatory data y.

Recall that if X and Y are bivariate random variables with means μ_X and μ_Y and standard deviations σ_X and σ_Y, respectively.

$$f(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{ -\frac{1}{2(1-\rho^2)} \left[\left(\frac{y-\mu_1}{\sigma_1}\right)^2 + \left(\frac{x-\mu_2}{\sigma_2}\right)^2 - 2\rho\left(\frac{y-\mu_1}{\sigma_1}\right)\left(\frac{x-\mu_2}{\sigma_2}\right) \right] \right\}$$

then the correlation coefficient between X and Y is defined as

$$\rho_{X,Y} = \frac{\operatorname{Cov}(X,Y)}{\sigma_X\,\sigma_Y} = \frac{\mathbb{E}[(X-\mu_X)(Y-\mu_Y)]}{\sqrt{\mathbb{E}[(X-\mu_X)^2]\,\mathbb{E}[(Y-\mu_Y)^2]}}. \tag{13.18}$$

Here is the correlation coefficient for a sample data set.

Definition 13.2 Sample Correlation Coefficient
Given a set of data $\{(x_i, y_i) \mid i = 1, 2, \ldots, n\}$, the sample correlation coefficient r is defined as

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{\sum_{i=1}^n (x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^n (x_i-\overline{x})^2 \sum_{i=1}^n (y_i-\overline{y})^2}}. \tag{13.19}$$

We can associate this sample data set into two vectors $\overrightarrow{x} = (x_1, x_2, \ldots, x_n)$ and $\overrightarrow{y} = (y_1, y_2, \ldots, y_n)$ in $\mathbb{R}^n$, the real n-dimensional Euclidean space. Let

$$\mathfrak{G} = \left\{ \lambda \underset{\sim}{\mathbf{e}} \mid \lambda \in \mathbb{R} \right\}$$

to be a subset of $\mathbb{R}^n$, where $\underset{\sim}{\mathbf{e}} = (1, 1, \ldots, 1)$ is a vector in $\mathbb{R}^n$ with all entries equal to 1.

Consider the quotient space $V = \mathbb{R}^n/\mathfrak{G}$, which is the set of all cosets of the form $\overrightarrow{x} + \mathfrak{G}$.

13.5 Regression Analysis — Analysis of Variance Approach

Beside t-test, we may also use an **ANOVA** to test significance of regression. The analysis of variance is based on a partitioning of total variability in the response variable y.

13.5.1 Decomposition of Sum of Squares

One useful aspect of SLR is that it can decomposite the variation into two parts: the variation of the predicted values, and the variation of prediction errors.

If all such deviations are squared, the squared deviations $(y_i - \overline{y})^2$ will provide the basis for measuring the spread of the data.

Squaring again both sides of the preceeding equation and sum over all $i = 1, 2, ..., n$, we can obtain the following

$$\sum (y_i - \overline{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \overline{y})^2 + 2\sum (y_i - \hat{y}_i)(\hat{y}_i - \overline{y}) \tag{13.20}$$

After squaring the right-hand side of the equation, we need some algebraic calculation to evaluate the cross-product term $(y_i - \hat{y}_i)(\hat{y}_i - \overline{y})$.

$$\begin{aligned}
\sum (y_i - \hat{y}_i)(\hat{y}_i - \overline{y}) &= \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)\,(\hat{\beta}_0 + \hat{\beta}_1 x_i - \overline{y}) \\
&= \sum [y_i - \overline{y} - \hat{\beta}_1 (x_i - \overline{x})]\,[\hat{\beta}_1 (x_i - \overline{x})] \\
&= \hat{\beta}_1 \,\mathrm{SS}_{xy} - \hat{\beta}_1^2 \,\mathrm{SS}_x \\
&= 0
\end{aligned}$$

this implies that the sum of the cross-products over all response variables is zero [1]. And hence, the finalized equation will be

$$\underbrace{\sum (y_i - \overline{y})^2}_{\text{TSS}} = \underbrace{\sum (y_i - \hat{y}_i)^2}_{\text{RSS}} + \underbrace{\sum (\hat{y}_i - \overline{y})^2}_{\text{RegSS}} \tag{13.21}$$

Let us interpret each SS one by one:

- **Total Sum of Squares (TSS)** represents the variation of Y, that is, the sum of squares in all of the responses. It measures the amount of variability inherent in the response prior to performing regression.
- **Residual Sum of Squares (RSS)**, also called the squared error loss or the L2-norm, is the variation of all the response values about the fitted regression line. It describes how well the model fits the data.

[1] In linear algebra, this implies the residuals are orthogonal to the regressor.

- **Regression Sum Of Squares (RegSS)**, the difference between TSS and RSS, or the total variation explained through knowledge of x. RegSS measure how effective the SLR model is in explaining the variation of y.

Since the TSS is fixed, and both RSS and RegSS always non-negative as they are sum of squares. The higher the RegSS of a regression model, the lower its RSS. In other word, when RSS is small, the prediction will be closer to actual value, and the model is a good fit. Else if RSS is large, the prediction is not accurate and this model is a poor fit.

You might be wondering why it should be the squared error? Why not the absolute error or the cubed error? In fact, they both can be used. But only the squared one maximized likelihood for betas.

13.5.2 Mean Square Error

An unbiased estimator of the unknown error variance σ^2, the mean square error (MSE) is derived as

$$s^2 = \text{MSE} = \frac{\text{RSS}}{n-2} = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{13.22}$$

Note that in defining MSE we have divided $n-2$ instead of $n-1$. This is because we need at least two observations to determine the fitting line.

13.5.3 Coefficient of determination

The coefficient of determination, denoted by R^2, represents the proportion of SS_T that is explained by use of the linear regression model.

Definition 13.3 R-square
The statistic

$$R^2 = \frac{SS_B}{SS_T} = 1 - \frac{SS_R}{SS_T} = \frac{S_{xy}}{SS_T} = \frac{\widehat{\beta}^2 S_{xx}}{S_{yy}}. \tag{13.23}$$

is called the R-square statistic or coefficient of determination, which measures the proportion of the total variation in Y_i's that is explained by the fitted line and, hence, it measures how well the simple linear regression fitted the sample points.

The *proportion* (or *percentage*) of the variation in Y that is explained by the linear relationship with predictor X is measured by R^2, the coefficient of determination. It is a measure of the variability in Y without considering the effect of the regressor variable X. Here are some facts need to take notes:

1. $R^2 \times 100\%$ of variation in Y can be "explained" by using X to predict Y, or the error in predicting Y can be reduced by $R^2 \times 100\%$ when the regressor is used instead of just $\overline{y}$.

2. $0 \leq R^2 \leq 1$ is a measure of "fit" for the regression line. The nearest the R^2 to 1, the better the regression line "fit".

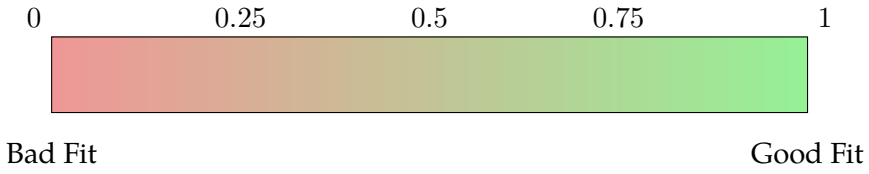


Figure 13.4: The diagram illustrated the spectrum of the quality of regression line based on R^2.

3. We can determine the *coefficient of correlation* with $r = \sqrt{R^2}$. In fact, The square of r is the coefficient of determination in simple linear regression.

Theorem 13.4
The F-statistic can be expressed as

$$\mathcal{F} = \frac{(n-2)R^2}{1-R^2}. \tag{13.24}$$

The analysis of variance consists of calculations that provide information about levels of variability within a regression model and form a basis for testing the

Source of Variation	Sums of squares	Degree of freedom	Mean squares	$\mathcal{F}$-statistic
Residual	SS_R	1	$MS_R = \frac{SS_R}{1}$	$\mathcal{F} = \frac{MS_R}{MS_E}$
Error	SS_E	$n-2$	$MS_E = \frac{SS_E}{n-2}$	
Total	SS_T	$n-1$		

Table 13.1: ANOVA table for simple linear regression

The final $\mathcal{F}$-value provides a statistic for testing the null hypothesis that there is no linear relationship between X and Y in the population. If the null hypothesis is rejected, it suggests that there is a significant linear relationship between the variables. That is, we are testing the hypotheses

$$H_0 : \beta_1 = 0 \quad \text{against} \quad H_1 : \beta_1 \neq 0$$

This test statistic is the ratio SS_R/SS_E, the mean square residual divided by the mean square error. Under the null hypothesis, this statistic has an $\mathcal{F}$-distribution with 1 and $n-2$ degrees of freedom. When the SS_R term is large relative to the SS_E term, then the ratio will be large and indicates that there is evidence against the null hypothesis.

Tutorials

Exercise 13.1 Let $Y_1, Y_2, \ldots, Y_n$ be independent random variables of size n, such that for each $Y_i \overset{\text{iid}}{\sim} EXP(\beta x_i)$, where β is an unknown parameter. If

$$\{(x_i, y_i) \mid i = 1, 2, \ldots, n\}$$

is a data set where $y_1, y_2, \ldots, y_n$ are the observed values based on $x_1, x_2, \ldots, x_n$. Find the least squares estimator of $\hat{\beta}$.

Exercise 13.2 Let $Y_1, Y_2, \ldots, Y_n$ be independent random variables of size n, such that for each $Y_i \overset{\text{iid}}{\sim} POI(\beta x_i)$, where β is an unknown parameter. If

$$\{(x_i, y_i) \mid i = 1, 2, \ldots, n\}$$

is a data set where $y_1, y_2, \ldots, y_n$ are the observed values based on $x_1, x_2, \ldots, x_n$. Find the least squares estimator of $\hat{\beta}$.

Exercise 13.3 Given the five pairs of data points (x_i, y_i) shown below:

x	10	20	30	40	50
y	50.071	0.078	0.112	0.120	0.131

What is the curve of the form $y = a + bx + cx^2$ best fits the data by method of least squares?

Answers for some questions

Answer of exercise 8.10

A parameter is a number that describes the population, while a statistic is a number that describes the sample.

Answer of exercise 10.1

The sample mean is

$$\overline{X} = \frac{1}{100}\sum_{i=1}^{100} x_i = \frac{1040.0}{100} = 10.4.$$

and the sample variance is

$$\begin{aligned} s^2 &= \frac{1}{100-1}\left(\sum_{i=1}^{100} x_i^2 - 100\overline{X}^2\right) \\ &= \frac{1}{99}\left(11102.11 - 100(10.4)^2\right) \\ &= 2.895. \end{aligned}$$

Since the population variance σ^2 is unknown, we use the sample variance s^2 to estimate σ^2. The unbiased estimate of σ^2 is

$$\widehat{\sigma}^2 = \frac{n}{n-1}s^2 = \frac{100(8.8209)}{99} = 2.895.$$