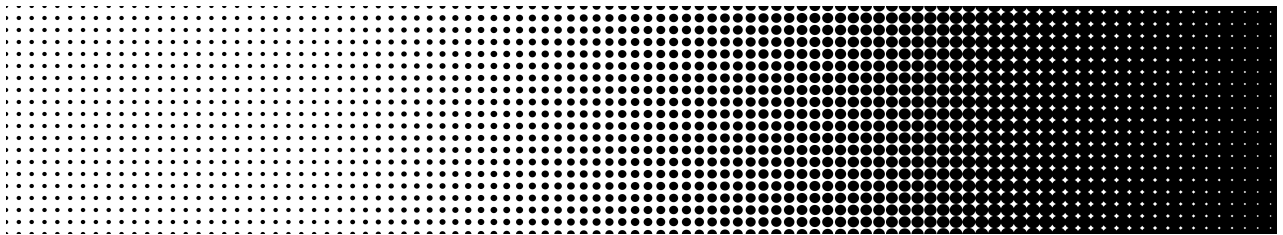


Mathematical Statistics

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



Based on lectures UECM 3363 Mathematical Statistics, UECM 3253 Nonparametric Statistics,
UCCM 2263 Applied Statistical Modelling in Universiti Tunku Abdul Rahman in 2020
Notes taken by MurphyShark.

Random Variables

1.1 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 student's GPA, or a student's height. The notation $X(s) = x$ means that x is the value associated with the outcome s by the random variable X .

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

Example 1.1.1. 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
$p(x)$	$\frac{5}{210}$	$\frac{50}{210}$	$\frac{100}{210}$	$\frac{50}{210}$	$\frac{5}{210}$

□

1.2 Cumulative Distribution

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

Definition 1.1 Increasing and Decreasing sequence of events

A sequence of sets $\{E_n\}_{n=1}^{\infty}$ is said to be increasing if

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

whereas it is said to be a decreasing sequence if

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

Lemma 1.1

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

$$\lim_{n \rightarrow \infty} \mathbb{P}[E_n] = \mathbb{P}[\lim_{n \rightarrow \infty} E_n]. \quad (1.1)$$

that is

$$\mathbb{P} \left[\bigcup_{n=1}^{\infty} E_n \right] = \lim_{n \rightarrow \infty} \mathbb{P}[E_n] \quad \text{for increasing sequence,} \quad (1.2)$$

and

$$\mathbb{P} \left[\bigcap_{n=1}^{\infty} E_n \right] = \lim_{n \rightarrow \infty} \mathbb{P}[E_n] \quad \text{for decreasing sequence,} \quad (1.3)$$

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 $F_i \cap F_j = \emptyset$. 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

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

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

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

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

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

Equivalently,

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

or

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

□

Theorem 1.1 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. $F_X(+\infty) = \lim_{x \rightarrow +\infty} F_X(x) = 1$.
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 $\lim_{x \downarrow -\infty} F(x) = \lim_{n \rightarrow \infty} F(x_n)$ 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 $E_1 \supseteq E_2 \supseteq \dots$. Moreover,

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

By previous proposition, we find

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

2. In the other hand, suppose that $\lim_{x \rightarrow \infty} F(x) = \lim_{n \rightarrow \infty} F(x_n)$ where $\{x_n\}$ is a increasing sequence such that $x_n \rightarrow \infty$. 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 $E_1 \subseteq E_2 \subseteq \dots$. Moreover,

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

By previous proposition, we find


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

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 1.2.1.** 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 $(-\frac{1}{2}, \frac{1}{2})$, we have $\mathbb{E}[X] = 0$. Therefore

$$\begin{aligned} \text{Var}[X] &= \mathbb{E}[X^2] - 0 = \int_{-1/2}^0 x^2(2 + 4x) dx + \int_0^{1/2} x^2(2 - 4x) dx \\ &= \frac{1}{24}. \end{aligned}$$

2. The cumulative function is


$$\begin{aligned} F(x) &= \begin{cases} 0 & \text{if } x < -\frac{1}{2} \\ \int_{-1/2}^x (2 + 4t) dt & \text{if } -\frac{1}{2} \leq x \leq 0 \\ \int_{-1/2}^0 (2 + 4t) dt + \int_0^x (2 - 4t) dt & \text{if } 0 \leq x \leq \frac{1}{2} \\ 1 & \text{if } x > \frac{1}{2} \end{cases} \\ &= \begin{cases} 0 & \text{if } x < -\frac{1}{2} \\ 2x^2 + 2x + \frac{1}{2} & \text{if } -\frac{1}{2} \leq x \leq 0 \\ -2x^2 + 2x + \frac{1}{2} & \text{if } 0 \leq x \leq \frac{1}{2} \\ 1 & \text{if } x > \frac{1}{2} \end{cases} \end{aligned}$$

□


1.3 Percentiles and Quantiles

1.3.1 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 1.3.1.** 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 .

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

$$\begin{aligned}\frac{d^2f}{dx^2} &= \frac{(1+x^3)^2\frac{d}{dx}(2cx - cx^4) - (2cx - cx^4)\frac{d}{dx}(1+x^3)^2}{(1+x^3)^4} \\ &= \frac{(1+x^3)^2(2c - 4cx^3) - 6x^2(2cx - cx^4)(1+x^3)}{(1+x^3)^4} \\ &= \frac{(1+x^3)(2c - 4cx^3) - 6x^2(2cx - cx^4)}{(1+x^3)^3} \\ &= \end{aligned}$$

□

By inspection, $\frac{d^2f}{dx^2} < 0$. And so $\frac{df}{dx} = 0$ 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.

1.4 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 1.2 Expectation value

Let (Ω, \mathbb{P}) be a probability space, let $E \subseteq \mathbb{R}$ be countable, and let X be a E -valued random variable on (Ω, \mathbb{P}) . The expectation of X , if it exists, is defined by

$$\mathbb{E}[X] := \sum_{e \in E} e f_X(e). \quad (1.4)$$

Lemma 1.2

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

 **Example 1.4.1.** Let X be a random variable representing the value shown a fair six-sided die is rolled. Then $X \sim \text{discreteU}(\{1, 2, 3, 4, 5, 6\})$, 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 1.4.2.** If $f(x) = (k+1)x^2$ for $0 < x < 1$, find the moment generating function.

 **Solution** Since f is a density function, thus $\int_0^1 f(x) dx = 1$, 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

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

\square

1.4.1 Variance

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

Definition 1.3 Variance

The variance of distribution X is sum of squared loss

$$\text{Var}[X] := \mathbb{E}_X(X_i - \mu_X)^2 = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \quad (1.5)$$

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 1.2 Chebyshev's Theorem

Let X be a random variable with mean μ_X and finite variance σ^2 . Then,

$$\mathbb{P}[|X - \mu_X| < k\sigma] \geq 1 - \frac{1}{k^2} \quad (1.6)$$

or

$$\mathbb{P}[|X - \mu_X| \geq k\sigma] \leq \frac{1}{k^2} \quad (1.7)$$

for some constant $k > 0$.

1.4.2 The Coefficient of Variation

Definition 1.4 Coefficient of Variation

The coefficient of variation is

$$CV = \frac{\sigma_X}{\mu_X} = \frac{\sqrt{Var[x]}}{\mathbb{E}[X]}. \quad (1.8)$$

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.

1.5 Discrete Random Variables

1.5.1 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 1.3

Let (Ω, \mathbb{P}) be a probability space, let $p \in [0, 1]$ and let $X_1, X_2, \dots, X_n : \Omega \rightarrow \{0, 1\}$ be

independent random variables such that each $X_i \sim \text{Bernoulli}(p)$. Then

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

1.5.2 Geometric distribution

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

$$p_X(n) = \mathbb{P}(X = n) = p(1 - p)^{n-1}, \quad n = 1, 2, \dots \quad (1.9)$$


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

$$\sum_{n=1}^{\infty} p(1 - p)^{n-1} = \frac{p}{1 - (1 - p)} = 1. \quad (1.10)$$

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 1.5.1.** 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

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

□

1.6 Continuous Probability Distributions

1.6.1 Uniform Probability Distribution

Definition 1.5 Uniform Distribution

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

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & \text{if } a \leq x \leq b \\ 0 & \text{elsewhere.} \end{cases} \quad (1.11)$$

Theorem 1.4 Mean and variance of Uniform Distribution

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

$$\mu_X = \mathbb{E}[X] = \frac{a + b}{2} \quad (1.12)$$

and

$$\sigma_X^2 = \text{Var}[X] = \frac{(b - a)^2}{12} \quad (1.13)$$

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

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

which is what we expected. □

1.7 Normal Distribution

1.8 Gamma Distribution

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

Definition 1.6 Gamma Distribution

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

$$f_X(x) = \begin{cases} \frac{x^{\alpha-1} e^{-x/\beta}}{\beta^\alpha \Gamma(\alpha)}, & \text{if } x \geq 0 \\ 0 & \text{elsewhere.} \end{cases} \quad (1.14)$$

Theorem 1.5 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

$$\mu_X = \alpha\beta \quad (1.15)$$

$$\sigma^2 = \alpha\beta^2. \quad (1.16)$$

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,

$$M_X(t) = \frac{\left(\frac{\beta}{1-\beta t}\right)^{\alpha}}{\beta^{\alpha}} \times 1 = \frac{1}{(1-\beta t)^{\alpha}}. \quad (1.17)$$

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. □

1.8.1 Chi-square distribution

Definition 1.7

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 $\nu > 0$)

Limiting Distribution

2.1 Convergence in distribution


Convergence in distribution state that the sequence of random variables X_1, X_2, \dots, 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.

Definition 2.1 Convergence 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

$$\lim_{n \rightarrow \infty} F_{X_n}(x) = F_X(x) \quad (2.1)$$

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

 **Example 2.1.1.** Let X_1, X_2, X_3, \dots be a sequence of random variable such that

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

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

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

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


 **Solution** The cdf of Y_n is

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

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

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

which is the cdf of exponential distribution with parameter λ . \square


 **Example 2.1.2.** Let $\{X_n\}_{n \geq 1}$ be a sequence of random variables with probability mass function

$$f_{X_n}(x) = \mathbb{P}[X_n = x] = \begin{cases} 1 & \text{if } x = 2 + \frac{1}{n} \\ 0 & \text{otherwise.} \end{cases}$$

Find the limiting distribution of X_n .

2.1.1 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 2.1.3.** 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 $\{X_n, n = 1, 2, 3, \dots\}$ 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 $n \gg \infty$ we have

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

However, if $s = 0$ then

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

Thus, we conclude that $\lim_{n \rightarrow \infty} X_n(s) = X(s)$ 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. \square

2.2 Law of Large Numbers

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

We will first state the Weak Law of Large Numbers (WLLN),

Theorem 2.1 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

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

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

$$\lim_{n \rightarrow \infty} \mathbb{P}[|\bar{X}_n - \mu| > \epsilon] = 0. \quad (2.2)$$

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

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

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 \bar{X}_n to get, for all $\epsilon > 0$,

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

So that \square

Theorem 2.2 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

$$\mathbb{P}[\lim_{n \rightarrow \infty} \bar{X}_n = \mu] = 1. \quad (2.3)$$

Proof. By Markov's inequality that

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

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

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

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

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

□


Theorem 2.3 Central Limit Theorem


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

$$Z = \frac{\sqrt{n}(\bar{X}_i - \mu)}{\sigma} \xrightarrow{d} \mathcal{N}(0, 1) \quad (2.4)$$

or

$$Z = \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma} \xrightarrow{d} \mathcal{N}(0, 1) \quad (2.5)$$

 **Example 2.2.1.** Let \bar{X}_n be the sample mean from a random sampling of size $n = 100$ from χ_{50}^2 . Compute approximate value of $\mathbb{P}(49 < \bar{X} < 51)$.

 **Solution** Because \bar{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,

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

□

Theorem 2.4 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

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

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 $\mathbb{P}[|Y_n - c| > \epsilon] \rightarrow 0$ as $n \rightarrow \infty$. 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 $\phi_{X_n}(t) \rightarrow \phi_X(t)$, and that convergence in probability preserves the limit of expectations for bounded random variables, we get:

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

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

 **Example 2.2.2.** If the random variable $X \sim \text{Gamma}(\mu, 1)$, show that

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

 **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

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

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

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

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

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sqrt{\bar{X}_n}} \xrightarrow{d} \mathcal{N}\left(\frac{0}{\mu}, \frac{\mu}{\mu}\right) \implies \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sqrt{\bar{X}_n}} \xrightarrow{d} \mathcal{N}(0, 1)$$

and we are done with the proof. \square

Theorem 2.5

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

$$\sqrt{X_n} \xrightarrow{p} \sqrt{c}, \quad c > 0. \quad (2.6)$$

Theorem 2.6

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 $n \rightarrow \infty$.

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 $n \rightarrow \infty$.

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 $\mathbb{P}[|X_n - c| > \epsilon] \rightarrow 0$.

In particular, $\mathbb{P}[|X_n - c| > \delta/2] \rightarrow 0$, which implies $\mathbb{P}[|X_n| > \delta/2] \rightarrow 1$. 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 $n \rightarrow \infty$, both terms approach 0, completing the proof. \square

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

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

In this notion, the maximum random variable is

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

and the minimum random variable is

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

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

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

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

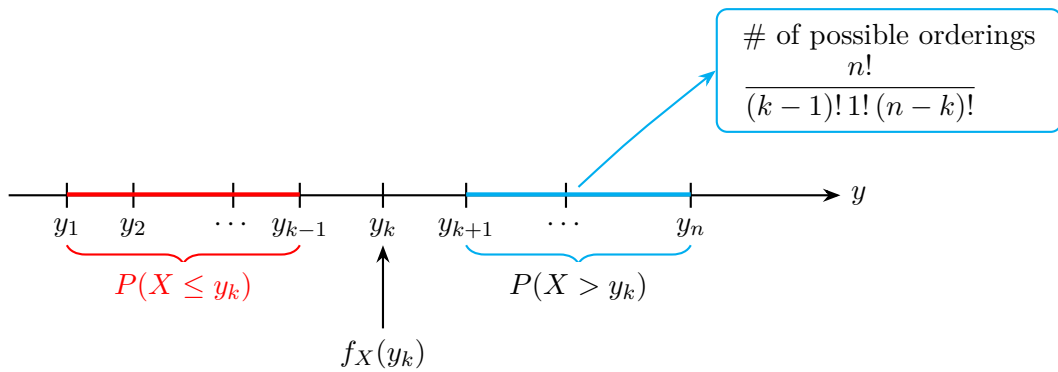
$$\begin{aligned} g_{(n)}(y) &= \frac{d}{dy} [F_X(y)]^n \\ &= n[F_X(y)]^{n-1} \frac{d}{dy} F_X(y) && \text{By Chain rule of derivative} \\ &= n[F_X(y)]^{n-1} f_X(y) \end{aligned}$$

Now we get the minimum 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, \dots, X_n , and the event $Y_i > y$ can be occurs for $i = 1, 2, 3, \dots, n$. In other words, any X_i in X_1, X_2, \dots, X_n can be the minimum variable. Hence


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




Theorem 2.7 k-th order statistics

Let X_1, X_2, \dots, 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

$$g_{(n)}(y) = \frac{n!}{(k-1)!(n-k)!} [F_X(y)]^{k-1} [1 - F_X(y)]^{n-k} f_X(y), \quad -\infty < y < \infty. \quad (2.7)$$

 **Example 2.3.1.** 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 $X_{(1)}$,

$$\mathbb{E}[Y_{(1)}] = \int_0^\theta y \left[\frac{5(\theta - y)^4}{\theta^5} \right] dy = \int_0^\theta \frac{5y(\theta - y)^4}{\theta^5} dy \quad (\clubsuit)$$

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

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

substitute back into (\clubsuit) and we have

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

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 than $\frac{2\theta}{3}$.

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

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

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

□

Decision theory

For the given observation \mathcal{X} , we decide to take an action $a \in \mathcal{A}$. An action is a map $a : \mathcal{X} \rightarrow \mathcal{A}$ 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.

$$L : \Theta \times \mathcal{A} \rightarrow \mathbb{R}. \quad (3.1)$$

3.1 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

$$\mathbb{P}(B | A) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}. \quad (3.2)$$

 **Example 3.1.1.** 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 .

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

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

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

□

 **Example 3.1.2 (Two-sample mean problems).** Consider the observations $X_1, X_2, \dots, X_m \sim \mathcal{N}(\mu, \sigma^2)$ response under control treatment. And $Y_1, Y_2, \dots, Y_n \sim \mathcal{N}(\mu + \Delta, \sigma^2)$ 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


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

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,

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

 **Example 3.1.3 (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

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

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

✎ **Example 3.1.4** (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

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

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

3.2 Value-at-risk

✎ **Example 3.2.1** (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: $\mathcal{A} = \{\mathbf{a} = [\underline{a}, \bar{a}] : \underline{a} < \bar{a} \in \mathfrak{U}\}$.
- ❖ Interval Estimator: define a map $\hat{\mu}(X) : \mathcal{X} \rightarrow \mathcal{A}$, 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

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

We define the zero-one loss function

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

The risk function under zero-one loss is

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

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

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


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

3.3 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

$$R(\theta, \delta') \leq R(\theta, \delta) \quad \forall \theta \in \Theta \tag{3.3}$$

with strict inequality for some θ .

 **Example 3.3.1.** Suppose, for $n \geq 2$, the observations X_1, X_2, \dots, X_n be i.i.d with mean $g(\theta) := \mathbb{E}_\theta[X_i] = \mu$, and $\text{Var}[X_i] = 1$ for all i . We take quadratic loss

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

Consider the decision

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

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

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

Therefore δ is inadmissible.


Review of Statistics

4.1 Distinguishing Between a Population and a Sample

Think of a **population** like all the cookies in a giant cookie jar - it's the entire group we're interested in studying. A **sample** is like taking a handful of cookies from that jar to taste and learn about all the cookies without eating them all.

Population: The complete set of all individuals, objects, or measurements that we want to study or make conclusions about.

Sample: A subset or portion of the population that we actually observe, measure, or collect data from. When the sample subset contained every member of the population, it is called a **census**.

 **Example 4.1.1.** Suppose you want to know the average height of all students at your university (that's about 25,000 students).

✱ The **population** would be all 25,000 students at the university.

✱ The **sample** might be 500 students that you randomly select and actually measure.

You use the average height from your sample of 500 students to estimate the average height of the entire population of 25,000 students. This saves time and money compared to measuring every single student!

4.2 Sampling Techniques

4.2.1 Simple Random Sampling

The probability sampling can be done in specifying the probability of a member being chosen into the sample. You can think of this as a weight assigned to each member of the population.

One caution before we proceed to other sampling techniques: Random sampling is totally different from random assignment. Random sampling is used to select a sample from a population, while random assignment is a random process of locating participants into different experimental groups or labels.

4.2.2 Stratified Random Sampling

To stratify means to classify or separate the individuals in a population into different groups or strata based on some common characteristic.

Sampling Technique	Example	Advantages	Limitations
Simple random sampling	The names of all 1,000 children are placed into a computer database. The computer is then instructed to randomly select 100 names. These children and their parents are then contacted.	Representative of the population	May be difficult to obtain the list; May be more expensive
Stratified random sampling	The names of all 1,000 children are placed into a computer database and organized by grade (sixth, seventh, eighth). The computer is then instructed to randomly select 35 names from each of the three grades. These children and their parents are then contacted.	Representative of the population	May be difficult to obtain the list; May be more expensive
Convenience sampling	The researcher knows one of the middle-school teachers, and the teacher volunteers her 35 students for the study. These children and their parents are then contacted.	Simple; Easy; Convenient; No complete member list needed	May not be representative of the population
Quota sampling	Using the middle-school directory, the researcher selects the first 20 sixth-grade boys, the first 20 sixth-grade girls, the first 20 seventh-grade boys, the first 20 seventh-grade girls, the first 20 eighth-grade boys, and the first 20 eighth-grade girls. These children and their parents are then contacted.	Simple; Easy; Convenient; No complete member list needed	May not be representative of the population

Estimation

Definition 5.1 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

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

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

Theorem 5.1 Mean Squared Error

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

$$MSE(\hat{\mu}) = Var[\hat{\mu}(X) | \theta] + Bias(\hat{\mu} | \theta)^2 \quad (5.1)$$

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

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), $\hat{\mu}(X) : \mathcal{X} \rightarrow \mathcal{A}$.
- ❖ 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,

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


□

5.1 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	$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$	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	$\bar{X} - \bar{Y}$	$\mu_1 - \mu_2$	$\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$
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 5.1.1.** 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 5.1.2.** 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 $\bar{X} - \bar{Y}$ 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

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

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

2. The mean of the point estimate is

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

The variance of the difference in means is

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

And

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

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

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

□

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 .

5.2 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?

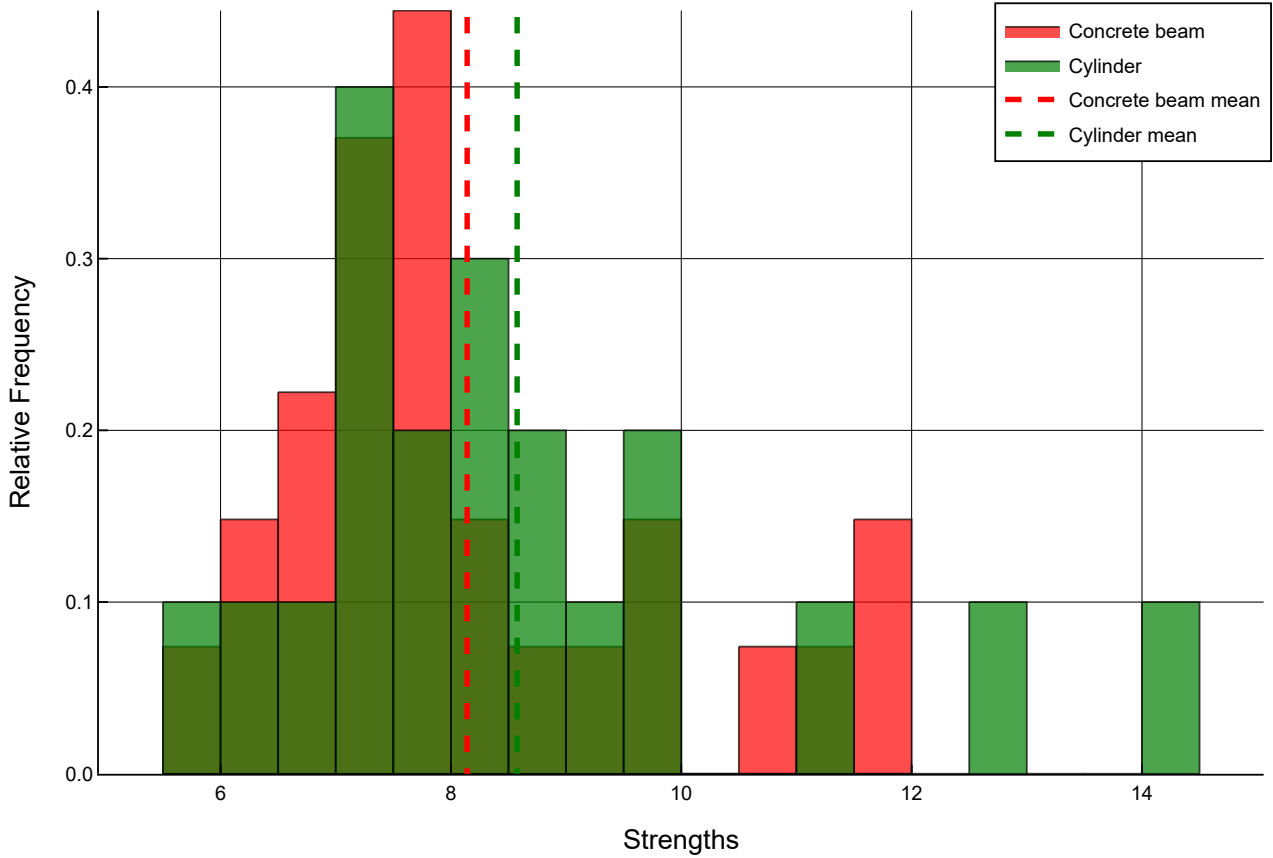
The resulting $\hat{\theta}$ is called the **minimum variance unbiased estimator (MVUE)** of θ .

Definition 5.2 Unbiased estimator

The estimator $\hat{\mu}$ is unbiased if $\text{Bias}(\hat{\mu} | \theta) = 0$

📌 **Example 5.2.1.** Let X_1, X_2, X_3 be a random sample of size 3 from a population with pmf

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



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 $X_i \sim \text{Poisson}(\lambda)$, we have

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

It is easy to see that

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

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

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

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

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

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

Example 5.2.2. Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be unbiased estimators of θ . Suppose $Var(\hat{\theta}_1) = 1$, $Var(\hat{\theta}_2) = 2$ and $Cov(\hat{\theta}_1, \hat{\theta}_2) = \frac{1}{2}$. 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,

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

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

$$\frac{d}{dc_1} Var[c_1\hat{\theta}_1 + c_2\hat{\theta}_2] = 6c_1 - 3 = 0 \implies c_1 = \frac{1}{2}.$$

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

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

\square

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.

5.2.1 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 5.2.3.** Let X_1, X_2, \dots, 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

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

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

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

that is

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

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

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


Now let say we have the following sample data:


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

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

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

□

 **Example 5.2.4.** Let $X \sim \mathcal{N}(\mu, \sigma^2)$, and X_1, X_2, \dots, 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

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

The first sample moment is

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

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

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

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

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

The last line follows from the fact that

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

Thus, the estimator of the parameter σ^2 is

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

□

Theorem 5.2

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

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

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

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

is an *unbiased estimator* of σ^2 .

Proof. From previous example, we can see that

$$\sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n X_i^2 - n(\bar{X})^2 \dots\dots\dots (\star)$$

Hence, we use this and the fact that

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

and

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

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

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

It follows that

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

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

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

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

5.3 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 5.3 Likelihood function and MLE

Let X_1, X_2, \dots, 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 states 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, \dots, 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 5.3.1.** 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, \dots, 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 $X = (X_1, X_2, \dots, X_n)$. Note that X_1, X_2, \dots, 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:

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

The sample mean (proportion) \bar{X} is the maximum likelihood estimator of p . □

5.3.1 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 5.4 MLE based on grouped data

If the data is grouped into k intervals with counts n_1, n_2, \dots, 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 5.3.2.** 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 $c = \frac{3^6}{50^6 \times 25^5}$ is a constant. To find the value of θ that maximizes $L(\theta)$, we take the logarithm of the likelihood function:

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

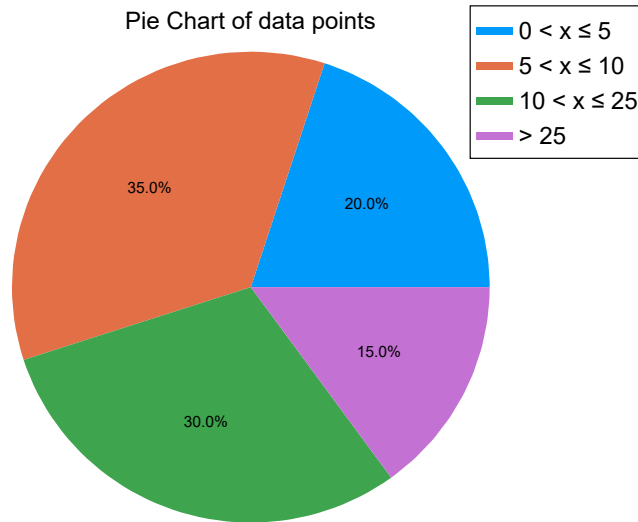
Next, we differentiate $\ell(\theta)$ with respect to θ and set it to zero:

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

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

 **Example 5.3.3.** 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:

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


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

□

Sufficiency


Definition 6.1 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 6.0.1.** If X_1, X_2, \dots, X_n are i.i.d. random samples from the Bernoulli distribution with parameter p , then the sum of the samples with density 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:

$$Y = \sum_{i=1}^n X_i \sim \text{Bin}(n, \theta),$$

Thus, the joint density function can be expressed as:

$$Y \sim g(y) = \binom{n}{y} \theta^y (1-\theta)^{n-y}, \quad y = 0, 1, \dots, n.$$

□

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 $h(x) = 1$.

Definition 6.2 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

$$S(X) = S(Y) \implies T(X) = T(Y).$$

Lemma 6.1 Rao Blackwell Theorem

If $T(X)$ is a sufficient statistic for parameter θ , and $\hat{\theta}$ is an unbiased estimator of θ with $\mathbb{E}[\hat{\theta}] < \infty$ for all $\theta \in \Theta$. Let $\hat{\theta}^* = \mathbb{E}[\hat{\theta}|T(X)]$, then $\hat{\theta}^* = \mathbb{E}[\hat{\theta}|T]$, then

$$\mathbb{E}[(\hat{\theta}^* - \theta)^2] \leq \mathbb{E}[(\hat{\theta} - \theta)^2]. \quad (6.1)$$

The inequality is strict unless $\hat{\theta}$ is a function of $T(X)$.

Proof. By the law of conditional expectation, we have

$$\mathbb{E}[\hat{\theta}^*] = \mathbb{E}[\mathbb{E}[\hat{\theta}|T]] = \mathbb{E}[\hat{\theta}] = \theta,$$

so $\hat{\theta}$ and $\hat{\theta}^*$ are having the same bias. By the conditional variance formula, we have

$$\mathbb{E}[\hat{\theta}] = \mathbb{E}[\text{Var}(\hat{\theta}|T)] + \text{Var}(\mathbb{E}[\hat{\theta}|T]) = \mathbb{E}[\text{Var}(\hat{\theta}|T)] + \text{Var}(\hat{\theta}^*).$$

Hence $\text{Var}[\hat{\theta}^*] \geq \text{Var}[\hat{\theta}]$, and so $\text{MSE}[\hat{\theta}^*] \geq \text{MSE}[\hat{\theta}]$. □

Theorem 6.1 Cramér-Rao lower bound

Let X_1, X_2, \dots, 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 $T = T(X_1, \dots, X_n)$ 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:

$$1 = \frac{d}{d\theta} \mathbb{E}[T] = \mathbb{E} \left[\frac{\partial T}{\partial \theta} \right] = \mathbb{E} \left[T \cdot \frac{\partial}{\partial \theta} \ln f(X_1, \dots, X_n | \theta) \right] = \mathbb{E}[T \cdot S(\theta)].$$

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.

6.1 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 6.3 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.,

$$I_T(\theta) = I_X(\theta).$$

6.1.1 Delta method – Variance of functions of estimators

Theorem 6.2

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

$$\text{Var}(g(\hat{\theta})) \approx [\dot{g}(\theta)]^2 \text{Var}[\hat{\theta}]. \quad (6.2)$$

Example 6.1.1. Given $g(s, t) = \frac{s}{t}$, $h(s, t) = \ln s$ and $\hat{\theta}_1$ and $\hat{\theta}_2$ 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 $\hat{\theta}_1 = 3.2$ and $\hat{\theta}_2 = 11.8$, 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:

$$\begin{aligned} \hat{\Sigma} &= \frac{1}{12\hat{\theta}_2^2 - 8\hat{\theta}_1^2} \begin{bmatrix} 3\hat{\theta}_2 & -2\hat{\theta}_1 \\ -2\hat{\theta}_1 & 2\hat{\theta}_2 \end{bmatrix} = \frac{1}{12(11.8)^2 - 8(3.2)^2} \begin{bmatrix} 3(11.8) & -2(3.2) \\ -2(3.2) & 2(11.8) \end{bmatrix} \\ &= \begin{bmatrix} 0.0222787 & -0.00402779 \\ -0.00402779 & 0.0148525 \end{bmatrix} \\ &= \begin{bmatrix} \text{Var}(\hat{\theta}_1) & \text{Cov}(\hat{\theta}_1, \hat{\theta}_2) \\ \text{Cov}(\hat{\theta}_1, \hat{\theta}_2) & \text{Var}(\hat{\theta}_2) \end{bmatrix}. \end{aligned}$$

1. Now take partial derivatives of $g(s, t)$ with respect to s and t :

$$g_s(s, t) = \frac{\partial g}{\partial s} = \frac{1}{t} \implies g_s(\hat{\theta}_1, \hat{\theta}_2) = \frac{1}{11.8},$$

$$g_t(s, t) = \frac{\partial g}{\partial t} = -\frac{s}{t^2} \implies g_t(\hat{\theta}_1, \hat{\theta}_2) = -\frac{3.2}{(11.8)^2}.$$

Hence, let $\mathbf{w} = [g_s(\hat{\theta}_1, \hat{\theta}_2) \quad g_t(\hat{\theta}_1, \hat{\theta}_2)] = \left[\frac{1}{11.8} \quad -\frac{3.2}{11.8^2} \right]$. the approximate variance of $g(\hat{\theta}_1, \hat{\theta}_2)$

is

$$\begin{aligned}
\text{Var}(g(\hat{\theta}_1, \hat{\theta}_2)) &\approx \mathbf{w} \hat{\Sigma} \mathbf{w}^T \\
&= \begin{bmatrix} \frac{1}{11.8} & -\frac{3.2}{11.8^2} \end{bmatrix} \begin{bmatrix} 0.0222787 & -0.00402779 \\ -0.00402779 & 0.0148525 \end{bmatrix} \begin{bmatrix} \frac{1}{11.8} \\ -\frac{3.2}{11.8^2} \end{bmatrix} \\
&= 0.000208785.
\end{aligned}$$

2. Continue with $h(s, t)$: take partial derivatives of $h(s, t)$ with respect to s and t :

$$h_s(s, t) = \frac{\partial h}{\partial s} = \frac{1}{s} \implies h_s(\hat{\theta}_1, \hat{\theta}_2) = \frac{1}{3.2},$$

$$h_t(s, t) = \frac{\partial h}{\partial t} = 0 \implies h_t(\hat{\theta}_1, \hat{\theta}_2) = 0.$$

The estimated covariance between $h(\hat{\theta}_1, \hat{\theta}_2)$ and $g(\hat{\theta}_1, \hat{\theta}_2)$ is

$$\begin{aligned}
\text{Cov}(h(\hat{\theta}_1, \hat{\theta}_2), g(\hat{\theta}_1, \hat{\theta}_2)) &= \begin{bmatrix} \frac{1}{\hat{\theta}_1} & -\frac{\hat{\theta}_1}{\hat{\theta}_2^2} \end{bmatrix} \begin{bmatrix} 0.0222787 & -0.00402779 \\ -0.00402779 & 0.0148525 \end{bmatrix} \begin{bmatrix} \frac{1}{\hat{\theta}_1} \\ 0 \end{bmatrix} \\
&= \begin{bmatrix} \frac{1}{11.8} & -\frac{3.2}{11.8^2} \end{bmatrix} \begin{bmatrix} 0.0222787 & -0.00402779 \\ -0.00402779 & 0.0148525 \end{bmatrix} \begin{bmatrix} \frac{1}{3.2} \\ 0 \end{bmatrix} \\
&= 6.2 \times 10^{-3}
\end{aligned}$$

□