

# STATISTICS WITH APPLIED PROBABILITY

Custom eBook for STA258

Nishan Mudalige

Nurlana Alili

Bryan Su

CANADA

# Statistics with Applied Probability

## Custom eBook for STA258

Nishan Mudalige

*Department of Mathematical and Computational Sciences  
University of Toronto Mississauga*

Nurlana Alili

*University of Toronto Mississauga*

Bryan Xu

*University of Toronto Mississauga*

© 2025 N. Mudalige, N. Alili, B. Xu  
All rights reserved.

This work may not be copied, translated, reproduced or transmitted in any form or by any means — graphic, electronic or mechanical including but not limited to photocopying, scanning, recording, microfilming, electronic file sharing, web distribution or information storage systems — without the explicit written permission of the authors.

Every effort has been made to trace ownership of all copyright material and to secure permission from copyright holders. In the event of any question arising as to the use of copyright material, we will be pleased to make necessary corrections in future publications.

First edition: August 2025

Mudalige, M.; Alili, N.; Xu, B.  
University of Guelph Bookstore Press

University of Toronto Mississauga,  
Mississauga, Ontario, Canada

# Contents

<b>0</b>	<b>Overview</b>	<b>1</b>
<b>1</b>	<b>Descriptive Statistics and an Introduction to R</b>	<b>2</b>
1.1	Introduction	2
1.2	Descriptive Statistics	4
1.3	Graphical Techniques	10
1.3.1	Histograms	10
1.3.2	Box-Plots	12
1.4	Introduction to R	13
<b>2</b>	<b>Sampling Distributions Related to a Normal Population</b>	<b>16</b>
2.1	Normal Distribution	16
2.2	Gamma and Chi-square Distribution	17
2.3	Student's t-Distribution and F-Distribution	20
<b>3</b>	<b>The Central Limit Theorem</b>	<b>23</b>
<b>4</b>	<b>Normal Approximation to the Binomial Distribution</b>	<b>26</b>
4.1	Introduction	26
4.2	Bernoulli Distribution	27
4.3	Sampling Distribution of the Sum and MGF Derivation	29
4.4	Binomial Distribution	29
4.4.1	Visualizing the PMF of Binomial Distributions	30
4.5	Sampling Distribution of a Sample Proportion and the Normal Approximation	33
4.6	Continuity Correction	37
<b>5</b>	<b>Law of Large Numbers</b>	<b>40</b>
5.1	Convergence in Probability	40
5.2	Weak Law of Large Numbers (WLLN)	41
<b>6</b>	<b>One Sample Confidence Intervals on a Mean When the Population Variance is Known</b>	<b>45</b>
6.1	Introduction	45
6.2	Interpretation	47
6.3	Confidence Interval for $\mu$ (Known Variance)	48
6.4	APPENDIX	54

<b>7</b>	<b>One-Sample Confidence Intervals on a Mean When the Population Vari- ance is Unknown</b>	<b>55</b>
7.1	CIs for $\mu$ . . . . .	55
<b>8</b>	<b>One Sample Confidence Intervals On a Proportion</b>	<b>65</b>
<b>10</b>	<b>Two Sample Confidence Interval</b>	<b>68</b>
10.1	Two Sample Confidence Interval on a Difference of Mean . . . . .	68
10.2	Two Sample Confidence Interval on Paired Data . . . . .	73
10.3	Two Sample Confidence Interval on Proportions . . . . .	76
10.4	Two Sample Confidence Interval on Variances . . . . .	78
<b>12</b>	<b>One Sample Hypothesis Test on a Proportion and Variance</b>	<b>80</b>
12.1	One Sample Hypothesis Test on a Proportion . . . . .	80
<b>14</b>	<b>Two Sample Hypothesis Test</b>	<b>83</b>
14.1	Two Sample Hypothesis Test on Paired Data . . . . .	83
14.2	Two Sample Hypothesis Test on Proportions . . . . .	85
14.3	Two Sample Hypothesis Test on Variances . . . . .	88
<b>15</b>	<b>Introduction to Simple Linear Regression</b>	<b>89</b>
15.1	Measures of Linear Relationship . . . . .	90
15.2	Least Squares Method . . . . .	94
15.3	Simple Linear Regression . . . . .	96
15.4	SST, SSE and SSR . . . . .	99



# Chapter 0

## Overview

Uncertainty is an inherent part of everyday life. We all face questions regarding uncertainty such as whether classes will go ahead as planned on any given day; will a flight leave on time; will a student pass a certain course? Uncertainties might also change depending on other factors, such as whether classes will still go ahead as planned when there is a snow warning in effect; if a flight is delayed can a person still manage to make their connection; will a student pass their course considering that the instructor is known to be a tough grader?

The ability to quantify uncertainty using rigorous mathematics is a powerful and useful tool. Calculating uncertainty on an intuitive level is something that is hard-wired in our DNA, such as the decision to fight or flight depending on a given set of circumstances. However we cannot always make such intuitive decisions based purely on hunches and gut feelings. Fortunes have been lost based on someone having a good feeling about something. If we have information available, we should make the best prediction possible using this information. For instance if we wanted to invest a lot of money in a company, we should use all available data such as past sales, market and industry trends, leadership ability of the CEO, forward looking statements etc. and with all this information we can then predict whether our investment will be profitable.

In order for companies to survive and remain competitive in today's environment it is essential to monitor industry trends and read markets properly. Companies that don't adapt and stick to an outdated business model tend to pay the price. At the other end of the spectrum, companies that understand the needs of the consumer, build their product around the consumer and keep evolving their product offerings based on consumer trends tend to perform well and remain competitive.

Statistics is the science of uncertainty and it is clearly a very useful subject for business. In this book you will be given an introduction to statistics and you will learn the framework as well as the language required at the introductory level. The material may be daunting at times, but the more you get familiar with the subject the more comfortable you will become with it. As business students, doing well in a statistics course will give you a competitive edge since the ability to interpret and perform quantitative analytics are skills that are highly desired by many employers.

# Chapter 1

## Descriptive Statistics and an Introduction to R

### 1.1 Introduction

Intuitively, statistics can be considered the science of uncertainty. Formally,

---

**Definition 1.1** (Statistics).

*Statistics is the science of collecting, classifying, summarizing, analyzing and interpreting data.*

---

#### Population, Sample, Parameter

In statistics, researchers need to observe behavior, pattern, trends and other types of data to give a conclusion. To make the conclusion more persuasive, researchers require huge amount of data to support them, that's why study statistics need population.

---

**Definition 1.2** (Population).

*In statistics, a population is a set of similar observations which is of interest for some experimental questions. It can be a set of existing objects such as all people in Canada, or hypothetical group of existing objects such as the set of all possible hands in a game of poker.*

---

However, data collection from population is a lot work. Usually, researchers select a finite number of observations to study.

---

**Definition 1.3** (Sample).

*It refers to a selection of a subset from population that researchers use it to estimate population characteristics.*

---



Now, we have already chosen a sample, but how do we use it to estimate population characteristics? This is the point where parameter comes to play.

---

**Definition 1.4** (Parameter Statistics).

*A parameter is a quantity of statistical population which summerizes characteristics of the population. For example, mean, variance and standard deviation.*

---

**Descriptive and Inferential Statistics**

Now, we have set everything we need. A population, a chosen sample in that population with its parameters. Next step is studying. There are two major types of analysis: descriptive and Inferential statistics. In this section, we are only going to give you a rough idea about what they are, more detailed materials will be introduced in later chapters.

---

**Definition 1.5** (Descriptive Statistics).

*It refers to the summation of all quantitive values that describe characteristics of the population. Usually, we use descriptive statistics to summerize characteristics of a data set.*

---

Furthermore, we use inferential statistics to do statistical analysis.

---

**Definition 1.6** (Inferential Statistics).

*It refers to the process of using data analysis to indicate properties of a population. For example, testing hypothesis and confidence interval (both will be introduced in later chapters).*

---

**Qualitative and Quantitative Data**

At this point, assume that we have finished all procedures such as obtaining parameters and analyzing properties. Now, another important thing is illustrating all the discovery.

---

**Definition 1.7** (Qualitative Data).

*This type of illustration refers to showing categorical data. For example, lecture notes from a course, open-question survey.*

---

To illustrate numerical data, we use quantitative data.

---

**Definition 1.8** (Quantitative Data).

*Unless the previous type of illustration, quantitative data is represented numerically, including anything that can be counted, measured, or given a numerical value. For example, STA258 final mark score range from 100 different students who have taken this course.*

---

## 1.2 Descriptive Statistics

Previously, we defined descriptive statistics. Now, let's introduce what exact they are.

### Sample Mean, Variance and Standard Deviation

Sample mean (or sample average) is the average value of a sample which is selected from an interested population of an experiment. Usually, the sample mean is used to estimate population mean. In other words, we say that the sample mean is an estimator of population mean.

---

#### Definition 1.9 (Sample Mean).

Let  $x_1, x_2, x_3, \dots, x_n$  be a sample of data points. We define sample mean of the sample data points ( $\bar{x}$ ) as the following:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i.$$

Also, we define sample variance of the sample data points ( $s^2$ ) as:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2.$$

Moreover, the standard deviation of the sample of data points ( $s$ ) is:

$$s = \sqrt{s^2}, \quad \text{for } s > 0.$$


---

Now, let's move to variance. It refers to the expected value of the squared deviation from the mean of a random variable in a population. Similarly, we do have sample variance as well, which is the expected value of the squared deviation from the mean of a random variable in a selected sample. At this point, we can still use sample variance to estimate population variance with adjustment, because the sample variance may differ significantly based on what data points are chosen from that population.

---

#### Definition 1.10 (Sample Variance).

Let  $x_1, x_2, x_3, \dots, x_n$  be a sample of data points, we define sample variance of the sample data points ( $s^2$ ) as:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2, \quad \text{where } \bar{x} \text{ is the sample mean of the data points.}$$


---

Next is standard deviation. It is a measure of the amount of variation of the values of a variable about its mean. If standard deviation is relatively larger, then data points are

widely spread out from the mean. Otherwise, data points stay close from the mean. Also, standard deviation is obtained by taking squared root from variance which is dependent on the choices of data points as well. To use sample standard deviation as an estimator to population standard deviation, we still need to adjust it.

---

**Definition 1.11** (Sample Standard Deviation).

*Let  $x_1, x_2, x_3, \dots, x_n$  be a sample of data points. The standard deviation of the sample of data points ( $s$ ) is:*

$$s = \sqrt{s^2}, \quad \text{for } s > 0.$$


---

## Median and Mode

The median and mode are two important measures of central tendency used in statistics to summarize and understand data. The median represents the middle value in a sorted dataset, giving a sense of the center that is not affected by extreme values or outliers. In contrast, the mode is the value that appears most frequently in a dataset, making it useful for identifying common or repeated observations.

---

**Definition 1.12** (Median).

*Let:  $x_1, x_2, x_3, \dots, x_n$  be a collection of data points which is arranged in ascending order from the smallest value to the largest value (or descending order from the largest value to the smallest value in that collection). The median of the given collection of data points is the middle value in that collection, which equally spreads the collection into two parts. Half of all the collection values are above the median value and the rest of the values in the collection is below the median value.*

- *Case 1: when  $n$  is an odd number. (i.e. 1, 3, 11, 237, ...). Then, the median  $M$  is defined as:*

$$M = \frac{n+1}{2}, \text{ where } n \text{ represents the } n^{\text{th}} \text{ position.}$$

- *Case 2: when  $n$  is an even number (i.e. 2, 6, 100, 500, ...). Then, the median  $M$  is: the average value of  $\frac{n}{2}$ 's and  $\frac{n+2}{2}$ 's position, where  $n$  represents the  $n^{\text{th}}$  position.*
- 

Now, let's introduce mode.

---

**Definition 1.13** (Mode).

*It refers to a value that appears the most frequent than the appearance of all other values in a given dataset.*

---

## Percentile and Quartile

Percentiles and quartiles are statistical measures used to describe the distribution of data. A percentile indicates the value below which a given percentage of observations fall, helping to understand relative standing within a dataset. Quartiles, a specific type of percentile, divide the data into four equal parts (Q1, Q2/median, and Q3), providing insights into the spread and central tendency.

---

**Definition 1.14** (Percentile and Quartile).

*Let:  $x_1, x_2, \dots, x_n$  be a collection of data points in either ascending order. Percentile is denoted as:  $p^{th}$ , which indicates  $p\%$  of observations are below to a such value. Quartiles, are special cases of percentile which equally spread the collection of data into four parts. Each part contains 25% of the entire collection. More specifically, we define quartiles as the following:*

- $Q_1$ : the 25 percentile (or  $25^{th}$ ), which shows that 25% of the data points are below the value  $Q_1$ .
- $Q_2$ : the 50 percentile (or  $50^{th}$ ), which shows that 50% of the data points are below the value  $Q_2$ .
- $Q_3$ : the 75 percentile (or  $75^{th}$ ), which shows that 75% of the data points are below the value  $Q_3$ .
- $Q_2$  is qual to median.

*Moreover, we use  $Q_3 - Q_1$  to calculate interquartile range (I.P.R), which shows the spread of the whole data set.*

---

### Skewness and Symmetry

The two terms 'skewness' and 'symmetry' are used to describe the shape of probability distribution. There are two types of skewness: left (or negative) skew and right (or positive) skew. In real life, a famous distribution highly used in hypothesis testing which is  $\chi_n^2$  with  $n$  degrees of freedom, is right skewed probability distribution function. Another example regarding to symmetry is normal distribution such that its probability under its curve greater than  $\mu$  is same as the probability below than  $\mu$ . Now let's introduce the proper definition of skewness and symmetry.

---

**Definition 1.15** (Skewness).

*Skewness refers to such a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. The skewness value can be positive, zero, negative or undefined.*

---

Now, let's break down the main definition of skewness and symmetry:

**Definition 1.16** (Left (or Negative) Skew). 

---

*By observing given probability distribution curve, if the left tail of the curve is longer than the right tail the mass of the distribution is concentrated on the right of the figure, then we say that probability distribution is left skew or negative skew. (See figure below)*

---



Figure 1.1: Visualization of left skew probability distribution

**Definition 1.17** (Right (or Positive) Skew). 

---

*By observing given probability distribution curve, if the right tail of the curve is longer than the left tail the mass of the distribution is concentrated on the left of the figure, then we say that probability distribution is right skew or positive skew. (See figure below)*

---



Figure 1.2: Visualization of right skew probability distribution

Symmetry is a special case of skewness when the value of skewness is 0.

**Definition 1.18** (Symmetry). —————

*In statistics, symmetry is a probability distribution is reflected around a vertical line at some value of the random variable represented by the distribution. Probability under the curve below that value is equal to probability under the curve greater than that value. (see figure below)*

Since symmetry is a special case, so that it has a unique property as the following:

**Theorem 1.1** (Empirical Rule (or 68 – 95 – 99.7 Rule)). *For any symmetric (bell-shaped) curve, let  $\mu$  be its mean and  $\sigma$  be its standard deviation, the following probability set function is true:*

- 1.  $P(\mu - \sigma < X < \mu + \sigma) = 68.27\%$ ;
- 2.  $P(\mu - 2\sigma < X < \mu + 2\sigma) = 95.45\%$ ;
- 3.  $P(\mu - 3\sigma < X < \mu + 3\sigma) = 99.73\%$ .

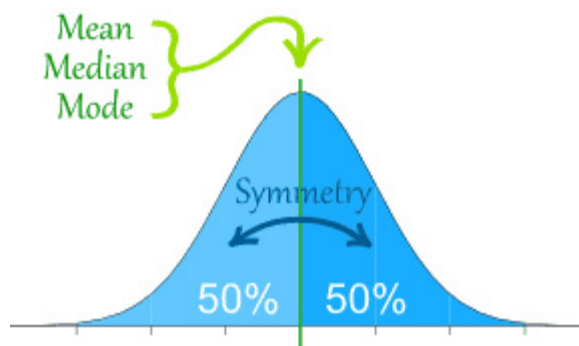


Figure 1.3: Visualization of symmetric probability distribution

### Practice Example

#### Example 1.1.

[Calculating Sample Mean, Variance and Standard Deviation] Let:  $x_1 = 1, x_2 = 3$  and  $x_3 = 7$ . Calculate the sample mean, sample variance and sample standard deviation for this collection of data points.

Solution (all results are kept in four digits):

By Definition 1.9, 1.10, 1.11, sample mean:

$$\bar{x} = \frac{1 + 3 + 7}{3} \approx 3.6667.$$

Then, we use sample mean to calculate sample variance:

$$s^2 = \frac{1}{3-1} \times [(1 - 3.6667)^2 + (3 - 3.6667)^2 + (7 - 3.6667)^2] \approx 9.3333.$$

Finally, we take the square root of sample variance to get sample deviation, and remember that  $s > 0$ :

$$s = \sqrt{s^2} \approx 3.0551.$$

#### Example 1.2.

[Median Calculation] Given two distinct collections of data points:  $S_1 = \{2, 4, 6\}$  and  $S_2 = \{1, 5, 16, 28\}$ . Calculate the median of both two sets.

Solution:

For  $S_1$ , since  $n = 3$  which is an odd number, so by *Definition 1.3*,  $M_{S_1} = 4$ . For  $S_2$ ,  $n = 4$  in this case, so that we need to calculate the average of  $\frac{n}{2}$  and  $\frac{n+1}{2}$ . Then,

$$M_{S_2} = \frac{5 + 16}{2} = 10.5.$$

**Example 1.3.**

Consider the data set  $S = \{4, 25, 30, 30, 30, 32, 32, 35, 50, 50, 50, 55, 60, 74, 110\}$ . Calculate its median and  $Q_1$  ( $25^{th}$ ).

Solution:

Simply counting the number of data points,  $n = 15$ , such that  $M_S = \frac{15+1}{2} = 8$ . Thus, the  $8^{th}$  value in the set which is 35.

Since we know the median of this collection of data points, we just need to find the median of the lower half of this data, which is exactly going to be 25 percentile ( $25^{th}$ ). In the lower half of the given collection (all values below the median),  $n_{lower} = 7$ . By *Definition 1.3*, then median of the lower half ( $25^{th}$ ) is going to be:

$$25^{th} = \frac{7+1}{2} = 4, \text{ the } 4^{th} \text{ position in the data set.}$$

Thus,  $Q_1$  ( $25^{th}$ ) = 30. To find  $Q_3$  ( $75^{th}$ ), apply the same strategy will guide you to find the correct answer, and we leave this as an exercise to you.

## 1.3 Graphical Techniques

In statistics, there are lots of types of graph to illustrate data, for example histograms and box-plots. This technique is used in the field of statistics for data visualization. Our objective is to both be able to identify some classical types of graph and interpret key statistical values (descriptive statistical values) from it.

### 1.3.1 Histograms

#### Introduction to Histograms

Histogram is a graphical representation of data that uses bars to display the frequency distribution of a dataset. Unlike bar graphs, which represent categorical data, histograms group numerical data into intervals (bins) and show how many values fall into each range. This makes histograms ideal for visualizing the shape, spread, and central tendency of continuous data, helping identify patterns such as symmetry, skewness, and outliers.

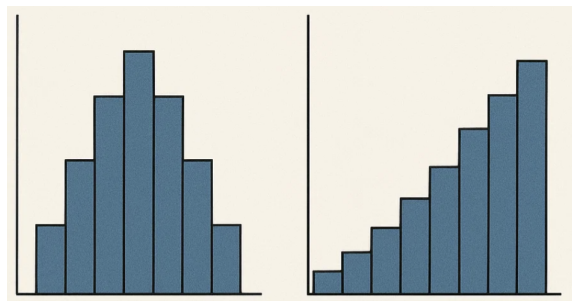


Figure 1.4: Visualization of histograms



## Advantages and Disadvantages of Histograms

- Advantages of Histograms:
  1. Histograms are easily to used for visualise data (relatively). It allows us to get the idea of the "shape" of distribution (i.e. skewness which will be discussed late in this section).
  2. It is also flexible that people are able to modify bin widths.
- Disadvantages of Histograms:
  1. It is not suitable for small data sets.
  2. The values from histograms close to breaking points are likely similar, in fact they need to be classified into different bins (i.e. Student A and B scores 79 and 80 respectively in STA258, we consider a breaking point between 79 and 80. The two students have similar score, but student A is *B+* and student B is *A-* in GPA from).

### Histograms with Skewness and Symmetry

A histogram visually represents the distribution of numerical data, making it a useful tool for assessing skewness and symmetry. It is quite straightforward to estimate the skewness of histograms by simply drawing a curve above bins on the histogram.

For a histogram to have a left (or negative) skew probability distribution:



Figure 1.5: Visualization of a histogram has a left (or negative) skew probability distribution

For a histogram to have a right (or positive) skew probability distribution:



Figure 1.6: Visualization of a histogram has a right (or positive) skew probability distribution

For a histogram to have a symmetric probability distribution:



Figure 1.7: Visualization of a histogram has a symmetric probability distribution

### 1.3.2 Box-Plots

A boxplot (or box-and-whisker plot) is a standardized way to display data distribution based on a five-number summary: minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum. The box represents the interquartile range (IQR), while the whiskers show variability outside Q1 and Q3. Outliers are plotted as individual points. Boxplots efficiently compare distributions and highlight skewness, spread, and outliers. (See figure below)

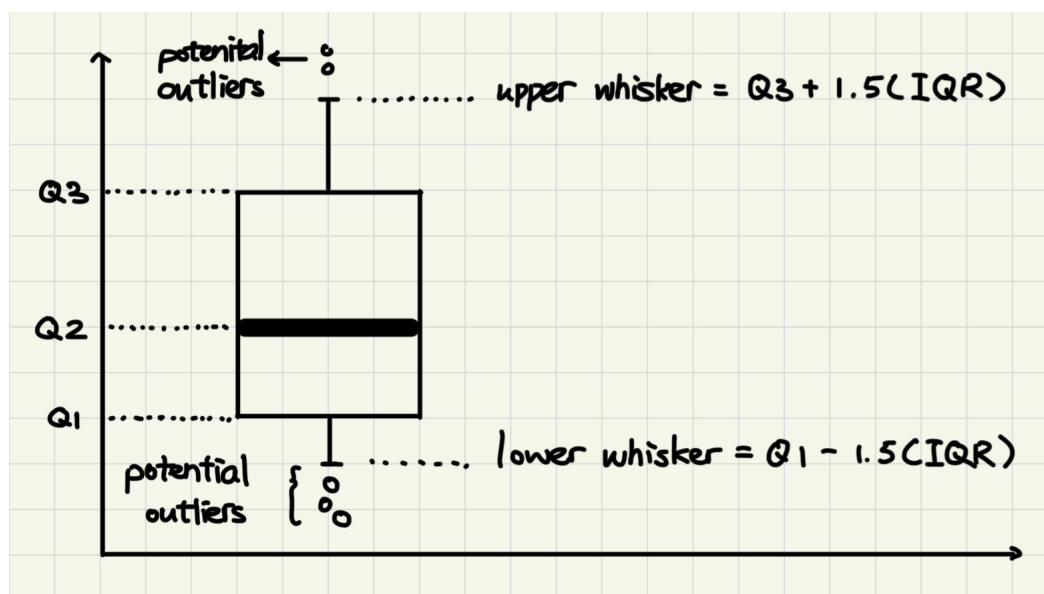


Figure 1.8: Visualization of a box-plot

Similar to histograms, we can still obtain information about skewness and symmetry, by observing the cut from the line of  $Q_2$ .

If the median ( $Q_2$ ) cuts the box with upper area smaller than lower area, then we say that box-plot with left skew probability distribution. Or, if the median ( $Q_2$ ) cuts the box with upper area larger than lower area, then we say that box-plot with right skew probability distribution.

Otherwise, if the median ( $Q_2$ ) cuts the box with upper area equal to lower area, then we say that box-plot with symmetric probability distribution.



Figure 1.9: Visualization of a box-plot with skew and symmetric probability distribution

## 1.4 Introduction to R

R is used for data manipulation, statistics, and graphics. It is made of: operations ( $+$ ,  $-$ ,  $<$ ) which is for calculations on vectors, arrays and matrices; a huge collection of functions; fa-

cilities for making unlimited types quality graphs; user contributed packages (sets of related functions); the ability to interface with procedures written in C, C+, or FORTRAN and to write additional primitives. R is also an open-source computing package which has seen a huge growth in popularity in the last few years (Please use this website: <https://cran.r-project.org>, to download R).

### What is R-studio?

RStudio is a relatively new editor specially targeted at R. RStudio is cross-platform, free and open-source software (Please use: <https://www.rstudio.com>, to download Rstudio).

### Make a Histogram Using R-studio

This is just a demonstration of how to start and use R-studio.

1. First of all, we need to know which dataset are we going to make into a histogram. In this case, as an example, we are going to use the waiting time in faithful in R-studio.
2. For any dataset, use the code: `names(faithful)` to get it. (inside the parentheses, type the names of variables you want in faithful dataset)
3. Then, we proceed with the code: `hist(faithful$waiting)` to get a basic plot.

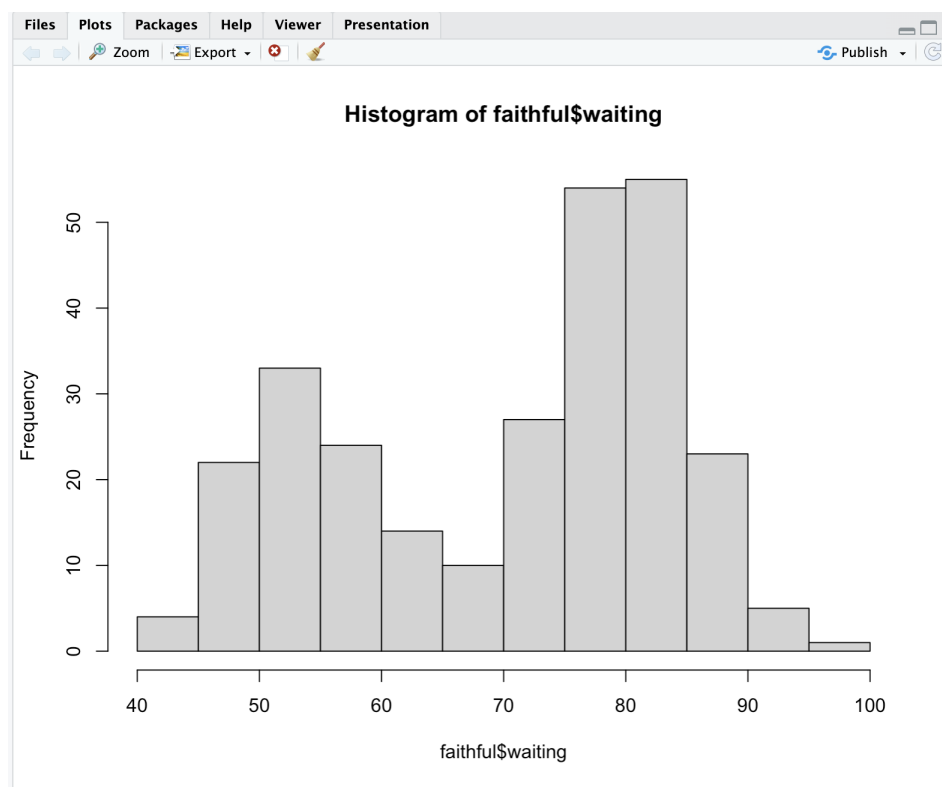


Figure 1.10: R-studio first three steps (by following the instructions, you should get this histogram)

4. Furthermore, we can also get more information. For example, by keep proceeding with the code: `hist(faithful$waiting,plot=FALSE)$breaks`, R-studio will show you all the breaking points between histogram cells.

```
> hist(faithful$waiting)
> hist(faithful$waiting,plot=FALSE)$breaks
[1] 40 45 50 55 60 65 70 75 80 85 90 95 100
```

Figure 1.11: R-studio the forth step(by following the instructions, you should get this histogram)

## Chapter 2

# Sampling Distributions Related to a Normal Population

Previously, we have introduced lots of definitions and given you a rough idea about what really statistics it and what people do in statistics. Now, we are going to proceed statistical distributions.

### 2.1 Normal Distribution

In probability theory and statistics, normal distribution also called Gaussian distribution which is discovered by a famous German mathematician Johann Carl Friedrich Gauss in 1809. It is one of the most important distribution that used to approximate other types of probability distribution, such as binomial, hypergeometric, inverse (or negative) hypergeometric, negative binomial and Poisson distribution. Generally, it is denote as  $N(\mu, \sigma^2)$  with probability density function as the following:

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

Formally, let's begin with its definition:

**Definition 2.1** (Normal Distribution).

Suppose a random variable  $X \sim N(\mu, \sigma^2)$ , then  $E(X) = \mu$  and  $Var(X) = \sigma^2$ . And  $-\infty < \mu < \infty, \sigma^2 > 0$ . Moreover,  $X$  has probability density function as:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \text{ for } -\infty < x < \infty \text{ (same as above).}$$

The only special case of normal distribution is standard normal distribution, such that a random variable  $Y \sim N(\mu = E(Y) = 0, \sigma^2 = Var(Y) = 1)$ , then  $Y$  has probability density function as:

$$f(y) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{y^2}{2}}.$$

## 2.2 Gamma and Chi-square Distribution

The Chi-square and Gamma distributions are two fundamental probability distributions widely used in statistical theory and applications. The Gamma distribution is a continuous distribution characterized by its shape and scale parameters, making it versatile for modeling waiting times and various positively skewed data. The Chi-square distribution, a special case of the Gamma distribution, arises naturally in the context of hypothesis testing and confidence interval estimation, especially in tests involving variance and categorical data.

### Gamma Distribution

---

**Definition 2.2** (Gamma Distribution).

Suppose a random variable  $X$  is Gamma distributed with  $\alpha > 0$  (shape parameter) and  $\beta > 0$  (scale parameter) if and only if the probability density function of  $X$  is

$$f(x) = \frac{x^{\alpha-1} e^{-\frac{x}{\beta}}}{\beta^\alpha \Gamma(\alpha)}, \text{ for } 0 < x < \infty.$$

Then,  $E(X) = \alpha\beta$ ,  $Var(X) = \alpha\beta^2$  and its moment generating function is  $M_X(t) = \frac{1}{(1-\beta t)^\alpha}$ , for  $t < \frac{1}{\beta}$ .

---

Now, let's introduce some properties of Gamma function:

- Gamma function (**not a distribution**):

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt, \text{ for } x > 0.$$

- Properties

- 1.  $\Gamma(x) = x \cdot \Gamma(x-1)$ ;
- 2. For all  $n \in \mathbb{N}$ ,  $\Gamma(n) = (n-1)!$ ;
- 3.  $\Gamma(\frac{1}{2}) = \sqrt{\pi}$ .

### Chi-square Distribution

Here is its formal definition:

---

**Definition 2.3** (Chi-square Distribution).

A random variable  $X$  has a Chi-squared distribution with  $n$  degrees of freedom ( $\chi_n^2$ ) if and only if  $X$  is a random variable with a Gamma distribution with parameters  $\alpha = \frac{n}{2}$  and  $\beta = 2$ . Then, the probability density function of  $X$  is given by

$$f(x) = \frac{1}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}.$$

Moreover,  $E(X) = n$ ,  $Var(X) = 2n$  and moment generating function of  $X$  is  $M_X(t) = (1 - 2t)^{-\frac{n}{2}}$ , for  $t < \frac{1}{2}$ .

We claim that Chi-square distribution is a special case of Gamma distribution with  $\alpha = \frac{n}{2}$  and  $\beta = 2$ . Now, let's prove it by using moment generating function.

The proof is quite straightforward as the following shows:

*Proof.* Suppose  $X \sim \text{Gamma}(\alpha = \frac{n}{2}, \beta = 2)$ .

Then the following moment generating function holds for  $X$ :

$$M_X(t) = (1 - 2t)^{-\frac{n}{2}}, \text{ for } t < \frac{1}{2}.$$

Compare the moment generating function of  $X$  under Gamma distribution with Chi-square distribution, we can conclude that  $X \sim \chi_n^2$ .  $\square$

### Obtaining Chi-square Distribution by Normal Distribution

Previously, we showed how to use Gamma distribution to get Chi-square distribution by moment generating function method. Now, let's do something interestingly, to use normal distribution to get Chi-square distribution. We will begin with a theorem, then prove it.

#### Theorem 2.1.

Suppose a random variable  $Z$  is standard normally distributed, such that  $Z \sim N(0, 1)$ . Then,  $Z^2$  is Chi-square distributed with 1 degree of freedom, so that  $Z^2 \sim \chi_1^2$ .

The proof of Theorem 2.1 isn't that trivial to see. We still need moment generating function, but in a different way. Before we get into the proper proof, let's grab everything we need:

- 1. Recall STA256 about how to get moment generating function for a given continuous random variable that:

$$M_Z(t) = \int_{-\infty}^{\infty} e^{tx} f_X(x) dx.$$

- 2. We also need Gaussian integral:

$$\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}; \tag{2.2.1}$$

$$\int_{-\infty}^{\infty} e^{-kx^2} dx = \sqrt{\frac{\pi}{k}}, \text{ for } k > 0; \tag{2.2.2}$$

$$\int_{-\infty}^{\infty} e^{kx^2} dx = \sqrt{\frac{\pi}{-k}}, \text{ for } k < 0. \tag{2.2.3}$$



*Proof.* Suppose that  $Z \sim N(0, 1)$ , then  $f_Z(z) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{z^2}{2}}$ .

The moment generating function (MGF) of  $Z^2$  is:

$$\begin{aligned} M_{Z^2}(t) &= \mathbb{E} \left( e^{tZ^2} \right) \\ &= \int_{-\infty}^{\infty} e^{tz^2} f_Z(z) dz \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tz^2} e^{-\frac{z^2}{2}} dz \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\left(\frac{1}{2}-t\right)z^2} dz. \end{aligned}$$

Apply substitution with  $u = z\sqrt{\frac{1}{2}-t}$ ,  $dz = \frac{du}{\sqrt{\frac{1}{2}-t}}$ :

$$\begin{aligned} M_{Z^2}(t) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-u^2} \cdot \frac{1}{\sqrt{\frac{1}{2}-t}} du \\ &= \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\sqrt{\frac{1}{2}-t}} \int_{-\infty}^{\infty} e^{-u^2} du \\ &= \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\sqrt{\frac{1}{2}-t}} \cdot \sqrt{\pi} \\ &= \frac{1}{\sqrt{1-2t}}. \end{aligned}$$

This is the MGF of a chi-squared distribution with 1 degree of freedom,  $Z^2 \sim \chi_1^2$ . □

Now, we can do another proof by using Theorem 2.1.

---

**Theorem 2.2.**

Suppose  $Z_1, Z_2, \dots, Z_n \stackrel{i.i.d.}{\sim} N(0, 1)$ , then the sum of  $n$  independent  $Z^2$  is going to be Chi-square distributed with  $n$  degrees of freedom, as the following:

$$\sum_{i=1}^n Z_i^2 \sim \chi_n^2.$$


---

We need Theorem 2.1 to prove this, but it going to be easier.

*Proof.* Suppose  $Z \sim \mathcal{N}(0, 1)$ , then its probability density function is:

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}.$$

Let  $\delta = \sum_{i=1}^n Z_i^2$ , where  $Z_1, \dots, Z_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$ . The moment generating function (MGF) of  $\delta$  is:

$$\begin{aligned} M_\delta(t) &= \mathbb{E} \left[ e^{t\delta} \right] \\ &= \mathbb{E} \left[ e^{t(Z_1^2 + \dots + Z_n^2)} \right] \\ &= \mathbb{E} \left[ \prod_{i=1}^n e^{tZ_i^2} \right]. \end{aligned}$$

Since  $Z_1, \dots, Z_n$  are independent and identically distributed:

$$\begin{aligned} M_\delta(t) &= \prod_{i=1}^n \mathbb{E} \left[ e^{tZ_i^2} \right] \\ &= \prod_{i=1}^n M_{Z_i^2}(t). \end{aligned}$$

From Theorem 2.1, we know  $Z_i^2 \sim \chi_1^2$  with MGF  $(1 - 2t)^{-1/2}$ , therefore:

$$\begin{aligned} M_\delta(t) &= \prod_{i=1}^n (1 - 2t)^{-1/2} \\ &= (1 - 2t)^{-n/2}. \end{aligned}$$

This is exactly the MGF of a chi-squared distribution with  $n$  degrees of freedom, proving that  $\delta \sim \chi_n^2$  as required.  $\square$

Here is the last theorem for Chi-square and normal distribution, but we won't show you the proof due to its complexity. For people who are interested in that, please see STA260 lecture notes or power point slide to figure out.

---

**Theorem 2.3.**

Let  $n$  be sample size,  $s^2$  be sample variance and  $\sigma^2$  be population variance, then  $\frac{(n-1)s^2}{\sigma^2}$  is Chi-square distributed with  $n - 1$  degrees of freedom. As the following:

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi_{n-1}^2.$$


---

## 2.3 Student's t-Distribution and F-Distribution

The t-distribution and F-distribution are essential tools in inferential statistics, particularly in the context of hypothesis testing and variance analysis. The t-distribution, which resembles the normal distribution but with heavier tails, is primarily used when estimating

population means in situations where the sample size is small and the population standard deviation is unknown. On the other hand, the F-distribution is used to compare variances between two populations and plays a central role in analysis of variance (ANOVA) and regression analysis.

### Student's t-Distribution

#### Definition 2.4 (Student's t-Distribution).

Suppose  $X$  is  $t$ -distributed with  $n$  degrees of freedom, then the probability density function of  $X$  is given by:

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

Alternatively, define a new variable  $T$  is the following:

$$T = \frac{W}{\sqrt{\frac{V}{r}}}, \text{ for } W \sim N(0, 1) \text{ and } V \sim \chi_r^2.$$

Or suppose  $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} N(\mu, \sigma^2)$ , then  $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ . Thus,

$$T = \frac{\bar{x} - \mu}{\left(\frac{s}{\sqrt{n}}\right)}.$$

Same as normal distribution, student's  $t$ -distribution is also symmetric. Also, as the degrees of freedom of  $t$ -distribution getting larger, the curve of student's  $t$ -distribution getting closer to standard normal distribution.

### F-Distribution

#### Definition 2.5.

We define a new variable  $F$  as the following shows:

$$F = \frac{\left(\frac{W_1}{v_1}\right)}{\left(\frac{W_2}{v_2}\right)} \sim F_{v_1, v_2}; \text{ for } W_1 \sim \chi_{v_1}^2 \text{ and } W_2 \sim \chi_{v_2}^2; \text{ also both } W_1 \text{ and } W_2 \text{ are independent.}$$

Alternatively, we select two samples (with same population variance) with size  $n$  and  $m$ , and also sample variance  $s_x$  and  $s_y$  respectively. Then,  $F$ -distribution is:

$$F = \frac{\left[\frac{\left(\frac{n-1}{2}\right)s_x^2}{n-1}\right]}{\left[\frac{\left(\frac{m-1}{2}\right)s_y^2}{m-1}\right]} \sim F_{n-1, m-1}.$$

Both student's t-distribution and F-distribution are highly used in inferential statistics, until confidence interval, testing hypothesis and ANOVA analysis, these two distributions will come to play a lot. At this point, just guarantee that you know how to obtain those distribution from random given information is sufficient.

## Chapter 3

# The Central Limit Theorem

The Central Limit Theorem (CLT) is one of the most important results in probability and statistics. It states that, given a sufficiently large sample size, the distribution of the sample mean of independent and identically distributed (i.i.d.) random variables approaches a normal distribution, regardless of the shape of the original distribution. Real-life Application of Central Limit Theorem in Financial Analysis. The CLT is often used by financial experts to examine stock market results.

Now, let's discuss Central Limit Theorem with more details. Suppose we have a finite number of populations and each population follows a distribution with population mean  $\mu$  and population variance  $\sigma^2$ . Then we take samples of same size  $n$  from each population, such that we have  $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_m$  from population group 1 to  $m$ , respectively. Next, we make a histogram using the large collection of sample taken from each population group. Then, what we are doing right now is sampling distribution of  $\bar{x}$ . As a result,  $\bar{x}$  follows a normal distribution with mean  $\mu_{\bar{x}} = \mu$  and variance  $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$ , which is denoted as the following:

$$\bar{x} \sim N(\mu_{\bar{x}} = \mu, \sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}).$$

Figure 3.1 below shows the entire procedure about the Central Limit Theorem.

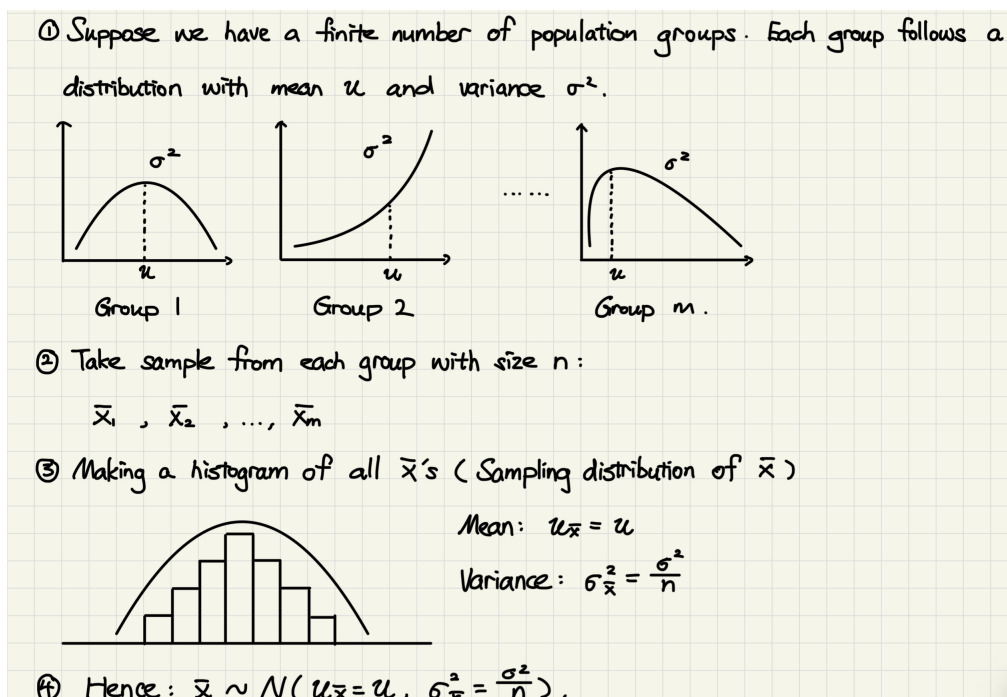


Figure 3.1: Procedure of the Central Limit Theorem

Now, let's begin with the proper definition of Central Limit Theorem.

**Definition 3.1** (Central Limit Theorem). —————

Let  $X_1, X_2, \dots, X_n$  be independent and identically distributed random variables with  $E(X_i) = \mu$  and  $\text{Var}(X_i) = \sigma^2 < \infty$ . Then, we define the following:

$$U_n = \frac{\bar{X} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)} \sim N(\mu = 0, \sigma^2 = 1), \text{ where } \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i.$$

Then the distribution function of  $U_n$  converges to the standard Normal distribution function as  $n \rightarrow \infty$ . That is,

$$\lim_{n \rightarrow \infty} P(U_n \leq u) = \int_{-\infty}^u \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt; \text{ for all } u.$$

For this course in particular, we do not need to pay that much attention to the proving part of the definition above. However, we use Central Limit Theorem to approximate distributions. Here are the two important approximations:

- $\bar{X}_n \approx N\left(\mu, \frac{\sigma^2}{n}\right);$
- $T = \sum_{i=1}^n X_i \approx N(n\mu, n\sigma^2).$

A reminder that the distribution of  $U_n$  in definition 3.1 and the two approximation of distribution above are extremely important in this course, until later chapters you may see some materials that are similar.

## Chapter 4

# Normal Approximation to the Binomial Distribution

### 4.1 Introduction

**Definition 4.1** (Statistic). 

---

*A statistic is a function of the observable random variables in a sample and known constants. Since statistics are functions of the random variables observed in a sample, they themselves are random variables. As such, all statistics have a corresponding probability distribution, which we refer to as their sampling distribution.*

---

**Review from STA256****Bernoulli Distribution:**

A Bernoulli trial is a single experiment with two outcomes:

- Success:  $X = 1$  with probability  $p$
- Failure:  $X = 0$  with probability  $1 - p$

$X = x$	0	1
$P(X = x)$	$1 - p$	$p$

The probability mass function (PMF) is:

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

**Binomial Distribution:**

A binomial distribution arises from  $n$  independent Bernoulli trials. Let:

$X$  = number of successes in  $n$  trials



Then:

$$X \sim \text{Binomial}(n, p)$$

where:

- Each trial results in either success (with probability  $p$ ) or failure (with probability  $1 - p$ )
- $X \in \{0, 1, \dots, n\}$

The PMF is:

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

### Moment Generating Function (MGF):

The moment generating function (MGF) of a random variable  $X$  is defined as:

$$M_X(t) = \mathbb{E}[e^{tX}]$$

The MGF uniquely characterizes the distribution of  $X$  (if it exists in an open interval around 0), and it can be used to compute moments such as the mean and variance.

## 4.2 Bernoulli Distribution

Bernoulli random variable is a discrete random variable that has exactly two possible outcomes which are either a **success** or a **failure**. An experiment in which there are exactly 2 outcomes (which are success or failure) is called a **Bernoulli trial**.

When  $x = 1$  we have a success and when  $x = 0$  we have a failure. The term success and failure are relative to the problem being studied.

### TIP: “success” need not be something positive

We chose to label a person who refuses to administer the worst shock a “success” and all others as “failures”. However, we could just as easily have reversed these labels. The mathematical framework we will build does not depend on which outcome is labeled a success and which a failure, as long as we are consistent.

Consider the random experiment of rolling a die once. Define the random variable:

$$X_i = \begin{cases} 1 & \text{if the } i\text{-th roll is a six,} \\ 0 & \text{otherwise} \end{cases}$$

Then  $X_i \sim \text{Bernoulli}(p)$ , where  $p = P(\text{rolling a six})$ .

Let  $X \sim \text{Bernoulli}(p)$ . The mass function of  $X$  is

$$P(X = x) = p^x(1 - p)^{1-x}, \quad x = 0, 1$$

where  $p$  represents the probability of success.

**Definition 4.2** (Mean and Variance of a Bernoulli Random Variable). 

---

Let  $X \sim \text{Bernoulli}(p)$ . The mean of  $X$  is

$$E(X) = \mu = p$$

and the variance of  $X$  is

$$\text{Var}(X) = \sigma^2 = p(1 - p)$$


---

To support the earlier result, we now provide a derivation of the mean, variance, and standard deviation of a Bernoulli random variable.

Let  $X$  be a Bernoulli random variable with the probability of a success as  $p$ . Then

$$\begin{aligned} E[X] &= \mu = \sum_{i=1}^n x_i \cdot P(X = x_i) \\ &= 0 \cdot P(X = 0) + 1 \cdot P(X = 1) \\ &= 0 \cdot (1 - p) + 1 \cdot p \\ &= p \end{aligned}$$

Similarly, the variance of  $X$  can be computed:

$$\begin{aligned} V(X) &= \sigma^2 = \sum_{i=1}^k (x_i - \mu)^2 \cdot P(X = x_i) \\ &= (0 - p)^2 \cdot P(X = 0) + (1 - p)^2 \cdot P(X = 1) \\ &= p^2(1 - p) + (1 - p)^2 p \\ &= p(1 - p) \end{aligned}$$

The standard deviation is

$$\begin{aligned} \sigma &= \sqrt{\sigma^2} \\ &= \sqrt{p(1 - p)} \end{aligned}$$

### 4.3 Sampling Distribution of the Sum and MGF Derivation

Consider determining the sampling distribution of the sample total:

$$T_n = X_1 + X_2 + \cdots + X_n$$

Suppose  $X_i \stackrel{iid}{\sim} \text{Bernoulli}(p)$ . Then the moment-generating function of  $T_n$  is:

$$\begin{aligned} M_{T_n}(t) &= \mathbb{E}[e^{tT_n}] \\ &= \mathbb{E}\left[e^{t(X_1+X_2+\cdots+X_n)}\right] \\ &= \mathbb{E}\left[e^{tX_1}e^{tX_2}\cdots e^{tX_n}\right] \quad (\text{independence}) \\ &= \mathbb{E}[e^{tX_1}] \cdot \mathbb{E}[e^{tX_2}] \cdots \mathbb{E}[e^{tX_n}] \\ &= M_{X_1}(t) \cdot M_{X_2}(t) \cdots M_{X_n}(t) \\ &= \left[pe^t + (1-p)\right]^n \end{aligned}$$

Since this is the MGF of a binomial random variable with parameters  $n$  and  $p$ , we conclude:

$$T_n \sim \text{Binomial}(n, p)$$

#### Example: Binomial Distribution from Die Rolls

We can think of rolling a die  $n$  times as an example of the binomial setting. Each roll gives either a six (a “success”) or a number different from six (a “failure”).

Knowing the outcome of one roll doesn’t tell us anything about the others, so the  $n$  rolls are independent.

If we call a six a success, then:

- The probability of success on each trial is  $p = P(\text{rolling a six}) = \frac{1}{6}$
- The probability of failure is  $1 - p = \frac{5}{6}$

Let  $Y$  be the number of sixes rolled in  $n$  trials. Then  $Y \sim \text{Binomial}(n, p)$ , and the distribution of  $Y$  is called a **binomial distribution**.

### 4.4 Binomial Distribution

In section 4.2 we learnt about Bernoulli random variables in which we were interested in the outcome of just a single trial. A **binomial random variable** is a generalization of several independent Bernoulli trials. Instead of performing just a single Bernoulli trial and observing whether we have a success or not, we are now performing several Bernoulli trials and observing whether we have a certain number of successes and failures. The **binomial distribution** describes the probability of having exactly  $k$  successes in  $n$  independent Bernoulli

trials with probability of a success  $p$ .

*Let  $X \sim \text{Bin}(n, p)$ . The probability of observing  $x$  successes in these  $n$  independent trials is given by*

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

*where*

- $n$  represents the number of trials,
- $x$  represents the number of successes,
- $p$  represents the probability of success on any given trial,

$$\binom{n}{x} = \frac{n!}{x!(n-x)!} \quad \text{is the binomial coefficient.}$$

**Definition 4.3** (Mean and Variance of a Binomial Random Variable). 

---

*Let  $X \sim \text{Bin}(n, p)$ . The mean of  $X$  is*

$$E(X) = \mu = np$$

*and the variance of  $X$  is*

$$\text{Var}(X) = \sigma^2 = np(1 - p)$$


---

#### 4.4.1 Visualizing the PMF of Binomial Distributions

**R code:**

```
## Pmf of Binomial with n=10 and p=1/6.

x <- seq(0, 10, by=1);
y <- dbinom(x, 10, 1/6);

plot(x, y, type= "p", col="blue", pch=19);
```

#### Probability Mass Functions (PMFs) for increasing $n$ :

The following plots display the probability mass functions (PMFs) for a binomial distribution with  $p = \frac{1}{6}$  and increasing values of  $n$ . As  $n$  increases, the binomial distribution begins to resemble a normal distribution.



Figure 4.1: PMF of Binomial distribution with  $n = 10$  and  $p = \frac{1}{6}$ .



Figure 4.2: PMF of Binomial distribution with  $n = 50$  and  $p = \frac{1}{6}$ .



Figure 4.3: PMF of Binomial distribution with  $n = 100$  and  $p = \frac{1}{6}$ .



Figure 4.4: PMF of Binomial distribution with  $n = 300$  and  $p = \frac{1}{6}$ .

## 4.5 Sampling Distribution of a Sample Proportion and the Normal Approximation

When studying categorical data, we are often interested not just in individual outcomes, but in the proportion of successes observed in a sample. Understanding how this proportion behaves across repeated samples is crucial for making inferences about a population. In this section, we explore the sampling distribution of a sample proportion and how it can be approximated by a normal distribution under certain conditions.

Draw a *Simple Random Sample (SRS)* of size  $n$  from a large population that contains proportion  $p$  of “successes”. Let  $\hat{p}$  be the **sample proportion** of successes:

$$\hat{p} = \frac{\text{number of successes in the sample}}{n}$$

Then:

- The **mean** of the sampling distribution of  $\hat{p}$  is  $p$ .
- The **standard deviation** of the sampling distribution is  $\sqrt{\frac{p(1-p)}{n}}$ .

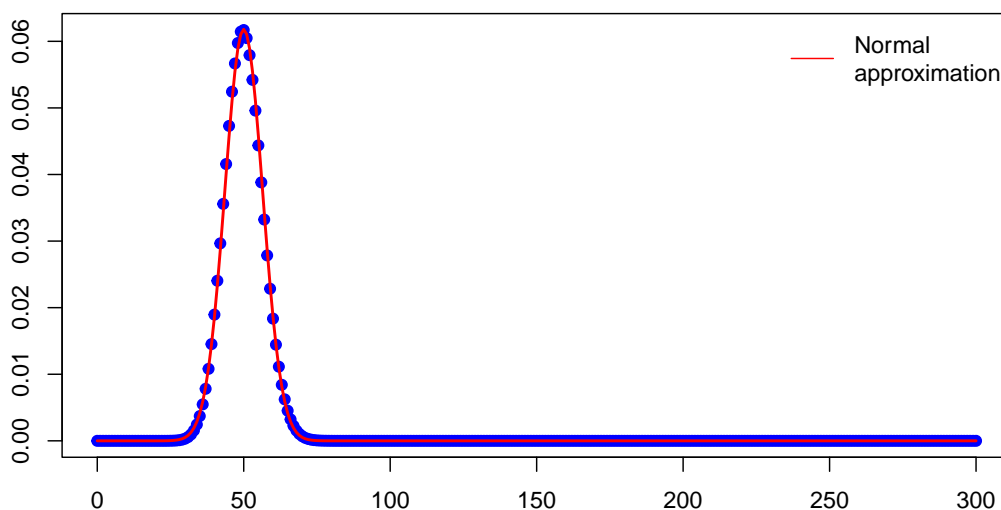


Figure 4.5: Binomial distribution with  $n = 300$ ,  $p = \frac{1}{6}$  and its Normal approximation.

According to the Central Limit Theorem (CLT), the sampling distribution of a sample proportion becomes approximately normal as the sample size increases.

That is:

$$\hat{p} \sim \mathcal{N}\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

This approximation is most accurate when both  $np \geq 10$  and  $n(1 - p) \geq 10$ . These are called the **success-failure conditions**.

*Key Point:* When the success-failure conditions are met, the normal approximation to the sampling distribution of  $\hat{p}$  can be used for probability calculations.

### Conditions for Using the Normal Approximation

Suppose  $X \sim \text{Binomial}(n, p)$ . Then:

$$\mu = np, \quad \sigma^2 = np(1 - p)$$

**Binomial probabilities can be approximated by the normal distribution:**

$$X \approx \mathcal{N}(np, np(1 - p))$$

This approximation is *useful for large n* and valid under the following conditions:

#### Standard Conditions

The binomial setting holds (i.e., independent trials, fixed  $n$ , same probability  $p$ ) and

$$np \geq 10 \quad \text{and} \quad np(1 - p) \geq 10$$

Alternatively, a more conservative criterion for using the normal approximation is:

$$n > 9 \cdot \left( \frac{\max(p, 1 - p)}{\min(p, 1 - p)} \right)$$

These ensure that the binomial distribution is sufficiently symmetric and smooth to approximate with the normal distribution.

We derive the sampling distribution of  $\hat{p}$  using properties of the Bernoulli distribution.

### Bernoulli Distribution (Binomial with $n = 1$ )

$$X_i = \begin{cases} 1 & \text{if the } i\text{-th roll is a six} \\ 0 & \text{otherwise} \end{cases}$$

$$\mu = \mathbb{E}(X_i) = p, \quad \sigma^2 = \text{Var}(X_i) = p(1 - p)$$

Let  $\hat{p}$  be our estimate of  $p$ . Note that  $\hat{p} = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}$ . Let  $\hat{p} = \frac{\# \text{ successes } (X)}{\text{sample size } (n)}$ . Recall that for  $X \sim \text{Binomial}(n, p)$ :

$$X \dot{\sim} \mathcal{N}(np, np(1 - p))$$

Let  $\hat{p} = \frac{X}{n}$

**Mean of  $\hat{p}$ :**

$$\mathbb{E}(\hat{p}) = \mathbb{E}\left(\frac{X}{n}\right) = \frac{1}{n} \cdot \mathbb{E}(X) = \frac{1}{n} \cdot np = p$$



**Variance of  $\hat{p}$ :**

$$\text{Var}(\hat{p}) = \text{Var}\left(\frac{X}{n}\right) = \frac{1}{n^2} \cdot \text{Var}(X) = \frac{1}{n^2} \cdot np(1-p) = \frac{p(1-p)}{n}$$

By the Central Limit Theorem (CLT), for sufficiently large  $n$ :

$$\hat{p} \sim \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$$

**Standardization of  $\hat{p}$ :**

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

If  $n$  is large, then by the Central Limit Theorem:

$$\bar{X} \approx \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \Rightarrow \hat{p} \sim \mathcal{N}\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

---

**Example 4.1.**

[Normal Approximation for Proportions] In the last election, a state representative received 52% of the votes cast. One year after the election, the representative organized a survey that asked a random sample of 300 people whether they would vote for him in the next election. If we assume that his popularity has not changed, what is the probability that more than half the sample would vote for him?

**Solution 1 (using Normal Approximation)**

We want to determine the probability that the sample proportion is greater than 50%. In other words, we want to find  $P(\hat{p} > 0.50)$ .

We know that the sample proportion  $\hat{p}$  is roughly Normally distributed with mean  $p = 0.52$  and standard deviation

$$\sqrt{p(1-p)/n} = \sqrt{(0.52)(0.48)/300} = 0.0288.$$

Thus, we calculate

$$\begin{aligned} P(\hat{p} > 0.50) &= P\left(\frac{\hat{p} - p}{\sqrt{p(1-p)/n}} > \frac{0.50 - 0.52}{0.0288}\right) \\ &= P(Z > -0.69) = 1 - P(Z < -0.69) \quad (Z \text{ is symmetric}) \\ &= P(Z > -0.69) = 1 - P(Z > 0.69) \\ &= 1 - 0.2451 = 0.7549. \end{aligned}$$

If we assume that the level of support remains at 52%, the probability that more than half the sample of 300 people would vote for the representative is 0.7549.

**R code (Normal approximation)** Just type in the following:

```
1 - pnorm(0.50, mean = 0.52, sd = 0.0288)
## [1] 0.7562982
```

Recall that, `pnorm` will give you the area to the left of 0.50, for a Normal distribution with mean 0.52 and standard deviation 0.0288.

### Solution 2 (using Binomial)

We want to determine the probability that the sample proportion is greater than 50%. In other words, we want to find  $P(\hat{p} > 0.50)$ . We know that  $n = 300$  and  $p = 0.52$ .

Thus, we calculate

$$\begin{aligned} P(\hat{p} > 0.50) &= P\left(\frac{\sum_{i=1}^n x_i}{n} > 0.50\right) \\ &= P\left(\sum_{i=1}^{300} x_i > 150\right) \\ &= 1 - P\left(\sum_{i=1}^{300} x_i \leq 150\right) \end{aligned}$$

(it can be shown that  $Y = \sum_{i=1}^{300} x_i$  has a Binomial distribution with

$$\begin{aligned} n &= 300 \text{ and } p = 0.52) \\ &= 1 - F_Y(150) \end{aligned}$$

**R code (using Binomial distribution)** Just type in the following:

```
1- pbinom(150, size = 300, prob = 0.52);
## [1] 0.7375949
```

Recall that, `pbinom` will give you the CDF at 150, for a Binomial distribution with  $n = 300$  and  $p = 0.52$ .

### Solution 3 (using continuity correction)

We have that  $n = 300$  and  $p = 0.52$ . Thus, we calculate

$$\begin{aligned} P(\hat{p} > 0.50) &= P\left(\frac{\sum_{i=1}^n x_i}{n} > 0.50\right) \\ &= P\left(\sum_{i=1}^{300} x_i > 150\right) \\ &= 1 - P\left(\sum_{i=1}^{300} x_i \leq 150\right) \end{aligned}$$

$$\begin{aligned}
& \text{(it can be shown that } Y = \sum_{i=1}^{300} x_i \text{ has a Binomial distribution with} \\
& n = 300 \text{ and } p = 0.52). \\
& \approx 1 - P\left(\sum_{i=1}^{300} x_i \leq 150.5\right) \quad \text{(continuity correction)} \\
& = 1 - P\left(\frac{\sum_{i=1}^{300} x_i}{n} \leq \frac{150.5}{300}\right) \\
& = 1 - P(\hat{p} \leq 0.5017) \\
& = 1 - P(Z \leq -0.6354) \quad \text{(Why?)}
\end{aligned}$$

**R code (Normal approximation with continuity correction)** Just type in the following:

```
1 - pnorm(0.5017, mean = 0.52, sd = 0.0288)
## [1] 0.7374216
```

Recall that, `pnorm` will give you the area to the left of 0.5017, for a Normal distribution with mean 0.52 and standard deviation 0.0288.

## 4.6 Continuity Correction

The normal distribution is continuous, while the binomial distribution is discrete. When we approximate a binomial probability using the normal distribution, this mismatch can lead to inaccuracy—especially near the boundaries of discrete values. A continuity correction improves the approximation by adjusting for this difference. In this section, we explore how and why this correction is applied.

Suppose that  $Y$  has a Binomial distribution with  $n = 20$  and  $p = 0.4$ . We will find the exact probabilities that  $Y \leq y$  and compare these to the corresponding values found by using two Normal approximations. One of them, when  $X$  is Normally distributed with  $\mu_X = np$  and  $\sigma_X = \sqrt{np(1-p)}$ . The other one,  $W$ , a shifted version of  $X$ .

For example,

$$P(Y \leq 8) = 0.5955987$$

As previously stated, we can think of  $Y$  as having approximately the same distribution as  $X$ .

$$P(Y \leq 8) \approx P(X \leq 8) = P\left[\frac{X - np}{\sqrt{np(1-p)}} \leq \frac{8 - 8}{\sqrt{20(0.4)(0.6)}}\right] = P(Z \leq 0) = 0.5$$

$$P(Y \leq 8) \approx P(W \leq 8.5) = P\left[\frac{W - np}{\sqrt{np(1-p)}} \leq \frac{8.5 - 8}{\sqrt{20(0.4)(0.6)}}\right] = P(Z \leq 0.2282) = 0.5902615$$

### Formula Sheet: Continuity Correction

Binomial Probability	Continuity Correction	Normal Approximation
$P(X = x)$	$P(x - 0.5 \leq X \leq x + 0.5)$	$P\left(\frac{x - 0.5 - \mu}{\sigma} \leq Z \leq \frac{x + 0.5 - \mu}{\sigma}\right)$
$P(X \leq x)$	$P(X \leq x + 0.5)$	$P\left(Z \leq \frac{x + 0.5 - \mu}{\sigma}\right)$
$P(X < x)$	$P(X \leq x - 0.5)$	$P\left(Z \leq \frac{x - 0.5 - \mu}{\sigma}\right)$
$P(X \geq x)$	$P(X \geq x - 0.5)$	$P\left(Z \geq \frac{x - 0.5 - \mu}{\sigma}\right)$
$P(X > x)$	$P(X \geq x + 0.5)$	$P\left(Z \geq \frac{x + 0.5 - \mu}{\sigma}\right)$
$P(a \leq X \leq b)$	$P(a - 0.5 \leq X \leq b + 0.5)$	$P\left(\frac{a - 0.5 - \mu}{\sigma} \leq Z \leq \frac{b + 0.5 - \mu}{\sigma}\right)$

#### Example 4.2.

Fifty-one percent of adults in the U. S. whose New Year's resolution was to exercise more achieved their resolution. You randomly select 65 adults in the U. S. whose resolution was to exercise more and ask each if he or she achieved that resolution. What is the probability that exactly forty of them respond yes?

We are given that  $p = 0.51$ ,  $n = 65$ , and we want to find  $P(X = 40)$  where  $X \sim \text{Binomial}(n = 65, p = 0.51)$ .

**Use Normal Approximation** We use normal approximation to the binomial. First, com-

pute the mean and standard deviation:

$$\begin{aligned}\mu &= np = 65 \times 0.51 = 33.15 \\ \sigma^2 &= np(1-p) = 65 \times 0.51 \times 0.49 = 16.485 \\ \sigma &= \sqrt{16.485} \approx 4.06\end{aligned}$$

We apply continuity correction:

$$\begin{aligned}P(X = 40) &= P(39.5 \leq X \leq 40.5) \\ &= P\left(\frac{39.5 - 33.15}{4.06} \leq Z \leq \frac{40.5 - 33.15}{4.06}\right) = P(1.56 \leq Z \leq 1.81)\end{aligned}$$

From the standard normal table:

$$= P(Z \leq 1.81) - P(Z \leq 1.56) = 0.0594 - 0.0352 = 0.0242$$

So the approximate probability is:

$$P(X = 40) \approx 0.0242$$


---

#### Normal Approximation to Binomial

Let  $X = \sum_{i=1}^n Y_i$  where  $Y_1, Y_2, \dots, Y_n$  are iid Bernoulli random variables. Note that  $X = n\hat{p}$ .

1.  $n\hat{p}$  is approximately Normally distributed provided that  $np \geq 10$  and  $n(1-p) \geq 10$ .
2. Another criterion is that the Normal approximation is adequate if

$$n > 9 \left( \frac{\text{larger of } p \text{ and } q}{\text{smaller of } p \text{ and } q} \right)$$

3. The expected value:  $E(\hat{p}) = np$ .
4. The variance:  $V(\hat{p}) = np(1-p) = npq$ .

## Chapter 5

# Law of Large Numbers

### 5.1 Convergence in Probability

**Definition 5.1** (Convergence in Probability). 

---

The sequence of random variables  $X_1, X_2, X_3, \dots, X_n, \dots$  is said to **converge in probability** to the constant  $c$ , if for every  $\epsilon > 0$ ,

$$\lim_{n \rightarrow \infty} P(|X_n - c| \leq \epsilon) = 1$$

or equivalently,

$$\lim_{n \rightarrow \infty} P(|X_n - c| > \epsilon) = 0$$

**Notation:**  $X_n \xrightarrow{P} c$

---

This concept plays a key role in the Law of Large Numbers, where the sample mean of independent and identically distributed random variables converges in probability to the population mean as the sample size grows.

**Definition 5.2** (Chebyshev's Inequality). 

---

Let  $X$  be a random variable with finite mean  $\mu$  and variance  $\sigma^2$ . Then, for any  $k > 0$ ,

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

Using complements:

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

---

## 5.2 Weak Law of Large Numbers (WLLN)

**Definition 5.3** (Weak Law of Large Numbers (WLLN)).

Let  $X_1, X_2, \dots$  be a sequence of independent and identically distributed random variables, each having finite mean  $E(X_i) = \mu$  and variance  $\text{Var}(X_i) = \sigma^2$ . Then, for any  $\epsilon > 0$ ,

$$P\left(\left|\frac{X_1 + X_2 + \dots + X_n}{n} - \mu\right| \geq \epsilon\right) \rightarrow 0 \quad \text{as } n \rightarrow \infty$$

**Notation:**  $\bar{X}_n \xrightarrow{P} \mu$

### Proof of the Weak Law of Large Numbers (WLLN)

We aim to show that for every  $\epsilon > 0$ ,

$$\lim_{n \rightarrow \infty} P(|\bar{X}_n - \mu| > \epsilon) = 0$$

where  $\bar{X}_n$  is the sample mean of  $n$  independent and identically distributed (i.i.d.) random variables with

$$E(X_i) = \mu, \quad \text{and} \quad \text{Var}(X_i) = \sigma^2.$$

Let

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

By the Central Limit Theorem (CLT), we know that

$$\bar{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

Now, applying **Chebyshev's Inequality**, which states that for any random variable  $X$  with mean  $\mu$  and variance  $\sigma^2$ ,

$$P(|X - \mu| > k) \leq \frac{\sigma^2}{k^2} \quad \text{for } k > 0,$$

to  $\bar{X}_n$ , we set  $k = \epsilon$ , and obtain:

$$P(|\bar{X}_n - \mu| > \epsilon) \leq \frac{\text{Var}(\bar{X}_n)}{\epsilon^2} = \frac{\sigma^2/n}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}.$$

Taking the limit as  $n \rightarrow \infty$ , we have:

$$\lim_{n \rightarrow \infty} P(|\bar{X}_n - \mu| > \epsilon) \leq \lim_{n \rightarrow \infty} \frac{\sigma^2}{n\epsilon^2} = 0.$$

Since probabilities are always non-negative, we conclude:

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

By the definition of convergence in probability,

$$\bar{X}_n \xrightarrow{P} \mu.$$

□

---

**Example 5.1.**

[Poisson Convergence via WLLN]

Let  $X_i$ , for  $i = 1, 2, 3, \dots$ , be independent Poisson random variables with rate parameter  $\lambda = 3$ . Prove that:

$$\bar{X}_n \xrightarrow{P} 3$$

**Properties of Poisson Distribution:**

$$E(X_i) = \lambda, \quad \text{Var}(X_i) = \lambda$$

In this case,  $\lambda = 3$ , so:

$$E(X_i) = \text{Var}(X_i) = 3$$

**Proof:**

We know:

$$E\left(\frac{X_1 + X_2 + \dots + X_n}{n}\right) = 3, \quad \text{and} \quad \text{Var}\left(\frac{X_1 + X_2 + \dots + X_n}{n}\right) = \frac{3}{n}$$

Applying Chebyshev's Inequality:

$$P\left(\left|\frac{X_1 + X_2 + \dots + X_n}{n} - 3\right| \geq \epsilon\right) \leq \frac{3}{n\epsilon^2}$$

Taking the limit as  $n \rightarrow \infty$ :

$$P\left(\left|\frac{X_1 + X_2 + \dots + X_n}{n} - 3\right| \geq \epsilon\right) \rightarrow 0$$

**Conclusion:**

$$\bar{X}_n \xrightarrow{P} 3$$


---





Figure 5.1: Simulation of running sample mean of Bernoulli( $p = 0.5$ ) trials over time.

### R Simulation Code (Single Sample Path)

```
n = 10
trial = seq(1, n, by = 1)
sample = rbinom(n, 1, 1/2)

plot(trial, cumsum(sample)/trial, type = "l", ylim = c(0,1), col = "blue")
points(trial, cumsum(sample)/trial, col = "red")
abline(h = 0.5, lty = 2, col = "black")
```



Figure 5.2: Simulation of 10 running sample means of Bernoulli( $p = 0.5$ ) trials converging over 100 trials.

### R Simulation Code (Multiple Sample Paths)

```

n = 100
trial = seq(1, 100, by = 1)

sample1 = rbinom(n, 1, 1/2)
sample2 = rbinom(n, 1, 1/2)
sample3 = rbinom(n, 1, 1/2)
sample4 = rbinom(n, 1, 1/2)
sample5 = rbinom(n, 1, 1/2)
sample6 = rbinom(n, 1, 1/2)
sample7 = rbinom(n, 1, 1/2)
sample8 = rbinom(n, 1, 1/2)

colors = rainbow(8)

plot(trial, cumsum(sample1)/trial, type = "l", col = colors[1], ylim = c(0,1))
lines(trial, cumsum(sample2)/trial, col = colors[2])
lines(trial, cumsum(sample3)/trial, col = colors[3])
lines(trial, cumsum(sample4)/trial, col = colors[4])
lines(trial, cumsum(sample5)/trial, col = colors[5])
lines(trial, cumsum(sample6)/trial, col = colors[6])
lines(trial, cumsum(sample7)/trial, col = colors[7])
lines(trial, cumsum(sample8)/trial, col = colors[8])
abline(h = 0.5, lty = 2, col = "black")

```

### Empirical Probability Insight

The Law of Large Numbers gives us empirical probabilities. Consider tossing a fair coin. Define the random variable  $X$  as:

$$X = \begin{cases} 1 & \text{heads up} \\ 0 & \text{tails up} \end{cases}$$

Then as we sample more and more values of  $X$ , the sample mean  $\bar{X}_n$  converges in probability to  $P(\text{heads up})$ , that is:

$$\bar{X}_n \xrightarrow{P} P(\text{heads up})$$

## Chapter 6

# One Sample Confidence Intervals on a Mean When the Population Variance is Known

### 6.1 Introduction

Statistical inference is concerned primarily with understanding the quality of parameter estimates. For example, a classic inferential question is, “How sure are we that the estimated mean,  $\bar{x}$ , is near the true population mean,  $\mu$ ?” While the equations and details change depending on the setting, the foundations for inference are the same throughout all of statistics. We introduce these common themes by discussing inference about the population mean,  $\mu$ , and set the stage for other parameters and scenarios. Some advanced considerations are discussed. Understanding this chapter will make the rest of this book, and indeed the rest of statistics, seem much more familiar.

---

**Definition 6.1** (Key Terms).

**Population:** A group of interest (typically large).

**Sample:** A subset of a population.

**Parameter (of population):** A numerical characteristic of a population. These are usually *unknown* in real-life settings.

$\mu$ : population mean

$\sigma^2$ : population variance

$\sigma$ : population standard deviation

*Note: Different from a parameter of a distribution.*

**Statistic (of sample):** A numerical characteristic of a sample, which is calculated and known (i.e., a function of the data).

$\bar{x}$ : sample mean

$s^2$ : sample variance

$s$ : sample standard deviation

**Statistical Inference:** Use statistics (known) to make conclusions on parameters (un-

*known) and quantify the degree of certainty of statements made.*

---

The sample mean,  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ , is a number we use to estimate the population mean,  $\mu$ . This is called a **point estimate**.

But, we know it's not equal to  $\mu$ . Then, we'd rather estimate the population mean using an **interval estimate** that gives a *range of real numbers* that we hope contains the population mean,  $\mu$ .

### Example 6.1.

---

- $\bar{x}$  is a point estimate of  $\mu$
- $s^2$  is a point estimate of  $\sigma^2$
- $s$  is a point estimate of  $\sigma$

*(All calculated with data from a sample)*

---

Due to the nature of randomness and calculating based on a subset, statistics are not guaranteed to be exactly equal to parameters.

Therefore, we create intervals around statistics which we believe capture the parameter.

### Definition 6.2 (Confidence Interval).

---

*A confidence interval is a plausible range of values that captures a parameter with a quantified degree of confidence.*

---



Suppose we are interested in the average mark for STA258 for the current semester. We are 100% confident that the average mark is between 0 and 100; however, this is not useful

information as we already know that the average mark must lie between 0 and 100. Using the marks of previous years, we can construct a 95% interval for the average mark. If it is determined that the average mark lies within 70% and 80%, this is much more meaningful as we can state with a high degree of certainty that the average mark is going to lie within a substantially narrow range.

In this course, all confidence intervals have the same basic skeleton:

$$\text{estimator} \pm \underbrace{(\text{value from reference distribution}) \times (\text{standard error of estimate})}_{\text{margin of error}}$$

The value from the reference distribution in the skeleton above will be either a value from the standard normal distribution or the Student  $t$ -distribution. The margin of error ( $MOE$ ) can be considered as the distance around our estimator in which the true value of the parameter of interest will be found, with a specified level of confidence.

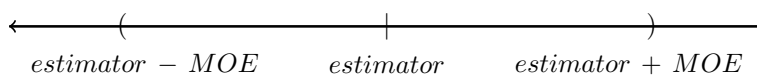


Figure 6.1: Visualization of a confidence interval on the real number line. The margin of error is abbreviated as  $MOE$ . The estimator is the centre of the interval. The confidence interval consists of all values between the estimator- $MOE$  and the estimator+ $MOE$ .

## 6.2 Interpretation

We use very specific language when we interpret a confidence interval.

*Suppose we construct a  $C\%$  confidence interval for some parameter such that  $C$  is between 0 and 100. In repeated sampling, we are  $C\%$  confident that approximately  $C\%$  of the intervals will capture the true value of the parameter.*

By this we mean that if we constructed several  $C\%$  confidence intervals using different samples (with or without replacing the units), then we should expect approximately  $C\%$  of these intervals to capture the parameter of interest. For example suppose we construct 1000 95% confidence intervals for the population mean  $\mu$ . We would expect approximately 95% of these 1000 intervals (i.e.  $95\% \times 1000 = 950$ ) to actually capture  $\mu$ .

### Note 6.1.

*A more intuitive but equivalent interpretation is to state that we are  $C\%$  confident that our target parameter is inside the interval constructed.*

It is incorrect to state that there is a  $C\%$  probability that the interval we constructed contains the parameter of interest. We assume that the value of a parameter is fixed. Therefore when we construct a confidence interval, the interval either contains the parameter or it does not.

### 6.3 Confidence Interval for $\mu$ (Known Variance)

When we know the population standard deviation  $\sigma$ , we can construct a confidence interval for  $\mu$  in the following manner.

#### Confidence Interval 6.1 (Confidence Interval on $\mu$ when $\sigma$ is Known)

A  $(100 - \alpha)\%$  confidence interval on  $\mu$  when  $\sigma$  is known is

$$\bar{x} \pm z_{\alpha/2} \left( \frac{\sigma}{\sqrt{n}} \right)$$

The  $z_{\alpha/2}$  value is obtained from standard normal tables. The standard error is  $\frac{\sigma}{\sqrt{n}}$  and the margin of error is  $z_{\alpha/2} \left( \frac{\sigma}{\sqrt{n}} \right)$ .

Let  $X_1, X_2, \dots, X_n$  be iid  $N(\mu, \sigma^2)$ , where  $\mu$  is unknown and  $\sigma$  is known. We know that:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

and

$$P(-1.96 < Z < 1.96) = 0.95$$

Therefore:

$$P\left(-1.96 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right) = 0.95 \Rightarrow P\left(\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}\right) = 0.95$$

#### Interpretation of Confidence Interval:

- This is a random interval  $\bar{X} \pm 1.96 \frac{\sigma}{\sqrt{n}}$
- The interval is random since  $\bar{X}$  is random due to sampling.
- The population mean  $\mu$  is a fixed, but unknown, number.
- The probability  $\mu$  is inside the random interval is 0.95 (success rate of the method).
- 95% of all samples give an interval that captures  $\mu$ , and 5% do not.

Once we observe our sample:

- This is **not** a random interval  $\bar{X} \pm 1.96 \frac{\sigma}{\sqrt{n}}$
- The probability  $\mu$  is inside this interval is either 1 or 0

### Confidence Interval Isn't Always Right:

Not all CIs contain the true value of the parameter. This can be illustrated by plotting many intervals simultaneously and observing.

---

#### R Output:

```
## Step 1. Generate random samples;
set.seed(2017)
m = 50;          # m = number of samples;
n = 25;          # n = number of obs in sample;
mu.i = 0;        # mu.i = mean of obs;
sigma.i = 5;     # sigma.i = std. dev. of obs;

mu.total = n * mu.i;          # mean of Total;
sigma.total = sqrt(n) * sigma.i; # std. dev. of Total;
```

#### R Output:

```
## Step 2. Construct CIs;
xbar = rnorm(m, mu.total, sigma.total) / n;
SE = sigma.i / sqrt(n);

alpha = 0.10;
z.star = qnorm(1 - alpha / 2);
```

#### R Output:

```
## Step 3. Graph CIs;
matplot(rbind(xbar - z.star * SE, xbar + z.star * SE),
        rbind(1:m, 1:m),
        type = "l", lty = 1,
        xlab = "-", ylab = "-");
abline(v = 0, lty = 2);
```

## Confidence Interval for the Mean of a Normal Population

Draw an SRS (Simple Random Sample) of size  $n$  from a Normal population having unknown mean  $\mu$  and **known** standard deviation  $\sigma$ . A level  $C$  confidence interval for  $\mu$  is:

$$\bar{x} \pm z_* \cdot \frac{\sigma}{\sqrt{n}}$$

The critical value  $z_*$  is illustrated in a Figure below and depends on  $C$ .

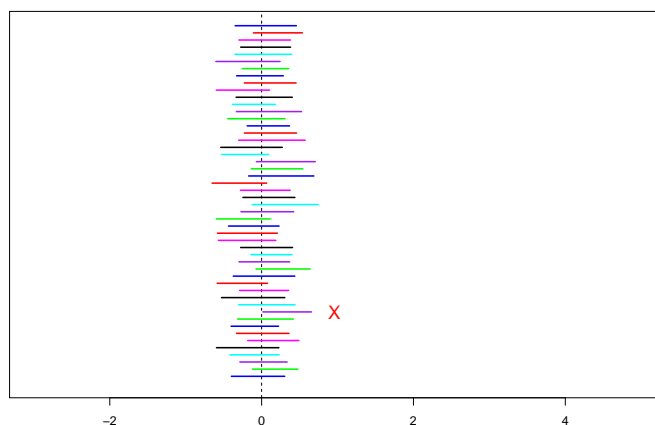


Figure 6.2: Simulated 95% confidence intervals for the population mean. Red “X” marks indicate intervals that do not contain the true mean ( $\mu = 0$ ).

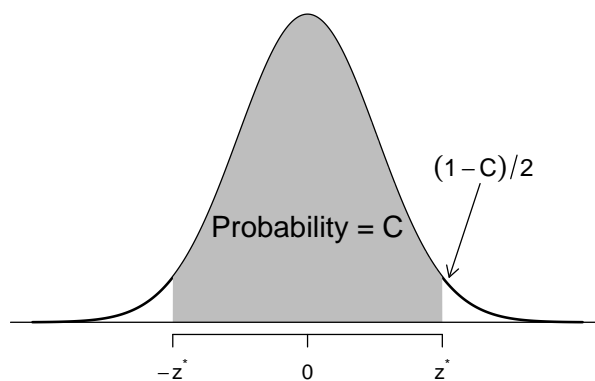


Figure 6.3: The central area under the standard normal curve with confidence level  $C$ .

## Large Sample CI for $\mu$ (Normal data)

$$\bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Valid if:

- $n$  large
- random sample from a Normal distribution
- independent observations

Some definitions:



- $1 - \alpha$  is the confidence coefficient
- $100(1 - \alpha)\%$  is the confidence level

## One Sample CI on the Population Mean $\mu$

- When population standard deviation  $\sigma$  is **known**
- Formula:  $\bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$
- Margin of error comes from standard normal and standard error

How to find  $z_{\alpha/2}$ ?

Example: Find  $z_{\alpha/2}$  for a 95% CI on  $\mu$ :

$$1 - \alpha = 0.95, \quad \alpha = 0.05, \quad \alpha/2 = 0.025$$

$$z_{\alpha/2} = 1.96 \quad (\text{from table or R: } \text{qnorm}(0.975))$$

## Table of Common $z$ -values

Confidence coefficient	Confidence level	$z$
0.90	90%	1.645
0.95	95%	1.96
0.99	99%	2.576

### Example 6.2.

Playbill magazine reported that the mean annual household income of its readers is \$119,155. Assume this estimate is based on a sample of 80 households, and that the population standard deviation is known to be  $\sigma = 30,000$ .

- $\bar{x} = 119,155$
- $n = 80$
- $\sigma = 30,000$

### Tasks:

- Develop a 90% confidence interval estimate of the population mean.
- Develop a 95% confidence interval estimate of the population mean.
- Develop a 99% confidence interval estimate of the population mean.

### 90% CI Calculation

$$\bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} = 119,155 \pm 1.645 \cdot \frac{30,000}{\sqrt{80}}$$

$$= 119,155 \pm 5,500.73$$

$$= (113,654.27, 124,655.73)$$

### 95% CI Calculation

$$\bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} = 119,155 \pm 1.96 \cdot \frac{30,000}{\sqrt{80}}$$

$$= 119,155 \pm 6,574.04$$

$$= (112,580.96, 125,729.04)$$

### 99% CI Calculation

$$\bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} = 119,155 \pm 2.576 \cdot \frac{30,000}{\sqrt{80}}$$

$$= 119,155 \pm 8,620.04$$

$$= (110,534.96, 127,775.04)$$

### Interpretation

We are 99% confident the mean household income of magazine readers is between \$110,534.96 and \$127,775.04.

---

## Example 6.3.

---

### Scenario:

The number of cars sold annually by used car salespeople is known to be **normally distributed**, with a population standard deviation of  $\sigma = 15$ . A random sample of  $n = 15$  salespeople was taken, and the number of cars each sold is recorded below. Construct a **95% confidence interval** for the population mean number of cars sold, and provide an interpretation.

### Raw data:

79	43	58	66	101
63	79	33	58	71
60	101	74	55	88

The sample mean is:

$$\bar{x} = \frac{79 + 43 + \cdots + 55 + 88}{15} = 68.6$$

### R function:

```
simple.z.test = function(x, sigma, conf.level = 0.95) {
  n = length(x);
  xbar = mean(x);
  alpha = 1 - conf.level;
  zstar = qnorm(1 - alpha/2);
  SE = sigma / sqrt(n);
  xbar + c(-zstar * SE, zstar * SE);
}
```

### R output:

```
# Step 1. Entering data;
cars = c(79, 43, 58, 66, 101, 63, 79,
         33, 58, 71, 60, 101, 74, 55, 88)

# Step 2. Finding CI;
simple.z.test(cars, 15)

## [1] 61.00909 76.19091
```

**Interpretation:** We estimate that the mean number of cars sold annually by all used car salespeople lies between 61 and 76, approximately. This type of estimate is correct 95% of the time.

### Cases Where Valid

- Large samples where population is **normal**.
- Large samples where population is **not normal** (By CLT).
- Small samples where population is **normal**.

*Note: A sample is considered large if  $n \geq 30$ .*

### Example 6.4.

Suppose a student measuring the boiling temperature of a certain liquid observes the readings (in degrees Celsius) 102.5, 101.7, 103.1, 100.9, 100.5, and 102.2 on 6 different samples of the liquid. He calculates the sample mean to be 101.82. If he knows that the distribution of boiling points is Normal, with standard deviation 1.2 degrees, what is the confidence interval for the population mean at a 95% confidence level?

A **confidence interval** uses sample data to estimate an unknown population parameter with an indication of how accurate the estimate is and of how confident we are that the result is correct.

The **interval** often has the form  
estimate  $\pm$  margin of error

The **confidence level** is the success rate of the method that produces the interval. A level  $C$  **confidence interval for the mean**  $\mu$  of a Normal population with **known** standard deviation  $\sigma$ , based on an SRS of size  $n$ , is given by

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$$

The **critical value**  $z^*$  is chosen so that the standard Normal curve has area  $C$  between  $-z^*$  and  $z^*$ .

Other things being equal, the **margin of error** of a confidence interval gets smaller as

- the confidence level  $C$  decreases,
- the population standard deviation  $\sigma$  decreases, and
- the sample size  $n$  increases.

## 6.4 APPENDIX

Interval estimators are commonly called **confidence intervals**. The upper and lower endpoints of a confidence interval are called the **upper** and **lower confidence limits**, respectively. The probability that a (random) confidence interval will enclose  $\theta$  (a fixed quantity) is called the **confidence coefficient**.

Suppose that  $\hat{\theta}_L$  and  $\hat{\theta}_U$  are the (random) lower and upper confidence limits, respectively, for a parameter  $\theta$ . Then, if

$$P(\hat{\theta}_L \leq \theta \leq \hat{\theta}_U) = 1 - \alpha,$$

the probability  $(1 - \alpha)$  is the **confidence coefficient**.

### Pivotal quantities

One very useful method for finding confidence intervals is called the **pivotal method**. This method depends on finding a pivotal quantity that possesses two characteristics:

- It is a function of the sample measurements and the unknown parameter  $\theta$ , where  $\theta$  is the **only** unknown quantity.
- Its probability distribution does not depend on the parameter  $\theta$ .

## Chapter 7

# One-Sample Confidence Intervals on a Mean When the Population Variance is Unknown

### 7.1 CIs for $\mu$

**Definition 7.1** (Large-Sample Confidence Interval for  $\mu$ ). 

---

*Let  $\mu$  be the population mean. When the population standard deviation  $\sigma$  is known and the sample size is large, a confidence interval for  $\mu$  is given by:*

$$\bar{Y} \pm z_{\alpha/2} \left( \frac{\sigma}{\sqrt{n}} \right)$$

*This interval is valid under the following conditions:*

- *The sample is random.*
  - *The observations are independent and identically distributed (i.i.d).*
  - *The sample size  $n$  is large enough for the Central Limit Theorem (CLT) to apply.*
- 

**Definition 7.2** (Small-Sample Confidence Interval for  $\mu$ ). 

---

*Let  $\mu$  be the population mean. When the population standard deviation is unknown and the sample size is small, a confidence interval for  $\mu$  is given by:*

$$\bar{Y} \pm t_{\alpha/2} \left( \frac{S}{\sqrt{n}} \right), \quad \nu = n - 1$$

*This interval is valid under the following conditions:*

- The observations are independent and identically distributed (*i.i.d.*).
- The sample is random.
- The population **must** follow a Normal distribution (*CLT does not apply*).

### Independence Assumption

The data values should be independent. There's really no way to check independence of the data by looking at the sample, but we should think about whether the assumption is reasonable.

### Randomization Condition

The data arise from a random sample or suitably randomized experiment. Randomly sampled data — especially data from a Simple Random Sample — are ideal.

#### Normal Population Assumption

- For very small samples ( $n < 15$  or so), the data should follow a Normal model pretty closely. If you do find outliers or strong skewness, don't use this method.
- For moderate samples ( $n$  between 15 and 40 or so), the t-method will work well as long as the data is unimodal and reasonably symmetric. Make a histogram, boxplot, or Q-Q plot to check.
- When the sample size is larger than 40 or 50, the t-method is safe to use unless the data are extremely skewed. Make a histogram, boxplot, or Q-Q plot to check.

### Standard Error

When the standard deviation of a statistic is estimated from data, the result is called the *standard error* of the statistic. The standard error of the sample mean  $\bar{x}$  is

$$\frac{s}{\sqrt{n}}.$$

#### The $t$ Distributions

- The density curves of the  $t$  distributions are similar in shape to the Standard Normal curve. They are symmetric about 0, single-peaked, and bell-shaped.
- The spread of the  $t$  distributions is a bit greater than that of the Standard Normal distribution. The  $t$  distributions have more probability in the tails and less in the center than the Standard Normal. This is because substituting the estimate  $s$  for the fixed parameter  $\sigma$  introduces more variation into the statistic.
- As the degrees of freedom increase, the  $t$  density curve approaches the  $N(0, 1)$  curve more closely. This happens because  $s$  estimates  $\sigma$  more accurately as the

sample size increases. So using  $s$  in place of  $\sigma$  causes little extra variation when the sample is large.

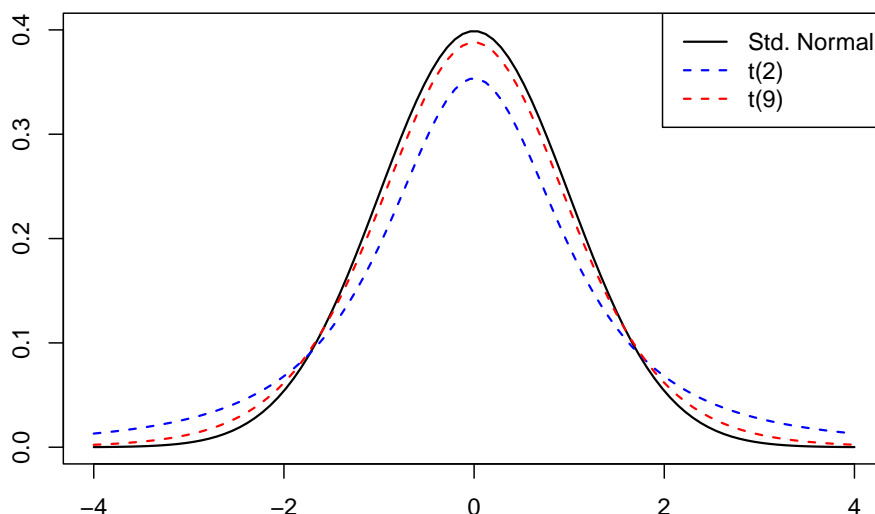


Figure 7.1: Comparison of the standard normal distribution and t-distributions with 2 and 9 degrees of freedom.

### Example 7.1.

[Ancient Air]

The composition of the Earth's atmosphere may have changed over time. To study the nature of the atmosphere long ago, scientists examined the gas in air bubbles trapped in ancient amber. Amber is fossilized tree resin that preserved the atmospheric gases at the time it was formed.

Measurements on amber specimens from the late Cretaceous era (75 to 95 million years ago) give the following percent values of nitrogen:

63.4, 65.0, 64.4, 63.3, 54.8, 64.5, 60.8, 49.1, 51.0

Assume these observations are a simple random sample (SRS) from the population of all ancient air bubbles. Construct a **90% confidence interval** to estimate the mean percent of nitrogen in ancient air. (Today's atmosphere contains about 78.1% nitrogen.)

#### Solution:

Let  $\mu$  represent the true mean percent of nitrogen in ancient air. We compute a 90% confidence interval for  $\mu$  using the sample data.

Given:

$$\bar{x} = 59.5888, \quad s = 6.2552, \quad n = 9, \quad t^* = 1.860 \quad (\text{df} = 8)$$

$$59.5888 \pm 1.860 \left( \frac{6.2552}{\sqrt{9}} \right) = 59.5888 \pm 3.8782$$

$$\boxed{55.7106 \text{ to } 63.4670}$$

### R Code:

Students can verify the result in R using:

```
# Step 1: Entering the data
nitrogen <- c(63.4, 65.0, 64.4, 63.3, 54.8, 64.5, 60.8, 49.1, 51.0)

# Step 2: Constructing the 90% confidence interval
t.test(nitrogen, conf.level = 0.90)
```

### R Output:

One Sample t-test

```
data:  nitrogen
t = 28.578, df = 8, p-value = 2.43e-09
alternative hypothesis: true mean is not equal to 0
90 percent confidence interval:
 55.71155 63.46622
sample estimates:
mean of x
 59.58889
```

---

### Example 7.2.

[Digital Camera Storage]

Most owners of digital cameras store their pictures on the camera. Some will eventually download these to a computer or print them using their own printers or a commercial printer. A film-processing company wanted to know how many pictures were stored on cameras. A random sample of 10 digital camera owners produced the following data:

25, 6, 22, 26, 31, 18, 13, 20, 14, 2

Estimate with **95% confidence** the mean number of pictures stored on digital cameras.

**Solution:**



We are given raw data with  $n = 10$  and no information about the population standard deviation, so we construct a confidence interval for the population mean using the t-distribution.

### Step 1: Compute sample statistics

- Sample mean:  $\bar{x} = \frac{177}{10} = 17.7$
- Sample variance (method 1 – direct):

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1} = \frac{742}{9} = 82.4556$$

- Sample standard deviation:

$$s = \sqrt{82.4556} = 9.081$$

### Step 2: Find the critical value

For a 95% confidence interval with  $n = 10$ , degrees of freedom = 9. From the t-distribution table:

$$t_{(9,0.025)} = 2.262$$

### Step 3: Construct the confidence interval

$$\bar{x} \pm t^* \cdot \frac{s}{\sqrt{n}} = 17.7 \pm 2.262 \cdot \frac{9.081}{\sqrt{10}} = 17.7 \pm 6.495$$

$$\boxed{11.205 \text{ to } 24.195}$$

### Interpretation:

We are 95% confident that the mean number of images stored on digital cameras is between 11.205 and 24.195.

### R Code:

```
# Step 1: Entering data
dataset <- c(25, 6, 22, 26, 31, 18, 13, 20, 14, 2)

# Step 2: Finding 95% confidence interval
t.test(dataset, conf.level = 0.95)
```

### R Output:

One Sample t-test

```
data: dataset
t = 6.164, df = 9, p-value = 0.0001659
alternative hypothesis: true mean is not equal to 0
```

95 percent confidence interval:

11.2042 24.1958

sample estimates:

mean of x

17.7

---

### Example 7.3.

[Electric Insulators]

A manufacturing company produces electric insulators. If the insulators break when in use, a short circuit is likely. To test the strength of the insulators, destructive testing is performed to determine how much force (in pounds) is required to break them.

The following dataset consists of force values (in pounds) recorded for a random sample of 30 insulators:

```
1870, 1728, 1656, 1610, 1634, 1784, 1522, 1696, 1592, 1662,
1866, 1764, 1734, 1662, 1734, 1774, 1550, 1756, 1762, 1866,
1820, 1744, 1788, 1688, 1810, 1752, 1680, 1810, 1652, 1736
```

Construct a **95% confidence interval** for the population mean force required to break the insulators.

#### Solution:

We want a confidence interval for the population mean  $\mu$ , where  $\mu$  = mean force required to break electric insulators. The population standard deviation is unknown, so we use a *one-sample* t-interval.

#### R Code:

```
# Step 1. Entering data;
dataset <- c(1870, 1728, 1656, 1610, 1634, 1784, 1522, 1696, 1592, 1662,
            1866, 1764, 1734, 1662, 1734, 1774, 1550, 1756, 1762, 1866,
            1820, 1744, 1788, 1688, 1810, 1752, 1680, 1810, 1652, 1736)

# Step 2. Finding CI;
t.test(dataset, conf.level = 0.95)
```

#### R Output:

One Sample t-test

```
data: dataset
t = 105.41, df = 29, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
```

95 percent confidence interval:

1689.961 1756.839

sample estimates:

mean of x

1723.4

### Interpretation:

We are 95% confident that the average force required to break an electric insulator is between **1689.961 pounds** and **1756.839 pounds**.

---

### Example 7.4.

---

[Assembly Time Estimation]

The operations manager of a production plant would like to estimate the mean amount of time a worker takes to assemble a new electronic component. After observing 120 workers assembling similar devices, she noticed that their average time was 16.2 minutes (with a standard deviation of 3.6 minutes).

Construct a **92% confidence interval** for the mean assembly time. State all necessary assumptions.

---

### Example 7.5.

---

[Tax Collected from Audited Returns]

In 2010, 142,823,000 tax returns were filed in the United States. The Internal Revenue Service (IRS) examined 1.107%, or 1,581,000, of them to determine if they were correctly done. To evaluate auditor performance, a random sample of these returns was drawn and the additional tax was recorded.

Estimate with 95% confidence the mean additional income tax collected from the 1,581,000 files audited.

### Solution:

We use a one-sample confidence interval for the population mean additional tax collected.

**Step 1: Import the data from taxes.txt**

```
# url of taxes;
url <- "https://mcs.utm.utoronto.ca/~nosedal/data/taxes.txt"
taxes_data <- read.table(url, header = TRUE)

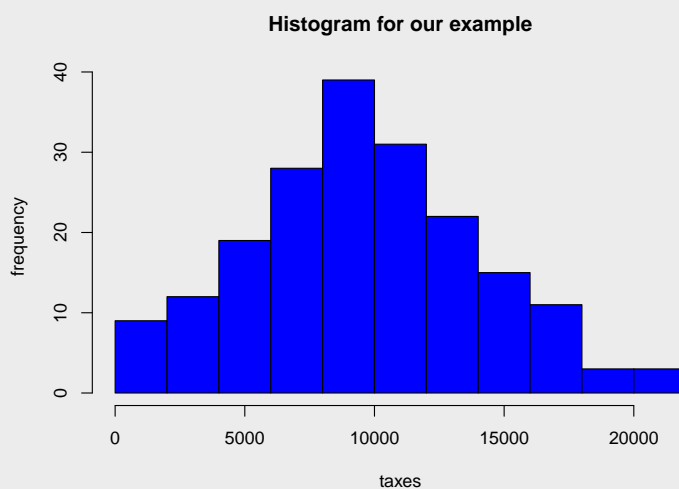
# inspect structure
names(taxes_data)
head(taxes_data)
```

```
# isolate the tax values
taxes <- taxes_data$Taxes
```

## Step 2: Plot the data

*Histogram:*

```
hist(taxes ,
     main = "Histogram for our example" ,
     xlab = "taxes" , ylab = "frequency" ,
     col = "blue")
```



*Boxplot:*

```
boxplot(taxes ,
        main = "Additional Income Tax" ,
        col = "blue")
```



*Q-Q Plot:*

```
qqnorm(taxes, col = "blue", pch = 19)
qqline(taxes)
```



### Step 3: Construct 95% CI

```
t.test(taxes, conf.level = 0.95)
```

**Interpretation:** Based on the t-test, we are 95% confident that the average additional tax collected lies within the interval calculated from the sample.

#### Assumptions:

- The sample is a random sample from the population of interest.

- Observations are independent.
- The population is approximately Normal, or the sample size is large enough (justified by the histogram, boxplot, and Q-Q plot).

```
##
## One Sample t-test
##
## data:  taxes
## t = 29.345, df = 191, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  8886.932 10167.721
## sample estimates:
## mean of x
##  9527.326
```

### Interpretation:

We estimate that the mean additional tax collected lies between \$8,887 and \$10,168 (with 95% confidence).

---

### A few final comments

When we introduced the Student  $t$ -distribution, we pointed out that the  $t$ -statistic is Student  $t$ -distributed if the population from which we've sampled is Normal. However, statisticians have shown that the mathematical process that derived the Student  $t$ -distribution is **robust**, which means that if the population is non-Normal, the results of the confidence interval estimate are still valid provided that the population is **not extremely non-Normal**. Our histogram, boxplot, and Q-Q plot suggest that our variable of interest is not extremely non-Normal, and in fact, may be Normal.

## Chapter 8

# One Sample Confidence Intervals On a Proportion

Perviously, we have introduced two types of confidence interval based on known and unknown variance. Moreover, confidence intervals are also applied to an unknown population proportion. For example, suppose we are interested the proportion of total number of left-handed students among all students who are currently studying at University of Toronto Mississauga. The question is: how do we know such the parameter which estimates the proportion of left-handed students at UTM? While, it is impossible to proceed it directly by counting both the total number of students and all left-handed students at UTM, due to the complexity and the total workload of that task. Then, we have to work with confidence intervals.

Firstly, we take a random sample of students at UTM, then we calculate how many students are left-handed by dividing total number of left-handed students in that sample with total number of students in it, and denote the proportion as  $\hat{p}$ . Next we begin our confidence interval calculation to get a range of number with a certain level of confidence.

Now let's begin with the proper definition of confidence interval on proportion.

**Definition 8.1** (One Sample Confidence Intervals On a Proportion). —————  
*We select a random sample of size  $n$  from a population with **unknown** proportion  $p$  of success. An approximate confidence interval for  $p$  is:*

$$p = \hat{p} \pm z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}, \text{ where } \hat{p} = \frac{\text{number of observations satisfying the criteria}}{n}.$$

*In addition,  $n$  is the sample size.*

*To apply this confidence interval, there are 3 conditions that we need to guarantee:*

- 1. Random sample;
- 2. Independent and identically distributed Bernoulli trails;
- 3. We have a large chosen sample size ( $n\hat{p} \geq 10$  and  $n(1 - \hat{p}) \geq 10$ ).

A good way to understand confidence interval is visualization. Now, suppose we have a valid estimation  $\hat{p}$ . After the entire procedure of confidence interval, our population proportion ( $p$ ) should be as the following number line shows:



Remember that your final answer of the range of  $p$  must be between 0 and 1, since we are working with proportion.

### Summary about One Sample Confidence Intervals

We have introduced one sample confidence interval under three different cases: given population variance, unknown population variance and unknown population proportion. All the material of one sample confidence interval comes from chapter 6, 7, 8, which seems like you to remember a lot. However, the reason why we give this summary is to help you to remember the basic skeleton of one sample confidence interval. Let's revisit the three distinct types of confidence interval:

- Known variance:  $\hat{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$  (Chapter 7).
- Unknown variance:  $\bar{x} \pm t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$
- Proportion:  $\hat{p} \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

The question is: what is the similarity between the three distinct types of confidence interval? You may have already noticed that the confidence intervals above are all follow such a skeleton that  $\bar{x}$  plus or minus its margin of error (different between each type of C.I.).

If we keep questioning ourselves that how the margin of error comes from, you will catch the pattern. The margin of error contains reference distribution and the standard deviation of  $\bar{x}$  under its reference distribution. Let's analyze each type of confidence interval to prove my statement is true:

The first type (given population variance) confidence interval is quite easy to recognize. Recall chapter 3: The Central Limit Theorem, we state that  $\bar{x} \sim N(\mu, \frac{\sigma^2}{n})$ , which is how reference distribution comes from with given population variance. To get the standard deviation of  $\bar{x}$ , we simply take the square root of the variance, then  $s_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ . Finally, we multiply reference distribution at the point  $\frac{\alpha}{2}$  with the standard deviation of  $\bar{x}$  under normal distribution to get the margin of error.

The second type (unknown population variance) confidence interval is similar to the first type, but the reference distribution is t-distribution instead of normal distribution. In chapter 7, we introduced the calculation of sample variance ( $s^2$ ) to estimate population variance ( $\sigma^2$ ), and  $s^2$  is an unbiased estimator of  $\sigma^2$  (the proof of this statement is in STA260, in



this course we can assume it freely). In chapter 2, we defined that  $T = \frac{\bar{x} - \mu}{(\frac{s}{\sqrt{n}})} \sim t_{n-1}$ , then the term  $\frac{s}{\sqrt{n}}$  (verify this by yourself as an extra exercise) is the standard deviation of  $\bar{x}$  under t-distribution with  $n - 1$  degrees of freedom. Finally, we multiply multiply reference distribution at the point  $\frac{\alpha}{2}$  with the standard deviation of  $\bar{x}$  under t-distribution to get the margin of error.

The third type (estimate population proportion) confidence interval is slightly harder to identify. Recall chapter 4 that we can approximate binomial distribution by normal distribution. Suppose that a random variable  $X \sim \text{Binomial}(n, p)$ , then the random variable  $X \sim N(np, np(1-p))$ . From chapter 4, we know that  $\hat{p} \sim N(\mu_{\hat{p}} = p, \sigma_{\hat{p}}^2 = \frac{p(1-p)}{n})$ . Trivially, the standard deviation of  $\hat{p}$  is  $\sqrt{\frac{p(1-p)}{n}}$ , since we don't know the value  $p$  and use  $\hat{p}$  to estimate that. Then, standard deviation of  $\hat{p}$  is  $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ . Finally, by multiplying the reference distribution and standard deviation, we get the margin of error of the confidence interval.

### Conclusion About One Sample Confidence Intervals

If you follow the skeleton below, one sample confidence interval will become easier:

- 1. Identify what type of one sample confidence interval to use from given information;
- 2. Construct your confidence interval that fits the circumstance which you are facing, it either going to be  $\bar{x} \pm M.E.(\bar{x})$  or  $\hat{p} \pm M.E.(\hat{p})$ ;
- 3. Check the validity of your final answer. For example, the range of  $p$  must between 0 and 1.
- 4. Clearly state your final conclusion: we have a certain percentage of confidence to guarantee that the value of the chosen sample is between its lower bound and upper bound.

## Chapter 10

# Two Sample Confidence Interval

We have discussed three distinct types of one sample confidence interval. Now, let's keep moving forward to see how confidence interval works for two sample. The aim of one sample confidence interval is giving a range of numbers to estimate population mean or proportion with a certain percentage of confidence. For two samples, the aim is comparing with sample has a relatively larger or smaller population mean or proportion with a certain percentage of confidence.

### 10.1 Two Sample Confidence Interval on a Difference of Mean

Suppose we are interested in the final mark of MAT135 from the same semester but with different campuses at the University of Toronto (let's use UTSG and UTM as the two independent population groups). We want to know which campus has a relatively higher average score, the question is: how do we determine that? It is going to be complicated if we proceed with the study directly by determining the sum of everyone's final marks and calculating the average for the two campuses. Similarly, as one sample confidence interval, we can select two groups of random sample from the two campuses (one group per each campus), and then calculate each sample mean. Finally, we apply a confidence interval to approximate which population has a higher mean (or average).

#### Two Sample Confidence Interval for Two Independent Groups of Population

We are going to introduce several definitions because two sample confidence interval has distinct cases. You need to be able to identify which exact case you are facing from given information. If you know how to solve one sample confidence interval, then two sample confidence interval is going to be easy, because all the techniques from one sample confidence interval are still usable.

**Case 1:** Two sample confidence interval with given population variance for both groups.

**Definition 10.1** (Two sample confidence interval with given population variances). ———  
*Suppose we are given the population variance for both two independent groups of population. The confidence interval of  $\mu_1 - \mu_2$  (difference of mean between population group 1 and 2) is*

given by the following:

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

For  $\sigma_1^2$ , which is population variance of population group 1;  $n_1$  is the sample size chosen from population group 1. Similarly for  $\sigma_2^2$ , which is population variance of population group 2;  $n_2$  is the sample size chosen from population group 2.

Additionally, case 1 is a bit unrealistic with other cases because the population variance ( $\sigma^2$ ) from both groups are rare to know.

**Case 2:** Two sample confidence interval with equal unknown population variance

Ideally, we have all the information about population variance from two chosen samples. However, that case does not usually happen. We may face the case with unknown variance.

#### Definition 10.2.

Suppose that the chosen two independent samples have same unknown population variance. Then the two sample confidence interval for  $\mu_1 - \mu_2$  is given by the following:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{n_1+n_2-2;\alpha/2} \cdot s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

In this case,  $n_1$  and  $n_2$  are sample size from the two chosen samples respectively;  $s_p$  is aggregated variance of both samples combined which accommodates samples of different sizes. Additionally,  $s_p$  is called pooled standard deviation which is calculated by the following equation:

$$s_p^2 = \frac{(n_1 - 1) \cdot s_1^2 + (n_2 - 1) \cdot s_2^2}{n_1 + n_2 - 2},$$

where  $s_1^2$  and  $s_2^2$  are sample variance of the two chosen samples respectively.

Then we take the square root  $s_p = \sqrt{s_p^2}$  to get pooled standard deviation.

Alternatively, we can write the equation for two sample confidence interval with equal unknown variance as:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{n_1+n_2-2;\alpha/2} \cdot \sqrt{s_p^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}, \text{ which is same as the one above.}$$

#### Pooled Variance and Standard Deviation

In statistics, pooled variance (also known as combined variance, composite variance, or overall variance) is a method to calculate such a value in order to estimate variance between several distinct populations. The mean of each population may or may not be the same, but

the variance of these populations are same. Pooled standard deviation does similar thing, we use that value to estimate standard deviation instead of variance.

### Case 3: Two sample confidence interval with unequal unknown population variance

At this point, you may wonder that what if the population variance is both unequal and unknown? Does two sample confidence interval still doable in this case? The answer is: Yes. We can still proceed with two sample confidence interval.

#### Definition 10.3.

Suppose that our chosen two independent samples with unequal and unknown population variance, then the confidence interval for  $\mu_1 - \mu_2$  is given by:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{df; \alpha/2} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \text{ where: } df = \min(n_1 - 1, n_2 - 1).$$

Moreover,  $s_1^2$  and  $s_2^2$  are sample variance of the two chosen groups; and  $n_1, n_2$  are the sample size of the two chosen groups respectively.

### Visualization of Two Sample Confidence Interval

Only with the equation seems hard to understand, the following number line helps you to visualize what we try to indicate:

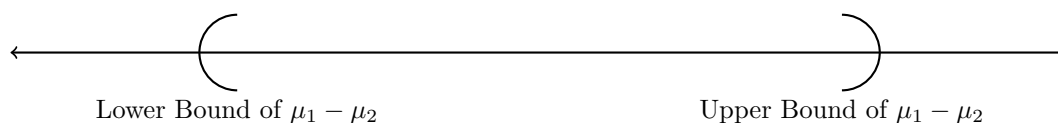


Figure 10.1: Visualization of two-sample confidence interval (Case 1, 2, 3)

Earlier in this chapter we said that two sample confidence interval aims to compare the mean between two populations. The number line above shows the result of difference of means between the two populations. Now, the question is, how do we know which population has a relatively larger mean? While, we can summarize it from that number line, with different cases:

#### 1. $\mu_1 < \mu_2$ :

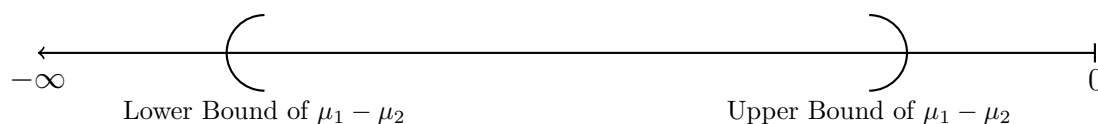


Figure 10.2: Visualization of the case when  $\mu_1 < \mu_2$

Now, we know that the difference between  $\mu_1 < \mu_2$  lies on the negative side on the number line, such that:  $\mu_1 - \mu_2 < 0$ . Hence, by solving the inequality above we get:  $\mu_1 < \mu_2$  trivially.

2.  $\mu_1 > \mu_2$ :

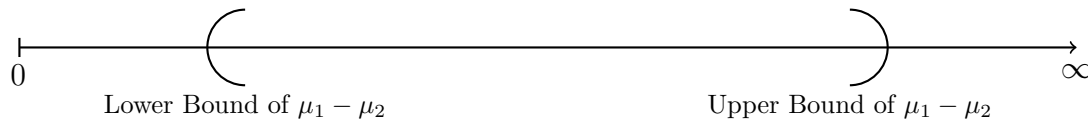


Figure 10.3: Visualization of the case when  $\mu_1 > \mu_2$

Similarly as the case above, we know that  $\mu_1 - \mu_2 > 0$ , by observing the number line. Thus, we have:  $\mu_1 > \mu_2$ .

3.  $\mu_1 = \mu_2$ :

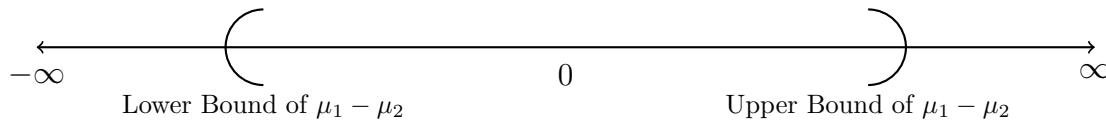


Figure 10.4: Visualization of the case when  $\mu_1 = \mu_2$

However, 0 is an element in the range of the difference between  $\mu_1 - \mu_2$ , such that there is a chance when the two means could be same. Hence, we conclude that  $\mu_1 = \mu_2$  in this case. While, if you prefer to use  $\mu_2 - \mu_1$ , this strategy is going to work as well. Just simply following the same steps, you will get the same conclusion.

### Conditions of Two Sample Confidence Interval

Same as all previous confidence intervals, we still need several conditions that guarantee the validity two sample confidence interval:

- 1. The two chosen sample is required to be independent and random;
- 2. If both sample size are small (both  $n_1 < 30$  and  $n_2 < 30$ ), then both sample should be from normal population;
- 3. If one of the sample has a small size (either  $n_1 < 30$  or  $n_2 < 30$ ), then the smaller sample must be from a normal population;

Note that if both  $n_1 \geq 30$  and  $n_2 \geq 30$ , then normality assumption is not required by the Central Limit Theorem.

### Example (Comparing Two Population Means Managerial Success Indexes for Two Groups)

**Example 10.1.**

Behavioural researchers have developed an index designed to measure managerial success. The index (measured on a 100- point scale) is based on the manager's length of time in the organization and their level within the term; the higher the index, the more successful the manager. Suppose a researcher wants to compare the average index for the two groups of managers at a large manufacturing plant. Managers in group 1 engage in high volume of interactions with people outside the managers' work unit (such interaction include phone and face-to-face meetings with customers and suppliers, outside meetings, and public relation work). Managers in group 2 rarely interact with people outside their work unit. Independent random samples of 12 and 15 managers are selected from groups 1 and 2, respectively, and success index of each is recorded.

Comparing Two Population Means Managerial Success Indexes for Two Group (With Equal Variances Assumed) Note: The response variable is "Managerial Success Indexes".

Managerial success indexes is a continuous quantitative variable, measured on 100-point scale.

The explanatory variable is "Type of group".

Type of group (Group 1: Interaction with outsiders, Group 2: Fewer interactions) is a nominal categorical variable.

Let's use R-code to demonstrate this example. The following lines of code helps you to get started.

```
# Importing data file into R;

success=read.csv(file="success.csv",header=TRUE);

# Getting names of variables;

names(success);

# Seeing first few observations;

head(success);

# Attaching data file; attach(success);
```

Now, you will get the following table by running the code above from R-studio.

```
## [1] "Success_Index" "Group"
## Success_Index Group
## 1 65 1
## 2 66 1
## 3 58 1
## 4 70 1
## 5 78 1
## 6 53 1
```

Then, we use R-studio to obtain some descriptive statistics.

```
## .group min Q1 median Q3 max mean sd n
## 1 1 53 62.25 65.5 69.25 78 65.33333 6.610368 12
## 2 2 34 42.50 50.0 54.50 68 49.46667 9.334014 15
## missing
## 1 0
## 2 0
```

Note that: Group 1 = “interaction with outsiders” and Group 2 = “fewer interactions”. Then, we can proceed with two sample confidence interval.

```
# 95\% CI for the difference between means;
# equal variances is assumed;

t.test(Success_Index~Group,
var.equal=TRUE, conf.level=0.95)$conf.int;
```

Finally, the output is:

```
## [1] 9.288254 22.445079
## attr(,"conf.level")
## [1] 0.95
```

Therefore, We are 95% confident that the mean success index is between 9.28 and 22.44 points higher for group 1 than group 2.

## 10.2 Two Sample Confidence Interval on Paired Data

It seems like two sample confidence interval only works on two independent samples, however what about two dependent samples? Suppose we are interested the growth of height from several distinct elementary students. We measure their height recently, then we will do it another time with five years later. The question is: how are we going to proceed with two

confidence interval? While, the answer is yes. We are able to do so by constructing two sample confidence interval, but with a different strategy. Now, let's introduce two sample confidence interval with paired data:

Sample Units	Measurement 1 ( $M_1$ )	Measurement 2 ( $M_2$ )	Difference ( $M_2 - M_1$ or $M_1 - M_2$ )
1	$x_{11}$	$x_{12}$	$x_{d1} = x_{12} - x_{11}$
2	$x_{21}$	$x_{22}$	$x_{d2} = x_{22} - x_{21}$
3	$x_{31}$	$x_{32}$	$x_{d3} = x_{32} - x_{31}$
.....			
n	$x_{n1}$	$x_{n2}$	$x_{dn} = x_{n2} - x_{n1}$

Figure 10.5: A table of paired data

The table shows how to get a paired data. The first column on the left is the sample size, the second column records the first time of measurement of objects, the third column records the second time of measurement of the same objects, the last column on the right is the difference between the second and the first measurement ( $M_2 - M_1$ ). Then we can use the fourth column to get the mean value, sample variance and sample standard deviation of the difference. Now, let's begin with the proper definition:

**Definition 10.4** (Two Sample Confidence Interval on Paired Data). —————

*Suppose we have two samples that are dependent with each other, the confidence interval on paired data's mean ( $\mu_d$ ) is given by:*

$$\bar{x}_d \pm t_{n-1, \alpha/2} \cdot \frac{s_d}{\sqrt{n}}.$$

*In this case, the reference distribution is  $t$ -distribution with  $n - 1$  degrees of freedom (sample size minus 1),  $\bar{x}_d$  represents the sample mean of difference between the two measurements on the paired data,  $s_d$  is the sample standard deviation of difference between the two measurements.*

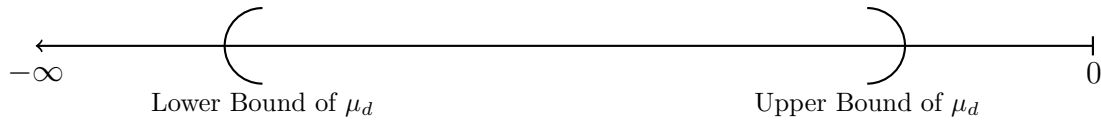
Note that: two sample confidence interval on paired data is to calculate a range of number on **the mean of difference between the two measurement**. Then, we can continue our analysis about the data.

**Visualization of Two Sample Confidence Interval on Paired Data**

Next, we need to state our final conclusion from the result of two sample confidence interval on paired data. Again, let's construct a number line for each case.

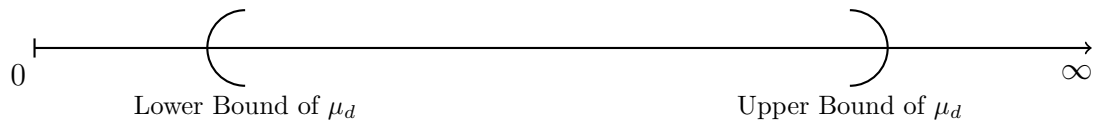
1.  $\bar{M}_1 < \bar{M}_2$ :



Figure 10.6: Visualization of the case when  $\bar{M}_1 < \bar{M}_2$ 

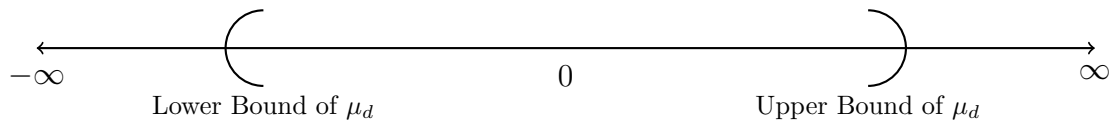
From the number line above, we know that the result of  $\mu_d$  lies only in the negative side of the number line, also we use  $M_1 - M_2$  to get the difference data then calculate the average of difference data. Now, we have:  $\bar{M}_1 - \bar{M}_2 = \mu_d < 0$ . Hence, we conclude that  $\bar{M}_1 < \bar{M}_2$ .

2.  $\bar{M}_1 > \bar{M}_2$ :

Figure 10.7: Visualization of the case when  $\bar{M}_1 > \bar{M}_2$ 

From the number line above, we know that the result of  $\mu_d$  lies only in the positive side of the number line, also we use  $M_1 - M_2$  to get the difference data then calculate the average of difference data. Now, we have:  $\bar{M}_1 - \bar{M}_2 = \mu_d > 0$ . Hence, we conclude that  $\bar{M}_1 > \bar{M}_2$ .

3.  $\bar{M}_1 = \bar{M}_2$ :

Figure 10.8: Visualization of the case when  $\bar{M}_1 = \bar{M}_2$ 

While, 0 is included in the range of  $\mu_d$ , then there is a chance that  $\bar{M}_1 - \bar{M}_2 = \mu_d = 0$ . Hence, we conclude that  $\bar{M}_1 = \bar{M}_2$ .

### Conditions of Two Sample Confidence Interval on Paired Data

We need the following conditions to make sure the validity of two sample confidence interval on paired data:

- 1. Units are independent (measurements are dependent on each unit)
- 2. Units must be random sample
- 3. If we have a small sample ( $n < 30$ ), then the population of difference should be normal (no restrictions on large samples).

Also, note that two sample confidence interval on paired data can only be applied with two dependent groups of data. For independent groups of data, you need to refer chapter 10.1.

### 10.3 Two Sample Confidence Interval on Proportions

Furthermore, two sample confidence intervals can approximate the proportion as well. Suppose we are interested in the proportion of left-handed students in UTSG and UTM, and we are asked to find the campus that has a relatively larger proportion of left-handed students. To begin with this task, it is impossible to complete it directly by calculation, due to its complexity and high workload. We can use select two independent groups (one group from each campus), then apply two sample confidence interval to approximate which campus has a larger proportion.

It may still be difficult to understand, now let's begin with figures:

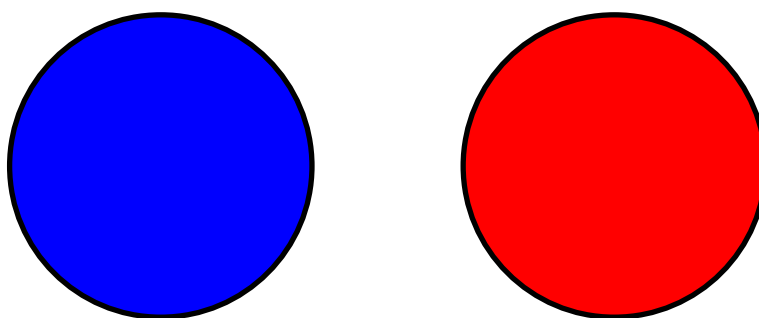


Figure 10.9: Visualization of two population of all students from UTSG (left) and UTM (right)

This figure represents the population of all students from two campuses. To begin with our task (find the proportion of left-handed students), we can select a group of random sample from each campus:



Figure 10.10: Visualization of two selected random sample from UTSG (left) and UTM (right)

As you can see, we have chosen our random sample from each campus with sample size  $n_1$  and  $n_2$ , respectively. Now we need estimators to construct our confidence interval:  $\hat{p}_1$  and  $\hat{p}_2$ . Those are the proportion of left-handed students from each random sample respectively. Since we have all the information we need, now we can apply our confidence interval from the two random sample.

**Definition 10.5** (Two Sample Confidence Interval on Proportions). 

---

Draw an SRS of size  $n_1$  from a population having proportion  $p_1$  of successes and draw an independent SRS of size  $n_2$  from another population having proportion  $p_2$  of successes. When  $n_1$  and  $n_2$  are large, an approximate level  $C$  confidence interval for  $p_1 - p_2$  is given by:

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}.$$

Now,  $n_1$  and  $n_2$  are sample size of selected random sample from each population;  $\hat{p}_1$  and  $\hat{p}_2$  are the proportion of success of each selected random sample respectively.

---

Note that,  $\hat{p}_1 = \frac{\text{number of successes in random sample 1}}{n_1}$  and  $\hat{p}_2 = \frac{\text{number of successes in random sample 2}}{n_2}$ .

**Visualization of  $p_1 - p_2$** 

Similarly as confidence interval on independent and dependent data, we are going to provide number lines, in order to help you to visualize the result easily.

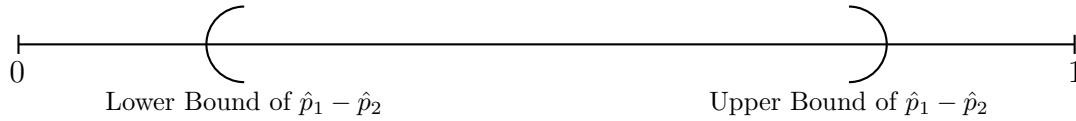
**1.  $p_1 > p_2$ :**

Figure 10.11: Visualization of the case when  $p_1 > p_2$

From the number line, the result of  $\hat{p}_1 - \hat{p}_2$  lies only on the positive side, such that  $\hat{p}_1 - \hat{p}_2 > 0$ . Hence,  $\hat{p}_1 > \hat{p}_2$ . Note that proportion is a number between 0 and 1, such that the difference of two proportions only between  $-1$  and  $1$ .

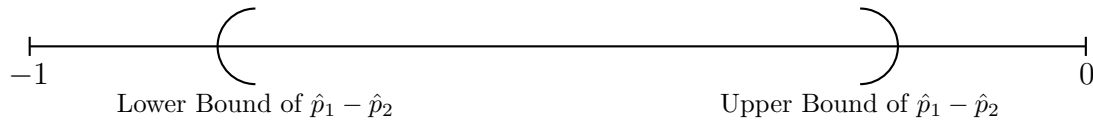
**2.  $\hat{p}_1 < \hat{p}_2$** 

Figure 10.12: Visualization of the case when  $p_1 < p_2$

Now, the result of  $\hat{p}_1 - \hat{p}_2$  lies only on the positive side, such that  $\hat{p}_1 - \hat{p}_2 < 0$ . Hence,  $\hat{p}_1 < \hat{p}_2$ .

**3.  $\hat{p}_1 = \hat{p}_2$**

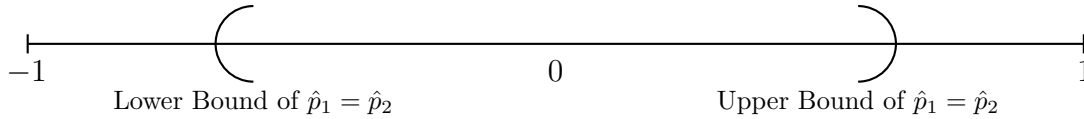


Figure 10.13: Visualization of the case when  $\hat{p}_1 = \hat{p}_2$

In this case, 0 lies in the range of  $\hat{p}_1 - \hat{p}_2$ , such that there is a chance when  $\hat{p}_1 - \hat{p}_2 = 0$ . Hence, we conclude that:  $\hat{p}_1 = \hat{p}_2$ .

### Conditions of Two Sample Confidence Interval on Proportion

- 1. Randomization Condition: The data in each group should be drawn independently and at random from a population or generated by a completely randomized designed experiment.
- 2. The 10% Condition: If the data are sampled without replacement, the sample should not exceed 10% of the population. If samples are bigger than 10% of the target population, random draws are no longer approximately independent.
- 3. Independent Groups Assumption: The two groups we are comparing must be independent from each other.
- 4. Sample size requirement: both selected sample size must greater than 70.

## 10.4 Two Sample Confidence Interval on Variances

Confidence interval is a strong technique in inferential statistics, we have discussed its application on population mean, proportion and dependent data. Now, let's move on to variance.

One simple method involves just looking at two sample variances. Logically, if two population variances are equal, then the two sample variances should be very similar. When the two sample variances are reasonably close, you can be reasonably confident that the homogeneity assumption is satisfied and proceed with, for example, Student t-interval. However, when one sample variance is three or four times larger than the other, then there is reason for a concern. The common statistical procedure for comparing population variances  $\sigma_1^2$  and  $\sigma_2^2$  makes an inference about the ratio of  $(\sigma_1^2)/(\sigma_2^2)$ .

To make an inference about the ratio of  $(\sigma_1^2)/(\sigma_2^2)$  we collect sample data and use the ratio of the sample variances  $(s_1^2)/(s_2^2)$ .

At this point, let's derive the confidence interval. We know that:  $\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \sim F_{n_1-1, n_2-1}$ . Then, we can construct our confidence interval as:

$$P[F_{n_1-1, n_2-1; 1-\alpha/2} < \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} < F_{n_1-1, n_2-1; \alpha/2}] = 1 - \alpha.$$

Now, the reference distribution of this confidence interval is F-distribution with  $n_1 - 1$  and  $n_2 - 1$  degrees of freedom, leaving areas of  $1 - \alpha/2$  and  $\alpha/2$ , respectively, to the right.

Rearranging gives us:

$$P\left[\frac{s_1^2}{s_2^2} \cdot \frac{1}{F_{n_1-1, n_2-1; \alpha/2}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2} \cdot \frac{1}{F_{n_1-1, n_2-1; 1-\alpha/2}}\right] = 1 - \alpha.$$

Using the fact that  $F_{n_1-1, n_2-1; 1-\alpha/2} = \frac{1}{F_{n_2-1, n_1-1; \alpha/2}}$ , we have:

$$P\left[\frac{s_1^2}{s_2^2} \cdot \frac{1}{F_{n_1-1, n_2-1, \alpha/2}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2} \cdot F_{n_2-1, n_1-1, \alpha/2}\right] = 1 - \alpha.$$

## Chapter 12

# One Sample Hypothesis Test on a Proportion and Variance

Inferential statistics is a powerful method for statistical analysis, because it allows people to analyze a lot parameters. Similarly to confidence interval, testing hypothesis can be applied to proportion and variance as well. Also, we use the exact same structure for one sample hypothesis test on a proportion and variance.

### 12.1 One Sample Hypothesis Test on a Proportion

Suppose we have assume the proportion of a criteria from a population  $p$  is equal to our parameter  $p_0$  (null hypothesis  $H_0 : p = p_0$ ). While, the question is: how do we know whether our assumption is correct or not? We need to use testing hypothesis on proportion to verify.

#### Step 1. Stating the Structure of Testing Hypothesis

First of all, let's proceed with a table to see all the cases:

Cases	Null Hypothesis	Alternative Hypothesis
1	$H_0 : p = p_0$	$H_a : p > p_0$
2	$H_0 : p = p_0$	$H_a : p < p_0$
3	$H_0 : p = p_0$	$H_a : p \neq p_0$

Figure 12.1: All possible cases of one sample hypothesis test on a proportion ( $p$  represents the actual proportion of a population)

We are not going to proceed with all three cases in a single question. You need to be able to identify which case of testing hypothesis are going to be applied from question.

#### Step 2. Computing Test Statistics

After that we need to compute our test statistics, as the following definition provides:

**Definition 12.1** (Test statistics of one sample hypothesis test on a proportion). —————

The test statistics of one sample hypothesis test on a proportion is given by:

$$Z^* = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}.$$

In this case,  $n$  means the sample size,  $\hat{p}$  is the parameter of the proportion of the population, which is calculated by  $\hat{p} = \frac{\text{number of successes in the sample}}{n}$ . Also, the reference distribution is standard normal distribution:  $N(0, 1)$ .

Note that be careful while you are computing the test statistics, because it directly affects the final answer.

### Step 3. Finding the $p$ - value

i. When is structure of testing hypothesis is  $H_0 : p = p_0$ ,  $H_a : p > p_0$ :

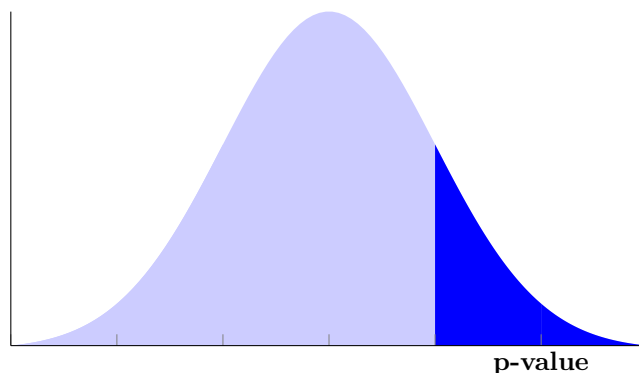


Figure 12.2: An illustration of p-value when  $H_a : p > p_0$  for one hypothesis test on a proportion.

In case (i), calculating  $P[Z > Z_*]$  as your p-value. Then, comparing with significant level:  $\alpha$ .

ii. When is structure of testing hypothesis is  $H_0 : p = p_0$ ,  $H_a : p < p_0$ :

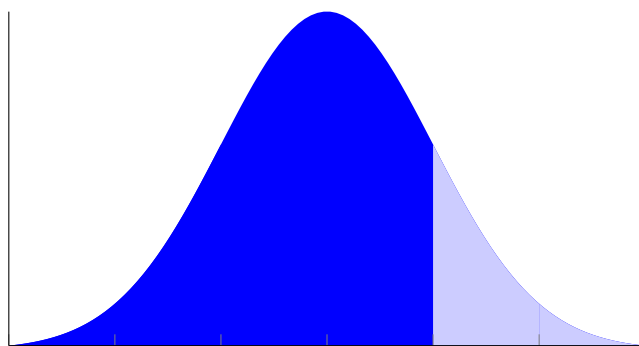


Figure 12.3: An illustration of p-value when  $H_a : p < p_0$  for one hypothesis test on a proportion.

In case (ii), calculating  $P[Z < Z_*]$  as your p-value. Then, comparing with significant level:  $\alpha$ .

iii. When is structure of testing hypothesis is  $H_0 : p = p_0$ ,  $H_a : p \neq p_0$ :

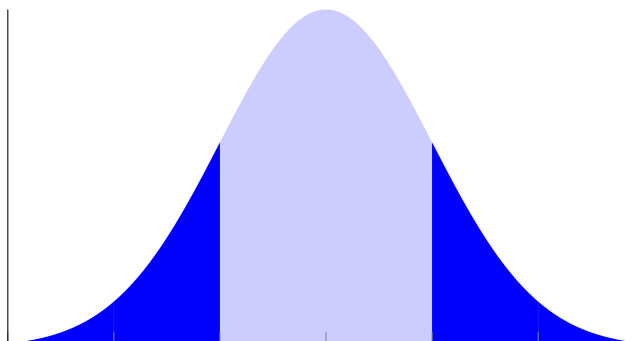


Figure 12.4: An illustration of p-value when  $H_a : p \neq p_0$  for one hypothesis test on a proportion.

In case (iii), calculating  $2 \cdot P[Z > |Z_*|]$  as your p-value. Then, comparing with significant level:  $\alpha$ .

#### Step 4: Comparing P-value with $\alpha$ -level

If p-value is less than  $\alpha$ -level, then we reject the null hypothesis ( $H_0$ ) and accept the alternative hypothesis ( $H_a$ ). Otherwise, If p-value is greater than  $\alpha$ -level, then we do not reject the null hypothesis ( $H_0$ ) and reject the alternative hypothesis ( $H_a$ ).

**Step 5: Final Conclusion** If we reject the null hypothesis, then we conclude that: there is sufficient evidence to reject the null hypothesis. If we do not reject the null hypothesis, then we conclude that: there is insufficient evidence to reject the null hypothesis.

#### Conditions on One Sample Test Hypothesis on a Proportion

- 1. Random sample;
- 2. Independent sample: each observations are independent to others;
- 3. Sufficient sample.



## Chapter 14

# Two Sample Hypothesis Test

### 14.1 Two Sample Hypothesis Test on Paired Data

When observations in sample 1 matches with an observation in sample 2. Observations in sample 1 are, usually, highly, correlated with observations in sample 2, these data are often called matched pairs. For each pair (the same cases), we form: Difference = observation in sample 2 - observation in sample 1. Thus, we have one single sample of differences scores. For example, in longitudinal studies: Pre- and post-survey of attitudes towards statistics (Same student is measured twice: Time 1 (pre) and Time 2 (post)). We measure change in the attitudes: Post - Pre (for each student). Often these types of studies are called, repeated measures.

Paired Data Condition: the data must be quantitative and paired.

Independence Assumption:

- If the data are paired, the groups are not independent. For this methods, it is the differences that must be independent of each other.
- The pairs may be a random sample.
- In experimental design, the order of the two treatments may be randomly assigned, or the treatments may be randomly assigned to one member of each pair.
- In a before-and-after study, we may believe that the observed differences are representative sample of a population of interest. If there is any doubt, we need to include a control group to be able to draw conclusions.
- If samples are bigger than 10 % of the target population, we need to acknowledge this and note in our report. When we sample from a finite population, we should be careful not to sample more than 10 % of that population. Sampling too large a fraction of the population calls the independence assumption into question.

Recall chapter 10 when we first introduced paired data, now we are going to use it again:

Sample Units	Measurement 1 ( $M_1$ )	Measurement 2 ( $M_2$ )	Difference ( $M_2 - M_1$ or $M_1 - M_2$ )
1	$x_{11}$	$x_{12}$	$x_{d1} = x_{12} - x_{11}$
2	$x_{21}$	$x_{22}$	$x_{d2} = x_{22} - x_{21}$
3	$x_{31}$	$x_{32}$	$x_{d3} = x_{32} - x_{31}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	$x_{n1}$	$x_{n2}$	$x_{dn} = x_{n2} - x_{n1}$

Figure 14.1: A table of paired data

From that table, we can get  $\bar{X}_d$ , which is the mean, variance and standard deviation of the difference. We need these values to continue our analysis.

### Step 1: Stating the Structure of Testing Hypothesis

The idea of two sample hypothesis test on pair data is transforming it to one sample hypothesis data, so that our analysis based on  $\mu_d$ .

Cases	Null Hypothesis	Alternative Hypothesis
1	$\mu_d = 0$	$\mu_d > 0$
2	$\mu_d = 0$	$\mu_d < 0$
3	$\mu_d = 0$	$\mu_d \neq 0$

Figure 14.2: All possible cases of two sample hypothesis test on paired data

### Step 2: Computing Test Statistics

**Definition 14.1** (Test statistics of two sample hypothesis test on paired data). —————

*The test statistics of two sample hypothesis test on paired data is given by:*

$$t_* = \frac{\bar{x}_d - 0}{s_d/\sqrt{n}} \sim t_{n-1; \alpha/2}.$$

$\bar{x}_d$  is the mean of sample difference and  $s_d$  is the standard deviation of difference data.

### Step 3: Finding the P-value

Two sample hypothesis test is quite similar to one sample hypothesis test on a mean. Notes that we are working within 2 measurement. You can refer one sample hypothesis test to get the idea about find the p-value in this case.

### Step 4: Comparing P-value with $\alpha$ -level

If p-value is less than  $\alpha$ -level, then we reject the null hypothesis ( $H_0$ ) and accept the alternative hypothesis ( $H_a$ ). Otherwise, If p-value is greater than  $\alpha$ -level, then we do not reject the null hypothesis ( $H_0$ ) and reject the alternative hypothesis ( $H_a$ ).

**Step 5: Final Conclusion** If we reject the null hypothesis, then we conclude that: there is sufficient evidence to reject the null hypothesis. If we do not reject the null hypothesis,

then we conclude that: there is insufficient evidence to reject the null hypothesis.

Note that your final conclusion is based on the value of  $\mu_d$ , but we still need more. We also can know which group has larger means from the result ( $\mu_d$ ).

- If  $\mu_d > 0$ : from the table above we get  $M_2 - M_1 > 0$ , then  $M_2 > M_1$ .
- If  $\mu_d < 0$ : from the table above we get  $M_2 - M_1 < 0$ , then  $M_2 < M_1$ .
- If  $\mu_d \neq 0$ : from the table above we get  $M_2 - M_1 \neq 0$ , then  $M_2 \neq M_1$ .

## 14.2 Two Sample Hypothesis Test on Proportions

Moreover, we can use hypothesis test to compare proportions from two independent groups as well. We are going to start all these steps again.

### Step 1: Stating the Structure of Testing Hypothesis

Cases	Null Hypothesis	Alternative Hypothesis
1	$p_1 - p_2 = 0$	$p_1 - p_2 > 0$
2	$p_1 - p_2 = 0$	$p_1 - p_2 < 0$
3	$p_1 - p_2 = 0$	$p_1 - p_2 \neq 0$

Figure 14.3: All possible cases of two sample hypothesis test on proportions

In this case, we are interested in the difference of proportions.

### Step 2: Computing Test Statistics

**Definition 14.2** (Test statistics of two sample hypothesis test on proportions). —————  
*To test the hypothesis  $H_0 : p_1 = p_2$  first find the pooled proportion  $\hat{p}$  of successes in both samples combined. Then compute the  $Z^*$  statistic:*

$$z_* = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}} \sim N(0, 1).$$

*In this case,  $\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$ , which is the number of success with two groups combined.  $n_1$  and  $n_2$  represent sample size of each group respectively.*

### Step 3: Finding the P-value

In terms of a variable  $Z$  having the standard Normal distribution, the approximate P-value for a test of  $H_0$  against:

- $H_a : p_1 - p_2 > 0$  is  $P(Z > Z_*)$ ;
- $H_a : p_1 - p_2 < 0$  is  $P(Z < Z_*)$ ;

- $H_a : p_1 - p_2 \neq 0$  is  $2 \cdot P(Z > |Z_*|)$ .

#### Step 4: Comparing P-value with $\alpha$ -level

If p-value is less than  $\alpha$ -level, then we reject the null hypothesis ( $H_0$ ) and accept the alternative hypothesis ( $H_a$ ). Otherwise, If p-value is greater than  $\alpha$ -level, then we do not reject the null hypothesis ( $H_0$ ) and reject the alternative hypothesis ( $H_a$ ).

#### Step 5: Final Conclusion

If we reject the null hypothesis, then we conclude that: there is sufficient evidence to reject the null hypothesis. If we do not reject the null hypothesis, then we conclude that: there is insufficient evidence to reject the null hypothesis. Moreover:

- If  $p_1 - p_2 > 0$ : if we reject the null hypothesis under this alternative test, then  $p_1 > p_2$ .
- If  $p_1 - p_2 < 0$ : if we reject the null hypothesis under this alternative test, then  $p_1 < p_2$ .
- If  $p_1 - p_2 \neq 0$ : if we reject the null hypothesis under this alternative test, then  $p_1 \neq p_2$ .

#### Conditions of Two Sample Hypothesis Test on Proportions

i. Independent Response Assumption: Within each group, we need independent responses from the cases. We cannot check that for certain, but randomization provides evidence of independence. So, we need to check the following:

- Randomization Condition: The data in each group should be drawn independently and at random from a population or generated by a completely randomized designed experiment.
- The 10 % Condition: If the data are sampled without replacement, the sample should not exceed 10 % of the population. If samples are bigger than 10 % of the target population, random draws are no longer approximately independent.
- Independent Groups Assumption: The two groups we are comparing must be independent from each other.

ii. Sample Size Condition Each of the groups must be big enough. As with individual proportions, we need larger groups to estimate proportions that are near 0% and 100%. We check the success / failure condition for each group.

- Success / Failure Condition: Both groups are big enough that at least 10 successes and at least 10 failures have been observed in each group or will be expected in each (when testing hypothesis).

Note: Two-sided significance tests (later we will discuss this concept) are robust against violations of this condition. In this case, we can conduct significance tests with smaller sample sizes. In practice, the two-sided significance test works well if there are at least five successes and five failures in each sample.

#### Example 14.1. —

Nicotine patches are often used to help smokers quit. Does giving medicine to fight depression

help? A randomized double-blind experiment assigned 244 smokers who wanted to stop to receive nicotine patches and another 245 to receive both a patch and the anti-depression drug bupropion. Results: After a year, 40 subjects in the nicotine patch group had abstained from smoking, as had 87 in the patch-plus-drug group. How significant is the evidence that the medicine increases the success rate? State hypotheses, calculate a test statistic, use Table 6 to give its P-value, and state your conclusion. (Use  $\alpha = 0.01$ )

Solution:

**Step 1: State Hypothesis**

$$H_0 : p_1 = p_2 \text{ and } H_a : p_1 < p_2$$

**Step 2: Find test statistics**

$$\hat{p}_1 = \frac{40}{244} = 0.1639 \text{ and } \hat{p}_2 = \frac{87}{245} = 0.3551. \text{ Then } \hat{p} = \frac{40+87}{244+245} = 0.2597.$$

$$\text{Now, } z_* = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = -4.82$$

**Step 3: Compute p-value**

$$\text{P-value} = P(Z < z_*) = P(Z < -4.82) < 0.0003$$

**Step 4: Conclusion**

Since  $p\text{-value} < 0.0003 < \alpha = 0.01$ , we reject the null hypothesis that  $p_1 = p_2$ . The data provide very strong evidence that bupropion increases success rate.

**R-code:**

Input

```
successes=c(87, 40);

totals=c(245, 244);

prop.test(successes, totals, alternative="greater",correct=FALSE);
```

Output

```
##
## 2-sample test for equality of proportions without
## continuity correction
##
## data: x and n
## X-squared = 23.237, df = 1, p-value = 7.161e-07
## alternative hypothesis: greater
## 95 percent confidence interval:
## 0.1275385 1.0000000
## sample estimates:
##      prop 1      prop 2
## 0.3551020 0.1639344
```

### 14.3 Two Sample Hypothesis Test on Variances

Let's begin this type of hypothesis test with a case. The question is: how do you know whether the homogeneity of variance assumption is satisfied? One simple method involves just looking at two sample variances. Logically, if two population variances are equal, then the two sample variances should be very similar. When the two sample variances are reasonably close, you can be reasonably confident that the homogeneity assumption is satisfied and proceed with, for example, Student t-interval. However, when one sample variance is three or four times larger than the other, then there is reason for a concern. The common statistical procedure for comparing population variances  $\sigma_1^2$  and  $\sigma_2^2$  makes an inference about the ratio of  $\sigma_1^2/\sigma_2^2$ .

#### Step 1: Stating the Structure of Testing Hypothesis

Cases	Null Hypothesis	Alternative Hypothesis
1	$\sigma_1^2 = \sigma_2^2$	$\sigma_1^2 > \sigma_2^2$
2	$\sigma_1^2 = \sigma_2^2$	$\sigma_1^2 < \sigma_2^2$
3	$\sigma_1^2 = \sigma_2^2$	$\sigma_1^2 \neq \sigma_2^2$

Figure 14.4: All possible cases of two sample hypothesis test on variances

#### Step 2: Computing Test Statistics

**Definition 14.3** (Test statistics of two sample hypothesis test on variances). —————  
*The test statistics of two sample hypothesis test on variances is given by:*

$$F_* = \frac{s_1^2}{s_2^2} \sim F_{n_1-1, n_2-1}.$$

#### Decision Rules

- $H_0 : \sigma_1^2 = \sigma_2^2$  and  $H_a : \sigma_1^2 \neq \sigma_2^2$ . If  $F^* > F_{n_1-1, n_2-1, \alpha/2}$  or  $F^* < F_{n_1-1, n_2-1, 1-\alpha/2}$ , then we reject  $H_0$ . Otherwise, we do not reject it.
- $H_0 : \sigma_1^2 = \sigma_2^2$  and  $H_a : \sigma_1^2 > \sigma_2^2$ . If  $F^* > F_{n_1-1, n_2-1, \alpha}$  or  $P(F_{n_1-1, n_2-1} > F^*)$  is too small, then we reject  $H_0$ . Otherwise, we do not reject it.
- $H_0 : \sigma_1^2 = \sigma_2^2$  and  $H_a : \sigma_1^2 < \sigma_2^2$ . If  $F^* < F_{n_1-1, n_2-1, 1-\alpha}$  or  $P(F_{n_1-1, n_2-1} < F^*)$  is too small, then we reject  $H_0$ . Otherwise, we do not reject it.

## Chapter 15

# Introduction to Simple Linear Regression

In statistics, simple linear regression (SLR) is a linear regression model with a single explanatory variable. In other words, we use linear functions to illustrate the relationship of variables (ie. time and one's height). The goal of simple linear regression is to find the best-fitting straight line, known as the regression line, that predicts the dependent variable based on the independent variable. For example we are interested a people's height within 10 months. Then, we use coordinate system to draw each data point and use simple linear regression to find a function which perfectly describes the relationship between height and time.

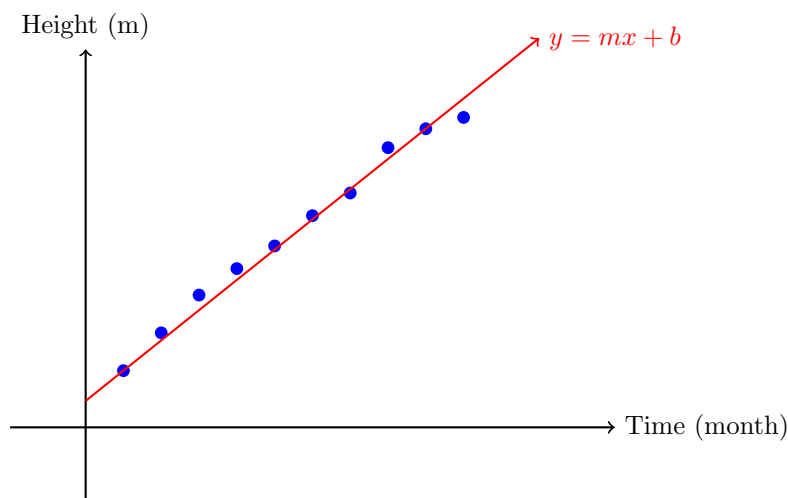


Figure 15.1: An illustration of simple linear regression. The blue points are measures of height monthly, and the red line is our SLR model. In this case  $m$  is the slope which tells you the rate of change,  $b$  is the intercept which may have a special meaning depending on the case.

Now, you may wonder the accuracy of this model. In statistics, we do have parameters that approximate the slope and intercept of the function  $y = mx + b$ . The model we are going to use is:  $\hat{y} = \hat{\beta}_1 x + \hat{\beta}_0 + \epsilon$ . From this model, the slope and the intercept are  $\hat{\beta}_1$ ,  $\hat{\beta}_0$  respectively ( $\hat{\beta}_1$  and  $\hat{\beta}_0$  are unbiased estimators). Moreover, the  $\epsilon$ -term is called error term, which we will discuss it later.

## 15.1 Measures of Linear Relationship

Before we formally introduce simple linear regression, there are some measures of SLR that should be discussed.

### Covariance (Sample Covariance)

In probability theory and statistics, covariance is a measure of the joint variability of two random variables. The covariance sign shows the direction of the linear relationship between two variables. If higher values of one variable tend to occur with higher values of the other (and lower with lower), the covariance is positive, meaning the variables move in the same direction. If higher values of one variable tend to occur with lower values of the other, the covariance is negative, meaning they move in opposite directions. The size (magnitude) of the covariance reflects how much the two variables vary together, based on the variances they share.

**Definition 15.1** (Covariance (sample covariance)).

*The formula of sample covariance is given by:*

$$S_{xy} = \frac{1}{n-1} \cdot \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \frac{\sum_{i=1}^n x_i \cdot y_i}{n-1} - \frac{n\bar{x}\bar{y}}{n-1}.$$

*These are the two ways to compute covariance. Both will give you the same answer.*

Basically, covariance indicates that how two variables move together.

- If covariance of two random variables is greater than 0 ( $cov(x, y) > 0$ ), then the two random variables show the same trend. That is: if one random variable is increasing, then the other one is also increasing; while if one random variable is decreasing, then the other one is also decreasing.
- If covariance of two random variables is less than 0 ( $cov(x, y) < 0$ ), then the two random variables show the opposite trend. That is: if one random variable is increasing, then the other one is decreasing; while if one random variable is decreasing, then the other one is increasing.
- If covariance of two random variables is equal to 0 ( $cov(x, y) = 0$ ), then we say that there is no relationship (systematically linear) between the two random variables.

Note that covariance is not standardized, so it can be difficult to interpret directly.

### Coefficient of Correlation

In statistics, correlation or dependence is any statistical relationship, whether causal or not, between two random variables or bivariate data. It helps us understand whether and how changes in one variable are associated with changes in another. A positive correlation means that as one variable increases, the other tends to increase as well, while a negative correlation means that one variable tends to decrease as the other increases. The degree of correlation is usually expressed with a correlation coefficient, which ranges from  $-1$  to  $+1$ .



**Definition 15.2** (Coefficient of correlation). 

---

The coefficient of correlation is given by:

$$r_{xy} = \frac{S_{xy}}{S_x \cdot S_y}.$$

Now,  $r_{xy}$  = sample correlation coefficient,  $S_{xy}$  = sample covariance,  $S_x$  = sample standard deviation of  $x$ ,  $S_y$  = sample standard deviation of  $y$ . Also, remember that the range of coefficient of correlation is between  $-1$  and  $+1$ :  $r_{xy} \in [-1, +1]$ .

---

The correlation  $r$  measures the strength and direction of the linear association between two quantitative variables  $x$  and  $y$ . Although you calculate a correlation for any scatter plot,  $r$  measures only straight-line relationships. In short, coefficient of correlation is a measure of the strength of the linear relationship between two random variables.

- If  $r_{xy} \approx +1$ , then we say that the two random variables have a strong positive correlation. (See figure 15.2)

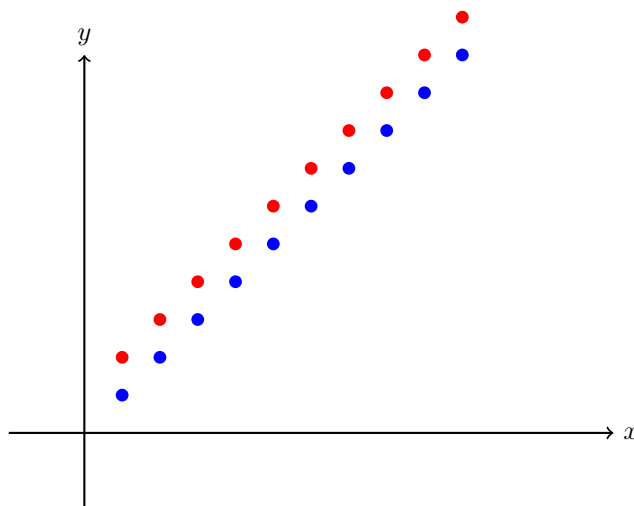


Figure 15.2: An illustration of strong positive correlation ( $r_{xy} \approx +1$ ).

- If  $r_{xy} \approx -1$ , then we say that the two random variables have a strong negative correlation. (See figure 15.3)

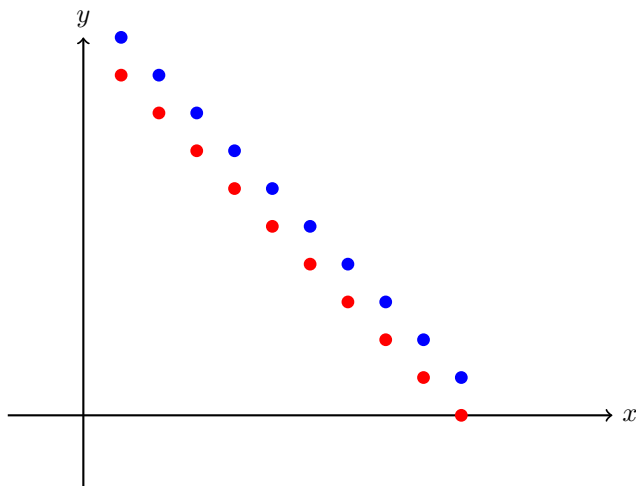


Figure 15.3: An illustration of strong negative correlation ( $r_{xy} \approx -1$ ).

- If  $r_{xy} \approx 0$ , then we say that there is essentially no correlation between the two random variables. Note that if  $r \approx 0$ , then it suggests a linear relationship doesn't exist but other relationship may exist. (See figure 15.4)

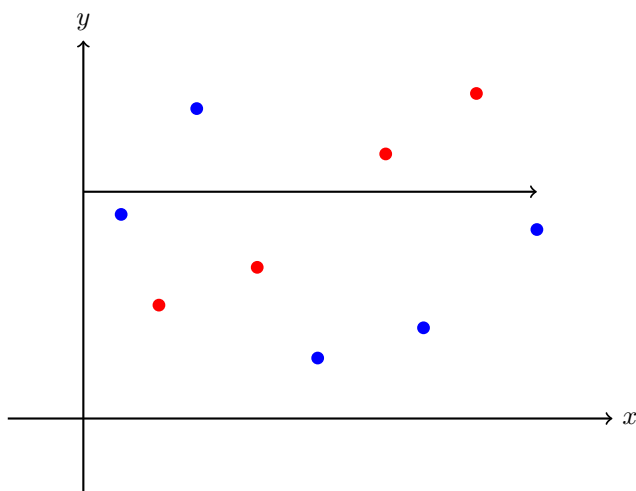


Figure 15.4: An illustration of no correlation ( $r_{xy} \approx 0$ ).

Note that correlation doesn't imply causation:

- $cor(x, y) \approx +1$  doesn't necessarily imply on increase in  $x$  causes increase in  $y$ .
- $cor(x, y) \approx -1$  doesn't necessarily imply on increase in  $x$  causes decrease in  $y$ .

### Properties of Covariance and Correlation

These two values are symmetric:

- $cov(x, y) = cov(y, x)$ ;
- $cor(x, y) = cor(y, x)$ .

**Example 15.1.**

Five observations taken for two variables follow.

$x_i$	$y_i$
4	50
6	50
11	40
3	60
16	30

Figure 15.5: Data of example 15.1

- (a) Compute the sample covariance.  
 (b) Compute and interpret the sample correlation coefficient.

**Solution:**

Step 1: Compute  $\bar{x}$  and  $\bar{y}$ ;  $\bar{x} = 8$  and  $\bar{y} = 46$  (check this by yourself).

Step 2: Find  $s_x$  and  $s_y$ .

$$s_x^2 = \frac{1}{5-1} \cdot \sum_{i=1}^5 (x_i - \bar{x})^2 = 29.5 \text{ and } s_y^2 = \frac{1}{5-1} \cdot \sum_{i=1}^5 (y_i - \bar{y})^2 = 130$$

Then:  $s_x = 5.4313$  and  $s_y = 11.4017$ .

Step 3: Find  $s_{xy}$  and  $r$ .

$$\sum_{i=1}^5 x_i \cdot y_i = 1600, \text{ then } s_{xy} = \frac{1600}{5-1} - \frac{5 \cdot 8 \cdot 46}{5-1} = -60 \text{ and } r_{xy} = \frac{s_{xy}}{s_x \cdot s_y} = 0.9688.$$

**R code**

Step 1: Entering data;

```
X=c(4,6,11,3,16);
Y=c(50,50,40,60,30);
```

Step 2: Finding means;

```
mean(X);
mean(Y);
```

Step 3: Finding variances;

```
var(X);
var(Y);
```

Step 4: Finding standard deviations;

```
sd(X);
```

```
sd(Y);
```

Step 5: Finding covariance and correlation;

```
cov(X,Y);
```

```
cor(X,Y);
```

## 15.2 Least Squares Method

The method of least squares is a mathematical optimization technique used to find the best-fitting function by minimizing the sum of the squared differences between the observed data points and the values predicted by the model. It is interested in a linear model of the form:  $y = \beta_0 + \beta_1 \cdot x_1 + \dots + \beta_p \cdot x_p + \epsilon$ , where  $\epsilon \sim N(0, \sigma^2)$ ,  $x_i$ 's ( $i = 1, \dots, p$ ) are independent predictors,  $\beta_j$ 's ( $j = 0, \dots, p$ ) are coefficients,  $y$  is dependent variable.

Using sample data, we get estimates of this model of the form:  $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_1 + \dots + \hat{\beta}_p \cdot x_p$ . Now,  $\hat{y}$  is predicted  $y$ ,  $x_i$ 's ( $i = 1, \dots, p$ ) are independent predictors,  $\hat{\beta}_j$ 's ( $j = 0, \dots, p$ ) are estimated coefficients. Moreover,  $\hat{\beta}_0$  is intercept;  $\hat{\beta}_1, \dots, \hat{\beta}_p$  are quantifiers how much  $y$  changes with a unit increase in  $x_i$ 's.

In this course, we focus on the following model a bit more:  $\hat{y} = b_0 + b_1 \cdot x$ , where  $b_0$  is the y-intercept, and  $b_1$  is the slope, and  $\hat{y}$  is the value of  $y$  determined by the line. The coefficients  $b_0$  and  $b_1$  are derived using Calculus so that we minimize the sum of squared deviations:  $\sum_{i=1}^n (y_i - \hat{y}_i)^2$ . Then the least squares line coefficients are  $b_1 = r \cdot \frac{s_y}{s_x}$  and  $b_0 = \bar{y} - b_1 \cdot \bar{x}$ .

### Facts about Least Squares Method

- The distinction between explanatory and response variables is essential in Least Squares Method.
- The least-squares line (trendline) always passes through the point  $(\bar{x}, \bar{y})$  on the graph of  $y$  against  $x$ .
- The square of the correlation,  $r^2$ , is the fraction of the variation in the values of  $y$  that is explained by the variation in  $x$ .

**Example 15.2.**

A tool die maker operates out of a small shop making specialized tools. He is considering increasing the size of his business and needs to know more about his costs. One such cost is electricity, which he needs to operate his machines and lights. He keeps track of his daily electricity costs and the number of tools that he made that day. These data are listed next. Determine the fixed and variable electricity costs using the Least Squares Method.

Day	Number of tools ( $X$ )	Electricity costs ( $Y$ )
1	7	23.80
2	3	11.89
3	2	15.89
4	5	26.11
5	8	31.79
6	11	39.93
7	5	12.27
8	15	40.06
9	3	21.38
10	6	18.65

Figure 15.6: Data of example 15.2

Solution:

Step 1: Entering Data;

```
tools=c(7,3,2,5,8,11,5,15,3,6);
cost=c(23.80,11.89,15.98,26.11,31.79, 39.93,12.27,40.06,21.38,18.65);
```

Step 2: Finding Slope;

```
Sx=sd(tools);
Sy=sd(cost);
r=cor(tools,cost);
b1=r*(Sy/Sx);
b1;
## [1] 2.245882
```

Step 3: Finding  $y$ -intercept;

```
x.bar=mean(tools);

y.bar=mean(cost);

b0=y.bar - b1*x.bar;

b0;

## [1] 9.587765
```

We can also use R-code to draw a graph:

```
plot(tools,cost,pch=19);

abline(least.squares$coeff,col="red");

# pch=19 tells R to draw solid circles;

# abline tells R to add trendline;
```

Interpretation:

The slope measures the marginal rate of change in the dependent variable. In this example, the slope is 2.25, which means that in this sample, for each one-unit increase in the number of tools, the marginal increase in the electricity cost is \$2.25 per tool.

The  $y$ -intercept is 9.57; that is, the line strikes the  $y$ -axis at 9.57. However, when  $x = 0$ , we are producing no tools and hence the estimated fixed cost of electricity is \$9.57 per day .

## 15.3 Simple Linear Regression

### Estimating Regression Model Parameters

The regression line which we are going to use is:  $E(Y) = \beta_0 + \beta_1 \cdot x$ . This is fitted to the data points  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  by finding the line that is closest to the data in some sense. There are many ways in which closeness can be defined, but the method most generally used is to consider the vertical deviations between the line and the data points:  $y_i - (\beta_0 + \beta_1 \cdot x_i), 1 \leq i \leq n$ .

The fitted line is chosen to be the line that minimizes the sum of the squares of these vertical deviations

$$Q = \sum_{i=1}^n [y_i - (\beta_0 + \beta_1 \cdot x_i)]^2$$

and this is referred to as the least squares fit. (The quantity  $Q$  is also called the **sum of squares for error**, SSE.)

The parameter estimates  $\hat{\beta}_0$  and  $\hat{\beta}_1$  are therefore the values that minimize the quantity  $Q$ . They are found taking partial derivatives of  $Q$  with respect to  $\hat{\beta}_0$  and  $\hat{\beta}_1$  and setting the resulting expressions equal to 0.

Now, you know the method to get the regression model  $E(Y) = \beta_0 + \beta_1 \cdot x$ , the following lines are the derivation:

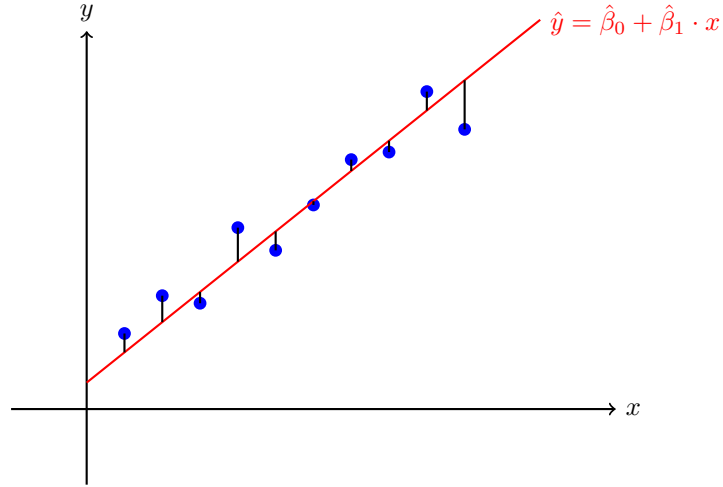


Figure 15.7: An illustration of a simple linear regression model. The red line is the fitted regression line, and the full black vertical lines represent the residuals (SSE). The data points deviate from the line to show errors clearly. Note that the sum of residuals is necessarily 0.

The following proof is the derive of  $\hat{\beta}_0$  and  $\hat{\beta}_1$  for simple linear regression model:  $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x + \epsilon$ .

*Proof.*

Firstly, we examine sum of residual squared:

$$\begin{aligned} \sum_{i=1}^n e_i^2 &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 = 0 \\ &= \sum_{i=1}^n [y_i - (\hat{\beta}_0 + \hat{\beta}_1 \cdot x_i)]^2 = 0 \\ &= \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot x_i)^2 = 0 \end{aligned}$$

To find  $\hat{\beta}_0$  and  $\hat{\beta}_1$  which minimizes  $\sum_{i=1}^n e_i^2$ , we need partial derivative with respect to  $\hat{\beta}_0$

and  $\hat{\beta}_1$ :

$$\begin{aligned}
\frac{\partial}{\partial \hat{\beta}_0} \sum_{i=1}^n e_i^2 &= \frac{\partial}{\partial \hat{\beta}_0} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot x_i)^2 = 0 \\
&= \sum_{i=1}^n 2(\hat{\beta}_0 + \hat{\beta}_1 x_i - y_i) = 0 \\
&= \sum_{i=1}^n \hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^n x_i - \sum_{i=1}^n y_i = 0; \text{ (consider: } \sum_{i=1}^n x_i = n\bar{x} \text{)} \\
&= n \cdot \hat{\beta}_0 + \hat{\beta}_1 \cdot n\bar{x} - n\bar{y} = 0
\end{aligned}$$

Hence:  $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \cdot \bar{x}$ . (Equ 1)

$$\begin{aligned}
\frac{\partial}{\partial \hat{\beta}_1} \sum_{i=1}^n e_i^2 &= \frac{\partial}{\partial \hat{\beta}_1} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot x_i)^2 = 0 \\
&= \sum_{i=1}^n 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-x_i) = 0 \\
&= \sum_{i=1}^n 2(\hat{\beta}_0 x_i + \hat{\beta}_1 x_i^2 - y_i x_i) = 0 \\
&= \hat{\beta}_0 \sum_{i=1}^n x_i + \hat{\beta}_1 \sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i y_i = 0 \\
&= \hat{\beta}_1 \sum_{i=1}^n x_i^2 + \sum_{i=1}^n (\bar{y} - \hat{\beta}_1 \bar{x}) x_i - \sum_{i=1}^n x_i y_i = 0; \text{ (sub Equ 1 into this line)} \\
&= \hat{\beta}_1 \sum_{i=1}^n x_i^2 + \bar{y} \sum_{i=1}^n x_i - \hat{\beta}_1 \bar{x} \sum_{i=1}^n x_i - \sum_{i=1}^n x_i y_i = 0 \\
&= \hat{\beta}_1 \sum_{i=1}^n x_i^2 + n\bar{x}\bar{y} - n\hat{\beta}_1(\bar{x})^2 - \sum_{i=1}^n x_i y_i = 0 \\
&= \hat{\beta}_1 [\sum_{i=1}^n x_i^2 - n(\bar{x})^2] + n\bar{x}\bar{y} - \sum_{i=1}^n x_i y_i = 0
\end{aligned}$$

$$\text{Hence: } \hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n(\bar{x})^2} = \frac{S_{xy}}{S_{xx}}.$$

□

## Introduction to Simple Linear Regression

At this point, we are going to provide the definition of simple linear regression model as the following:



**Definition 15.3** (Simple Linear Regression). 

---

Let  $x$  be independent variable and  $y$  be dependent variable, then the model of simple linear regression is:  $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x + \epsilon$ , where  $\hat{\beta}_0$  represents the  $y$ -intercept,  $\hat{\beta}_1$  represents the slope and  $\epsilon$  is the error term that  $\epsilon \sim N(0, \sigma^2)$ . Moreover,  $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \cdot \bar{x}$  and  $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$ , which is also equal to:

$$\frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - n(\bar{x})^2}.$$


---

**Example 15.3.** 

---

Suppose an appliance store conducts a 5-month experiment to determine the effect of advertising on sales revenue. The results are shown in a table below. The relationship between sales revenue,  $y$ , and advertising expenditure,  $x$ , is hypothesized to follow a first-order linear model, that is,  $y = \beta_0 + \beta_1 \cdot x + \epsilon$ , where  $y$  = dependent variable,  $x$  = independent variable,  $\beta_0$   $y$ -intercept,  $\beta_1$  = slope of the line and  $\epsilon$  = error variable.

---

## 15.4 SST, SSE and SSR

Early in Section 15.3, we introduced a value called the sum of residual squared (SSE). There are two more values that are important in simple linear regression, which are total sum of squares (SST) and sum of squares for regression (SSR). We will introduce all these three values with figures, so that you may have a better understanding of what they measure.

**SST (Total Sum of Squares)**

It is defined as the sum over all squared differences between the observations and their overall mean  $\bar{y}$ .

**Definition 15.4** (SST (Total sum of squares)). 

---

For any simple linear regression model, SST (Total sum of squares) measures the sum over all squared differences between the observations and their overall mean  $\bar{y}$  is given by:

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2.$$


---

**SSE (Sum of Residual Squared)**

It is the sum of the squares of residuals (deviations predicted from actual empirical values of data). It is a measure of the discrepancy between the data and an estimation model, such

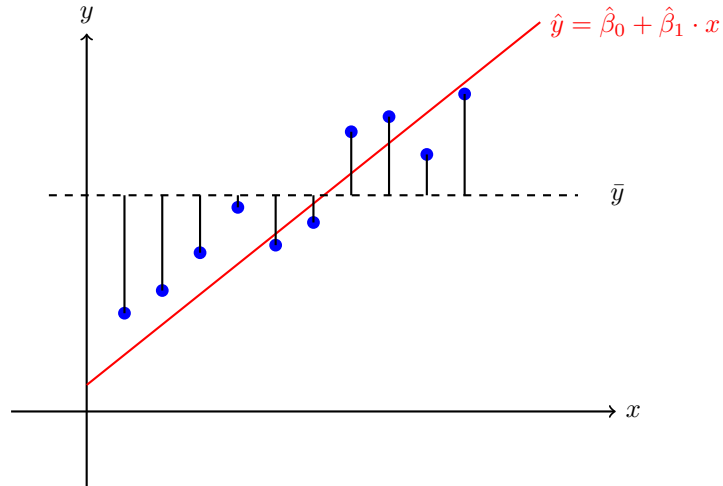


Figure 15.8: An illustration of Total Sum of Squares (SST). The blue points represent height measurements over time, the dashed line is the mean height  $\bar{y}$ , and the solid black vertical lines represent the squared deviations from the mean (SST components).

as a linear regression. A small SSE indicates a tight fit of the model to the data.

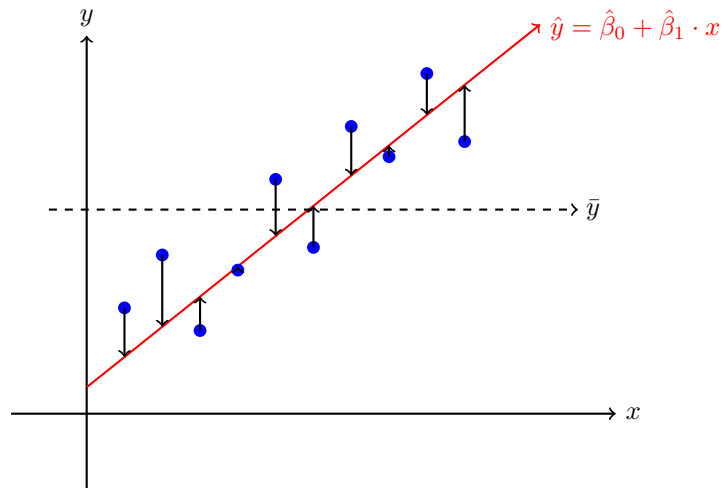


Figure 15.9: A simple linear regression illustration with residuals shown. Blue points are observed data, red line is the model, dashed line is the average of  $y$ , and black lines represent residuals  $y_i - \hat{y}_i$ .

---

**Definition 15.5** (SSE (Sum of residual squared)).

*For any simple linear regression model, SSE (Sum of residual squared) measures the distance between observed data and estimated data, which is given by:*

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$


---

Note that SSE (sum of residual squared) is an explained variation.

### SSR (Sum Square Regression)

It measures the distance between estimated value (estimated dependent data) and the mean of dependent data ( $\bar{y}$ ).

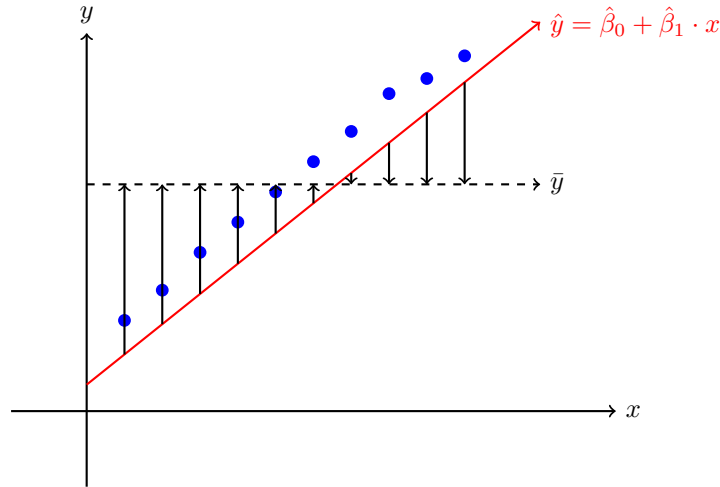


Figure 15.10: An illustration of simple linear regression. The blue points represent monthly height measurements with added variability. The red line is the fitted simple linear regression model. The dashed line shows the mean of the dependent variable,  $\bar{y}$ , and the solid black lines illustrate the deviation of the model's predictions from this mean.

#### Definition 15.6 (SSR (Sum square regression)).

*For any simple linear regression, the distance between the mean of dependent value and estimated dependent value is called sum square regression (SSR), which is given by*

$$SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2.$$

Note that sum square regression (SSR) is an explained variation.

### Summary

Now, let's zoom in to see SST, SSE and SSR. Note that the relationship of these measures is that total deviation is equal to unexplained deviation (error) plus explained deviation (regression), that is:  $(y_i - \bar{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$ .

We square all three deviations for each one of our data points, and sum over all  $n$  points. Here, cross terms drop out, and we are left with the following equation:

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2,$$

$$SST = SSE + SSR.$$

Total sum of squares = Sum of squares for error + Sum of squares for regression.

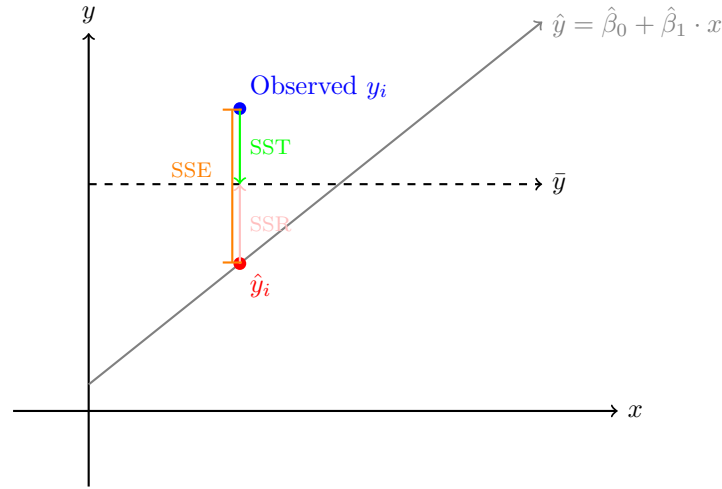


Figure 15.11: Visual representation of regression variability. **SST** (green) is total variability from the mean, **SSR** (pink) is explained variability, and **SSE** (orange curly brace) is the residual (unexplained) variability between the observed value  $y_i$  and the prediction  $\hat{y}_i$ .

### Coefficient of Determination ( $r^2$ )

Moreover, we can use SST, SSE and SSR to calculate another value which is important in simple linear regression, that is coefficient of determination. It is proportion of variability in  $y$  which is explained by  $x$ .

#### Definition 15.7 (Coefficient of determination).

We define the coefficient of determination as the sum of squares due to the regression divided by the total sum of squares.

$$r^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}.$$

The coefficient of determination can be interpreted as the proportion of the variation in  $Y$  that is explained by the regression relationship of  $Y$  with  $X$  (or the proportion of the total corrected sum of squares explained by the regression). Note that:  $0 \leq r^2 \leq 1$ .

# Index

Introduction, [2](#)

Overview, [1](#)

Statistics

Definition, [2](#)

Introduction, [1](#)