STATISTICS WITH APPLIED PROBABILITY

Custom E-Book for STA258

Nishan Mudalige
Nurlana Alili
Bryan Su

CANADA

Statistics with Applied Probability
Custom E-Book 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

First edition: August 2025

Mudalige, M.; Alili, N.; Xu, B.

University of Toronto Mississauga,
Mississauga, Ontario, Canada

Contents

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 todays 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 Companies that don't adapt and stick to an outdated business model tend to pay the price.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 $x_1, x_2, x_3, ..., x_n$ 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 $x_1, x_2, x_3, ..., x_n$ be a sample of data points, we define sample variance of the sample data points (s^2) as:*

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

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 $x_1, x_2, x_3, ..., x_n$ 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: $x_1, x_2, x_3, ..., x_n$ 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:*
$$M = \frac{n+1}{2}, \text{ where } n \text{ represents the } n^{th} \text{ position.}$$
- *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^{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: $x_1, x_2, ..., x_n$ 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 χ^2_n 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 μ is same as the probability below than μ. 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)

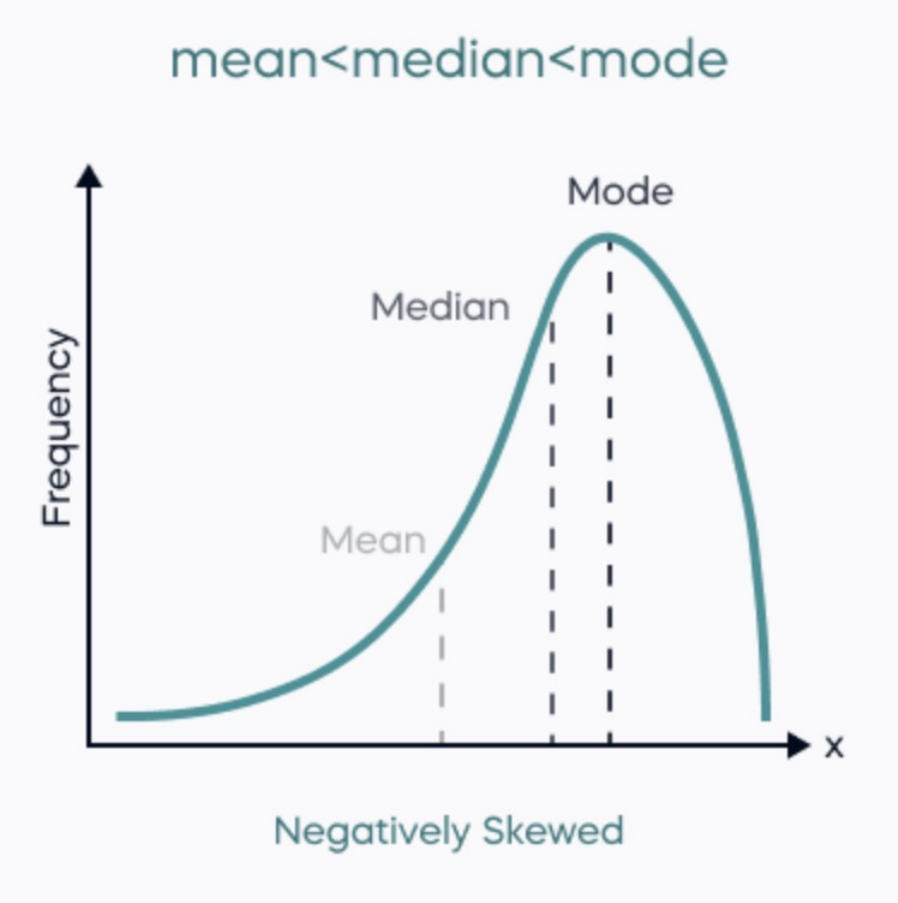


Figure 1.1: Visualization of left skew probability distribution

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)

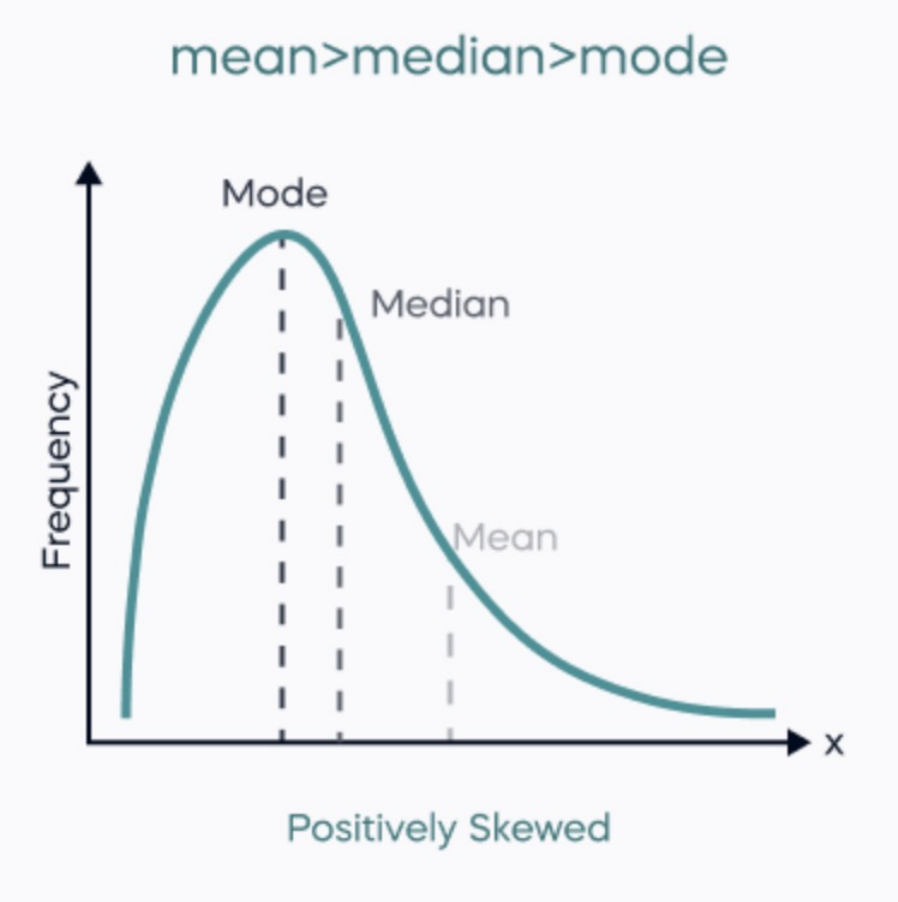


Figure 1.2: Visualization of right skew probability distribution

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

Definition 1.18 (Symmetry)**.** *In statistics, symmetry s 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 μ be its mean and σ 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. $P(\mu - 3\sigma < X < \mu + 3\sigma) = 99.73\%$.

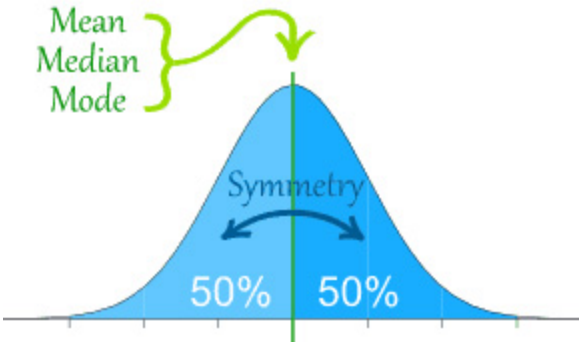


Figure 1.3: Visualization of symmetric probability distribution

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.

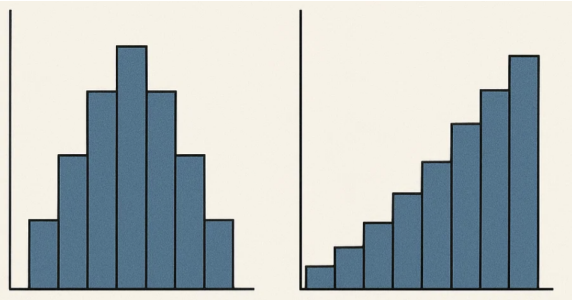


Figure 1.4: Visualization of histograms

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:

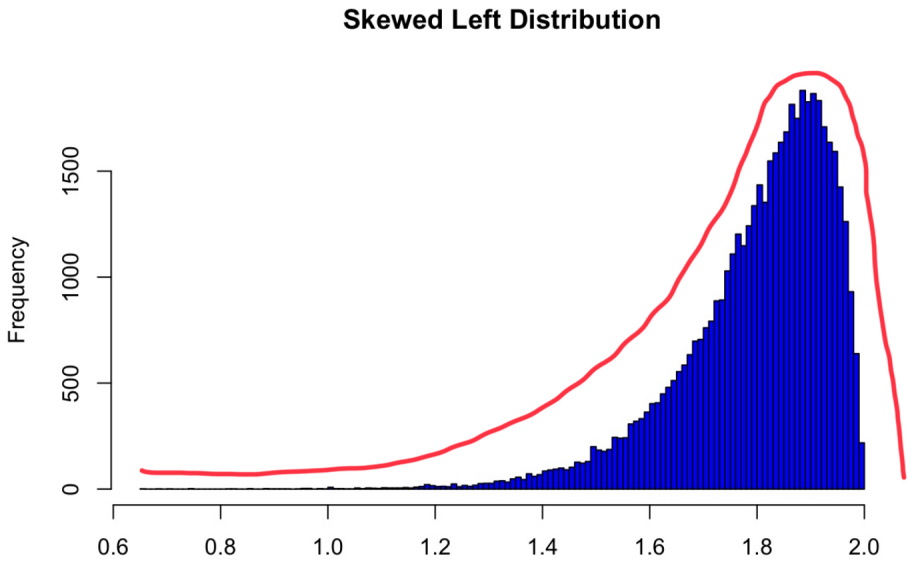


Figure 1.5: Visualization of a histogram has a left (or negative) skew probability distribution

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

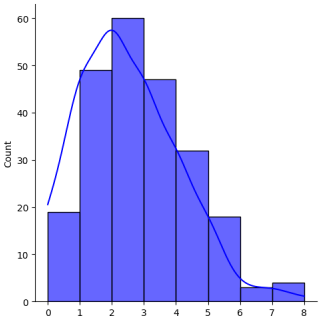


Figure 1.6: Visualization of a histogram has a right (or positive) skew probability distribution

For a histogram to have a symmetric probability distribution:

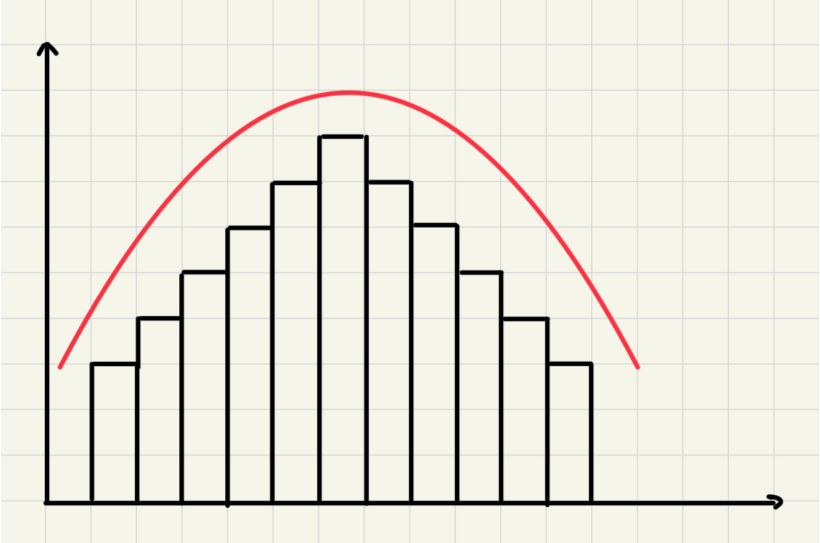


Figure 1.7: Visualization of a histogram has a symmetric probability distribution

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)

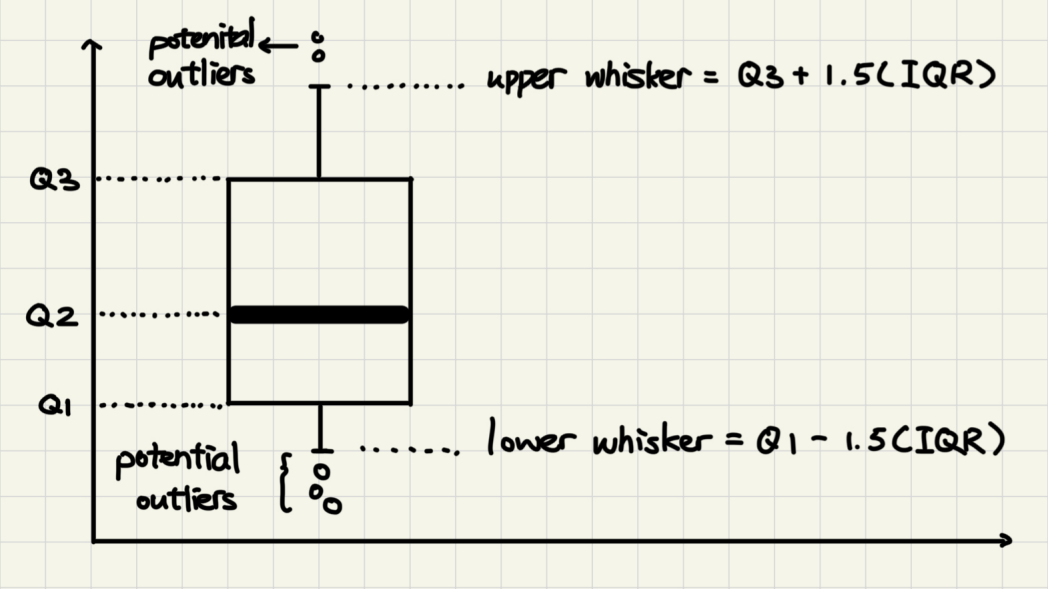


Figure 1.8: Visualization of a box-plot

Similar to histograms, we can still obtain information about skewness and symmetry, by observing the cut from the line of Q2.

If the median (Q2) 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 (Q2) 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 (Q2) cuts the box with upper area equal to lower area, then we say that box-plot with symmetric probability distribution.

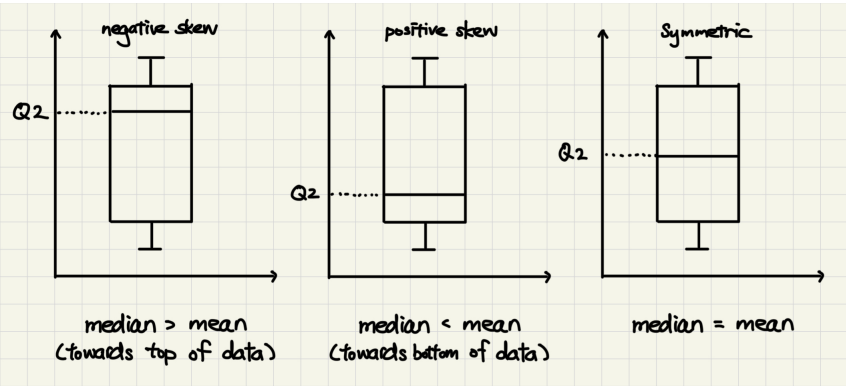


Figure 1.9: Visualization of a box-plot with skew and symmetric probability distribution

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.

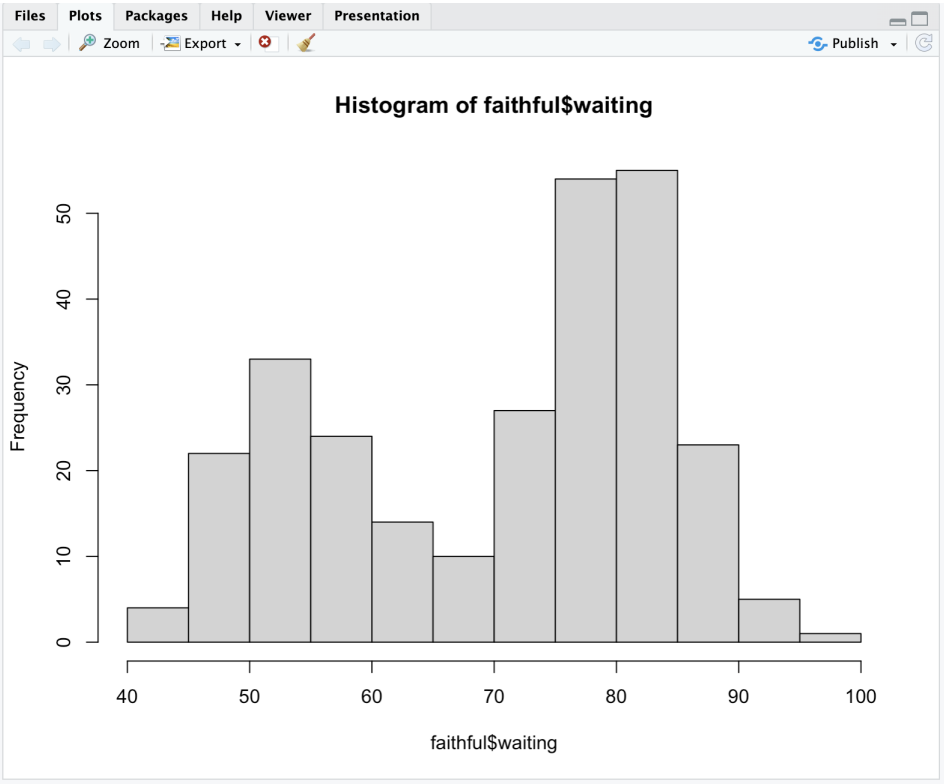


Figure 1.10: R-studio first three steps (by following the instructions, you should get this histogram)

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 (χ^2_n) 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*

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

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 $X \sim Gamma(\alpha = \frac{n}{2}, \beta = 2)$.
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 $X \sim \chi^2_n$. □

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* $Z^2 \sim \chi^2_1$.

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 $Z_1, Z_2, ..., Z_n \overset{i.i.d.}{\sim} N(0,1)$, 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 $Z_1, \ldots, Z_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. The moment generating function (MGF) of δ is:

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

Since $Z_1, \ldots, Z_n$ 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. □

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 σ^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:*

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

Alternatively, define a new variable T is the following:

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

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

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

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:*

$$F = \frac{(\frac{W_1}{v_1})}{(\frac{W_2}{v_2})} \sim F_{v_1,\, v_2};\ \textit{for } W_1 \sim \chi^2_{v_1}\ \textit{and } W_2 \sim \chi^2_{v_2};\ \textit{also both } W_1\ \textit{and } W_2\ \textit{are independent.}$$

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:

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

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 μ and population variance σ^2.. Then we take samples of same size n from each population, such that we have $\bar{x}_1, \bar{x}_2, ..., \bar{x}_m$ 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 row is sampling distribution of $\bar{x}$. As a result, $\bar{x}$ follows a normal distribution with mean $\mu_{\bar{x}} = \mu$ and variance $\sigma^2_{\bar{x}} = \frac{\sigma^2}{n}$, which is denoted as the following:

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

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

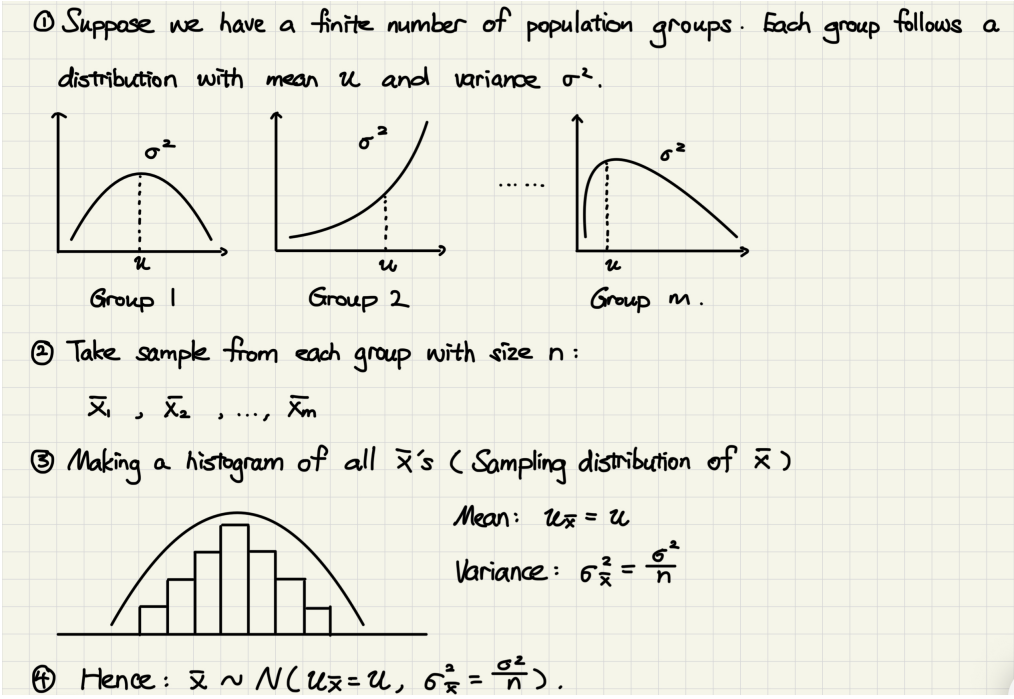


Figure 3.1: Procedure of the Central Limit Theorem

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

Definition 3.1 (Central Limit Theorem)**.** *Let $X_1, X_2, ..., X_n$ be independent and identically distributed random variables with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2 < \infty$. Then, we define the following:*

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

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

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

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:

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

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:

$$X = \text{number of successes in } n \text{ trials}$$

Then:

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

where:

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

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 $X \sim Bernoulli(p)$. 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 $X \sim Bernoulli(p)$. The mean of X is*

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

and the variance of X is

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

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

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

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

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

and the variance of X is

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

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.

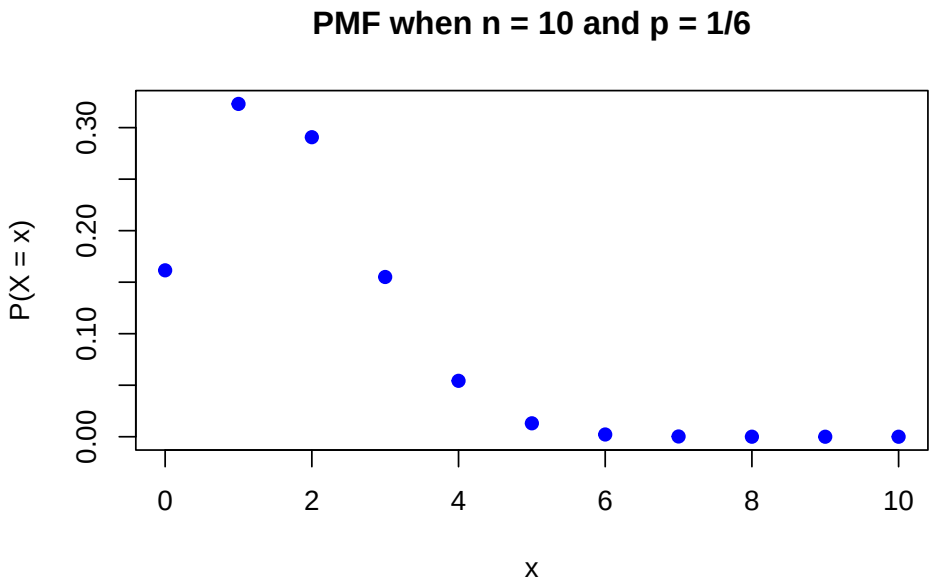


Figure 4.1: PMF of Binomial distribution with $n = 10$ and $p = \frac{1}{6}$.

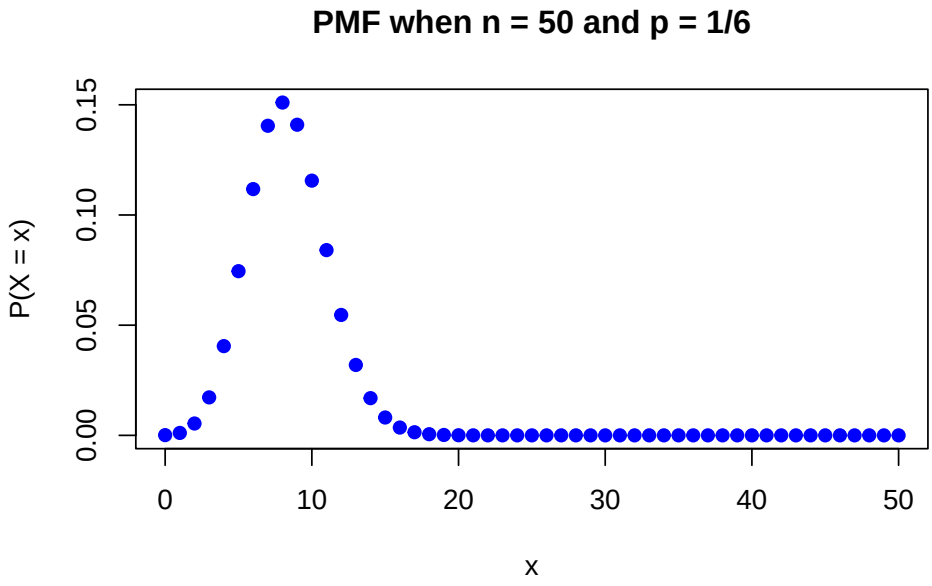


Figure 4.2: PMF of Binomial distribution with $n = 50$ and $p = \frac{1}{6}$.

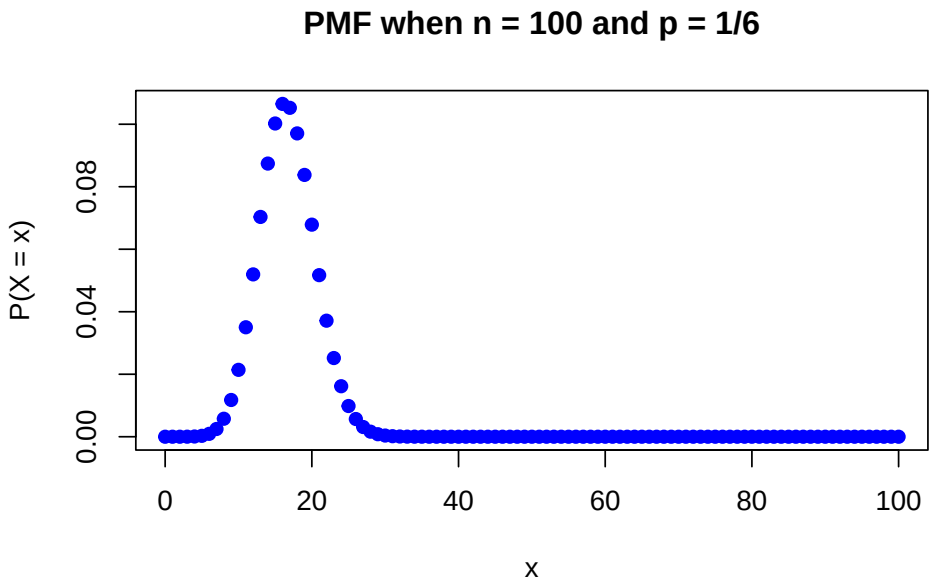


Figure 4.3: PMF of Binomial distribution with $n = 100$ and $p = \frac{1}{6}$.

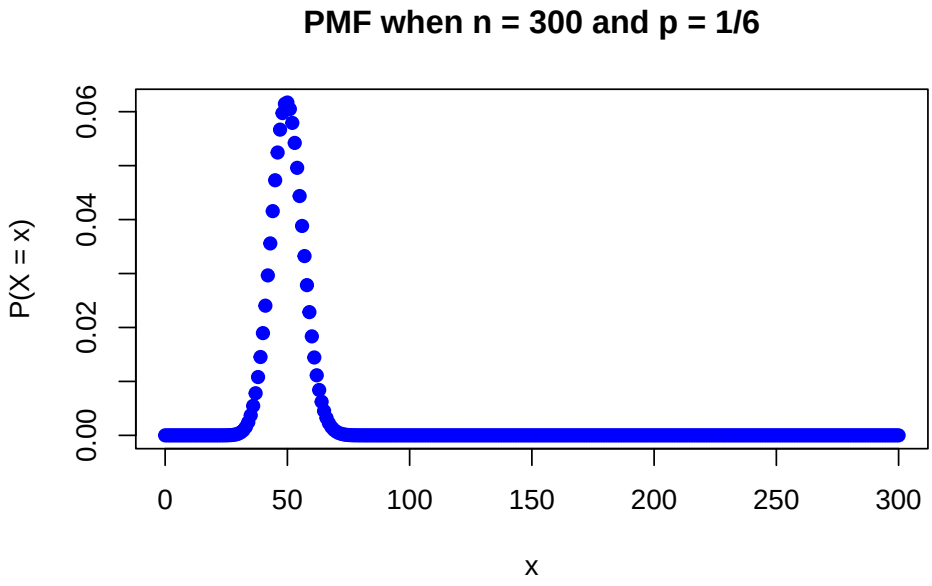


Figure 4.4: PMF of Binomial distribution with $n = 300$ and $p = \frac{1}{6}$.

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 $\sqrt{\frac{p(1-p)}{n}}$.

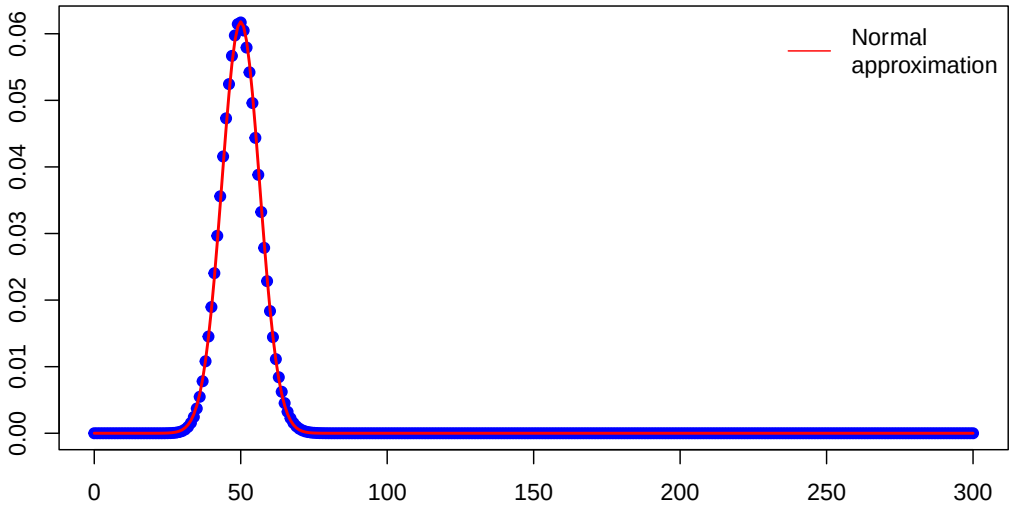


Figure 4.5: Binomial distribution with $n = 300$, $p = \frac{1}{6}$ and its Normal approximation.

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:

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

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

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

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):

```
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

$$\begin{aligned} P(\hat{p} > 0.50) &= P\left(\frac{\sum_{i=1}^{n} x_i}{n} > 0.50\right) \\ &= P\left(\sum_{i=1}^{300} x_i > 150\right) \\ &= 1 - P\left(\sum_{i=1}^{300} x_i \leq 150\right) \\ &\text{(it can be shown that } Y = \sum_{i=1}^{300} x_i \text{ has a Binomial distribution with} \\ &n = 300 \text{ and } p = 0.52) \\ &= 1 - F_Y(150) \end{aligned}$$

R code (using Binomial distribution):

```
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

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

(it can be shown that $Y = \sum_{i=1}^{300} x_i$ has a Binomial distribution with $n = 300$ and $p = 0.52$).

$$
\begin{aligned}
&\approx 1 - P\left(\sum_{i=1}^{300} x_i \leq 150.5\right) \quad \text{(continuity correction)} \\
&= 1 - P\left(\frac{\sum_{i=1}^{300} x_i}{n} \leq \frac{150.5}{300}\right) \\
&= 1 - P(\hat{p} \leq 0.5017) \\
&= 1 - P\left(Z \leq -0.6354\right) \quad \text{(Why?)}
\end{aligned}
$$

R code (Normal approximation with continuity correction):

```
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 Normal Approximation to Binomial

Let $X = \sum_{i=1}^{n} Y_i$ where $Y_1, Y_2, \ldots, Y_n$ 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: $V(\hat{p}) = np(1-p) = npq$.

4.7 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.

Continuity Correction Table

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

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.

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

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

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 $X \sim Binomial(n = 65, p = 0.51)$.

Use Normal Approximation We use normal approximation to the binomial. First, compute 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:

$$P(X = 40) = P(39.5 \leq X \leq 40.5)$$

$$= P\left(\frac{39.5 - 33.15}{4.06} \leq Z \leq \frac{40.5 - 33.15}{4.06}\right) = P(1.56 \leq Z \leq 1.81)$$

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:

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

Chapter 5

Law of Large Numbers

5.1 Convergence in Probability

Definition 5.1 (Convergence in Probability)**.** *The sequence of random variables $X_1, X_2, X_3, \ldots, X_n, \ldots$ is said to* ***converge in probability*** *to the constant c, if for every $\epsilon > 0$,*

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

or equivalently,

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

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 μ and variance σ^2. Then, for any $k > 0$,*

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

Using complements:

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

5.2 Weak Law of Large Numbers (WLLN)

Definition 5.3 (Weak Law of Large Numbers (WLLN)). *Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables, each having finite mean $E(X_i) = \mu$ and variance $\mathrm{Var}(X_i) = \sigma^2$. Then, for any $\epsilon > 0$,*

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

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

Proof of the Weak Law of Large Numbers (WLLN)

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

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

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

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

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 μ and variance σ^2,

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

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

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

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

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

Since probabilities are always non-negative, we conclude:

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

By the definition of convergence in probability,

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

□

Example 5.1.

[Poisson Convergence via WLLN]
Let X_i, for $i = 1, 2, 3, \ldots$, 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:

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

Applying Chebyshev's Inequality:

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

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

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

Conclusion:

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

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")
```

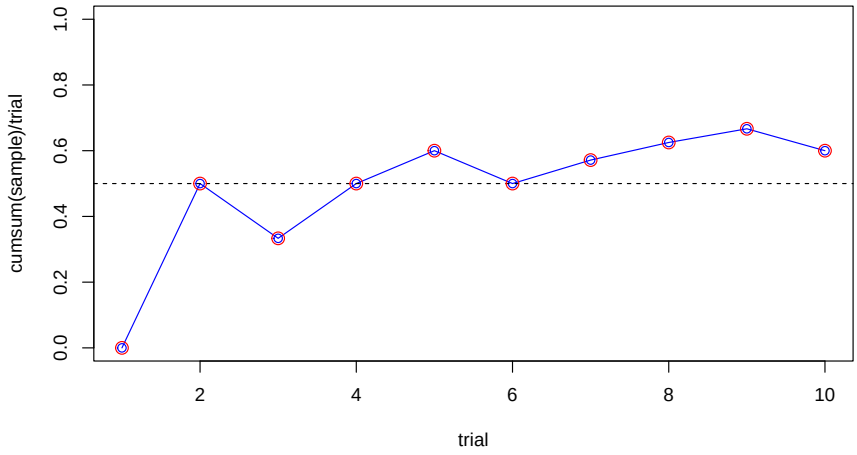


Figure 5.1: Simulation of running sample mean of Bernoulli($p = 0.5$) trials over time.

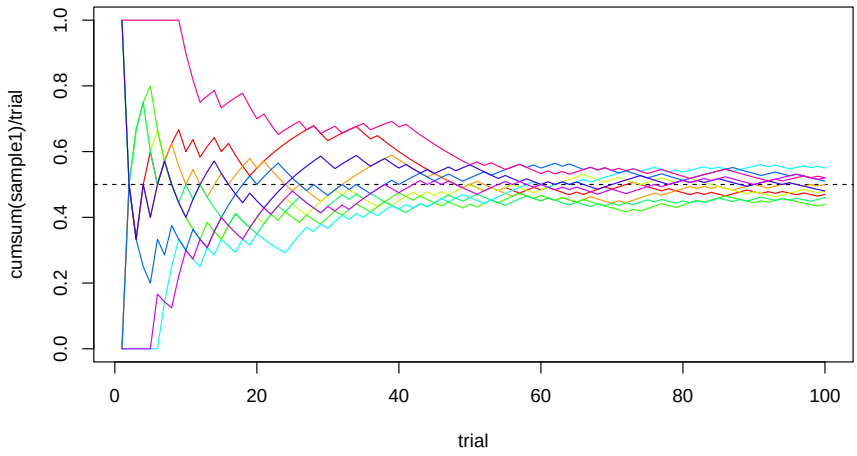


Figure 5.2: Simulation of 10 running sample means of Bernoulli($p = 0.5$) trials converging over 100 trials.

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, μ?" 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, μ, 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.*
μ*: population mean*
σ^2*: population variance*
σ*: 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, μ. This is called a **point estimate**.

But, we know it's not equal to μ. 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, μ.

Example 6.1.

- $\bar{x}$ is a point estimate of μ
- s^2 is a point estimate of σ^2
- s is a point estimate of σ

(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.*

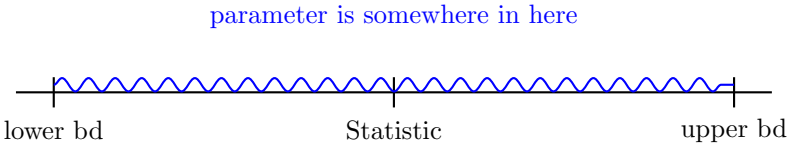


Figure 6.1: Parameter location within the interval.

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:

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

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.

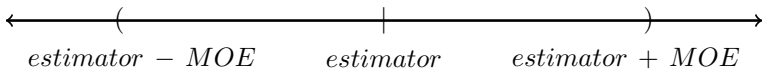


Figure 6.2: Visualization of a confidence interval on the real number line. The margin of error is abbreviated as MOE. The estimator is the centre of the interval. The confidence interval consists of all values between the estimator$-MOE$ and the estimator$+MOE$.

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 μ. We would expect approximately 95% of these 1000 intervals (i.e. $95\% \times 1000 = 950$) to actually capture μ.

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 μ (Known Variance)

When we know the population standard deviation σ, we can construct a confidence interval for μ in the following manner.

Confidence Interval 6.1 (Confidence Interval on μ when σ is Known)

A $(100-\alpha)\%$ *confidence interval on* μ *when* σ *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 $X_1, X_2, ..., X_n$ be iid $N(\mu, \sigma^2)$, where μ is unknown and σ 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 μ is a fixed, but unknown, number.
- The probability μ is inside the random interval is 0.95 (success rate of the method).
- 95% of all samples give an interval that captures μ, 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 μ 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;
```

```
## 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);
```

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

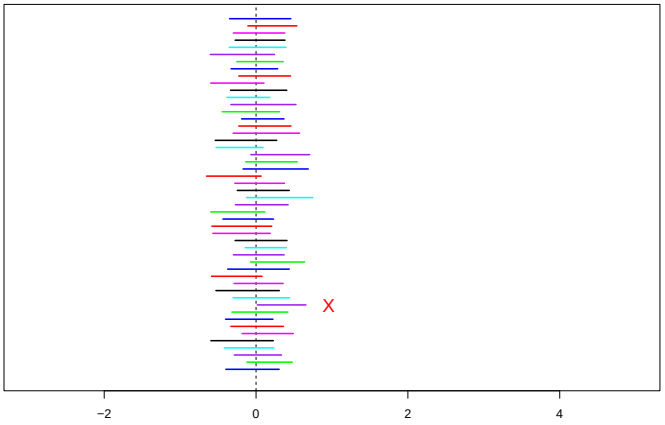


Figure 6.3: Simulated 95% confidence intervals for the population mean. Red "X" marks indicate intervals that do not contain the true mean ($\mu = 0$).

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 μ and **known** standard deviation σ. A level C confidence interval for μ is:

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

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

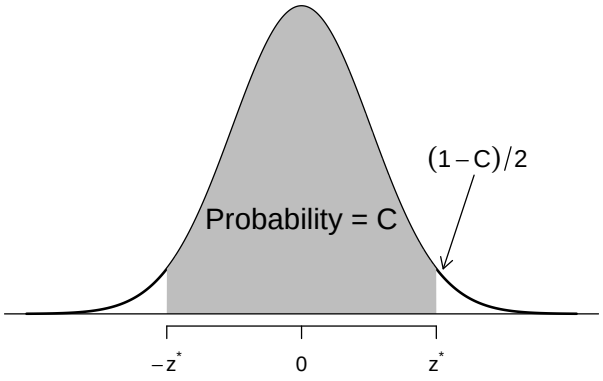


Figure 6.4: The central area under the standard normal curve with confidence level C.

Large Sample CI for μ (Normal data)

When we have a large sample from a Normal distribution, the confidence interval for the population mean μ can be approximated by the formula:

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

This formula is valid under the following assumptions:

- 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 μ

To construct a confidence interval for the population mean, we rely on several important assumptions:

- When population standard deviation σ 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 μ:

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

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

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:

(a) Develop a 90% confidence interval estimate of the population mean.

(b) Develop a 95% confidence interval estimate of the population mean.

(c) Develop a 99% confidence interval estimate of the population mean.

90% CI Calculation

$$\begin{aligned}\bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} &= 119{,}155 \pm 1.645 \cdot \frac{30{,}000}{\sqrt{80}} \\ &= 119{,}155 \pm 5{,}500.73 \\ &= (113{,}654.27,\ 124{,}655.73)\end{aligned}$$

95% CI Calculation

$$\begin{aligned}\bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} &= 119{,}155 \pm 1.96 \cdot \frac{30{,}000}{\sqrt{80}} \\ &= 119{,}155 \pm 6{,}574.04 \\ &= (112{,}580.96,\ 125{,}729.04)\end{aligned}$$

99% CI Calculation

$$\begin{aligned}\bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} &= 119{,}155 \pm 2.576 \cdot \frac{30{,}000}{\sqrt{80}} \\ &= 119{,}155 \pm 8{,}620.04 \\ &= (110{,}534.96,\ 127{,}775.04)\end{aligned}$$

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:

$$\begin{array}{ccccc} 79 & 43 & 58 & 66 & 101 \\ 63 & 79 & 33 & 58 & 71 \\ 60 & 101 & 74 & 55 & 88 \end{array}$$

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 ± 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** μ of a Normal population with **known** standard deviation σ, based on an SRS of size n, is given by

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

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

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

- the confidence level C decreases,
- the population standard deviation σ 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 θ (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 θ. 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 θ, where θ is the **only** unknown quantity.
- Its probability distribution does not depend on the parameter θ.

Chapter 7

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

7.1 CIs for μ

Definition 7.1 (Large-Sample Confidence Interval for μ)**.**

Let μ be the population mean. When the population standard deviation σ is known and the sample size is large, a confidence interval for μ 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 μ)**.**

Let μ be the population mean. When the population standard deviation is unknown and the sample size is small, a confidence interval for μ 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 σ 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 σ more accurately as the

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

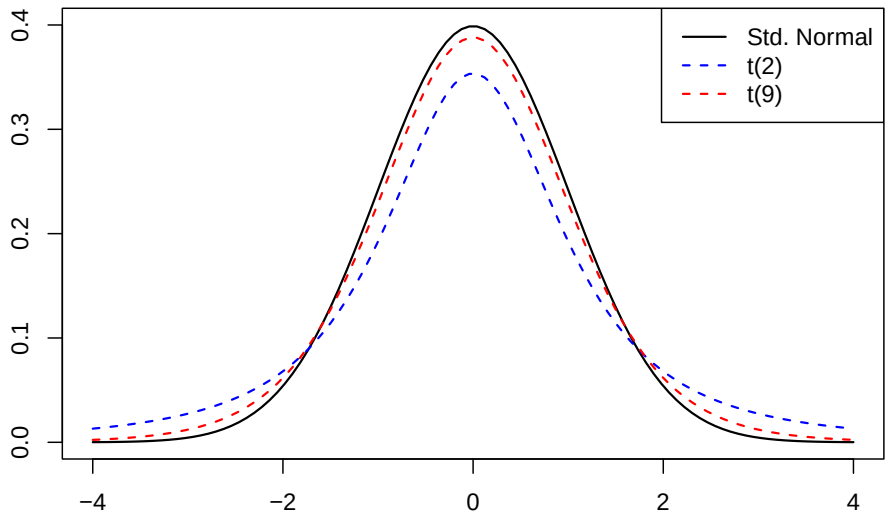


Figure 7.1: Comparison of the standard normal distribution and t-distributions with 2 and 9 degrees of freedom.

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 μ represent the true mean percent of nitrogen in ancient air. We compute a 90% confidence interval for μ 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:

```
# 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: $\bar{x} = \dfrac{177}{10} = 17.7$
- 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 μ, where μ = 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.

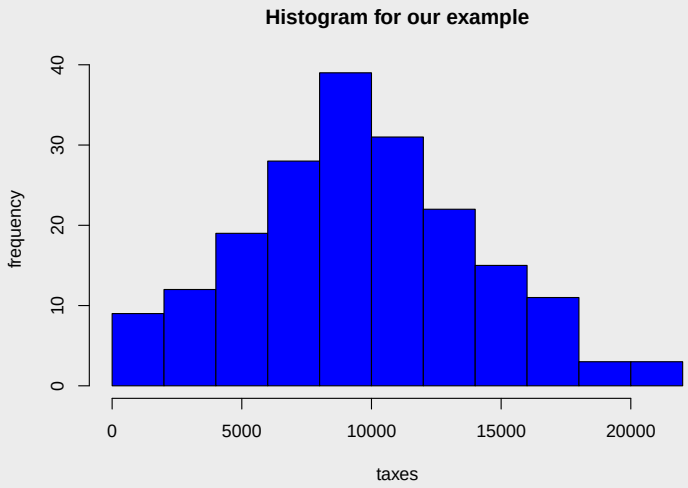
Solution:
We use a one-sample confidence interval for the population mean additional tax collected.
Step 1: Import the data from `taxes.txt`

```
# 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

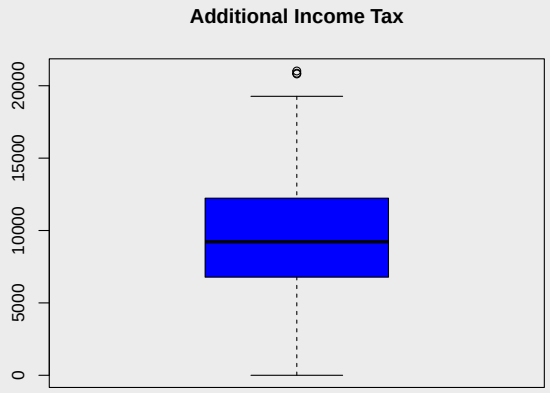
Histogram:

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



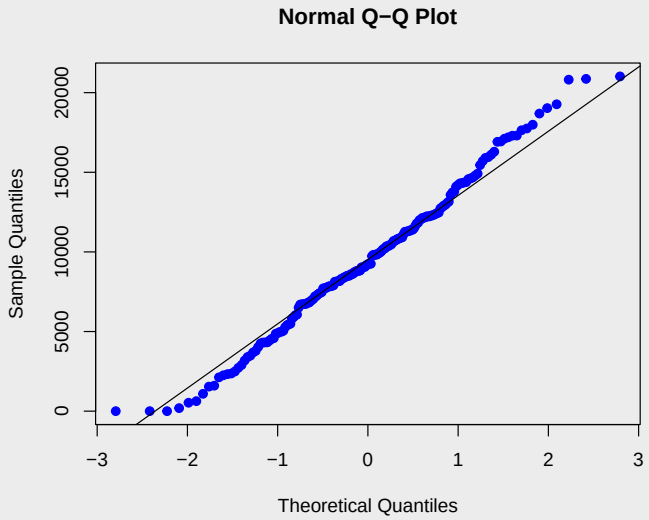
Boxplot:

```
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.

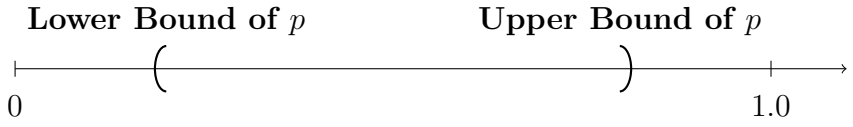
Definition 8.1 (One Sample Confidence Intervals On a Proportion)**.**
We select a random sample of size n from a population with ***unknown*** *proportion p of success. An approximate confidence interval for* $\boldsymbol{p}$ *is:*

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

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 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 summery 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 (σ^2), and s^2 is an unbiased estimator of σ^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 $X \sim Binomial(n, p)$, then the random variable $X \sim N(np, np(1-p))$. From chapter 4, we know that $\hat{p} \sim N(\mu_{\hat{p}} = p, \sigma^2_{\hat{p}} = \frac{p(1-p)}{n})$. 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 9

Sample Size Selection using Confidence Intervals

In this section we will examine techniques to calculate the minimum sample size required to obtain a confidence interval to be within a specified margin of error. Recall the one-sample confidence intervals we have constructed for a population mean μ are

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \qquad \text{(When } \sigma \text{ is known)}$$

$$\bar{x} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} \quad \text{(When } \sigma \text{ is not known)}$$

and the one-sample confidence interval for a proportion is

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

9.0.1 Empirical Rule

For any sample from a population that is close to Normally distributed:

- about 68% of all observations will lie in the interval $\mu \pm \sigma$
- about 95% of all observations will lie in the interval $\mu \pm 2\sigma$
- about 99.7% of all observations will lie in the interval $\mu \pm 3\sigma$

This suggests that for a sample drawn from a population that is approximately to Normally distributed, we can approximate the standard deviation using

$$\hat{\sigma} \approx \frac{\text{Sample Range}}{4}.$$

9.1 Calculating Sample Size for a Confidence Interval on a Mean

We will examine how to calculate the minimum sample size n for a confidence interval on a mean for a margin of error E at confidence level $1-\alpha$.

When σ is Known

For a desired margin of error E and confidence level $1 - \alpha$:

$$E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

We can rearrange E to calculate the sample size using

$$n = \left(\frac{z_{\alpha/2}\, \sigma}{E} \right)^2$$

where n is always *rounded up* to the next integer.

Example 9.1.

A manufacturer of pharmaceutical products analyzes a specimen from each batch of a product to verify the concentration of the active ingredient. The chemical analysis is not perfectly precise. Repeated measurements on the same specimen give slightly different results. Suppose we know that the results of repeated measurements follow a Normal distribution with mean μ equal to the true concentration and standard deviation $\sigma = 0.0068$ grams per liter. (That the mean of the population of all measurements is the true concentration says that the measurements process has no bias. The standard deviation describes the precision of the measurement.) The laboratory analyzes each specimen n times and reports the mean result.

Management asks the laboratory to produce results accurate to within ± 0.005 with 95% confidence. How many measurements must be averaged to comply with this request?

$$n = \left(\frac{z_{0.025}\, \sigma}{E} \right)^2 = \left(\frac{1.96 \times 0.0068}{0.005} \right)^2 \approx 7.1$$

Since the sample size should be a whole number, we round our result up to $n = 8$ measurements.

In Example **??**, we note that 7 measurements will give a slightly larger margin of error than desired, and 8 measurements a slightly smaller margin of error, the lab must take 8 measurements on each specimen to meet management's demand. Always round up to the next higher whole number when finding n.

Example 9.2.

Planning value $\sigma = 22.50$, desired margin $E = 2$.

1. 90% confidence, $z_{0.05} = 1.65$:

$$n = \left(\frac{1.65 \times 22.50}{2} \right)^2 \approx 344.6 \quad \Longrightarrow \quad n = 345.$$

2. 95% confidence, $z_{0.025} = 1.96$:

$$n = \left(\frac{1.96 \times 22.50}{2}\right)^2 \approx 486.2 \quad \Longrightarrow \quad n = 487.$$

3. 99% confidence, $z_{0.005} = 2.58$:

$$n = \left(\frac{2.58 \times 22.50}{2}\right)^2 \approx 842.5 \quad \Longrightarrow \quad n = 843.$$

9.2 Calculating Sample Size for a Confidence Interval on a Proportion

We will examine how to calculate the minimum sample size n for a confidence interval on a proportion for a margin of error E at confidence level $1 - \alpha$. For sample of size n with unknown population proportion p:

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

where the margin of error is

$$E = z^* \sqrt{\frac{p^*(1-p^*)}{n}}$$

we can rearrange E to calculate the sample size using

$$n = \left(\frac{z^*}{E}\right)^2 p^*(1-p^*),$$

where p^* can be a *planning value*, which is value obtained from prior information such as a pilot study. If we do not have any prior information on p^*, we can use $p^* = 0.5$ which is the conservative value that maximizes the product $p^*(1-p^*)$.

Example 9.3.

Aisha Shariff and Yvette Ye are the candidates for mayor in a large city. You are planning a sample survey to determine what percent of the voters plan to vote for Shariff. This is a population proportion p. You will contact an SRS of registered voters in the city. You want to estimate p with 95% confidence and a margin of error no greater than 3%, or 0.03. How large a sample do you need?

For a 95% CI on p: $z_{0.025} = 1.96$. Margin of error = 0.03. Since no information on a good estimate of p, use $p^* = 0.5$.

$$1.96\sqrt{\frac{(0.5)(1-0.5)}{n}} \leq 0.03 \quad \Longrightarrow \quad n = \left(\frac{1.96}{0.03}\right)^2 (0.5)(0.5) \approx 1067.1.$$

Round up:

$$n = 1068.$$

Example 9.4.

The percentage of people not covered by health care insurance in 2007 in the USA was 15.6%. A congressional committee has been charged with conducting a sample survey to obtain more current information.

1. What sample size would you recommend if the committee's goal is to estimate the current proportion of individuals without health care insurance with a margin of error of 0.03? Use a 95% confidence level.
2. Repeat part (a) using a 99% confidence level.

$$\text{(a) } n = \left(\frac{z^*}{E}\right)^2 p^*(1-p^*), \quad z^* = 1.96,\ E = 0.03,\ p^* = 0.156.$$

$$n = \left(\frac{1.96}{0.03}\right)^2 (0.156)(1-0.156) \approx 563.$$

$$\text{(b) } n = \left(\frac{2.58}{0.03}\right)^2 (0.156)(1-0.156) \approx 974.$$

Example 9.5.

A consumer advocacy group would like to find the proportion of consumers who bought the newest generation of iPhone and were happy with their purchase. How large a sample should they take to estimate p with 2% margin of error and 90% confidence?

Parameters:

$$\begin{aligned} E &= 0.02, \quad \text{(margin of error)} \\ \text{Confidence level} = 0.90 &\implies \alpha = 0.10,\ \alpha/2 = 0.05, \\ z^* &= z_{0.05} = 1.645, \\ p^* &= 0.5, \quad p^*(1-p^*) = 0.5 \times 0.5 = 0.25. \end{aligned}$$

Sample-Size Formula:

$$n \ = \ \left(\frac{z^*}{E}\right)^2 \ p^*(1-p^*).$$

Substituting $z^* = 1.645$, $E = 0.02$, and $p^*(1 - p^*) = 0.25$:

$$n = \left(\frac{1.645}{0.02}\right)^2 \times 0.25 = \big(82.25\big)^2 \times 0.25 = 6{,}764.0625 \times 0.25$$

$$= 1{,}691.015625.$$

Since n must be a whole number and we always round up to ensure the margin of error is at most 2%, we take

$$n = 1{,}692.$$

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 σ_1^2, which is population variance of population group 1; n_1 is the sample size chosen form population group 1. Similarly for σ_2^2, which is population variance of population group 2; n_2 is the sample size chosen form population group 2.

Additionally, case 1 is a bit unrealistic with other cases because the population variance (σ^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:

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

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:

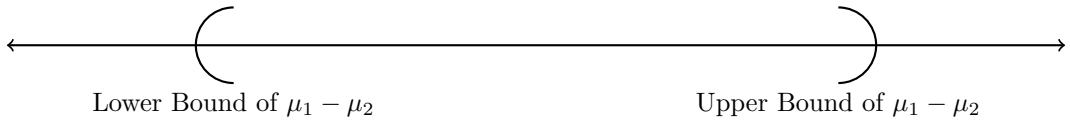


Figure 10.1: Visualization of two-sample confidence interval (Case 1, 2, 3)

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. $\mu_1 < \mu_2$**:**

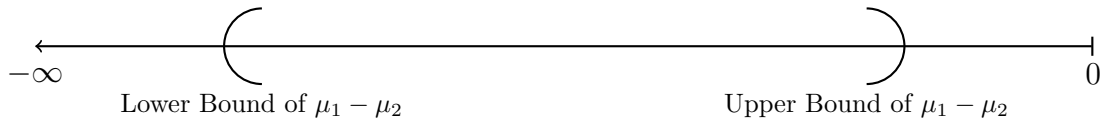


Figure 10.2: Visualization of the case when $\mu_1 < \mu_2$

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. $\mu_1 > \mu_2$**:**

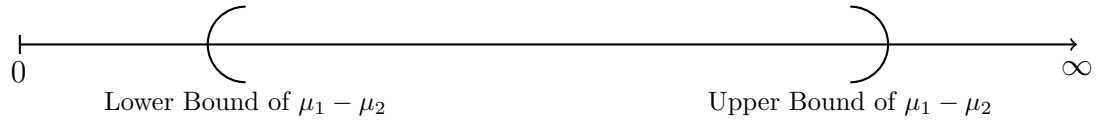


Figure 10.3: Visualization of the case when $\mu_1 > \mu_2$

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. $\mu_1 = \mu_2$**:**

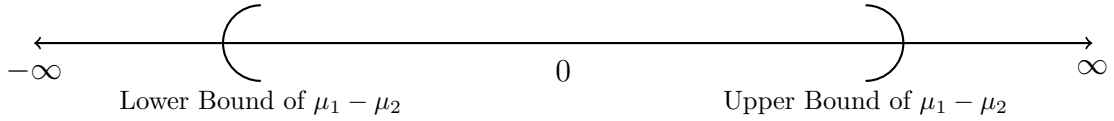


Figure 10.4: Visualization of the case when $\mu_1 = \mu_2$

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 p hone 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 on100-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}	$x_{d2} = x_{32} - x_{31}$
.....			
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 (μ_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. $\bar{M}_1 < \bar{M}_2$**:**

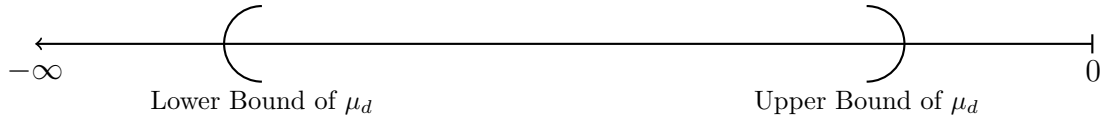


Figure 10.6: Visualization of the case when $\bar{M}_1 < \bar{M}_2$

From the number line above, we know that the result of μ_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. $\bar{M}_1 > \bar{M}_2$**:**

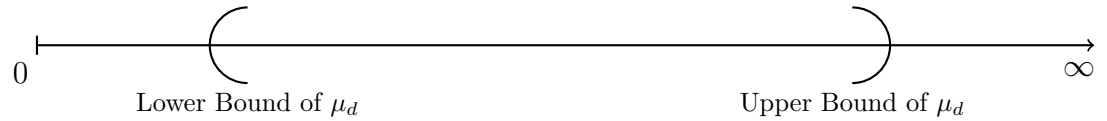


Figure 10.7: Visualization of the case when $\bar{M}_1 > \bar{M}_2$

From the number line above, we know that the result of μ_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. $\bar{M}_1 = \bar{M}_2$**:**

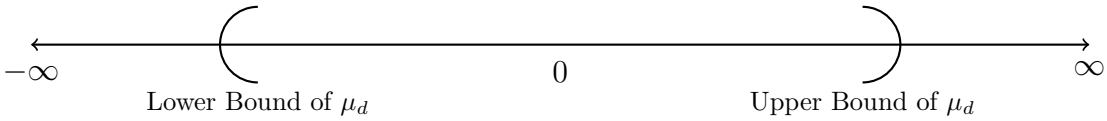


Figure 10.8: Visualization of the case when $\bar{M}_1 = \bar{M}_2$

While, 0 is included in the range of μ_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:

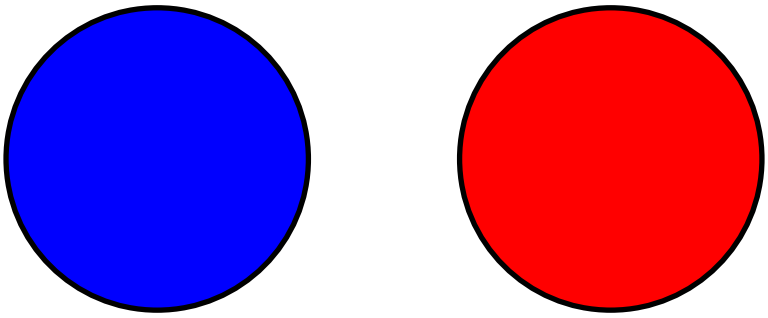


Figure 10.9: Visualization of two population of all students from UTSG (left) and UTM (right)

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. $p_1 > p_2$**:**

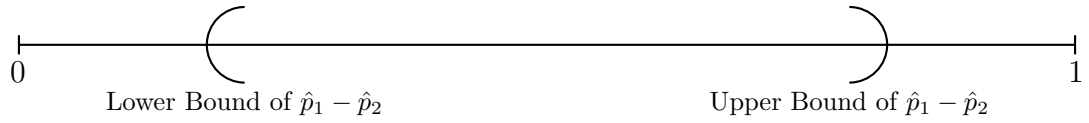


Figure 10.11: Visualization of the case when $p_1 > p_2$

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

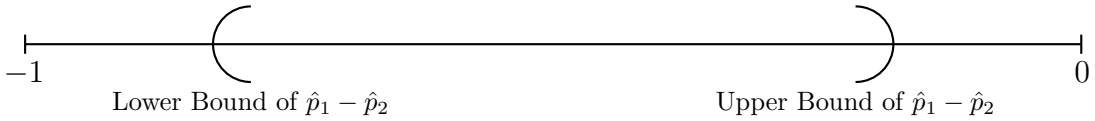


Figure 10.12: Visualization of the case when $p_1 < p_2$

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

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

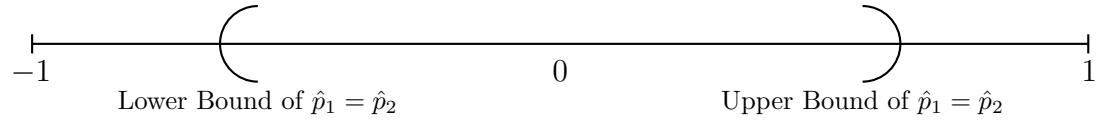


Figure 10.13: Visualization of the case when $\hat{p}_1 = \hat{p}_2$

- 1. Randomization Condition: The data in each group should be drawn independently and at random from a population or generated b y 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 σ_1^2 and σ_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 $(\sigma_1^2)/(\sigma_2^2)$.

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:

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

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

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

Chapter 11

Introduction to Hypothesis Testing

11.1 Test of Hypothesis for One Mean

Definition 11.1 (Hypothesis Tests)**.** *An inferential procedure to determine whether there is sufficient evidence to suggest a condition for a population parameter using statistics from a sample.*

Attach a probability to the conclusion of a hypothesis test.

Steps

1. Decide on a level of significance (α)
2. State the null hypothesis (H_0) and the alternative hypothesis (H_a) (H_1)
3. Calculate the appropriate test statistic.
4. Use the test statistic and a reference distribution to calculate a p-value. (Also refer back to H_a)
5. Compare p-value to α to make a conclusion.

Note:
The definition of a p-value can be confusing. We will define it later.

Step 1: Decide on a Level of Significance (α)

-Threshold for decision making.
-Depends on tolerance for consequences of errors, sample size, nature of the study, and variability.
-Common values: 0.10, 0.01, 0.05 (very common default)

Step 2: State the Null Hypothesis (H_0) and the Alternative Hypothesis (H_a)

Θ**:** parameter of interest

Θ_0**:** numerical value of the parameter of interest hypothesized under the null hypothesis.

$$\begin{array}{llll} H_0 : \Theta = \Theta_0 & (\Theta \leq \Theta_0) & H_a : \Theta > \Theta_0 & \text{one-sided (one-tailed)} \\ H_0 : \Theta = \Theta_0 & (\Theta \geq \Theta_0) & H_a : \Theta < \Theta_0 & \text{one-sided (one-tailed)} \\ H_0 : \Theta = \Theta_0 & & H_a : \Theta \neq \Theta_0 & \text{two-sided (two-tailed)} \end{array}$$

Null (H_0): Represents the current belief (status quo) or the safe belief.
Alternative (H_a): Represents the research hypothesis (or what you are asked to test)

Step 3: Calculate an appropriate test statistic

Depends on the hypothesis test conducted and the information available.

Definition 11.2 (Test Statistic Skeleton)**.**

$$\textit{test statistic} = \frac{\textit{(a statistic)} - \textit{(hypothesized value of parameters under } H_0\textit{)}}{\textit{standard error of statistic}}$$

The test statistic follows a reference distribution (Z, t, F, χ^2).

Step 4: Calculate the p-value

Use the test statistic, reference distribution, and refer back to H_a.

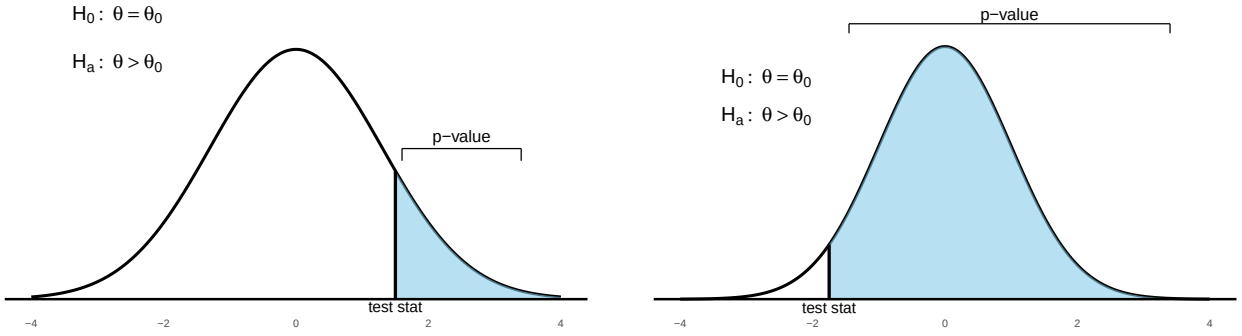


Figure 11.1: Right-tailed test: p-value is the area to the right of the test statistic.

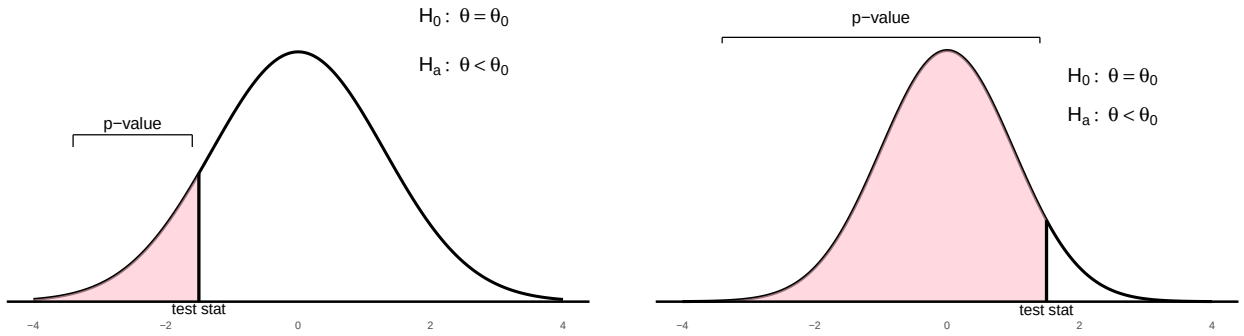


Figure 11.2: Left-tailed test: p-value is the area to the left of the test statistic.

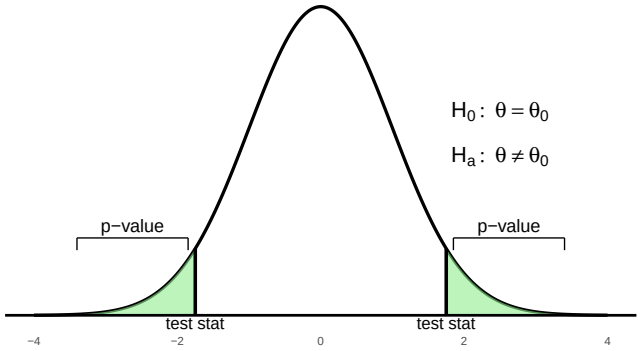


Figure 11.3: Two-tailed test: p-value is the total area in both tails beyond ± test statistic.

Step 5: Compare *p*-value to level of significance α and make a conclusion

- If p-value $< \alpha$:
  Sufficient evidence against H_0. The hypothesis test rejects H_0 in favor of H_a.
- If p-value $> \alpha$:
  Insufficient evidence against H_0. Do not reject H_0 (fail to reject H_0).

Note: It is not good practice to give conclusions in the context of stating we *accept* H_0 or *accept* H_a.

Example 11.1.
[Sweetening Colas] Diet colas use artificial sweeteners to avoid sugar. These sweeteners gradually lose their sweetness over time. Manufacturers therefore test new colas for loss of sweetness before marketing them. Trained tasters sip the cola along with drinks of standard sweetness and score the cola on a "sweetness score" of 1 to 10. The cola is then stored for a month at high temperature to imitate the effect of four months' storage at room temperature. Each taster scores the cola again after storage. This is a matched pairs experiment. Our data are the differences (score before storage minus score after storage) in the tasters' scores. The bigger these differences, the bigger the loss of sweetness.

Suppose we know that for any cola, the sweetness loss scores vary from taster to taster according to a Normal distribution with standard deviation $\sigma = 1$. The mean μ for all tasters measures loss of sweetness.
The following are the sweetness losses for a new cola as measured by 10 trained tasters:

2.0, 0.4, 0.7, 2.0, -0.4, 2.2, -1.3, 1.2, 1.1, 2.3

Are these data good evidence that the cola lost sweetness in storage?
Solution

μ = mean sweetness loss for the population of **all** tasters.
Step 1: State hypotheses.

$$H_0 : \mu = 0$$
$$H_a : \mu > 0$$

Step 2: Test statistic: $z_\star = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \dfrac{1.02 - 0}{1/\sqrt{10}} = 3.23$
Step 3: P-value. $P(Z > z_\star) = P(Z > 3.23) = 0.0006$
Step 4: Conclusion. We would rarely observe a mean as large as 1.02 if H_0 were true. The small p-value provides strong evidence against H_0, supporting $H_a : \mu > 0$. That is, the mean sweetness loss is likely positive.

R code (Simulation)

```
# n = sample size;
n<-10;
mu.zero<-0;
sigma<-1;
sigma.xbar<-sigma/sqrt(n);

# x bar = sample mean with 10 obs;
x.bar<-rnorm(1,mean=mu.zero,sd=sigma.xbar);
x.bar;

## [1] 0.3265859

# z.star = test statistic;
z.star<-(x.bar-mu.zero)/sigma.xbar;
z.star;

## [1] 1.032755
```

R code (10,000 Simulations)

```
n <- 10;
mu.zero <- 0;
sigma <- 1;
sigma.xbar <- sigma / sqrt(n);
# x bar = sample mean with 10 obs;
# m = number of simulations;
m <- 10000;
x.bar <- rnorm(m, mean = mu.zero, sd = sigma.xbar);

# z.star = test statistic;
z.star <- (x.bar - mu.zero) / sigma.xbar;
hist(z.star, xlab = "differences", col = "blue");
```

Histogram from simulation (see Section 7 for R code format)

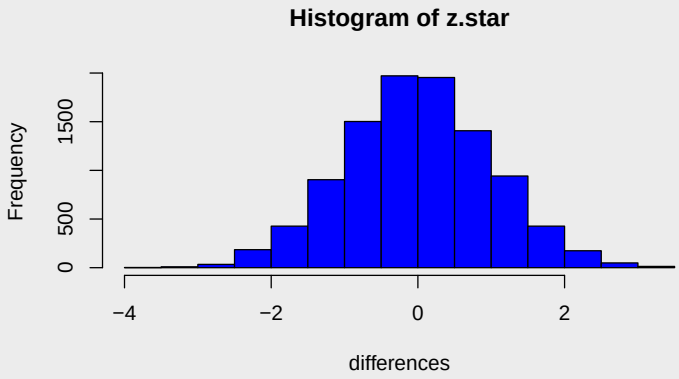


Figure 11.4: Histogram of $z_\star$ values from 10,000 simulations under H_0.

R code (Empirical p-value)

```
## P-value

p_value <- length(z.star[z.star > 3.23]) / m;

p_value

## [1] 8e-04
```

One-Sample Hypothesis Test for Population Mean (μ) (known σ)

When σ is known:

- $H_0: \mu = \mu_0$ (or $\mu \leq \mu_0$) $\quad H_a: \mu > \mu_0$
- $H_0: \mu = \mu_0$ (or $\mu \geq \mu_0$) $\quad H_a: \mu < \mu_0$
- $H_0: \mu = \mu_0 \quad H_a: \mu \neq \mu_0$

Test statistic: $z^* = \dfrac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$

Reference distribution: Standard normal (Z)

Example 11.2.

[UTM Deer] Deer are a common sight on the UTM campus. Suppose an ecologist is interested in the average mass of adult white-tailed does (female deer) around the Mississauga campus to determine whether they are healthy for the upcoming winter. The ecologist captures a sample of 36 adult females around the UTM and measures the average mass of this sample to be 42.53 kg.

From previous studies conducted in the area, the average mass of healthy does was reported to be 45 kg. Conduct a hypothesis test at the 5% significance level to determine whether the mass of does around UTM has decreased. Assume the standard deviation is known to be 5.25 kg.

1. Level of significance. $\alpha = 0.05$

2. State the null and alternative hypotheses.

$$H_0 : \mu = 45 \qquad H_a : \mu < 45$$

3. Calculate appropriate test statistic.

Given:

$$n = 36, \quad \bar{x} = 42.53, \quad \sigma = 5.25$$

Since σ is known, the test statistic is:

$$z^* = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{42.53 - 45}{5.25/\sqrt{36}} = -2.82$$

Reference distribution: standard normal

4. Calculate p-value

$$\text{p-value} = P(Z < -2.82) \approx 0.0024$$

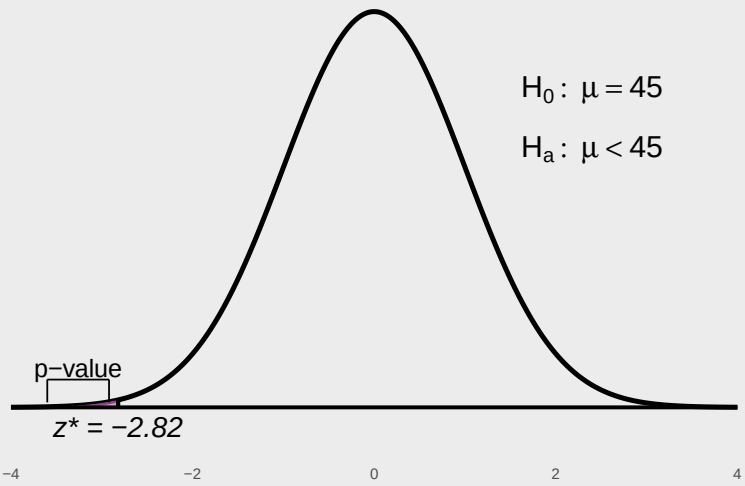


Figure 11.5: Left-tailed p-value for the test statistic $z^* = -2.82$

5. Compare p-value with level of significance α and make a conclusion:

$$0.0024 < 0.05 \Rightarrow \text{p-value} < \alpha$$

There is sufficient evidence at the 5% level of significance to reject the null that does this winter weigh the same as in the past and to conclude the alternative that does this winter weigh less than 45 kg.

R code:

```
# Find test stat
z_test_stat = (42.53 - 45) / (5.25 / sqrt(36))
z_test_stat
[1] -2.822857

# Find the p-value
# Since the alternative is Ha : mu < 45
p-value = pnorm(z_test_stat)
[1] 0.00237989
```

Note: The `pnorm()` function in R, by default, returns the cumulative probability (area) to the left of the given value.

R code: Using BSDA package

```
# Using the BSDA library. install BSDA if it is not already installed.
# install.packages("BSDA")
> library(BSDA)
> # Conduct the z-test with the zsum.test function
> zsum.test(mean.x = 42.53, sigma.x = 5.24, n.x = 36, mu = 45, alternative = "less")

        One-sample z-Test

data:  Summarized x
z = -2.8282, p-value = 0.00234
alternative hypothesis: true mean is less than 45
95 percent confidence interval:
 NA 43.96651
sample estimates:
mean of x
    42.53
```

Interpretation:
There is sufficient evidence at the 5% level of significance to reject the null hypothesis. We conclude that the average mass of does this winter is significantly less than 45 kg.

Example 11.3.

[Executives' Blood Pressures]
The National Center for Health Statistics reports that the systolic blood pressure for males 35 to 44 years of age has mean 128 and standard deviation 15.
The medical director of a large company looks at the medical records of 72 executives in this age group and finds that the mean systolic blood pressure in this sample is $\bar{x} = 126.07$. Is this evidence that the company's executives have a different mean blood pressure from the general population?
Suppose we know that executives' blood pressures follow a Normal distribution with standard deviation $\sigma = 15$.
Solution: Let μ be the mean systolic blood pressure of the executive population.

1. **State hypotheses:**

$$H_0 : \mu = 128$$
$$H_a : \mu \neq 128$$

2. **Test statistic:**

$$z_* = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{126.07 - 128}{15/\sqrt{72}} = -1.09$$

3. **P-value:**

$$2P(Z > |z_*|) = 2P(Z > 1.09) = 2(1 - 0.8621) = 0.2758$$

4. **Conclusion:**
   More than 27% of the time, a simple random sample of size 72 from the general male population would have a mean blood pressure at least as far from 128 as that of the executive sample. The observed $\bar{x} = 126.07$ is therefore not good evidence that executives differ from other men.

Tests for a Population Mean

There are four steps in carrying out a significance test:

1. State the hypotheses.
2. Calculate the test statistic.
3. Find the P-value.
4. State your conclusion in the context of your specific setting.

Once you have stated your hypotheses and identified the proper test, you or your computer can do Steps 2 and 3 by following a recipe.

Z Test for a Population Mean (μ)

Here is the recipe for the test we have used in our examples.
Draw a simple random sample of size n from a Normal population that has unknown mean μ and known standard deviation σ. To test the null hypothesis that μ has a specified value, $H_0 : \mu = \mu_0$, calculate the **one-sample z statistic**:

$$z_* = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

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

- $H_a : \mu > \mu_0$ is $P(Z > z_*)$
- $H_a : \mu < \mu_0$ is $P(Z < z_*)$
- $H_a : \mu \neq \mu_0$ is $2P(Z > |z_*|)$

Example 11.4.

Consider the following hypothesis test:

$$H_0 : \mu = 20$$

$$H_a : \mu < 20$$

A sample of 50 provided a sample mean of 19.4. The population standard deviation is 2.

(a) Compute the value of the test statistic.

(b) What is the p-value?

(c) Using $\alpha = 0.05$, what is your conclusion?

Solution:

(a) **Test statistic:**

$$z_* = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{19.4 - 20}{2/\sqrt{50}} = -2.1213$$

(b) **P-value:**

$$P(Z < z_*) = P(Z < -2.1213) = 0.0169$$

(c) **Conclusion:**
Since the P-value $= 0.0169 < \alpha = 0.05$, we reject $H_0 : \mu = 20$. We conclude that $\mu < 20$.

Example 11.5.

Consider the following hypothesis test:

$$H_0 : \mu = 25$$
$$H_a : \mu > 25$$

A sample of 40 provided a sample mean of 26.4. The population standard deviation is 6.

(a) Compute the value of the test statistic.

(b) What is the p-value?

(c) Using $\alpha = 0.01$, what is your conclusion?

Solution:

(a) **Test statistic:**

$$z_* = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{26.4 - 25}{6/\sqrt{40}} = 1.4757$$

(b) **P-value:**

$$P(Z > z_*) = P(Z > 1.4757) = 0.0700$$

(c) **Conclusion:**
Since P-value $= 0.0700 > \alpha = 0.01$, we **cannot reject** $H_0 : \mu = 25$.
We conclude that we don't have enough evidence to claim that $\mu > 25$.

Example 11.6.

Consider the following hypothesis test:

$$H_0 : \mu = 15$$
$$H_a : \mu \neq 15$$

A sample of 50 provided a sample mean of 14.15. The population standard deviation is 3.

(a) Compute the value of the test statistic.

(b) What is the p-value?

(c) Using $\alpha = 0.05$, what is your conclusion?

Solution:

(a) **Test statistic:**

$$z_* = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{14.15 - 15}{3/\sqrt{50}} = -2.0034$$

(b) **P-value:**

$$2P(Z > |z_*|) = 2P(Z > |-2.0034|) = 2P(Z > 2.0034) = 0.0451$$

(c) **Conclusion:**
Since P-value $= 0.0451 < \alpha = 0.05$, we reject $H_0 : \mu = 15$.
We conclude that $\mu \neq 15$.

Confidence Interval Interpretation:
The 95% confidence interval for μ is:

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

$$14.15 \pm 1.96 \left(\frac{3}{\sqrt{50}} \right) = (13.3184,\ 14.9815)$$

Since the hypothesized value $\mu_0 = 15$ falls **outside** this interval, we again reject $H_0 : \mu = 15$.

Note 11.1.
A level α two-sided significance test rejects a hypothesis $H_0 : \mu = \mu_0$ exactly when the value μ_0 falls outside a level $1 - \alpha$ confidence interval for μ.

One-Sample Hypothesis Test for Population Mean (μ) (unknown σ)

When σ is NOT known:

- $H_0 : \mu = \mu_0$ (or $\mu \geq \mu_0$) $\quad H_a : \mu < \mu_0$
- $H_0 : \mu = \mu_0$ (or $\mu \leq \mu_0$) $\quad H_a : \mu > \mu_0$
- $H_0 : \mu = \mu_0 \quad H_a : \mu \neq \mu_0$

Test statistic: $t^* = \dfrac{\bar{X} - \mu_0}{s/\sqrt{n}}$

Reference distribution: t distribution with $n - 1$ degrees of freedom

Example 11.7.
[Hypothesis Test: Laughter and Heart Rate, Unknown σ]
Researchers studied the physiological effects of laughter. They measured heart rates (in beats per minute) of $n = 25$ subjects (ages 18–34) while they laughed. They obtained:

$$\bar{x} = 73.5, \quad s = 6, \quad \alpha = 0.05$$

It is well known that the resting heart rate is 71 bpm. Is there evidence that the mean heart rate during laughter exceeds 71 bpm?

Step 1: State the hypotheses.

$$H_0 : \mu = 71$$
$$H_a : \mu > 71$$

Step 2: Check assumptions.

- The sample is an independent random sample of individuals aged 18–34.
- The population of heart rates during laughter is normally distributed.

Step 3: Compute the test statistic.
Since σ is unknown, we use the t statistic:

$$t^* = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{73.5 - 71}{6/\sqrt{25}} = 2.083$$

Reference distribution: t distribution with $n - 1 = 25 - 1 = 24$ degrees of freedom.

Step 4: Determine the p-value.
Using the t distribution with 24 df:

$$0.01 < \text{p-value} < 0.025$$

Step 5: Make a conclusion.
Since p-value $< \alpha = 0.05$, we reject H_0.
There is sufficient evidence at the 5% level of significance to reject the null that the mean is 71 bpm in favor of the alternative that the mean is greater than 71 bpm for people who are laughing.

Example 11.8.

A researcher is asked to test the hypothesis that the average price of a 2-star (CAA rating) motel room has decreased since last year. Last year, a study showed that the prices were Normally distributed with a mean of $89.50.
A random sample of twelve 2-star motels produced the following room prices:

$85.00, 92.50, 87.50, 89.90, 90.00, 82.50, 87.50, 90.00, 85.00, 89.00, 91.50, 87.50

At the 5% level of significance, can we conclude that the mean price has decreased?
Solution:
Let μ be the true average price of a 2-star motel room.

1. **State hypotheses.**
$$H_0 : \mu = 89.5$$
$$H_a : \mu < 89.5$$

2. **Compute test statistic.**
$$t^* = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{88.1583 - 89.5}{2.9203/\sqrt{12}} = -1.5915$$

3. **Find the P-value.**
With $df = 11$, the P-value (from the t-distribution table) is between 0.05 and 0.10.

4. **Conclusion.**
Since P-value > 0.05, we **fail to reject** H_0.
There is not sufficient evidence to conclude that the average price of 2-star motels has decreased this year.

R code (One Sample t-test)

```
# Step 1. Entering data;
prices=c(85.00, 92.50, 87.50, 89.90, 90.00, 82.50,
         87.50, 90.00, 85.00, 89.00, 91.50, 87.50);

# Step 2. Hypothesis test;
t.test(prices, alternative="less", mu=89.5);
```

R output

```
##
##  One Sample t-test
##
## data:  prices
## t = -1.5915, df = 11, p-value = 0.0699
## alternative hypothesis: true mean is less than 89.5
## 95 percent confidence interval:
##  -Inf 89.67229
## sample estimates:
## mean of x
##  88.15833
```

The one-sample t test

Draw an SRS of size n from a large population having unknown mean μ.
To *test the hypothesis* $H_0 : \mu = \mu_0$, compute the *one-sample t statistic*

$$t^* = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

In terms of a variable T having the t_{n-1} distribution, the P-value for a test of H_0 against

$$\begin{aligned} H_a : \mu > \mu_0 \quad &\text{is} \quad P(T \geq t^*) \\ H_a : \mu < \mu_0 \quad &\text{is} \quad P(T \leq t^*) \\ H_a : \mu \neq \mu_0 \quad &\text{is} \quad 2P(T \geq |t^*|) \end{aligned}$$

These P-values are exact if the population distribution is Normal and are approximately correct for large n in other cases.

Example 11.9.

We are conducting a two-sided one-sample t-test for the hypotheses:

$$\begin{aligned} H_0 &: \mu = 64 \\ H_a &: \mu \neq 64 \end{aligned}$$

based on a sample of $n = 15$ observations, with test statistic $t^* = 2.12$.
a) Degrees of freedom:

$$df = n - 1 = 15 - 1 = 14$$

b) Critical values and P-value bounds:
From the t-distribution table for $df = 14$:

- $t = 1.761$ corresponds to a two-tailed probability of 0.10
- $t = 2.145$ corresponds to a two-tailed probability of 0.05

Since $t^* = 2.12$ falls between these values, the two-sided P-value satisfies:

$$0.05 < \text{P-value} < 0.10$$

c) Significance:

- At the 10% level: **Yes**, since P-value < 0.10
- At the 5% level: **No**, since P-value > 0.05

d) Exact two-sided P-value using R:

```
# Compute exact two-sided P-value for t* = 2.12 with df = 14
2 * (1 - pt(2.12, df = 14))

## [1] 0.05235683
```

Thus, the exact two-sided P-value is approximately **0.0524**, confirming the bracketing result.

11.2 Test of Hypothesis for One Proportion

One-Sample Hypothesis Test for Population Proportion (p)

When sample size is large enough (np, n(1-p) $\geq$ 10):

- $H_0: p = p_0$ (or $p \geq p_0$) $H_a: p < p_0$
- $H_0: p = p_0$ (or $p \leq p_0$) $H_a: p > p_0$
- $H_0: p = p_0$ $H_a: p \neq p_0$

Test statistic: $z^* = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$

Reference distribution: Standard normal ($\mathcal{Z}$)

Example 11.10.
[100-Cup Challenge] A YouTuber goes to her nearest Tim Hortons and buys 100 empty cups.

After rolling up the rims, she ends up with 12 winning cups out of the 100 she bought, all of them were food prizes.
If the probability of winning a food prize is supposed to be $\frac{1}{6}$, does she have evidence to claim that the probability of winning a food prize is less than $\frac{1}{6}$?

Definition 11.3.

*The **sample space** S of a random phenomenon is the set of all possible outcomes.*
*An **event** is an outcome or a set of outcomes of a random phenomenon. That is, an event is a subset of the sample space.*
*A **probability model** is a mathematical description of a random phenomenon consisting of two parts: a sample space S and a way of assigning probabilities to events.*

Rolling a fair die (random phenomenon). There are 6 possible outcomes when we roll a die. The sample space for rolling a die and counting the pips is

$$S = \{1, 2, 3, 4, 5, 6\}$$

"Roll a 6" is an event that contains one of these 6 outcomes.

Definition 11.4.

*A random variable X has a **discrete uniform distribution** if each of the n values in its range, say, $x_1, x_2, \ldots, x_n$, has equal probability. Then,*

$$f(x_i) = \frac{1}{n}$$

R code:

```
# Define a die with values 1 through 6
die <- c(1, 2, 3, 4, 5, 6)

# Roll the die once
sample(die, 1, replace = TRUE)
## [1] 2

# Roll the die six times
sample(die, 6, replace = TRUE)
## [1] 1 3 2 6 3 1
```

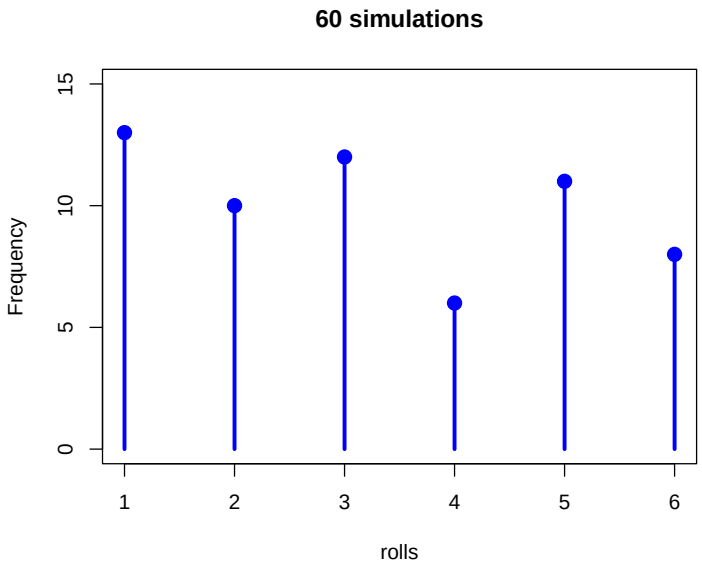


Figure 11.6: Plot of frequencies from 60 simulations of a fair six-sided die.

Definition 11.5.

*A **random variable** is a variable whose value is a numerical outcome of a random phenomenon.*
*The **probability distribution** of a random variable X tells us what values X can take and how to assign probabilities to those values.*

Note 11.2.

The Binomial setting

- *There are a fixed number n of observations.*
- *The n observations are all **independent**. That is, knowing the result of one observation tells you nothing about the other observations.*
- *Each observation falls into one of just two categories, which for convenience we call "success" and "failure".*
- *The probability of a success, call it p, is the same for each observation.*

A random variable Y is said to have a **binomial distribution** based on n trials with success probability p if and only if

$$p(y) = \frac{n!}{y!(n-y)!}\, p^y (1-p)^{n-y}, \quad y = 0, 1, 2, \ldots, n \quad \text{and} \quad 0 \le p \le 1.$$

Example 11.11.
Think of rolling a die n times as an example of the binomial setting. Each roll gives either a six or a number different from six. Knowing the outcome of one roll doesn't tell us anything about other rolls, so the n rolls are independent.
If we call six a success, then p is the probability of a six and remains the same as long as we roll the same die. The number of sixes we count is a random variable X. The distribution of X is called a **binomial distribution**.

R code (Binomial Simulations and PMF)

```
## Simulation: Binomial with n = 10 and p = 1/6.
rbinom(1, size = 10, prob = 1/6);
## [1] 3

rbinom(1, size = 10, prob = 1/6);
## [1] 1

rbinom(1, size = 10, prob = 1/6);
## [1] 0

## Pmf: 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);
```

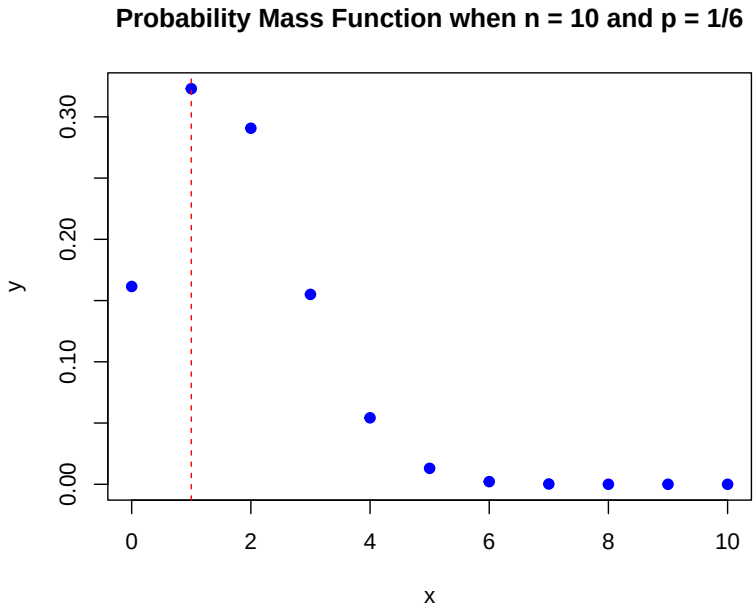


Figure 11.7: PMF when $n = 10$ and $p = 1/6$

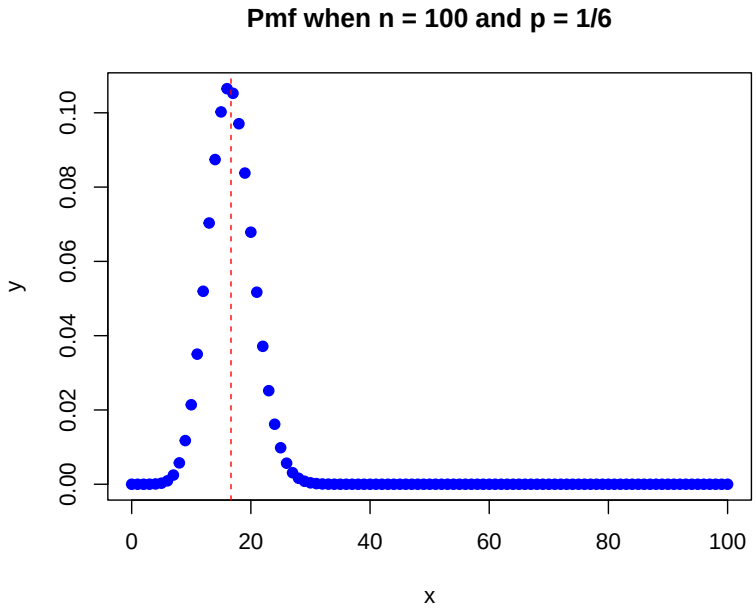


Figure 11.8: PMF when $n = 100$ and $p = 1/6$

R code (PMF values for selected x values)

```
dbinom(c(15, 16, 17, 18), size = 100, prob = 1/6);
## [1] 0.10023663 0.10650142 0.10524847 0.09706247
```

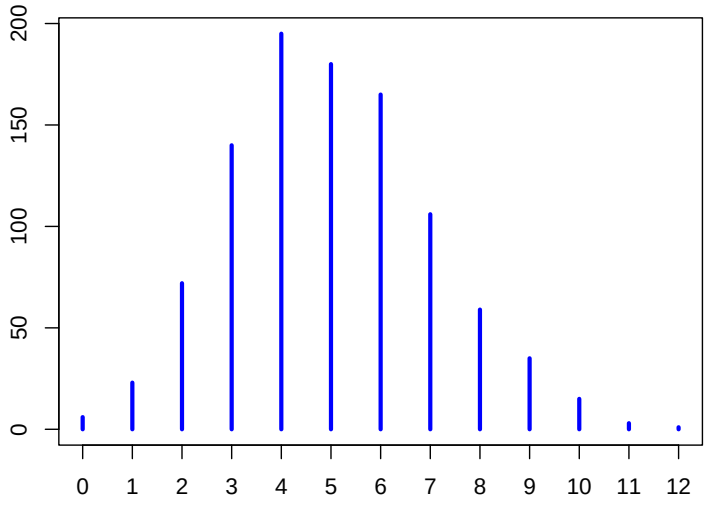


Figure 11.9: Simulation: 2000 YouTuber, $n = 100$, and $p = 1/6$

R code (A few values from our simulation)

```
## vec.prop
##  6  7  8  9 10 11 12
##  7  3  8 24 46 72 106
## [1] 266
## [1] 0.133
```

It turns out that our P-value for this simulation is:
0.133

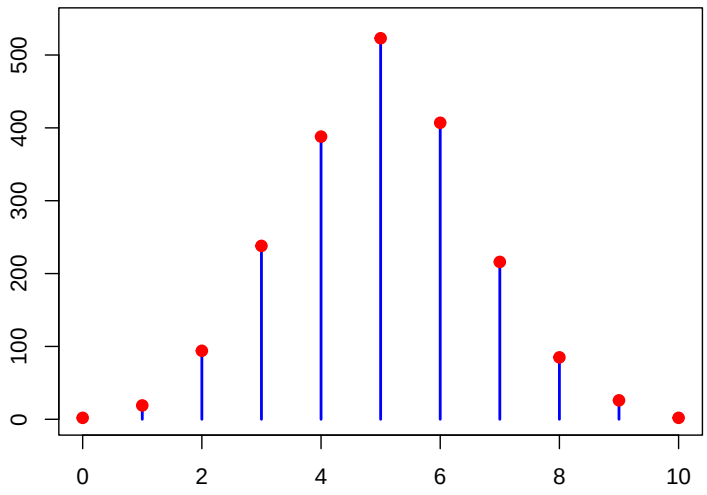


Figure 11.10: Simulation vs Theoretical pmf

Sampling Distribution of a Sample Proportion

Draw an 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

$$\sqrt{\frac{p(1-p)}{n}}.$$

- As the sample size increases, the sampling distribution of $\hat{p}$ becomes **approximately Normal**. That is, for large n, $\hat{p}$ has approximately the

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

distribution.

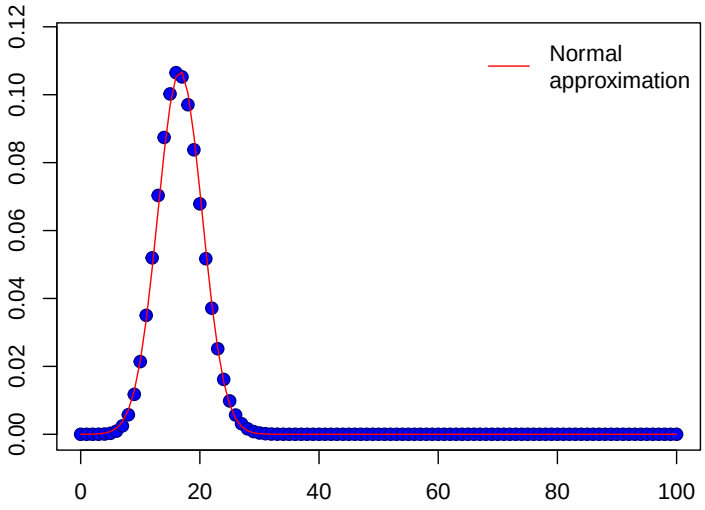


Figure 11.11: Binomial with Normal Approximation

Hypotheses Tests for a Proportion

To *test the hypothesis* $H_0 : p = p_0$, compute the z_* statistic:

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

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

$$\begin{aligned} H_a : p > p_0 \quad &\text{is} \quad P(Z > z_*) \\ H_a : p < p_0 \quad &\text{is} \quad P(Z < z_*) \\ H_a : p \neq p_0 \quad &\text{is} \quad 2P(Z > |z_*|) \end{aligned}$$

Introduction to Hypothesis Testing (Significance Test)

Consider the following problem: In 1980s, it was generally believed that congenital abnormalities affect 5% of the nation's children. Some people believe that the increase in the number of chemicals in the environment in recent years has led to an increase in the incidence of abnormalities. A recent study examined 384 children and found that 46 of them showed signs of abnormality. Is this strong evidence that the risk has increased?

- The above statement serves as a hypothesis, moreover it is a Research Hypothesis.

A hypothesis is:

- a statement about a population.
- a prediction that a parameter describing some characteristics of a variable (e.g., true proportion, p) takes a particular numerical value or falls in a certain range of values.

For conducting a Significance Test:

- Researchers (you) use data to summarize the evidence about a hypothesis.
- With data, you can compare the point estimates of parameters to the values predicted by the hypothesis.

Important Ideas about Hypothesis Testing

- All the hypothesis tests boil down to the same question: "Is an observed difference or pattern too large to be attributed to chance?"
- We measure "how large" by putting our sample results in the context of a sampling distribution model (e.g., Normal model, t distribution).

To plan a statistical hypothesis test, specify the model you will use to test the null hypothesis and the parameter of interest.

- All models require assumptions, so you will need to state them and check any corresponding conditions.
- For example, if the conditions are satisfied, we can model the sampling distribution of the proportion with a Normal model. Otherwise, we cannot proceed with the test (we need to stop and reconsider).

Steps in conducting Hypothesis Testing

1. State the null and the alternative hypothesis.
2. Check the necessary assumptions.
3. Identify the test-statistic. Find the value of the test-statistic.
4. Find the p-value of the test-statistic.
5. State (if any) a conclusion.

Example 11.12.
Example of Hypothesis Testing for a Proportion
In 1980s, it was generally believed that congenital abnormalities affect 5% of the nation's children. Some people believe that the increase in the number of chemicals in the environment in recent years has led to an increase in the incidence of abnormalities. A recent study

examined 384 children and found that 46 of them showed signs of abnormality. Is this strong evidence that the risk has increased?

Step 1. Set up the null and alternative hypothesis:

- The null hypothesis is the current belief: $H_0 : p = p_0$

In our example it would have a form: $H_0 : p = 0.05$

- The Alternative hypothesis is what the researcher(s) [you] want to prove: $H_a : p > p_0$

In our example it would have a form: $H_a : p > 0.05$
This means a one-sided test.

- The goal here is to provide evidence against H_0 (e.g., suggest H_a).

You want to conclude H_a.
Try a Proof by Contradiction: Assume H_0 is true ... and hope your data contradicts it.

Step 2. Check the Necessary Assumptions:

- **Independence Assumption:** There is no reason to think that one child having genetic abnormalities would affect the probability that other children have them.
- **Randomization Condition:** This sample may not be random, but genetic abnormalities are plausibly independent. The sample is probably representative of all children, with regards to genetic abnormalities.
- **10% Condition:** The sample of 384 children is less than 10% of all children.
- **Success/Failure Condition:** $np = (384)(0.05) = 19.2$ and $n(1-p) = (384)(0.95) = 364.8$ are both greater than 10, so the sample is large enough.

Step 3. Identify the test-statistics. Find the value of the test-statistic:
Since the conditions are met, assume H_0 is true:
The sampling distribution of $\hat{p}$ becomes **approximately Normal**. That is, for large n, $\hat{p}$ has approximately the

$$N\left(p_0, \sqrt{\frac{p_0(1-p_0)}{n}}\right)$$

distribution.

$$z_* = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.1198 - 0.05}{\sqrt{\frac{(0.05)(0.95)}{384}}} \approx 6.28$$

Recall that

$$\hat{p} = \frac{46}{384} = 0.1198.$$

The value of z^* is approximately 6.28, meaning that the observed proportion of children with genetic abnormalities is over 6 standard deviations above the hypothesized proportion ($p_0 = 0.05$).

Step 4. Find the p-value of the test-statistic.
P-value = $P(Z > 6.28) \approx 0.000$ (better to report p-value < 0.0001)
Note: We find the area above $Z = 6.28$ since $H_a : p > 0.05$.

Meaning of this p-value:
If 5% of children have genetic abnormalities, the chance of observing 46 children with genetic abnormalities in a random sample of 384 children is almost 0.

Step 5. Give (if any) a conclusion.
p-value is less than 0.0001, which is less than $\alpha = 0.05$; We reject $H_0 : p = 0.05$, and conclude $H_a : p > 0.05$. Our result is statistically significant at $\alpha = 0.05$.
There is very strong evidence that more than 5% of children have genetic abnormalities.

R code (1-sample proportion test)

```
prop.test(x=46, n = 384 ,p=0.05,alternative="greater", correct=FALSE);

##
## 1-sample proportions test without continuity correction
##
## data:  46 out of 384, null probability 0.05
## X-squared = 39.377, df = 1, p-value = 1.747e-10
## alternative hypothesis: true p is greater than 0.05
## 95 percent confidence interval:
##  0.09516097 1.00000000
## sample estimates:
##         p
## 0.1197917
```

Note 11.3.
About the P-value of the Test-statistics

- *P-value is a conditional probability.*
- *It is not the probability that H_0 (null hypothesis: current belief) is true.*
- *It is: P(observed statistic value [or even more extreme] — H_0). Given H_0 (the null hypothesis), because H_0 gives the parameter values that we need to find required probability.*

- *P-value serves as a measure of the strength of the evidence against the null hypothesis (but it should not serve as a hard and fast rule for decision).*
- *If p-value = 0.03 (for example) all we can say is that there is 3% chance of observing the statistic value we actually observed (or one even more inconsistent with the null value).*
- *P-value is the chance (the proportion) of getting a, for instance, $\hat{p}$ as far as or further from H_0 than the value observed.*
- *P-value is the probability of getting at least something (e.g., sample proportion $\hat{p}$) more extreme (e.g., unusual, unlikely, or rare) than what we have already found (our observed value of $\hat{p}$) that provide even stronger evidence against H_0.*
- *The more extreme the z-score (large in absolute values) are the ones that denote farther departure of the observed value (e.g., our $\hat{p}$) from the parameter value (p_0) in H_0.*
- *In the one-sided test, e.g., $H_a : p > p_0$, p-value is one-tailed probability. This is the probability that sample proportion $\hat{p}$ falls at least as far from p_0 in one direction as the observed value of $\hat{p}$.*
- *In the two-sided test, e.g., $H_a : p \neq p_0$, p-value is two-tailed probability. This is the probability that sample proportion $\hat{p}$ falls at least as far from p_0 in either direction as the observed value of $\hat{p}$.*

The probability, computed assuming that H_0 is true, that the test statistic would take a value as extreme or more extreme than that actually observed is called the **P-value** of the test. The smaller the P-value, the stronger the evidence against H_0 provided by the data. Small P-values are evidence against H_0, because they say that the observed result is unlikely to occur when H_0 is true. Large P-values fail to give evidence against H_0.

The P-value Scale

- If P-value < 0.001, we have very strong evidence against H_0.
- If $0.001 \leq$ P-value < 0.01, we have strong evidence against H_0.
- If $0.01 \leq$ P-value < 0.05, we have evidence against H_0.
- If $0.05 \leq$ P-value < 0.075, we have some evidence against H_0.
- If $0.075 \leq$ P-value < 0.10, we have slight evidence against H_0.

Use p-value Method to Make a Decision (Reject or Fail to Reject H_0)
But how small is small p-value?
We would need to choose an α-level (significance-level): a number such that if:

- P–value $\leq \alpha$-level, we reject H_0; We can conclude H_a (we have evidence to support our claim). Often we phrase as a statistically significant result at that specified α-level.

- P–value $> \alpha$-level, we fail to reject H_0; We cannot conclude H_a (we have not enough evidence to support our claim; thus, H_0 is plausible - We do not accept H_0). Often we phrase as the result is not statistically significant at that specified α-level.
- The default α-level (significance-level) is typically $\alpha = 0.05$ (but it can be different based on the context of the study - it is usually not higher than 0.10).

The p-value in the previous example was extremely small (less than 0.0001). That is a strong evidence to suggest that more than 5% of children have genetic abnormalities. However, it does not say that the percentage of sampled children with genetic abnormalities was "a lot more than 5%". That is, the p-value by itself says nothing about how much greater the percentage might be. The confidence interval provides that information.
To assess the difference in practical terms, we should also construct a confidence interval:

$$0.1198 \pm (1.96 \times 0.0166)$$

$$0.1198 \pm 0.0324$$

$$(0.0874,\ 0.1522)$$

Interpretation: We are 95% Confident that the true percentage of children with genetic abnormalities is between 8.74% and 15.22%.
95% CI for p: (9.1%, 15.6%) – We are 95% confident that the true percentage of all children that have genetic abnormalities is between approximately 9.1% and 15.6%. Since both values of this CI are more than the hypothesized value of $p = 0.05$ (5%), we can further infer that this true percentage is more than 5%.
Do environmental chemicals cause congenital abnormalities?
We do not know that environmental chemicals cause genetic abnormalities. We merely have evidence that suggests that a greater percentage of children are diagnosed with genetic abnormalities now, compared to the 1980s.

Note 11.4.
More About P-values

- *Big p-values just mean that what we have observed is not surprising. It means that the results are in line with our assumption that the null hypothesis models the world, so we have no reason to reject it.*
- *A big p-value does not prove that the null hypothesis is true.*
- *When we see a big p-value, all we can say is: we cannot reject H_0 (we fail to reject H_0) – we cannot conclude H_a (We have no evidence to support H_a).*

Some Additional Examples

Example 11.13.

Consider the following hypothesis test:

$$H_0 : p = 0.75$$
$$H_a : p < 0.75$$

A sample of 300 items was selected. Compute the p-value and state your conclusion for each of the following sample results. Use $\alpha = 0.05$.

a. $\hat{p} = 0.68$

b. $\hat{p} = 0.72$

c. $\hat{p} = 0.70$

d. $\hat{p} = 0.77$

Solution a.

$$z_* = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{0.68 - 0.75}{\sqrt{0.75(1-0.75)/300}} = -2.80$$

Using Normal table, P-value $= P(Z < z_*) = P(Z < -2.80) = 0.0026$
P-value $< \alpha = 0.05$, reject H_0.

Solution b.

$$z_* = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{0.72 - 0.75}{\sqrt{0.75(1-0.75)/300}} = -1.20$$

Using Normal table, P-value $= P(Z < z_*) = P(Z < -1.20) = 0.1151$
P-value $> \alpha = 0.05$, do not reject H_0.

Solution c.

$$z_* = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{0.70 - 0.75}{\sqrt{0.75(1-0.75)/300}} = -2.00$$

Using Normal table, P-value $= P(Z < z_*) = P(Z < -2.00) = 0.0228$
P-value $< \alpha = 0.05$, reject H_0.

Example 11.14.

Consider the following hypothesis test:

$$H_0 : p = 0.20$$
$$H_a : p \neq 0.20$$

A sample of 400 provided a sample proportion $\hat{p} = 0.175$.

a. Compute the value of the test statistic.

b. What is the p-value?

c. At the $\alpha = 0.05$, what is your conclusion?

d. What is the rejection rule using the critical value? What is your conclusion?

Solution

a.
$$z_* = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{0.175 - 0.20}{\sqrt{(0.20)(0.80)/400}} = -1.25$$

b. Using Normal table, P-value =

$$2P(Z > |z_*|) = 2P(Z > |-1.25|) = 2P(Z > 1.25) = 2(0.1056) = 0.2112$$

c. P-value $> \alpha = 0.05$, we CAN'T reject H_0.

Example 11.15.

A study found that, in 2005, 12.5% of U.S. workers belonged to unions. Suppose a sample of 400 U.S. workers is collected in 2006 to determine whether union efforts to organize have increased union membership.

a. Formulate the hypotheses that can be used to determine whether union membership increased in 2006.

b. If the sample results show that 52 of the workers belonged to unions, what is the p-value for your hypothesis test?

c. At $\alpha = 0.05$, what is your conclusion?

Solution

a.
$$H_0 : p = 0.125$$

$$H_a : p > 0.125$$

b.

$$\hat{p} = \frac{52}{400} = 0.13$$

$$z_* = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{0.13 - 0.125}{\sqrt{(0.125)(0.875)/400}} = 0.30$$

Using Normal table, P-value =

$$P(Z > z_*) = P(Z > 0.30) = 1 - 0.6179 = 0.3821$$

c. P-value > 0.05, do not reject H_0. We cannot conclude that there has been an increase in union membership.

R code

```
prop.test(52, 400, p=0.125, alternative="greater", correct=FALSE);

##
## 1-sample proportions test without continuity correction
##
## data:  52 out of 400, null probability 0.125
## X-squared = 0.091429, df = 1, p-value = 0.3812
## alternative hypothesis: true p is greater than 0.125
## 95 percent confidence interval:
##  0.1048085 1.0000000
## sample estimates:
##       p
##    0.13
```

11.3 Test of Hypothesis for One Variance

In many practical situations, we are interested in testing whether the variability in a population (i.e., its variance) has changed. This is especially important in quality control, finance, and experimental science. When we have data from a single normal population and want to test a claim about the population variance, we use the chi-squared (χ^2) test for one variance. This method assumes that the underlying population is normally distributed and the sample observations are independent.

Hypothesis Tests for One Variance

- Data from a single normal population; independent observations

- Variance unknown
- Large or small sample

Hypothesis Test

$$H_0 : \sigma^2 = \sigma_0^2$$
$$H_a : \sigma^2 \neq \sigma_0^2 \quad (\text{or } \sigma^2 > \sigma_0^2 \text{ or } \sigma^2 < \sigma_0^2)$$

Assume H_0 is true, then:

$$\text{Test statistic:} \quad \chi_*^2 = \frac{(n-1)s^2}{\sigma_0^2} \sim \chi_{n-1}^2$$

Decision rules:
$H_a : \sigma^2 \neq \sigma_0^2$.
Reject H_0 if $\chi_*^2 > \chi_{n-1;\alpha/2}^2$ or if $\chi_*^2 < \chi_{n-1;1-\alpha/2}^2$.

$H_a : \sigma^2 > \sigma_0^2$.
Reject H_0 if $\chi_*^2 > \chi_{n-1;\alpha}^2$ or if $P[\chi_{n-1}^2 > \chi_*^2]$ is too small.

$H_a : \sigma^2 < \sigma_0^2$.
Reject H_0 if $\chi_*^2 < \chi_{n-1;1-\alpha}^2$ or if $P[\chi_{n-1}^2 < \chi_*^2]$ is too small.

Note. This is **NOT** robust to departures from Normality.

Example 11.16.
A company produces metal pipes of a standard length, and claims that the standard deviation of the length is at most 1.2 cm. One of its clients decides to test this claim by taking a sample of 25 pipes and checking their lengths. They found that the standard deviation of the sample is 1.5 cm. Does this undermine the company's claim? Use $\alpha = 0.05$.
Note: Assume length is Normally distributed.

Solution

$$H_0 : \sigma^2 \leq 1.2^2$$
$$H_a : \sigma^2 > 1.2^2$$

$$\chi_*^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{(25-1)\cdot 1.5^2}{1.2^2} = 37.5$$

$$\text{P-value} = P[\chi_{24}^2 > 37.5] \approx 0.0389$$

R Code

```
1 - pchisq(37.5, df = 24);
## [1] 0.0389818
```

Conclusion
We reject $H_0 : \sigma^2 \leq 1.2^2$. We have evidence to indicate that the variance of the length of metal pipes is more than 1.2^2.

One-Sample Hypothesis Test for Population Variance (σ^2)

Assumptions: $Y_1, Y_2, \ldots, Y_n$ constitute a random sample from a Normal distribution with $E(Y_i) = \mu$ and $V(Y_i) = \sigma^2$.

Hypotheses:

$$H_0 : \sigma^2 = \sigma_0^2$$

$$H_a : \begin{cases} \sigma^2 > \sigma_0^2 & \text{(upper-tailed alternative)} \\ \sigma^2 < \sigma_0^2 & \text{(lower-tailed alternative)} \\ \sigma^2 \neq \sigma_0^2 & \text{(two-tailed alternative)} \end{cases}$$

Test statistic: $\chi^2 = \dfrac{(n-1)S^2}{\sigma_0^2}$

Rejection Region:

$$\begin{cases} \chi^2 > \chi^2_{\alpha} & \text{upper-tailed RR} \\ \chi^2 < \chi^2_{1-\alpha} & \text{lower-tailed RR} \\ \chi^2 > \chi^2_{\alpha/2} \text{ or } \chi^2 < \chi^2_{1-\alpha/2} & \text{two-tailed RR} \end{cases}$$

Example 11.17.
[Car Battery Lifetime Variance Test] A manufacturer of car batteries claims that the life of his batteries is approximately Normally distributed with a standard deviation equal to 0.9 year. If a random sample of 10 of these batteries has a standard deviation of 1.2 years, do you think that $\sigma > 0.9$ year? Use a 0.05 level of significance.
Step 1. State hypotheses.

$$H_0 : \sigma^2 = 0.81$$
$$H_a : \sigma^2 > 0.81$$

Step 2. Compute test statistic.
$S^2 = 1.44$, $n = 10$, and

$$\chi^2 = \frac{(9)(1.44)}{0.81} = 16$$

Step 3. Find Rejection Region.
From the chi-squared table, the null hypothesis is rejected when $\chi^2 > 16.919$, where $\nu = 9$ degrees of freedom.

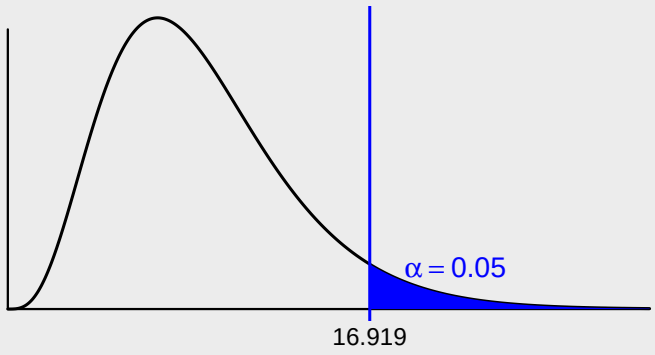


Figure 11.12: Right-tailed chi-squared distribution with critical value at 16.919.

Step 4. Conclusion.
The χ^2 statistic is not significant at the 0.05 level. We conclude that there is insufficient evidence to claim that $\sigma > 0.9$ year.

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 $\hat{p} = \frac{number\ of\ successes\ in\ the\ sample}{n}$. 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
Case 1: $H_0 : p = p_0$, $H_a : p > p_0$:
i. When is structure of testing hypothesis is $H_0 : p = p_0$, $H_a : p > p_0$:

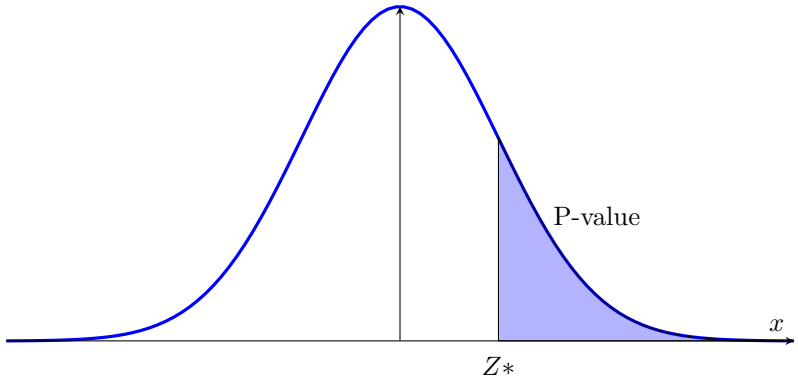


Figure 12.2: An illustration of hypothesis test on a proportion that $H_0 : p = p_0$, $H_a : p > p_0$.

Just like one same hypothesis test on a mean from previous chapter, the p-value in the case when $H_0 : p = p_0$, $H_a : p > p_0$ is the probability under the standard normal cure where the area greater than your test statistics: p-value $= P(X \geq Z^*)$.

Case 2: $H_0 : p = p_0$, $H_a : p < p_0$:

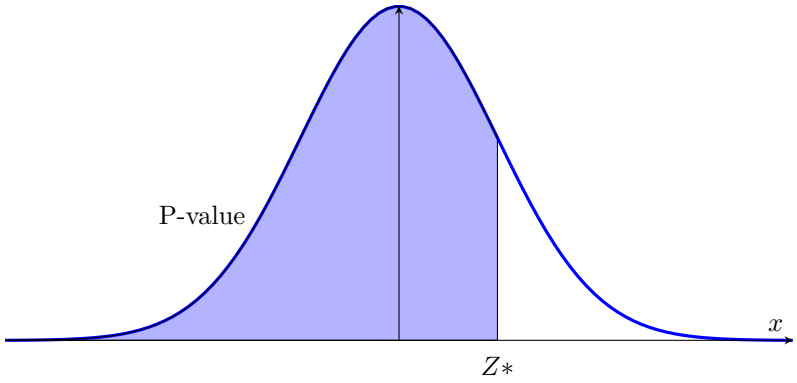


Figure 12.3: An illustration of hypothesis test on a proportion that $H_0 : p = p_0$, $H_a : p < p_0$.

The p-value in the case when $H_0 : p = p_0$, $H_a : p < p_0$ is the probability under the standard normal cure where the area less than your test statistics: p-value $= P(X \leq Z^*)$.

Case 3: $H_0 : p = p_0$, $H_a : p \neq p_0$

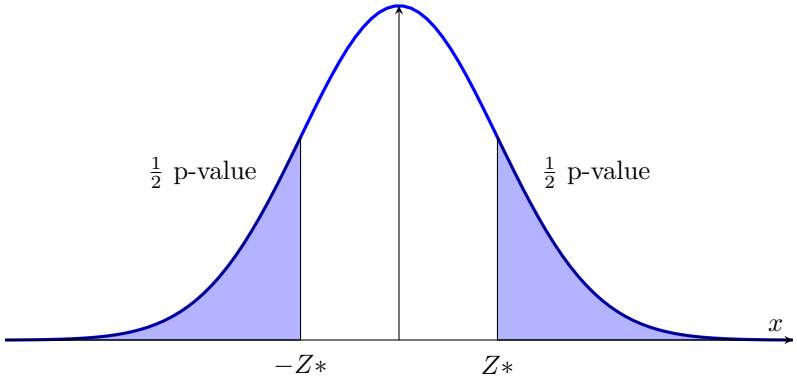


Figure 12.4: An illustration of hypothesis test on a proportion that $H_0 : p = p_0$, $H_a : p \neq p_0$.

Note that when the alternative test is $H_a : p \neq p_0$, then we are going to consider the test hypothesis as a two-tailed test. Now, let's consider that the test statistics is a positive value, then: p-value $= P(X \geq Z^*) + P(X \leq -Z^*) = 2 \cdot P(X \geq Z^*) = 2 \cdot P(X \leq -Z^*)$. Note that these different methods of calculation will give you same answer.

Step 4: Comparing p-value with α
Same as before, if $p > \alpha$, then we do not reject the null hypothesis. Otherwise, we reject the null hypothesis and take the alternative hypothesis as our final conclusion.

Step 5: Stating the final conclusion about the test
P-value $< \alpha$ level, we reject H_0; we can conclude H_a (we have evidence to support our claim). Often we phrase as a statistically significant result at that specified α-level. P-value $> \alpha$-level, we fail to reject H_0; We cannot conclude H_a (we have not enough evidence to support our claim); thus, H_0 is plausible, we do not accept H_a. Often we phrase as the result is not statistically significant at that specified α-level.

Conditions of One Sample Test Hypothesis on a Proportion

- 1. Random sample;
- 2. Independent sample;
- 3. If sample size $n < 30$, then population should be normal.

12.2 One Sample Hypothesis Tests for a Variance

Now, let's move to hypothesis tests for one variance, which follows the exact same idea from test hypothesis on proportion. We are going to introduce that by using steps as well.

Step 1: Stating the Structure of Testing Hypothesis

Cases	Null Hypothesis	Alternative Hypothesis
1	$H_0 : \sigma^2 = \sigma_0^2$	$H_a : \sigma^2 > \sigma_0^2$
2	$H_0 : \sigma^2 = \sigma_0^2$	$H_a : \sigma^2 < \sigma_0^2$
3	$H_0 : \sigma^2 = \sigma_0^2$	$H_a : \sigma^2 \neq \sigma_0^2$

Figure 12.5: All possible cases of one sample hypothesis test on a variance

Step 2: Computing Test Statistics

Definition 12.2 (Test statistics of one sample hypothesis test on a variance)**.** *The test statistics of one sample hypothesis test on a variance is given by:*

$$\chi_*^2 = \frac{(n-1)s^2}{\sigma_0^2} \sim \chi_{n-1}^2.$$

Note that n represents the sample size, s^2 is the sample variance of the chosen sample.

Step 3: Decision Rules

- 1. $H_0 : \sigma^2 = \sigma_0^2$ and $H_a : \sigma^2 \neq \sigma_0^2$ (teo tailed alternative). We reject H_0 if $\chi_*^2 > \chi_{n-1;\alpha/2}^2$ or if $\chi_*^2 < \chi_{n-1;1-\alpha/2}^2$.
- 2. $H_0 : \sigma^2 = \sigma_0^2$ and $H_a : \sigma^2 > \sigma_0^2$ (upper tailed alternative). We reject H_0 if $\chi_*^2 > \chi_{n-1;\alpha}^2$ or if $P[\chi_{n-1}^2 > \chi_*^2]$ is too small.
- 3. $H_0 : \sigma^2 = \sigma_0^2$ and $H_a : \sigma^2 < \sigma_0^2$ (lower tailed alternative). We reject H_0 if $\chi_*^2 < \chi_{n-1;1-\alpha}^2$ or if $P[\chi_{n-1}^2 < \chi_*^2]$ is too small.

Note that this is not robust to departures from normality.
In case (i), calculating $P[Z > Z_*]$ as your p-value. Then, comparing with significant level: α.
ii. When is structure of testing hypothesis is $H_0 : p = p_0$, $H_a : p < p_0$:
In case (ii), calculating $P[Z < Z_*]$ as your p-value. Then, comparing with significant level: α.
iii. When is structure of testing hypothesis is $H_0 : p = p_0$, $H_a : p \neq p_0$:
In case (iii), calculating $2 \cdot P[Z > |Z_*|]$ as your p-value. Then, comparing with significant level: α.

Step 4: Comparing P-value with α-level
If p-value is less than α-level, then we reject the null hypothesis (H_0) and accept the alternative hypothesis (H_a). Otherwise, If p-value is greater than α-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 13

Statistical Power

13.1 Statistical Power

Definition 13.1 (Power of a Test)**.** *The probability that a fixed level α significance test will reject H_0 when a particular alternative value of the parameter is true is called the* ***power*** *of the test against that alternative.*

The statistical power of a test is its ability to detect an effect if it exists in reality.
It is the probability of correctly rejecting H_0 when H_0 is false in reality.

$$\text{Power} = P(\text{reject } H_0 \mid H_0 \text{ false})$$

$$0 < \text{power} < 1$$

Power close to 1 (high power):
Test is good at detecting effects.

Power close to 0 (low power):
Test is not reliable (i.e., we expect the test will not reject H_0 when H_0 is false).

Power is affected by:

- **The effect**
  (larger differences between reality and the null are easier to detect)
- **Sample size**
  (larger samples increase power)
- **Significance level** (α)
  (as α increases, easier to reject H_0)
- **Variability in data**
  (lower variability, higher power)

Type I and II Errors

It is possible to make an incorrect conclusion on a hypothesis test.

Type I: Incorrectly reject H_0 when H_0 is true in reality.
Type II: Incorrectly fail to reject H_0 when H_0 is false in reality.

Reality vs Conclusion Table:

Conclusion		Reality: H_0 True	Reality: H_0 False
	Reject H_0	Type I (α)	No error ✓
	Fail to reject H_0	No error ✓	Type II (β)

Note: Type I errors are generally considered worse.

Let β be the probability of a Type II error. Then:

$$\text{Power} = 1 - \beta = 1 - P(\text{Type II})$$

Example 13.1. [Sweetening Colas: Power] The cola maker determines that a sweetness loss is too large to accept if the mean response for all tasters is $\mu = 1.1$. Will a 5% significance test detect this?
Hypotheses:

$$H_0 : \mu = 0$$
$$H_A : \mu > 0$$

Assume:

$$n = 10, \quad \sigma = 1, \quad \alpha = 0.05$$

Step 1: Determine the rejection region.
Since the test is one-sided with $\alpha = 0.05$, we find:

$$z_{\text{crit}} = 1.645 \quad \text{(from Z-table)}$$

We reject H_0 if:

$$Z^* > 1.645$$

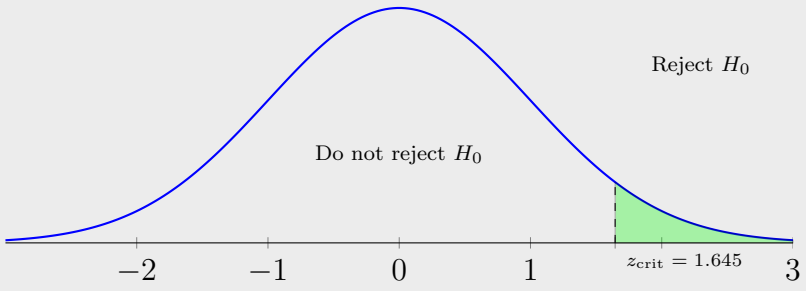


Figure 13.1: Rejection region for Z with $\alpha = 0.05$

Step 2: Find the equivalent critical value of $\bar{x}$.
Since σ is known,

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \quad \Rightarrow \quad 1.645 = \frac{\bar{x}_{\text{crit}} - 0}{1/\sqrt{10}} \Rightarrow \bar{x}_{\text{crit}} \approx 0.520$$

So we reject H_0 if $\bar{x} > 0.520$.
Step 3: Calculate the power when $\mu = 1.1$ is true.

$$P\left(Z > \frac{0.520 - 1.1}{1/\sqrt{10}}\right) \approx P(Z > -1.83) = 1 - 0.0336 = 0.9664$$

Interpretation: There is a 96.6% chance the test correctly detects $\mu = 1.1$.

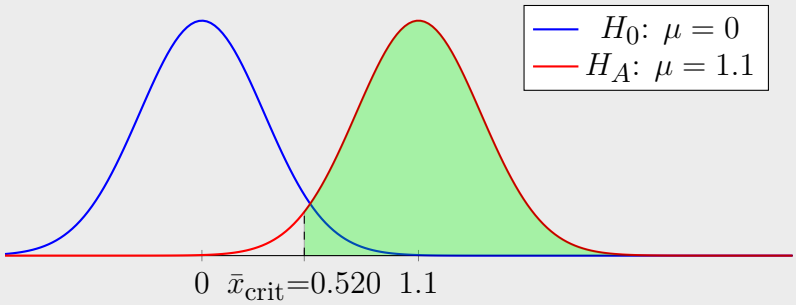


Figure 13.2: Power curve showing shaded rejection area under H_A

This result can also be visualized using a power curve, which shows how the probability of correctly rejecting H_0 increases with the true mean μ.

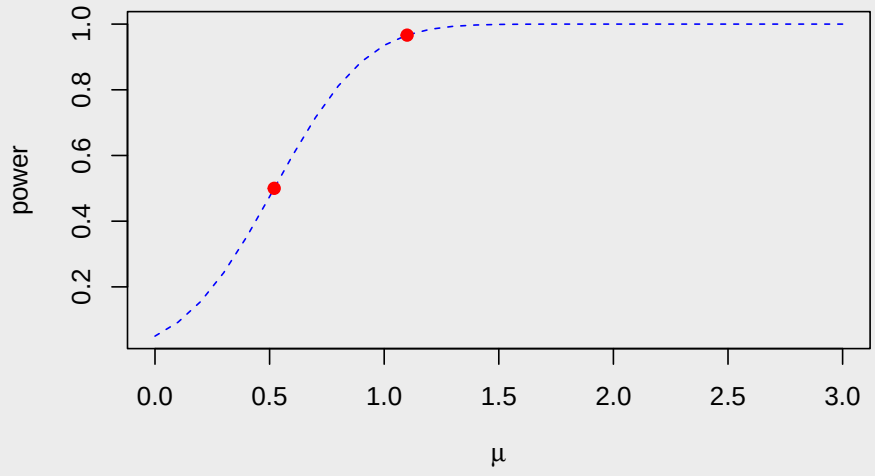


Figure 13.3: *Power curve for a one-sided test with points at $\mu = 0.52$ and $\mu = 1.1$*

13.2 Type I and Type II Errors

If we reject H_0 when in fact H_0 is true, this is a **Type I error.**

If we fail to reject H_0 when in fact H_a is true, this is a **Type II error.**

The **significance level** α of any fixed level test is the probability of a Type I error.

The **power** of a test against any alternative is 1 minus the probability of a Type II error for that alternative.

The probability of making a Type II error is denoted by β.

Decision Errors in Tests

Type I Error
H_0 is true, but sampling variation in the data leads you to reject H_0, you've made a Type I error.
When H_0 is true, a Type I error occurs if H_0 is rejected.

Type II Error
H_0 is false, but sampling variation in the data does not lead you to reject H_0, you've made a Type II error.
When H_0 is false, a Type II error occurs if H_0 is **NOT** rejected.

Example 13.2.
According to Access and Support to Education and Training Survey (2008), of 4,756 adult Canadians, 1,581 indicated that they worked at a job or business at anytime (between July 2007 and June 2008), regardless of the number of hours per week.
Is there evidence to suggest that the true proportion p is greater than 0.50?

$$H_0 : p = 0.50$$
$$H_a : p > 0.50$$

R Output

```
prop.test(x = 1581, n = 4756, p = 0.50,
          alternative = "greater", correct = FALSE)

##
##  1-sample proportions test without continuity
##  correction
##
## data:  1581 out of 4756, null probability 0.5
## X-squared = 534.24, df = 1, p-value = 1
## alternative hypothesis: true p is greater than 0.5
## 95 percent confidence interval:
##  0.3212845 1.0000000
## sample estimates:
##         p
## 0.3324222
```

P**-value** $> \alpha = 0.05$; we Fail to Reject H_0.
This means we could be making a Type II error. We indicated that there is no evidence to conclude that the true proportion of adult Canadians who worked at a job or business at anytime (between July 2007 and June 2008), regardless of the number of hours per week, was more than 0.50 — this conclusion implies that $H_0 : p = 0.50$ is plausible, but we could be wrong.

Example 13.3.

[Cola Bottles: Power Analysis]
Bottles of a popular cola are supposed to contain 300 milliliters (ml) of cola. There is some variation from bottle to bottle because the filling machinery is not perfectly precise. The distribution of contents is Normal with standard deviation $\sigma = 3$ ml. Will inspecting 6 bottles discover underfilling?
The hypotheses are:

$$H_0 : \mu = 300$$
$$H_a : \mu < 300$$

A 5% significance test rejects H_0 if $z_* \leq -1.645$, where the test statistic z_* is:

$$z_* = \frac{\bar{x} - 300}{3/\sqrt{6}}$$

Power calculations help us see how large a shortfall in the bottle contents the test can be expected to detect. Find the power of this test against the alternative $\mu = 299$.

Step 1. Write the rule for rejecting H_0 in terms of $\bar{x}$.

We know that $\sigma = 3$, so the z test rejects H_0 at the $\alpha = 0.05$ level when:

$$z = \frac{\bar{x} - 300}{3/\sqrt{6}} < -1.645$$

This is the same as:

$$\bar{x} < 300 - 1.645 \cdot \frac{3}{\sqrt{6}} \quad \Rightarrow \quad \bar{x} < 297.985$$

Step 2. The power is the probability of this event under the condition that the alternative $\mu = 299$ is true.
To calculate this probability, standardize $\bar{x}$ using $\mu = 299$:

$$\begin{aligned} \text{power} &= P(\bar{x} < 297.985 \mid \mu = 299) \\ &= P\left(Z < \frac{297.985 - 299}{3/\sqrt{6}}\right) \\ &= P(Z < -0.83) = 0.2033 \end{aligned}$$

13.3 Using Power to Determine Sample Size

When designing a study, one of the most important decisions is how large a sample to collect. If the sample size is too small, even meaningful effects may go undetected due to low statistical power. On the other hand, collecting an unnecessarily large sample can be inefficient and costly. By using power calculations, researchers can determine the minimum sample size needed to detect an effect of a given size with a specified probability (power), while controlling for Type I error. This section introduces how statistical power is used in planning and justifying sample sizes before conducting a hypothesis test.

Example 13.4.
Suppose an experimenter wishes to test

$$\begin{aligned} H_0 &: \mu = 100 \\ H_a &: \mu > 100 \end{aligned}$$

at the $\alpha = 0.05$ level of significance and wants $1 - \beta$ to equal 0.60 when $\mu = 103$. What is the smallest (i.e., cheapest) sample size that will achieve that objective? Assume that the variable being measured is Normally distributed with $\sigma = 14$.
Step 1. Write the rule for rejecting H_0 in terms of $\bar{x}_*$.
By definition,

$$\alpha = P(\text{we reject } H_0 \mid \mu = 100) = P\left(\bar{X} > \bar{x}_* \mid \mu = 100\right) = P\left(Z > \frac{\bar{x}_* - 100}{14/\sqrt{n}}\right) = 0.05$$

From the standard normal table, $P(Z > 1.645) = 0.05$, so:

$$\bar{x}_* = 100 + 1.645 \cdot \frac{14}{\sqrt{n}}$$

Step 2. The power is the probability of this event under the condition that the alternative $\mu = 103$ is true.
To calculate this probability, standardize $\bar{x}$ using $\mu = 103$:

$$\begin{aligned} \text{power} = 1 - \beta &= P(\bar{X} > \bar{x}_* \mid \mu = 103) \\ &= P\left(Z > \frac{\bar{x}_* - 103}{14/\sqrt{n}}\right) = 0.60 \end{aligned}$$

From the standard normal table, $P(Z > -0.25) \approx 0.5987 \approx 0.60$, so:

$$\frac{\bar{x}_* - 103}{14/\sqrt{n}} = -0.25 \Rightarrow \bar{x}_* = 103 - 0.25 \cdot \frac{14}{\sqrt{n}}$$

Step 3. Solving for n
From Steps 1 and 2:

$$100 + 1.645 \cdot \frac{14}{\sqrt{n}} = 103 - 0.25 \cdot \frac{14}{\sqrt{n}}$$

Solving for n:

$$\left(\frac{(1.645 + 0.25) \cdot 14}{103 - 100}\right)^2 \approx 78.2045$$

Therefore, a minimum of 79 observations must be taken to guarantee that the hypothesis test will have the desired power of at least 0.60.

Example 13.5.
A vending machine advertises that it dispenses 225 ml cups of coffee ($\sigma = 7$ ml). You believe the mean volume of coffee per cup is something less than 225 ml. You plan to sample 40 cups of coffee from this machine to test your hypothesis.

a) If the true mean volume of coffee per cup is 223 ml, what is the power of your test at $\alpha = 0.05$? *(Homework)*

b) How many coffee cups should you sample if you want to raise the power in part (a) to 0.80?

Solution (b):
Step 1. Write the rule for rejecting H_0 in terms of $\bar{x}_*$.

By definition:

$$\begin{aligned}\alpha &= P(\text{we reject } H_0 \mid \mu = 225)\\ &= P(\bar{X} < \bar{x}_* \mid \mu = 225)\\ &= P\left(Z < \frac{\bar{x}_* - 225}{7/\sqrt{n}}\right) = 0.05\end{aligned}$$

From the standard normal table, $P(Z < -1.645) = 0.05$, so:

$$\bar{x}_* = 225 - 1.645 \cdot \frac{7}{\sqrt{n}}$$

Step 2. The power is the probability of this event under the condition that the alternative $\mu = 223$ is true.
To calculate this probability, standardize $\bar{x}$ using $\mu = 223$:

$$\begin{aligned}\text{power} = 1 - \beta &= P(\bar{X} < \bar{x}_* \mid \mu = 223)\\ &= P\left(Z < \frac{\bar{x}_* - 223}{7/\sqrt{n}}\right) = 0.80\end{aligned}$$

From the table, $P(Z < 0.84) = 0.7995 \approx 0.80$, so:

$$\frac{\bar{x}_* - 223}{7/\sqrt{n}} = 0.84 \quad \Rightarrow \quad \bar{x}_* = 223 + 0.84 \cdot \frac{7}{\sqrt{n}}$$

Step 3. Solving for n
From Steps 1 and 2:

$$225 - 1.645 \cdot \frac{7}{\sqrt{n}} = 223 + 0.84 \cdot \frac{7}{\sqrt{n}}$$

Solving for n:

$$n = \left(\frac{(1.645 + 0.84) \cdot 7}{225 - 223}\right)^2 \approx 75.6465$$

Therefore, a minimum of 76 observations must be taken to guarantee that the hypothesis test will have the desired precision.

Example 13.6.
[Power and Sample Size for Milk Consumption Study]
A newsletter reports that 90% of adults drink milk. The researchers are interested in investigating if fewer than 90% of adults drink milk (at $\alpha = 0.05$). They collect a random sample of 200 adults in a certain region.

a. Calculate power of the test if the percentage of adults who drink milk is really 85%.

$$\text{Under } H_0 : \hat{p}^* = 0.90 - 1.645\sqrt{\frac{0.90(1-0.90)}{200}} = 0.8651$$

$$\begin{aligned}
\text{Power} &= P(\text{Reject } H_0 \mid p = 0.85) \\
&= P(\hat{p} < 0.8651 \mid p = 0.85) \\
&= P\left(Z < \frac{0.8651 - 0.85}{\sqrt{\frac{0.85(1-0.85)}{200}}}\right) \\
&= P(Z < 0.5980) \approx 0.7250
\end{aligned}$$

R Output

```
pnorm(0.5980, mean = 0, sd = 1)
# [1] 0.72508
```

b. Calculate beta if the percentage of adults who drink milk is really 85%.

$$\beta = 1 - \text{Power} = 1 - 0.725 = 0.275$$

c. How many adults should you sample if you want to raise the power in part (a) to 0.80?

Step 1: Determine rejection cutoff under $H_0 : p = 0.90$

$$\hat{p}^* = 0.90 - 1.645\sqrt{\frac{0.90(1-0.90)}{n}}$$

Step 2: Set power to 0.80 under $p = 0.85$

$$P\left(\frac{\hat{p} - 0.85}{\sqrt{\frac{0.85(1-0.85)}{n}}} < \frac{\hat{p}^* - 0.85}{\sqrt{\frac{0.85(1-0.85)}{n}}}\right) = 0.80$$

$$\frac{\hat{p}^* - 0.85}{\sqrt{\frac{0.85(1-0.85)}{n}}} = 0.8416$$

$$\hat{p}^* = 0.85 + 0.8416\sqrt{\frac{0.85(1-0.85)}{n}}$$

R Output

```
qnorm(0.80, mean = 0, sd = 1)
# [1] 0.8416212
```

Step 3: Equating expressions for $\hat{p}^*$

$$0.90 - 1.645\sqrt{\frac{0.90(1-0.90)}{n}} = 0.85 + 0.8416\sqrt{\frac{0.85(1-0.85)}{n}}$$

$$0.05 = \left(1.645\sqrt{0.90(0.10)} + 0.8416\sqrt{0.85(0.15)}\right) \cdot \frac{1}{\sqrt{n}}$$

$$\sqrt{n} = \frac{0.4935 + 0.3005}{0.05} = 15.87 \Rightarrow n \approx 253$$

Example 13.7.

A newsletter reports that 90% of adults drink milk. The researchers are interested in investigating if less than 90% of adults drink milk (at $\alpha = 0.05$). They collect a **random sample of 100 adults** in a certain region.

Calculate power of the test if the percentage of adults who drink milk is really 85%.
We test:

$$H_0 : p = 0.90$$
$$H_a : p < 0.90$$

The rejection region is determined by:

$$\alpha = 0.05 = P(\text{reject } H_0 \mid H_0 \text{ is true})$$
$$P(Z < -1.645) = 0.05 \quad \text{(Z critical value is -1.645)}$$

Now calculate power under the alternative $p = 0.85$:

$$\begin{aligned}
\text{Power} &= P(\text{reject } H_0 \mid H_0 \text{ is false}) \\
&= P\left(\frac{\hat{p} - 0.90}{\sqrt{\frac{0.90(1-0.90)}{100}}} < -1.645 \,\middle|\, p = 0.85\right) \\
&= P(\hat{p} < 0.85065 \mid p = 0.85) \\
&= P\left(Z < \frac{0.85065 - 0.85}{\sqrt{\frac{0.85(1-0.85)}{100}}}\right) \\
&= P(Z < 0.0182) \approx 0.5072
\end{aligned}$$

R Output

```
pnorm(0.0182, mean = 0, sd = 1)
# [1] 0.5072603
```

Example 13.8.
A newsletter reports that 90% of adults drink milk. The researchers are interested in investigating if less than 90% of adults drink milk (at $\alpha = 0.05$). They collect a **random sample of 50 adults** in a certain region.

Calculate power of the test if the percentage of adults who drink milk is really 85%.

$$H_0 : p = 0.90$$
$$H_a : p < 0.90$$

The rejection region is determined by:

$$\alpha = 0.05 = P(\text{reject } H_0 \mid H_0 \text{ is true})$$
$$P(Z < -1.645) = 0.05 \quad (\text{Z critical value is -1.645})$$

Power:

$$\begin{aligned}
\text{Power} &= P(\text{reject } H_0 \mid H_0 \text{ is false}) \\
&= P\left(\frac{\hat{p} - 0.90}{\sqrt{\frac{0.90(1-0.90)}{50}}} < -1.645 \,\middle|\, p = 0.85 \right) \\
&= P(\hat{p} < 0.85065 \mid p = 0.85) \\
&= P\left(Z < \frac{0.85065 - 0.85}{\sqrt{\frac{0.85(1-0.85)}{50}}} \right) \\
&= P(Z < -0.3921) \\
&\approx 0.3475
\end{aligned}$$

R Output

```
pnorm(-0.3921, mean = 0, sd = 1)
# [1] 0.3474922
```

If we keep α at the same size, larger sample sizes increase the power of the test because sampling variability (sampling distributions) are much narrower.

Power Formulas

If $H_a : \mu > \mu_0$, then

$$\text{Power}(\mu_a) = \pi(\mu_a) = P\left[Z \geq Z_\alpha + \frac{\sqrt{n}(\mu_0 - \mu_a)}{\sigma}\right]$$

where $\mu_a > \mu_0$.

If $H_a : \mu < \mu_0$, then

$$\text{Power}(\mu_a) = \pi(\mu_a) = P\left[Z \leq -Z_\alpha + \frac{\sqrt{n}(\mu_0 - \mu_a)}{\sigma}\right]$$

where $\mu_a < \mu_0$.

If $H_a : \mu \neq \mu_0$, then

$$\begin{aligned}\text{Power}(\mu_a) &= \pi(\mu_a) \\ &= 1 - P\left[-Z_{\alpha/2} + \frac{\sqrt{n}(\mu_0 - \mu_a)}{\sigma} \leq Z \leq Z_{\alpha/2} + \frac{\sqrt{n}(\mu_0 - \mu_a)}{\sigma}\right]\end{aligned}$$

where $\mu_a \neq \mu_0$.

You need to know:

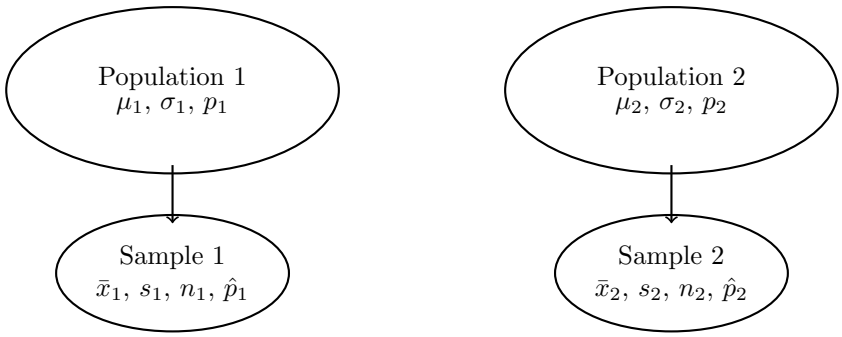
- Standard deviation, σ
- Significance level, α
- Effect size to detect, $\mu_0 - \mu_a$

Calculations of power (or of error probabilities) are useful for planning studies because we can make these calculations before we have any data. Once we actually have data, it is more common to report a P-value rather than a reject-or-not decision at a fixed significance level α. The P-value measures the strength of the evidence provided by the data against H_0. It leaves any action or decision based on that evidence up to each individual. Different people may require different strengths of evidence.

Chapter 14

Two Sample Hypothesis Tests

14.1 Comparing Means with Independent Samples



Parameters of Interest:

Difference in population means: $\mu_1 - \mu_2$

Difference in proportions: $p_1 - p_2$

Figure 14.1: Structure of Two-Sample Hypothesis Tests

Setting Up Hypotheses

Let θ_1 and θ_2 be the parameters of interest from populations 1 and 2, respectively.

1. Interested in whether $\theta_1 > \theta_2$:

$$H_0 : \theta_1 = \theta_2$$
$$H_a : \theta_1 > \theta_2$$
$$\text{Equivalent:} \quad H_0 : \ \theta_1 - \theta_2 = 0 \qquad H_a : \ \theta_1 - \theta_2 > 0$$

2. Interested in whether $\theta_1 < \theta_2$:

$$H_0 : \theta_1 = \theta_2$$
$$H_a : \theta_1 < \theta_2$$

$$\text{Equivalent:} \quad H_0 \colon \ \theta_1 - \theta_2 = 0 \qquad H_a \colon \ \theta_1 - \theta_2 < 0$$

3. Interested in whether $\theta_1 \neq \theta_2$:

$$H_0 \colon \theta_1 = \theta_2$$
$$H_a \colon \theta_1 \neq \theta_2$$
$$\text{Equivalent:} \quad H_0 \colon \ \theta_1 - \theta_2 = 0 \qquad H_a \colon \ \theta_1 - \theta_2 \neq 0$$

Structure of a Test Statistic

The general structure of a test statistic is:

$$\text{test statistic} = \frac{(\text{observed statistic}) - (\text{hypothesized value})}{\text{standard error}}$$

Common Cases:

- σ known: $z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$
- σ unknown: $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$
- Proportions: $z = \dfrac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$

Hypothesis Test on a Difference of Means ($\mu_1 - \mu_2$)

- When σ_1 and σ_2 are known:

$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_a : \mu_1 - \mu_2 > 0 \quad (\text{or } \ < 0, \ \neq 0)$$

Test Statistic:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$$

Reference distribution: standard normal (Z distribution).
Note: This scenario is rare in practice since both population standard deviations σ_1 and σ_2 are seldom known.

- When σ_1 and σ_2 are unknown:

Case 1: Assume equal variances $\sigma_1^2 = \sigma_2^2$

$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_a : \mu_1 - \mu_2 > 0 \quad (\text{or } \ < 0, \ \neq 0)$$

Test Statistic:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where the pooled standard deviation is defined as:

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

Reference distribution: t-distribution with $n_1 + n_2 - 2$ degrees of freedom.
(Use this only when equal variances can reasonably be assumed.)

Case 2: Assume Unequal Variances $\sigma_1^2 \neq \sigma_2^2$ (Welch's t-test)

$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_a : \mu_1 - \mu_2 > 0 \quad (\text{or } < 0, \neq 0)$$

Test Statistic:

$$t^* = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Reference distribution: approximately a t-distribution.

Degrees of freedom (by hand):

$$\min(n_1 - 1,\ n_2 - 1)$$

Note: R uses a more sophisticated approximation for degrees of freedom in this case.

Comparing Means of Independent Samples (Normal Population Assumptions)

We should check the assumption that the underlying populations of individual responses are each Normally distributed. Nearly Normal Condition:

- We must check this for both groups; a violation for either one violates the condition.
- The Normality assumption matters most when sample sizes are very small.
- For $n < 10$ in either group, this method should not be used if the histogram or Normality plots show clear skewness.
- For n's of 10 or so, a moderately skewed histogram is okay. But, for strongly skewed data or data containing outliers this method should be avoided.
- For larger samples $n \geq 20$, data skewness is less of an issue — but, we still need to check if there are any outliers in the data, extreme skewness, and multiple modes.

14.1.1 Comparing Two Populations Means: Independent Sampling (Equal Variances Assumed)

Consider two independent populations with unknown means μ_1 and μ_2, and unknown standard deviations σ_1 and σ_2 ($\sigma_1 = \sigma_2$), respectively. We can make an inference about their mean difference $\mu_1 - \mu_2$ by using the difference between their point estimates (sample means): $\bar{Y}_1 - \bar{Y}_2$. When the assumptions and conditions are met,

$$\frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}},$$

can be modelled by a $t(\nu)$ distribution; where $\nu = n_1 + n_2 - 2$ and

$$S_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

Comparing Two Populations Means: Independent Sampling (Equal Variances Assumed) (cont.)

Conditions Required for Valid Inference about $\mu_1 - \mu_2$:

1. The two samples are randomly selected in an independent manner from the two target populations.
2. Both sampled populations have distributions that are approximately Normal.
3. The population variances are equal (e.g., $\sigma_1 = \sigma_2$).

Small-Sample Confidence Interval for $\mu_1 - \mu_2$ (with equal variances)

Parameter: $\mu_1 - \mu_2$

Confidence interval ($\nu = \text{df}$):

$$(\bar{Y}_1 - \bar{Y}_2) \pm t_{\alpha/2}(\nu) S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

where $\nu = n_1 + n_2 - 2$ and

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

(requires that Normal samples are independent and the assumption that $\sigma_1^2 = \sigma_2^2$). The critical value $t_{\alpha/2}(\nu)$ depends on the particular confidence level and the number of degrees of freedom.

Example 14.1.
Comparing Two Population Means Managerial Success Indexes for Two Groups (With Equal Variances Assumed)

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.

Response variable: Managerial Success Indexes (quantitative, continuous, 0–100 scale)
Explanatory variable: Type of group (nominal categorical: *Group 1 = interaction with outsiders*, *Group 2 = fewer interactions*)

R Code

```
# 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);
```

R Code

```
## [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
```

R code (Descriptive Statistics)

```
# loading library mosaic
library(mosaic)

favstats(Success_Index ~ Group)
```

R Code (Descriptive Statistics)

```
## .group min   Q1 median   Q3 max     mean       sd  n
##       1  53 62.25  65.50 69.25  78 65.33333 6.610368 12
##       2  34 42.50  50.00 54.50  68 49.46667 9.334014 15
```

R Code (Descriptive Statistics)

```
summary(Success_Index[Group == 1]);
length(Success_Index[Group == 1]);
sd(Success_Index[Group == 1]);

summary(Success_Index[Group == 2]);
length(Success_Index[Group == 2]);
sd(Success_Index[Group == 2]);
```

Note: Group 1 = "interaction with outsiders" and Group 2 = "fewer interactions".

R Output

```
##  Min.  1st Qu.  Median    Mean  3rd Qu.    Max.
##  53.00   62.25   65.50   65.33   69.25   78.00
## [1] 12
## [1] 6.610368

##  Min.  1st Qu.  Median    Mean  3rd Qu.    Max.
##  34.00   42.50   50.00   49.47   54.50   68.00
## [1] 15
## [1] 9.334014
```

Nearly Normal Condition (Group 1: "interaction with outsiders"):

```
stem(Success_Index[Group == 1]);
```

R Output

```
##
##  The decimal point is 1 digit(s) to the right of the |
##
##  5 | 38
##  6 | 0335689
##  7 | 018
```

Nearly Normal Condition (Group 2: "fewer interactions"):

```
stem(Success_Index[Group == 2]);
```

R Output

```
##
##  The decimal point is 1 digit(s) to the right of the |
##
##  3 | 46
##  4 | 22368
##  5 | 023367
##  6 | 28
```

Nearly Normal Condition (Group 1: "interaction with outsiders"):

```
qqnorm(Success_Index[Group == 1]);
qqline(Success_Index[Group == 1]);
```

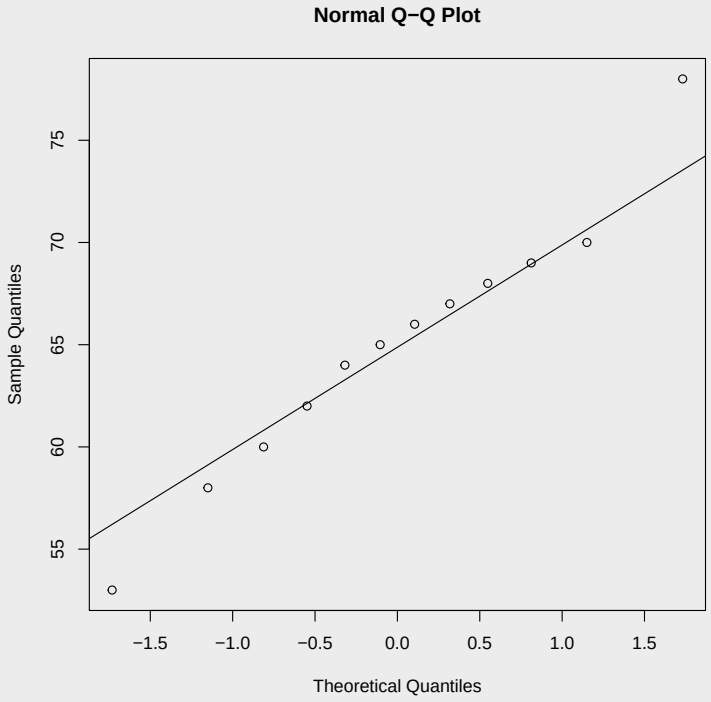


Figure 14.2: Q-Q Plot for Group 1: Interaction with Outsiders

Nearly Normal Condition (Group 2: "fewer interactions"):

```
qqnorm(Success_Index[Group == 2]);
qqline(Success_Index[Group == 2]);
```

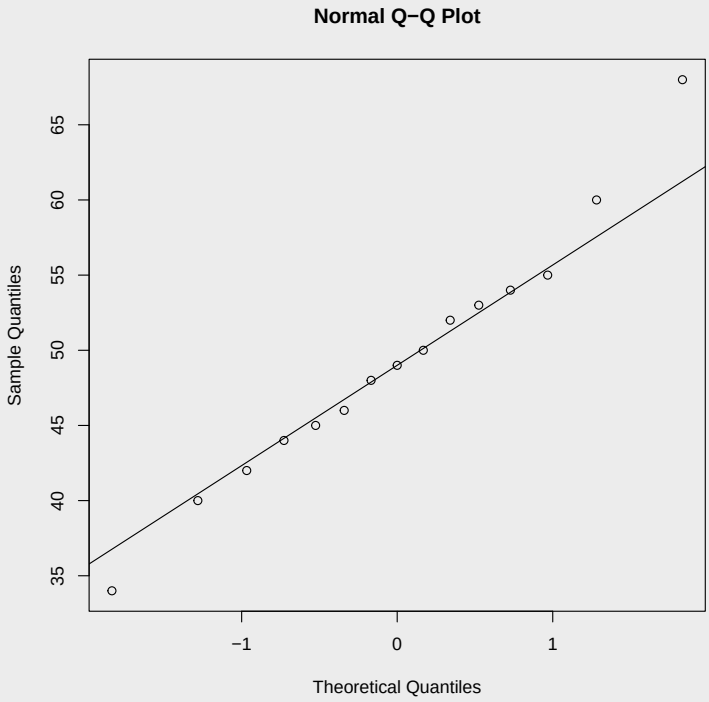


Figure 14.3: Q-Q Plot for Group 2: Fewer Interactions

Nearly Normal Condition:

```
boxplot(Success_Index ~ Group, col = c("red", "blue"))
```

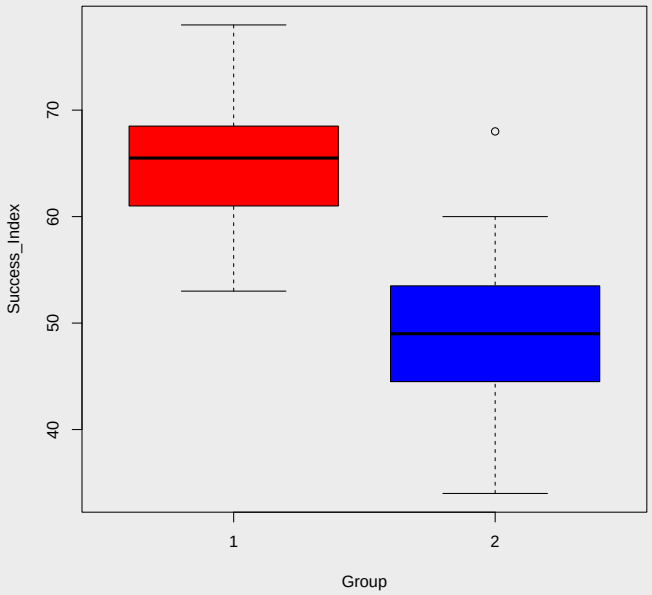


Figure 14.4: Boxplot of Success Index by Group

Boxplot with ggplot2:

```
# loading library;
library(ggplot2);

# converting a numeric variable into factor (categorical data)
group <- factor(Group);

# bp: just a name (not code) to store boxplots;
bp <- ggplot(success,
             aes(x = group, y = Success_Index, fill = group));

our.labs <- c("Interaction with Outsiders", "Fewer Interactions");

bp +
  geom_boxplot() +
  scale_x_discrete(labels = our.labs);
```

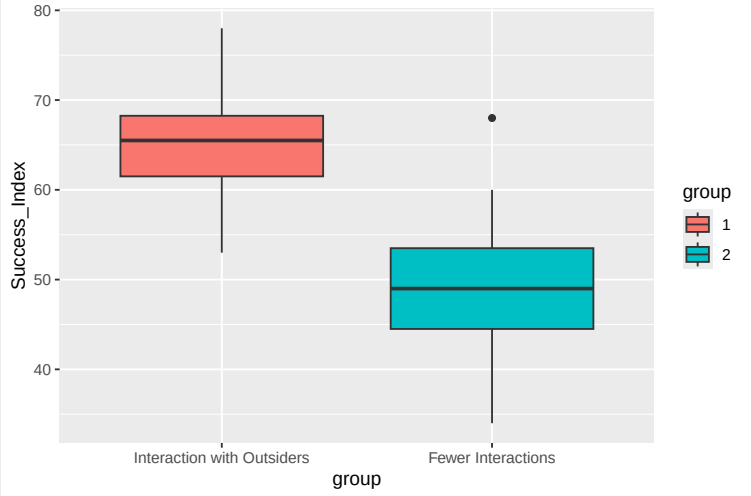


Figure 14.5: Boxplot of Success Index by Group using ggplot2

Checking the Assumptions and Conditions

Independent Group Assumption: The success index in group 1 is unrelated to the success index in group 2.
Randomization Condition: The 27 managers were randomly and independently selected (12 for group 1, and 15 for group 2).
Nearly Normal Condition: The two boxplots of success indexes do not show skewness; the two stemplots/histograms of success indexes are unimodal, fairly symmetric and approximately bell-shaped. Q–Q plots also suggest the normality assumption is reasonable.
Equal Variances Assumption: The two boxplots of success indexes appear to have the same spread; thus, the samples appear to have come from populations with approximately the same variance.

Since the conditions are satisfied, it is appropriate to construct a t confidence interval with $df = 12 + 15 - 2 = 25$.

From the data, the following statistics were calculated:

$$\begin{aligned} n_1 &= 12 & n_2 &= 15 \\ \bar{x}_1 &= 65.33 & \bar{x}_2 &= 49.47 \\ s_1^2 &= 6.61^2 & s_2^2 &= 9.33^2 \end{aligned}$$

The pooled variance estimator is:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(12 - 1)(6.61^2) + (15 - 1)(9.33^2)}{12 + 15 - 2} = 67.97$$

The number of degrees of freedom is:

$$\nu = n_1 + n_2 - 2 = 12 + 15 - 2 = 25$$

Two-sample t-test (Student's t-test) for the Difference Between Means $\mu_1 - \mu_2$: Is there evidence to suggest that the mean success index differs between the two groups?

1. State Hypotheses:

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 \quad \text{or} \quad H_0 : \mu_1 - \mu_2 = 0 \\ H_a &: \mu_1 \neq \mu_2 \quad \text{or} \quad H_a : \mu_1 - \mu_2 \neq 0 \end{aligned}$$

2. Test Statistic:

$$t^* = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} = \frac{65.33 - 49.47}{\sqrt{\frac{67.97}{12} + \frac{67.97}{15}}} = 4.97$$

3. P-value:
Using table:

$$\text{df} = 25 \quad \Rightarrow \quad \text{P-value} < 2(0.005)$$

Using R:

```
# One way:
2*(1 - pt(4.97, df = 25));

# Another way:
2 * pt(4.97, df = 25, lower.tail = FALSE);
```

Both methods return:

$$\text{P-value} \approx 4.03 \times 10^{-5}$$

4. Conclusion.

Since the P-value is very small, we have strong evidence to indicate that there is a difference in mean success index between group 1 and group 2.

Note. As a follow-up, we could find a 95% CI for $\mu_1 - \mu_2$ to estimate this difference. Then, we could provide an estimate of how much higher the mean success index for group 1 is.

14.1.2 Comparing Two Populations Means: Independent Sampling (Unequal Variances Assumed)

Small-Sample Confidence Interval for $\mu_1 - \mu_2$ (Unequal Variances)

Draw an SRS of size n_1 from a Normal population with unknown mean μ_1, and draw an independent SRS of size n_2 from another Normal population with unknown mean μ_2. A confidence interval for $\mu_1 - \mu_2$ is given by

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Here t^* is the critical value for the $t(k)$ density curve with area C between $-t^*$ and t^*. The degrees of freedom k are equal to the smaller of $n_1 - 1$ and $n_2 - 1$.

Degrees of Freedom

Option 1. With software, use the statistic t with accurate critical values from the approximating t distribution.
The distribution of the two-sample t statistic is very close to the t distribution with degrees of freedom df given by

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2\right) + \left(\frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2\right)}$$

This approximation is accurate when both sample sizes n_1 and n_2 are 5 or larger.

Option 2. Without software, use the statistic t with critical values from the t distribution with degrees of freedom equal to the smaller of $n_1 - 1$ and $n_2 - 1$. These procedures are always conservative for any two Normal populations.

Two-Sample t-Test (Unequal Variances)

To test the hypothesis $H_0 : \mu_1 = \mu_2$, calculate the two-sample t statistic:

$$t^* = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

and use P-values or critical values for the $t(k)$ distribution.

Example 14.2.

[Logging in the rain forest] "Conservationists have despaired over destruction of tropical rain forest by logging, clearing, and burning". These words begin a report on a statistical study of the effects of logging in Borneo. Here are data on the number of tree species in 12 unlogged forest plots and 9 similar plots logged 8 years earlier:
Unlogged: 22 18 22 20 15 21 13 13 19 13 19 15
Logged : 17 4 18 14 18 15 15 10 12
Does logging significantly reduce the mean number of species in a plot after 8 years? State the hypotheses and do a t test. Is the result significant at the 5% level?

Solution:

1. State hypotheses. $H_0 : \mu_1 = \mu_2$ vs. $H_a : \mu_1 > \mu_2$, where μ_1 is the mean number of species in unlogged plots and μ_2 is the mean number of species in plots logged 8 years earlier.
2. Test statistic.

$$t^* = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} = 2.1140$$

$$(\bar{x}_1 = 17.5,\ \bar{x}_2 = 13.6666,\ s_1 = 3.5290,\ s_2 = 4.5,\ n_1 = 12,\ n_2 = 9)$$

3. P-value. Using Table, we have $df = 8$, and $0.025 <$ P-value < 0.05.
4. Conclusion. Since P-value < 0.05, we reject H_0. There is strong evidence that the mean number of species in unlogged plots is greater than that for logged plots 8 years after logging.

Example 14.3.

A company that sells educational materials reports statistical studies to convince customers that its materials improve learning. One new product supplies "directed reading activities" for classroom use. These activities should improve the reading ability of elementary school pupils.
A consultant arranges for a third-grade class of 21 students to take part in these activities for an eight-week period. A control classroom of 23 third-graders follows the same curriculum without the activities. At the end of the eight weeks, all students are given a Degree of Reading Power (DRP) test, which measures the aspects of reading ability that the treatment is designed to improve. The data appear in the following table.

Treatment				Control			
24	61	59	46	42	33	46	37
43	44	52	43	43	41	10	42
58	67	62	57	55	19	17	55
71	49	54		26	54	60	28
43	53	57		62	20	53	48
49	56	33		37	85	42	

Because we hope to show that the treatment (Group 1) is better than the control (Group 2), the hypotheses are:

$$H_0 : \mu_1 = \mu_2$$

$$H_a : \mu_1 > \mu_2$$

```
# Step 1. Entering data
treatment = c(24, 61, 59, 46, 43, 44, 52, 43, 58, 67, 62, 57,
              71, 49, 54, 43, 53, 57, 49, 56, 33);
control = c(42, 33, 46, 37, 43, 41, 10, 42, 55, 19, 17, 55,
            26, 54, 60, 28, 62, 20, 53, 48, 37, 85, 42);
```

Nearly Normal Condition (treatment):

```
# Making stemplot;

stem(treatment);
```

```
##
## The decimal point is 1 digit(s) to the right of the |
##
## 2 | 4
## 3 | 3
## 4 | 3334699
## 5 | 23467789
## 6 | 127
## 7 | 1
```

Nearly Normal Condition (control):

```
# Making stemplot;

stem(control);
```

```
##
## The decimal point is 1 digit(s) to the right of the |
##
## 0 | 079
## 2 | 068377
## 4 | 12223683455
## 6 | 02
## 8 | 5
```

Nearly Normal Condition (treatment):

```
# Making Q-Q plot;
qqnorm(treatment, pch=19, col="red", main="Treatment");
qqline(treatment, lty=2);
```

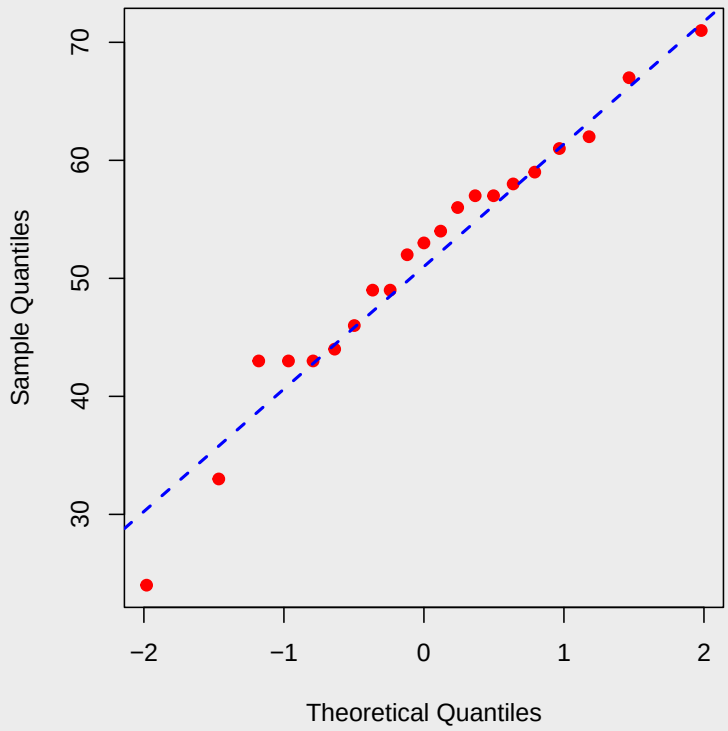


Figure 14.6: Q-Q Plot for Treatment Group

Nearly Normal Condition (control):

```
# Making Q-Q plot;
qqnorm(treatment, pch=19, col="red", main="Control");
qqline(control, lty=2);
```

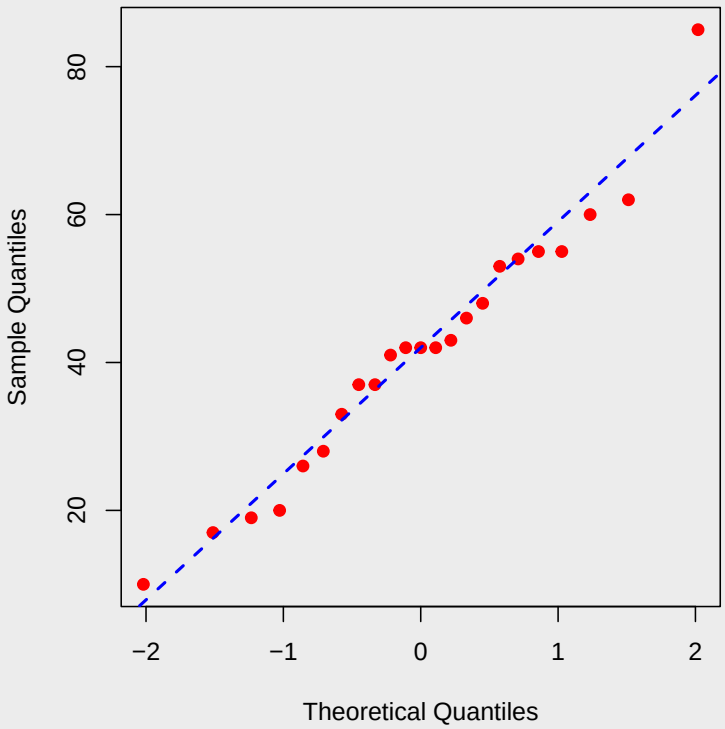


Figure 14.7: Q-Q Plot for Control Group

Stemplots suggest that there is a mild outlier in the control group but no deviation from Normality serious enough to prevent us from using t procedures. Normal Q-Q plots for both groups confirm that both are roughly Normal. The summary statistics are:

```
summary(treatment);

## Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 24.00   44.00   53.00   51.48   58.00   71.00

summary(control);

## Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 10.00   30.50   42.00   41.52   53.50   85.00
```

```
# Step 1. Entering data;

treatment = c(24, 61, 59, 46, 43, 44, 52, 43, 58, 67, 62, 57,
              71, 49, 54, 43, 53, 57, 49, 56, 33);

control = c(42, 33, 46, 37, 43, 41, 10, 42, 55, 19, 17, 55,
            26, 54, 60, 28, 62, 20, 53, 48, 37, 85, 42);

# Step 2. Hypothesis Test

t.test(treatment, control, alternative="greater")
```

HT (using R)

```
##
##  Welch Two Sample t-test
##
## data:  treatment and control
## t = 2.3106, df = 37.855, p-value = 0.01305
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  2.333784      Inf
## sample estimates:
## mean of x mean of y
##    51.47619 41.52174
```

HT (using table)

```
round (mean(treatment) ,2);

##[1] 51.48

round (sd(treatment) ,2);

##[1] 11.01

round (mean(control) ,2);

##[1] 41.52

round (sd(control) ,2);

##[1] 17.15
```

Test statistic.

$$t^* = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = 2.31$$

$$(\bar{x}_1 = 51.48,\ \bar{x}_2 = 41.52,\ s_1 = 11.01,\ s_2 = 17.15, \quad n_1 = 21 \text{ and } n_2 = 23)$$

The conservative approach uses the $t(20)$ distribution. The P-value for the one-sided test is

$$\text{P-value} = P\big(T \geq 2.31\big)$$

Comparing $t = 2.31$ with the entries in Table 5 for 20 degrees of freedom, we see that

$$0.01 < \text{P-value} < 0.025.$$

Since our P-value is "small", we reject the null hypothesis (Note that we would reject H_0 at the 2.5% significance level). The data strongly suggest that directed reading activity improves the DRP score.

The design of the DRP study is not ideal. Random assignment of students was not possible in a school environment, so existing third-grade classes were used. The effect of the reading programs is therefore confounded with any other differences between the two classes. The classes were chosen to be as similar as possible in variables such as the social and economic status of the students. Pretesting showed that the two classes were on the average quite similar in reading ability at the beginning of the experiment. To avoid the effect of two different teachers, the same teacher taught reading in both classes during the eight-week period of the experiment. We can therefore be somewhat confident that our two-sample procedure is detecting the effect of the treatment and not some other difference between the classes.

Note 14.1.

Why do we use hypothesis tests?

To conduct tests on parameters and to quantify the degree of certainty with probability. The smaller the p-value, the stronger the evidence against H_0 (provided assumptions are met).

Understanding Significance Levels

Suppose you have the following limited information:

- A test was statistically significant at the 5% level.

Can we also conclude this test is significant at the 1% level? (i.e., is the p-value < 0.01?) We **cannot** make this conclusion for certain. That is,

$$p\text{-value} < 0.05 \quad \not\Rightarrow \quad p\text{-value} < 0.01.$$

- A test was statistically significant at the 1% level.

Can we also conclude this test is significant at the 5% level? (i.e., is the p-value < 0.05?) Yes! Because

$$p\text{-value} < 0.01 < 0.05.$$

14.2 The Fold Rule

A rule which can be used to quickly determine whether population variances are equal or unequal using sample variances.
Note: only a rule

If

$$\frac{\max(s_1, s_2)}{\min(s_1, s_2)} < \sqrt{2} \quad \text{then we can consider } \sigma_1^2 = \sigma_2^2$$

and

$$\frac{\max(s_1^2, s_2^2)}{\min(s_1^2, s_2^2)} < 2 \quad \text{then we can consider } \sigma_1^2 = \sigma_2^2$$

Note: rather crude technique.
Hypothesis tests for equality of two variances do exist:

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad \text{vs} \quad H_a : \sigma_1^2 \neq \sigma_2^2$$

These tests involve a χ^2 test statistic.

14.3 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 p aired, 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 ma y 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}	$x_{d2} = x_{32} - x_{31}$
.....			
n	x_{n1}	x_{n2}	$x_{dn} = x_{n2} - x_{n1}$

Figure 14.8: 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 μ_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 14.9: 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 α-level
If p-value is less than α-level, then we reject the null hypothesis (H_0) and accept the alternative hypothesis (H_a). Otherwise, If p-value is greater than α-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 μ_d, but we still need more. We also can know which group has larger means from the result (μ_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 $M_2 \neq M_1$.

Example 14.4.
In an effort to determine whether a new type of fertilizer is more effective than the type currently in use, researchers took 12 two-acre plots of land scattered throughout the county. Each plot was divided into two equal-size sub plots, one of which was treated with the new fertilizer. Wheat was planted, and the crop yields were measured.

Plot	1	2	3	4	5	6	7	8	9	10	11	12
Current	56	45	68	72	61	69	57	55	60	72	75	66
New	60	49	66	73	59	67	61	60	58	75	72	68

Figure 14.10: Data of example 14.1

Can we conclude at the 5% significance level that the new fertilizer is more effective than the current one?
Solution:
You can verify that the mean and standard deviation of the twelve difference measurements are $\bar{d} = new - current = 1$ and $s_d = 3.0151$.
Step 1: State Hypothesis
$H_0 : \mu_d = 0$ and $H_a : \mu_d > 0$
Step 2: Find test statistics
$t* = \frac{\bar{d}-0}{s_d/\sqrt{n}} = \frac{1}{3.0151/\sqrt{12}} = 1.1489$
Step 3: Compute p-value
Using t-distribution table with df $= 11$, then $0.10 < p - value < 0.15$.

Step 4: Conclusion
Since p-value $> \alpha = 0.05$, we can't reject H_0. There is not enough evidence to infer that the new fertilizer is better.

14.4 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 14.11: 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 α-level
If p-value is less than α-level, then we reject the null hypothesis (H_0) and accept the alternative hypothesis (H_a). Otherwise, If p-value is greater than α-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 group s 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 14.5.
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 y ear, 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 y our conclusion. (Use $\alpha = 0.01$)
Solution:
Step 1: State Hypothesis
$H_0 : p_1 = p_2$ and $H_a : p_1 < p_2$
Step 2: Find test statistics
$\hat{p_1} = \frac{40}{244} = 0.1639$ and $\hat{p_2} = \frac{87}{245} = 0.3551$. Then $\hat{p} = \frac{40+87}{244+245} = 0.2597$.
Now, $z_* = \frac{\hat{p_1} - \hat{p_2}}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1}+\frac{1}{n_2})}} = -4.82$
Step 3: Compute p-value
P-value $= P(Z < z_*) = P(Z < -4.82) < 0.0003$
Step 4: Conclusion
Since p-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
```

14.5 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 σ_1^2 and σ_2^2 makes an inference about the ratio of σ_1^2/σ_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 14.12: 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.

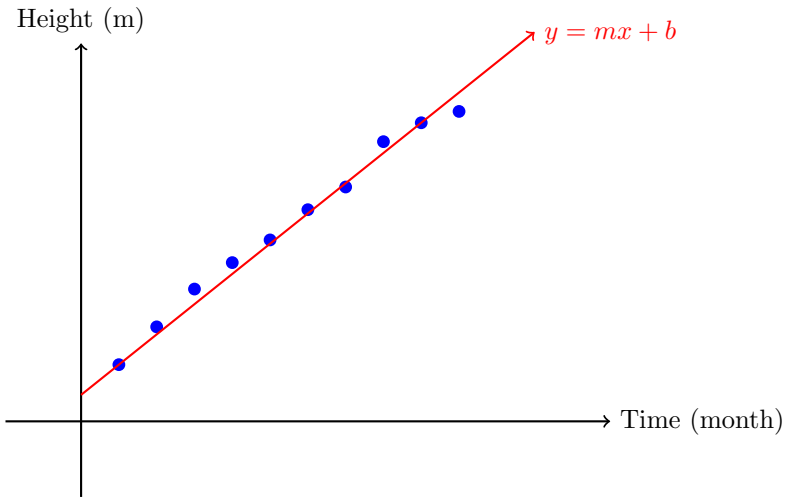


Figure 15.1: An illustration of simple linear regression. The blue points are measures of height monthly, and the red line is our SLR model. In this case m is the slope which tells you the rate of change, b is the intercept which may have a special meaning depending on the case.

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 ϵ-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)

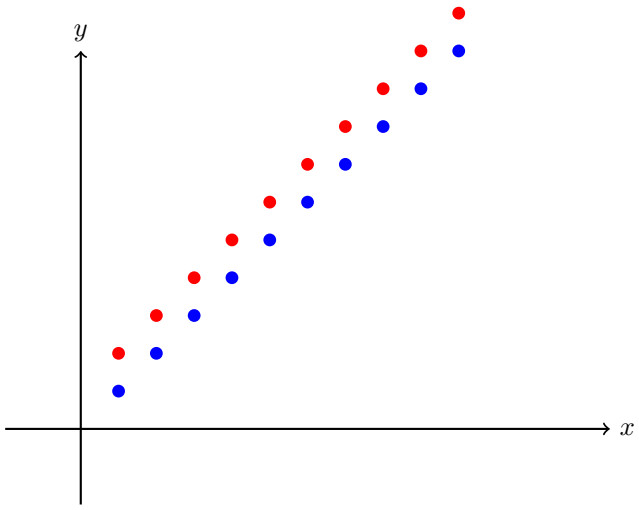


Figure 15.2: An illustration of strong positive correlation ($r_{xy} \approx +1$).

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

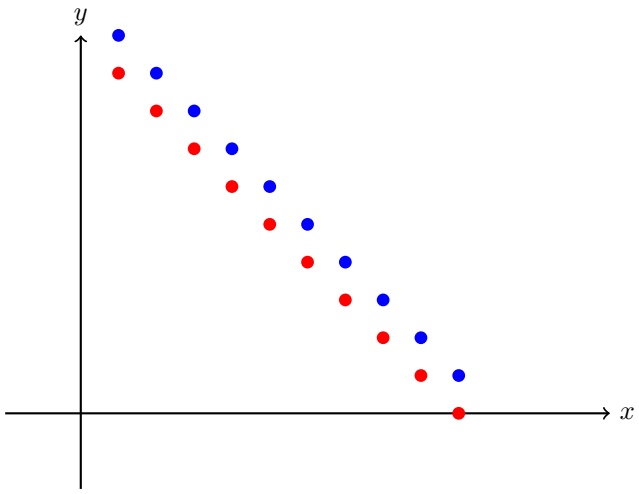


Figure 15.3: An illustration of strong negative correlation ($r_{xy} \approx -1$).

- 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)

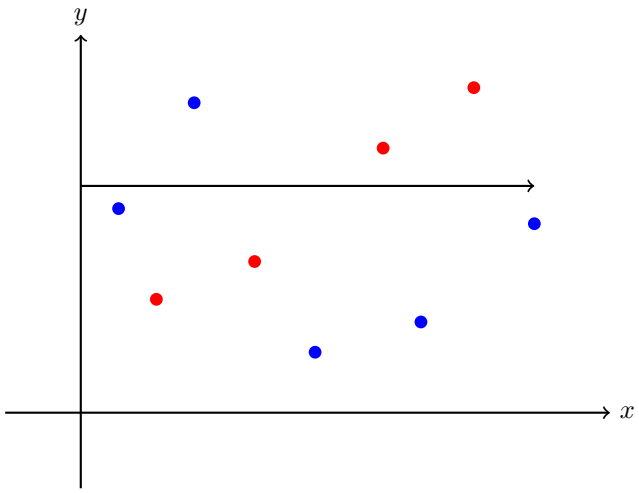


Figure 15.4: An illustration of no correlation ($r_{xy} \approx 0$).

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 interested in a linear model of the form: $y = \beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_p \cdot x_p + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$, x_i's $(i = 1, ..., p)$ are independent predictors, β_j's $(j = 0, ..., p)$ are coefficients, y is dependent variable.

Using sample data, we get estimates of this model of the form: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_1 + \cdots + \hat{\beta}_p \cdot x_p$. Now, $\hat{y}$ is predicated y, x_i's $(i = 1, ..., p)$ are independent predicators, $\hat{\beta}_j$'s $(j = 0, ..., p)$ are estimated coefficients. Moreover, $\hat{\beta}_0$ is intercept; $\hat{\beta}_1, ..., \hat{\beta}_p$ 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 $b_1 = r \cdot \frac{sy}{sx}$ 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 $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ 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:

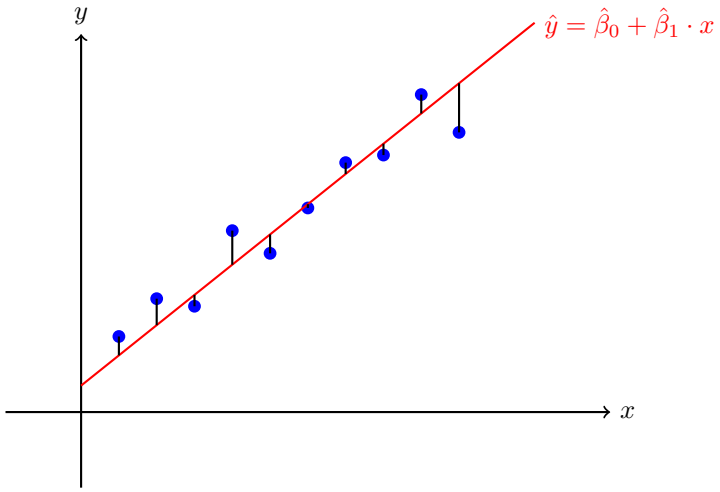


Figure 15.7: An illustration of a simple linear regression model. The red line is the fitted regression line, and the full black vertical lines represent the residuals (SSE). The data points deviate from the line to show errors clearly. Note that the sum of residuals is necessarily 0.

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 $\hat{\beta}_1$:

$$\begin{aligned}
\frac{\partial}{\partial\hat{\beta}_0}\sum_{i=1}^{n}e_i^2 &= \frac{\partial}{\partial\hat{\beta}_0}\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1\cdot x_i)^2=0\\
&=\sum_{i=1}^{n}2(\hat{\beta}_0+\hat{\beta}_1x_i-y_i)=0\\
&=\sum_{i=1}^{n}\hat{\beta}_0+\hat{\beta}_1\sum_{i=1}^{n}x_i-\sum_{i=1}^{n}y_i=0;\ (\text{consider: }\sum_{i=1}^{n}x_i=n\bar{x})\\
&=n\cdot\hat{\beta}_0+\hat{\beta}_1\cdot n\bar{x}-n\bar{y}=0\\
\text{Hence: }\hat{\beta}_0 &= \bar{y}-\hat{\beta}_1\cdot\bar{x}.\ (\text{Equ 1})\\
\frac{\partial}{\partial\hat{\beta}_1}\sum_{i=1}^{n}e_i^2 &= \frac{\partial}{\partial\hat{\beta}_1}\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1\cdot x_i)^2=0\\
&=\sum_{i=1}^{n}2(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)(-x_i)=0\\
&=\sum_{i=1}^{n}2(\hat{\beta}_0x_i+\hat{\beta}_1x_i^2-y_ix_i)=0\\
&=\hat{\beta}_0\sum_{i=1}^{n}x_i+\hat{\beta}_1\sum_{i=1}^{n}x_i^2-\sum_{i=1}^{n}x_iy_i=0\\
&=\hat{\beta}_1\sum_{i=1}^{n}x_i^2+\sum_{i=1}^{n}(\bar{y}-\hat{\beta}_1\bar{x})x_i-\sum_{i=1}^{n}x_iy_i=0;\ (\text{sub Equ 1 into this line})\\
&=\hat{\beta}_1\sum_{i=1}^{n}x_i^2+\bar{y}\sum_{i=1}^{n}x_i-\hat{\beta}_1\bar{x}\sum_{i=1}^{n}x_i-\sum_{i=1}^{n}x_iy_i=0\\
&=\hat{\beta}_1\sum_{i=1}^{n}x_i^2+n\bar{x}\bar{y}-n\hat{\beta}_1(\bar{x})^2-\sum_{i=1}^{n}x_iy_i=0\\
&=\hat{\beta}_1[\sum_{i=1}^{n}x_i^2-n(\bar{x})^2]+n\bar{x}\bar{y}-\sum_{i=1}^{n}x_iy_i=0\\
\text{Hence: }\hat{\beta}_1 &= \frac{\sum_{i=1}^{n}x_iy_i-n\bar{x}\bar{y}}{\sum_{i=1}^{n}x_i^2-n(\bar{x})^2}=\frac{S_{xy}}{S_{xx}}.
\end{aligned}$$

□

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 ϵ 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, β_0 y-intercept, β_1 = slope of the line and ϵ = error variable.

Month	Advertising Expenditure x (\$ hundreds)	Sales Revenue y (\$ thousands)
1	1	1
2	2	1
3	3	2
4	4	2
5	5	4

Figure 15.8: Data of example 15.3

a) Obtain the least squares estimates of β_0 and β_1, and state the estimated regression function.
b) Plot the estimated regression function and the data.
Solution:
a) $\bar{x} = 3$, $\bar{y} = 2$, $S_{xx} = 10$, $S_{xy} = 7$
Then, the slope of the least squares line is $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = 0.7$ and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = -0.1$. Thus, the least squares line is $\hat{y} = -0.1 + 0.7x$.
b) R-code

```
plot(x, y, main="Scatterplot: Simple Linear Regression",
xlab="x", ylab="y", pch=19,col="blue");

abline(coef(linear.reg), col="red",lty=2);
```

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 $\bar{y}$.

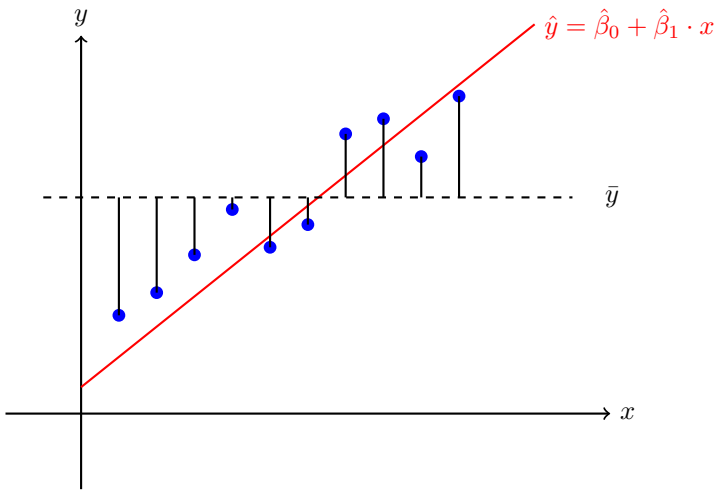


Figure 15.9: An illustration of Total Sum of Squares (SST). The blue points represent height measurements over time, the dashed line is the mean height $\bar{y}$, and the solid black vertical lines represent the squared deviations from the mean (SST components).

Definition 15.4 (SST (Toal sum of squares))**.** *For any simple linear regression model, SST (Toal 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 as a linear regression. A small SSE indicates a tight fit of the model to the data.

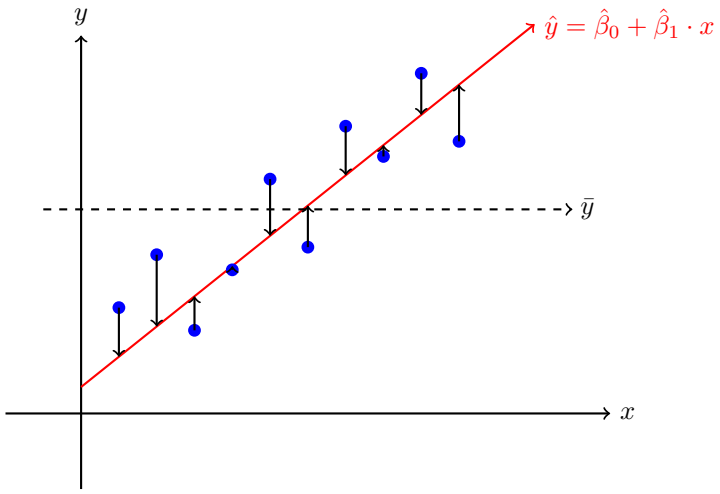


Figure 15.10: A simple linear regression illustration with residuals shown. Blue points are observed data, red line is the model, dashed line is the average of y, and black lines represent residuals $y_i - \hat{y}_i$.

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 ($\bar{y}$).

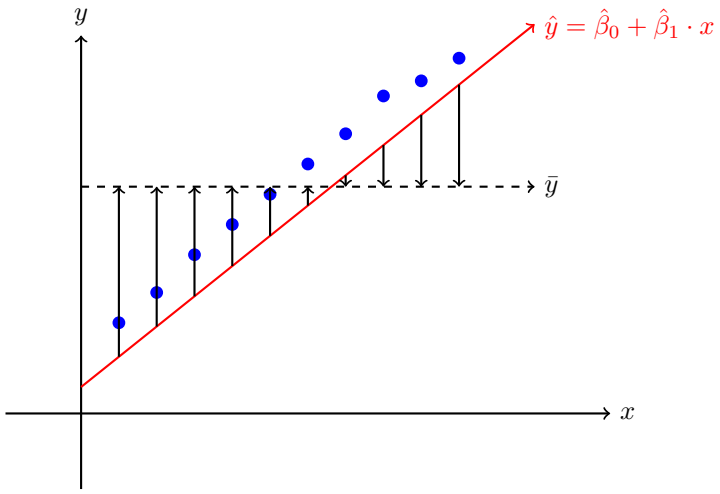


Figure 15.11: An illustration of simple linear regression. The blue points represent monthly height measurements with added variability. The red line is the fitted simple linear regression model. The dashed line shows the mean of the dependent variable, $\bar{y}$, and the solid black lines illustrate 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.

Summery

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.

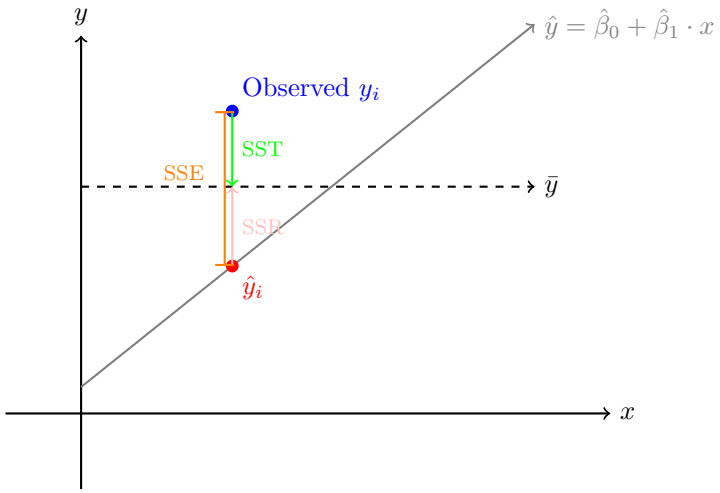


Figure 15.12: Visual representation of regression variability. **SST** (green) is total variability from the mean, **SSR** (pink) is explained variability, and **SSE** (orange curly brace) is the residual (unexplained) variability between the observed value y_i and the prediction $\hat{y}_i$.

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: $0 \leq r^2 \leq 1$.

Chapter 16

Inference for Simple Linear Regression

16.1 Inference on Regression

In previous chapters, we focused on estimating regression parameters and interpreting the fitted line. In this chapter, we take a step further by conducting formal inference on the slope and intercept of a simple linear regression model. We examine the distribution of errors, assess variability, and introduce the idea of using hypothesis tests and confidence intervals to evaluate whether the linear relationship observed in the data is statistically significant.

We begin by introducing the regression model and exploring the assumptions necessary to perform inference on the coefficients, particularly the slope.

$$Y = \beta_0 + \beta_1 X + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

Can perform inference on β_0 and β_1, however we are usually more interested in β_1.

What does the error term $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ mean?
At each value of X, the errors are distributed normally with a mean of zero and a constant variance.

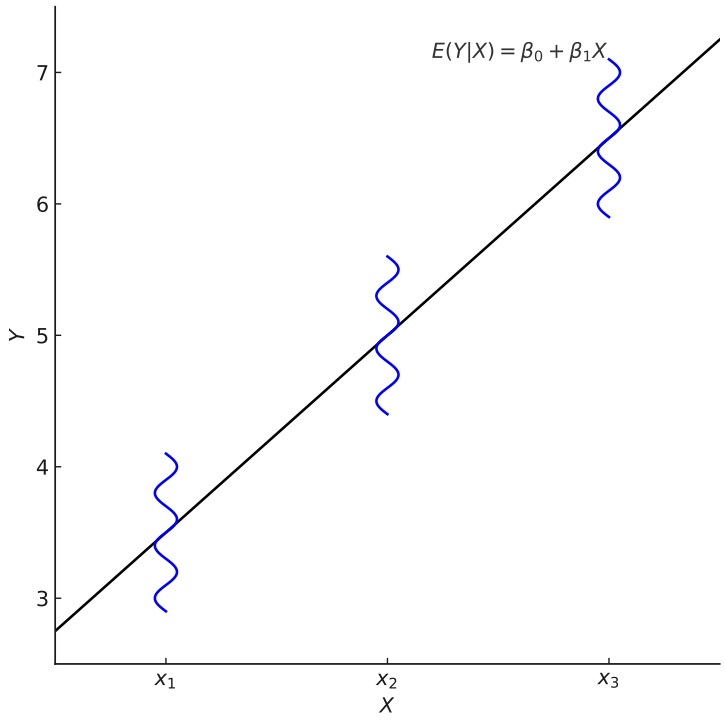


Figure 16.1: Regression line with normal errors at each X

Can verify with residual plots (assumptions).
We estimate σ^2 with a value we call S^2 and use S^2 for inference.

Estimating Variance in Linear Regression

$$Y = \beta_0 + \beta_1 X + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

Estimate σ^2 with S^2:

$$S^2 = \frac{\sum_{i=1}^n e_i^2}{n-2} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2} = \frac{SSE}{n-2}$$

notice similarity

$$S_x^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} \quad \text{(sample variance, } \bar{x} \text{ estimated)}$$

$$S = +\sqrt{S^2} \quad \text{(estimate of standard deviation)}$$

In calculating S^2, why do we divide by $n-2$?
Since we estimate 2 unknown parameters in the model (both β_0 and β_1), which are used in the calculation of S^2.

Equation of the Least-Squares Regression Line

We have data on an explanatory variable x and a response variable y for n individuals. From the data, calculate the means $\bar{x}$ and $\bar{y}$ and the standard deviations S_x and S_y of the two variables, and their correlation r. The least-squares regression line is the line:

$$\hat{y} = b_0 + b_1 x$$

with *slope*

$$b_1 = r\frac{S_y}{S_x}$$

and *intercept*

$$b_0 = \bar{y} - b_1\bar{x}$$

Definition 16.1 (Least-Squares Regression Line)**.** *The* ***least-squares regression line*** *of y on x is the line that makes the sum of the squares of the vertical distances of the data points from the line as small as possible.*

Example 16.1.
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,

$$y = \beta_0 + \beta_1 x + \varepsilon$$

where

$$\begin{aligned} y &= \text{dependent variable} \\ x &= \text{independent variable} \\ \beta_0 &= y\text{-intercept} \\ \beta_1 &= \text{slope of the line} \\ \varepsilon &= \text{error variable} \end{aligned}$$

Month	Advertising Expenditure x ($ hundreds)	Sales Revenue y ($ thousands)
1	1	1
2	2	1
3	3	2
4	4	2
5	5	4

The question is this: How can we best use the information in the sample of five observations in our table to estimate the unknown y-intercept β_0 and slope β_1?

We are given:

$$\bar{x} = 3, \quad \bar{y} = 2, \quad S_x = 1.5811, \quad S_y = 1.2247, \quad S_{xy} = 1.75$$

Then, the slope of the least squares line is

$$b_1 = r\frac{S_y}{S_x} = (0.9037)\left(\frac{1.2247}{1.5811}\right) = 0.7$$

and

$$b_0 = \bar{y} - b_1\bar{x} = 2 - (0.7)(3) = -0.1$$

The least squares line is thus:

$$\hat{y} = -0.1 + 0.7x$$

The Regression Model

We have n observations on an explanatory variable x and a response variable y. Our goal is to study or predict the behavior of y for given values of x.

- For any fixed value of x, the response y varies according to a Normal distribution. Repeated measures y are independent of each other.
- The mean response μ_y has a straight-line relationship with x: $\mu_y = \beta_0 + \beta_1 x$. The slope β_1 and intercept β_0 are **unknown** parameters.
- The standard deviation of y (call it σ) is the same for all values of x. The value of σ is **unknown**. The regression model has three parameters, β_0, β_1, and σ.

Thus, if

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

is the predicted value of the ith y value, then the deviation of the observed value y_i from $\hat{y}_i$ is the difference $y_i - \hat{y}_i$ and the sum of squares of deviations to be minimized is

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2.$$

The quantity *SSE* is also called the **sum of squares for error**.

$$\begin{aligned} \text{Fitted Value:} \quad & \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \\ \text{Residual:} \quad & \hat{\varepsilon}_i = y_i - \hat{y}_i \end{aligned}$$

The **regression standard error** is

$$s = \sqrt{\frac{1}{n-2}\sum \text{residual}^2} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} = \sqrt{\frac{SSE}{n-2}}$$

Use s to estimate the **unknown** σ in the regression model.

The standard error of $\hat{\beta}_1$ is the standard deviation of the sampling distribution of $\hat{\beta}_1$ (estimate of slope β_1):

$$SE(\hat{\beta}_1) = \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = \frac{s}{\sqrt{(n-1)s_x^2}}$$

Confidence Interval for the Slope

$$\hat{\beta}_1 \pm t_{(n-2,\,\alpha/2)} \cdot SE(\hat{\beta}_1) \quad = \quad \hat{\beta}_1 \pm t_{(n-2,\,\alpha/2)} \cdot \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Example 16.2.
[continued] Revisit the example on advertising and sales and construct a 95% confidence interval on the slope. Provide an interpretation of the CI.
From earlier:

$$\hat{y} = -0.1 + 0.7x$$

x	y	$\hat{y}$	$y - \hat{y}$	$(y - \hat{y})^2$	$(x - \bar{x})^2$
1	1	0.6	0.4	0.16	4
2	1	1.3	-0.3	0.09	1
3	2	2.0	0.0	0.00	0
4	2	2.7	-0.7	0.49	1
5	4	3.4	0.6	0.36	4

We are given:

$$\sum x_i = 15, \quad \bar{x} = \frac{15}{5} = 3, \quad SSE = 1.10, \quad \sum(x_i - \bar{x})^2 = 10$$

Step 1: Estimate variance and standard deviation

$$s^2 = \frac{SSE}{n-2} = \frac{1.10}{5-2} = 0.3667 \quad \Rightarrow \quad s = \sqrt{0.3667} = 0.6055$$

Step 2: Compute standard error of $\hat{\beta}_1$

$$SE(\hat{\beta}_1) = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}} = \frac{0.6055}{\sqrt{10}} = 0.1914$$

Step 3: Determine critical t-value

$$n - 2 = 3, \quad \alpha = 0.05, \quad \alpha/2 = 0.025 \Rightarrow \quad t_{(3,0.025)} = 3.182$$

Step 4: Construct CI for the slope

$$\hat{\beta}_1 \pm t_{(n-2,\alpha/2)} \cdot SE(\hat{\beta}_1) = 0.7 \pm 3.182 \cdot 0.1914 = 0.7 \pm 0.6092$$

$$\Rightarrow \text{CI: } (0.0908,\ 1.3092)$$

Interpretation: We are 95% confident the slope (β_1) for this model lies between 0.0908 and 1.3092.

Interpreting Confidence Intervals for β_1

Suppose CI: $(-, -)$
Suggests β_1 has a **negative** sign.
Suggests negative correlation, potentially good model.

Suppose CI: $(+, +)$
Suggests β_1 has a **positive** sign.
Suggests positive correlation, potentially good model.

Suppose CI: $(-, +)$
$\beta_1 = 0$ is plausible.
Suggests **no linear relationship** between x and y.

In cases where the CI does not contain zero, we can infer the sign of the slope (just not the steepness).

The **regression standard error** is

$$s = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{1.1}{3}} = 0.6055$$

Use s to estimate the **unknown** σ in the regression model.

A level C confidence interval for the slope β_1 of the true regression line is

$$b_1 \pm t^* SE_{b_1}$$

In this formula, the standard error of the least-squares slope b is

$$SE_{b_1} = \frac{s}{\sqrt{\sum(x_i - \bar{x})^2}} = \frac{s}{\sqrt{(n-1)S_x^2}}$$

and t^* is the critical value for the $t(n-2)$ density curve with area C between $-t^*$ and t^*.

Hypothesis Test on the Slope β_1

Hypotheses:

$$\begin{aligned} &H_0\colon \beta_1 = 0 \quad \text{vs.} \quad H_a\colon \beta_1 > 0 \\ &H_0\colon \beta_1 = 0 \quad \text{vs.} \quad H_a\colon \beta_1 < 0 \\ &H_0\colon \beta_1 = 0 \quad \text{vs.} \quad H_a\colon \beta_1 \neq 0 \quad \textit{(most common)} \end{aligned}$$

Test Statistic:

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{\frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}}$$

Reference distribution: t distribution with $n-2$ degrees of freedom.

Note: A test statistic always follows the form:

$$\text{test stat} = \frac{\text{statistic} - \text{hypothesized value}}{\text{SE(statistic)}}$$

Example 16.3.

[continued] For the advertising example, perform a two-sided hypothesis test on the slope.

Hypotheses:

$$H_0\colon \beta_1 = 0 \qquad H_a\colon \beta_1 \neq 0$$

Test Statistic:

$$t^* = \frac{\hat{\beta}_1 - 0}{\frac{s}{\sqrt{\sum(x_i - \bar{x})^2}}} = \frac{0.7 - 0}{0.6055/\sqrt{10}} = 3.6558$$

Reference Distribution: t distribution with $n-2=5-2=3$ degrees of freedom.

Decision Rule:
Using a two-tailed test:

$$\text{p-value} = 2 \cdot P(T_3 > 3.6558) < 0.01 \Rightarrow \text{p-value} < 0.05$$

Conclusion: Since p-value < 0.05, we reject H_0 and conclude $H_a : \beta_1 \neq 0$.
Interpretation: The slope should be included in the model. There is significant evidence of a linear relationship between advertising and sales revenue.

For the advertising-sales example, a 95% Confidence Interval for the slope β_1 is

$$0.7 \pm 3.182 \left(\frac{0.6055}{\sqrt{10}} \right)$$

$$0.7 \pm 0.6092$$

Thus, we estimate with 95% confidence that the interval from 0.0908 and 1.3092 includes the parameter β_1.

We can also test hypotheses about the slope β_1. The most common hypothesis is

$$H_0 : \beta_1 = 0.$$

A regression line with slope 0 is horizontal. That is, the mean of y does not change at all when x changes. So this H_0 says that there is no true linear relationship between x and y.

Testing the Hypothesis for β_1

To test the hypothesis $H_0 : \beta_1 = 0$, compute the t statistic

$$t = \frac{b_1}{SE_{b_1}}.$$

In terms of a random variable T having the $t(n-2)$ distribution, the P-value for a test of H_0 against:

$$\begin{aligned} H_a : \beta_1 \neq 0 &\quad \text{is } 2P(T > |t|). \\ H_a : \beta_1 > 0 &\quad \text{is } P(T > t). \\ H_a : \beta_1 < 0 &\quad \text{is } P(T < t). \end{aligned}$$

Example 16.4. ———————————————————

[continued]

$$\alpha = 0.05$$

1) $H_0 : \beta_1 = 0 \quad \text{vs} \quad H_a : \beta_1 > 0$

2) $t^* = \frac{b_1}{SE_{b_1}} = \frac{0.7}{0.1914} = 3.6572$

3) P-value $= P(T > t) = P(T > 3.6572) \quad$ d.f. $= n - 2 = 5 - 2 = 3$.
Using t-distribution table, $0.01 < P$ value < 0.025

4) Since P-value $< \alpha = 0.05$, we reject H_0.

Our example (different H_a)

$$\alpha = 0.05$$

1) $H_0 : \beta_1 = 0 \quad \text{vs} \quad H_a : \beta_1 \neq 0$

2) $t^* = \frac{b_1}{SE_{b_1}} = \frac{0.7}{0.1914} = 3.6572$

3) P-value $= 2P(T > |t|) = 2P(T > 3.6572) \quad$ d.f. $= n - 2 = 5 - 2 = 3$.
Using Table 3, $0.02 < P$ value < 0.05

4) Since P-value $< \alpha = 0.05$, we reject H_0.

R code

```
x = c(1, 2, 3, 4, 5);
y = c(1, 1, 2, 2, 4);
mod = lm(y~x);
summary(mod);
```

R Output

```
##
## Call:
## lm(formula = y~x)
##
## Residuals:
##       1        2        3        4        5
##  4.000e-01 -3.000e-01 -3.886e-16 -7.000e-01  6.000e-01
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.1000     0.6351  -0.157  0.8849
## x             0.7000     0.1915   3.656  0.0354 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6055 on 3 degrees of freedom
```

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = -0.10 + 0.70x$$
$$SE(\hat{\beta}_1) = 0.1915$$

By default, R conducts the following test for each coefficient:

$$H_0 : \beta_j = 0$$
$$H_a : \beta_j \neq 0 \quad \text{(two sided)}$$

Test statistic:

$$t^* = \frac{\hat{\beta}_j - 0}{SE(\hat{\beta}_j)}$$

For advertising and sales data:

$$H_0 : \beta_1 = 0$$
$$H_a : \beta_1 \neq 0$$

$$t^* = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} = \frac{0.70}{0.1915} = 3.656$$

Note 16.1.
Review

$$y = \beta_0 + \beta_1 x + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$
$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{s_{xy}}{s_{xx}}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

r: *coefficient of correlation (strength)*
r^2: *coefficient of determination (% variability)*

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} = \frac{SSE}{n-2}$$
$$s = \sqrt{s^2}$$
$$SE(\hat{\beta}_1) = \frac{s}{\sqrt{s_{xx}}}$$
$$CI : \hat{\beta}_1 \pm t_{n-2,\alpha/2} \cdot SE(\hat{\beta}_1)$$

Hypothesis test:

$$H_0 : \beta_1 = 0$$

$$\textit{Test stat:} \quad t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

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:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

$$\text{SST} = \text{SSE} + \text{SSR}$$

Total sum of squares = Sum of squares for error + Sum of squares for regression.

$$\begin{aligned} SSE &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n}(y_i - \bar{y})^2 - \hat{\beta}_1 \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \\ &= S_{YY} - \hat{\beta}_1 S_{XY} \end{aligned}$$

Notice that this provides an easier computational method of finding SSE.

R output (Additional example)

```
> summary(model);

Call:
lm(formula = camrys$Price ~ Odometer, data = camrys)

Residuals:
     Min       1Q   Median       3Q      Max
-0.68679 -0.27263  0.00521  0.23210  0.70071

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.248727   0.182093   94.72   <2e-16 ***
Odometer     -0.066861   0.004975  -13.44   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3265 on 98 degrees of freedom
Multiple R-squared:  0.6483,    Adjusted R-squared:  0.6447
F-statistic: 180.6 on 1 and 98 DF,  p-value: < 2.2e-16
```

16.2 ANOVA Table (ANalysis Of VAriance)

Analysis of Variance (ANOVA) is a statistical method used to assess whether variation in a response variable can be explained by predictor variables in a regression model. It summarizes sources of variation using sums of squares, degrees of freedom, and mean squares in a structured table format.

Source of Variation	**Sum of Squares**	**Degrees of Freedom**	**Mean Square**	**Computed F**
Regression	SSR	1	SSR	$\frac{SSR}{SSE/(n-2)}$
Error	SSE	$n-2$	$s^2 = \frac{SSE}{n-2}$	
Total	SST	$n-1$		

For the general multivariate regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon,$$
$$\varepsilon \sim N(0, \sigma^2)$$

with p predictors.

ANOVA can be used for testing:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$
$$H_a : \text{At least one } \beta_j \neq 0, \quad j = 1, \ldots, p$$

Test Statistic:

$$F = \frac{MSR}{MSE} = \frac{SSR/p}{SSE/(n-p-1)} \sim F(p, n-p-1)$$

Reference distribution: F with numerator df $= p$, denominator df $= n - p - 1$

Example 16.5.
[Interpreting ANOVA Table from R Output]
We fitted a simple linear regression model using:

```
x = c(1,2,3,4,5);
y = c(1,1,2,2,4);
mod = lm(y~x);
anova(mod);
```

The ANOVA output was:

```
## Analysis of Variance Table
##
## Response: y
##             Df  Sum Sq Mean Sq F value   Pr(>F)
## x            1     4.9    4.9000  13.364  0.03535 *
## Residuals    3     1.1    0.3667
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation:

- The regression model includes one predictor x, so the degrees of freedom for regression is 1.
- The sum of squares for regression is $SSR = 4.9$, and for residuals $SSE = 1.1$.
- Mean squares are calculated as:

$$MSR = \frac{SSR}{1} = 4.9, \quad MSE = \frac{SSE}{n-2} = \frac{1.1}{3} = 0.3667$$

- The F-statistic is:

$$F = \frac{MSR}{MSE} = \frac{4.9}{0.3667} \approx 13.364$$

- The p-value is ≈ 0.03535, indicating that the predictor is significant at the 5% level.

Conclusion: Since the p-value is less than 0.05, we reject H_0 and conclude that x has a statistically significant linear relationship with y.

Example 16.6.
[Apartments Around UTM] We consider data on apartments near UTM, with price (in thousands of dollars), area (in 100 square feet), and number of beds and baths.

Price (×1000)	**Area** (×100 sq ft)	**Beds**	**Baths**
620	11.0	2	2
590	6.5	2	1
620	10.0	2	2
700	8.4	2	2
680	8.0	2	2
500	5.7	1	1
760	12.0	2	2
800	14.0	3	1
660	7.3	2	1

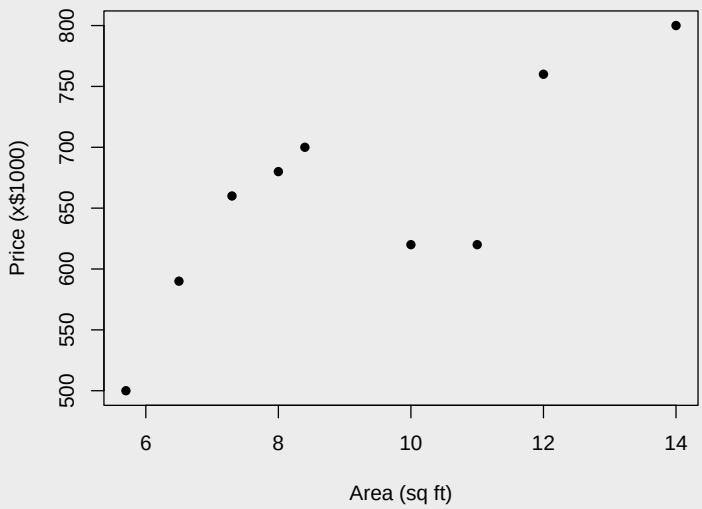


Figure 16.2: Plot of Price vs Area for Apartments near UTM

Price	**Area**	$(x - \bar{x})$	$(y - \bar{y})$	$(x - \bar{x})(y - \bar{y})$	$(x - \bar{x})^2$
620	11.0	1.8	-38.9	-70.02	3.24
590	6.5	-2.7	-68.9	186.03	7.29
620	10.0	0.8	-38.9	-31.12	0.64
700	8.4	-0.8	41.1	-32.88	0.64
680	8.0	-1.2	21.1	-25.32	1.44
500	5.7	-3.5	-158.9	556.15	12.25
760	12.0	2.8	101.1	283.08	7.84
800	14.0	4.8	141.1	677.28	23.04
660	7.3	-1.9	1.1	-2.09	3.61
Sum	82.9			$\sum(x - \bar{x})(y - \bar{y}) = 1541.11$	$\sum(x - \bar{x})^2 = 60.00$

Table 16.1: Deviation table for computing $\hat{\beta}_1$ and $\hat{\beta}_0$

The sample means are:

$$\bar{y} = \frac{\sum y}{n} = \frac{5930}{9} = 658.89, \qquad \bar{x} = \frac{\sum x}{n} = \frac{82.9}{9} = 9.21$$

Finding the Regression Coefficients

To compute the least squares regression line, we calculate the slope and intercept using the formulas:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{1541.11}{60} = 25.69$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 658.89 - (25.69)(9.21) = 422.28$$

Equation of the Regression Line

Using the values above, we write the estimated regression equation as:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 422.28 + 25.69x$$

This equation gives the predicted apartment price (in \$1000s) based on area (in 100 sq ft).

Interpretation of Coefficients

The slope $\hat{\beta}_1 = 25.69$ means that for every additional 100 sq ft in area, we expect the apartment price to increase by approximately \$25,690 on average.
The intercept $\hat{\beta}_0 = 422.28$ suggests the predicted price when the area is zero. While this has no practical interpretation in this context, it is a necessary component of the regression model.

Interpolation and Extrapolation

To estimate the price of an apartment with an area of 800 sq ft (i.e., $x = 8$), we compute:

$$\hat{y} = 422.28 + 25.69(8) = 627.8 \quad (\$1000)$$

Since 8 is within the range of observed values, this is an example of **interpolation.**
For an apartment with 2,500 sq ft ($x = 25$):

$$\hat{y} = 422.28 + 25.69(25) = 1064.53 \quad (\$1000)$$

This is an example of **extrapolation**, and such predictions should be treated with caution since they lie outside the data range.
We can create a simple linear regression model in R using the `lm` command:

R code

```
lm(y ~ x, data = data_source)
```

The data is available in the `apt_around_utm.csv` file.

R code

```
apt = read.csv(file.choose())
# apt = read.csv("~/PATH_TO_FILE/apt_around_utm.csv")

apt_model = lm(price ~ area, data = apt)
```

R output

```
> apt_model

Call:
lm(formula = price ~ area, data = apt)

Coefficients:
(Intercept)       area
     422.26       25.69
```

We can compute the residuals and the sum of squared errors (SSE) using the table below:

y	x	$\hat{y}$	$y - \hat{y}$	$(y - \hat{y})^2$
620	11.0	704.85	-84.85	7198.73
590	6.5	589.24	0.76	0.58
620	10.0	679.16	-59.16	3499.36
700	8.4	638.05	61.95	3837.62
680	8.0	627.78	52.22	2727.4
500	5.7	568.69	-68.69	4718.13
760	12.0	730.54	29.46	868.17
800	14.0	781.92	18.08	327.06
660	7.3	609.79	50.21	2520.79

Recall our fitted regression model:

$$\hat{y} = 25.69x + 422.26$$

$$\text{SSE} = \sum (y_i - \hat{y}_i)^2 = 25{,}697.83$$

We now conduct a hypothesis test on the slope β_1 at the 5% significance level.

Step 1: Hypotheses

$$H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_a : \beta_1 \neq 0$$

Step 2: Test statistic

$$s^2 = \frac{SSE}{n-2} = \frac{25{,}697.83}{7} = 3670.26$$

$$s = \sqrt{3670.26} = 60.58$$

$$SE(\hat{\beta}_1) = \frac{s}{\sqrt{S_{xx}}} = \frac{60.58}{\sqrt{60}} = 7.82$$

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} = \frac{25.69}{7.82} = 3.284$$

Step 3: Conclusion Using t-distribution with 7 degrees of freedom:

$$0.005 < p\text{-value} < 0.01$$

Since p-value < 0.05, we reject H_0 and conclude that there is sufficient evidence that $\beta_1 \neq 0$.This suggests there is a statistically significant relationship between price and area for apartments near UTM.

Final Check: Total Sum of Squares

y	x	$\hat{y}$	$(y-\hat{y})^2$	$(y-\bar{y})^2$	$(\hat{y}-\bar{y})^2$
620	11.0	704.85	7198.73	2112.00	1512.35
590	6.5	589.24	0.58	4850.88	4745.68
620	10.0	679.16	3499.36	410.73	1512.35
700	8.4	638.05	3837.62	434.20	1690.12
680	8.0	627.78	2727.4	968.04	445.68
500	5.7	568.69	4718.13	8136.08	2524.68
760	12.0	730.54	868.17	5133.21	10223.46
800	14.0	781.92	327.06	1513.47	19912.35
660	7.3	609.79	2520.79	2410.45	1.23

$$SSE + SSR = SST = 25{,}697.83 + 39{,}591.06 = 65{,}288.90$$

This confirms the ANOVA identity: Total = Explained + Residual

We previously estimated the model:

$$\hat{y} = 25.69x + 422.26$$

Coefficient of Determination and Correlation

$$r^2 = \frac{SSR}{SST} = \frac{39591.06}{65288.90} = 0.6063 = 60.63\%$$

Interpretation: Approximately 60.63% of the variability in price is explained by the regression model.

$$r = \pm\sqrt{r^2} = \pm\sqrt{0.6063} = \pm 0.779$$

Since $\hat{\beta}_1 > 0$, we choose the positive root:

$$r = 0.779$$

Interpretation: There is a strong positive correlation between apartment area and price.

R code

```
model = lm(y~x, data = data\_source)
summary(model)
```

R code

```
apt\_model = lm(price ~ area, data = apt)
summary(apt\_model)
```

R code

```
Call:
lm(formula = price ~ area, data = apt)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  422.256   74.834     5.643  0.00078 ***
area          25.690    7.823     3.284  0.01341 *

Residual standard error: 60.59 on 7 degrees of freedom
Multiple R-squared: 0.6064,    Adjusted R-squared: 0.5502
F-statistic: 10.78 on 1 and 7 DF,  p-value: 0.01341
```

Two-sided Test for Slope Coefficient

By default, R performs a two-sided test:

$$H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_a : \beta_1 \neq 0$$

Test statistic:

$$t^* = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} = \frac{25.690 - 0}{7.823} = 3.284$$

$$t^* \sim t_{(n-2)} \quad \text{with } df = 9 - 2 = 7$$

p-value is the total shaded area in both tails. From R output:

$$\text{p-value} = 0.01341$$

Definition 16.2 (Interpolation and Extrapolation)**.**

- ***Interpolation*** *is calculating predicted values of y using our linear model while working within the range of x in which data was available to construct our model.*
- ***Extrapolation*** *is calculating predicted values of y using our linear model outside the range of x used to obtain the linear model.*
- *Interpolation is usually safe if we have a good linear model.*
- *Extrapolation must be performed carefully since extrapolations that are done without any foresight can be very inaccurate.*

16.3 Residual Plots

Residual plots are used to verify assumptions related to the error terms in a regression model.

$$Y = \beta_0 + \beta_1 X + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

The assumption $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ implies:

- Mean of errors is 0
- Constant variance of errors (homoscedasticity)

We plot the residuals:

$$e_i = y_i - \hat{y}_i$$

against the fitted values $\hat{y}_i$ to assess these assumptions.

What to Look for in a Good Residual Plot

If the assumption $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ is satisfied, the residual plot should have the following features:

1. **Random scattering:** No obvious pattern in residuals.
   - A pattern (e.g., curve) may indicate a non-linear relationship.
   - Random scattering also suggests independence of errors.
2. **Constant variance:** Residuals should fall within a horizontal band, roughly half above and half below zero.
   - Suggests constant variance (homoscedasticity).
3. **No influential points or clustering:** The plot should not show isolated influential observations or clustering.

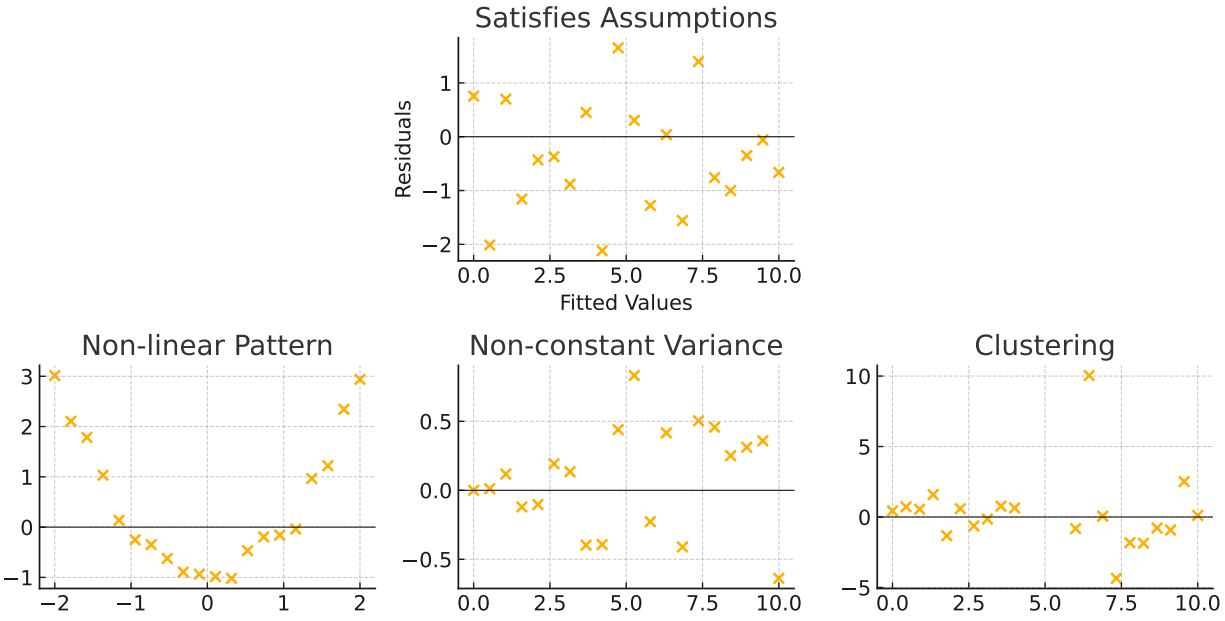


Figure 16.3: Residual Plots - Good and Bad Examples

Assumptions of Simple Linear Regression (SLR)

The model is: $Y = \beta_0 + \beta_1 X + \varepsilon, \quad$ where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$

- The relationship between X and Y is linear.
- Residuals:
  - are independent
  - have constant variance
  - are normally distributed

These assumptions can be verified using residual plots.

Inference on **one** sample (z or t)

↓

Inference on **two** samples (z or t)

↓

Inference on **3 or more** groups (F-tests with ANOVA)

Index