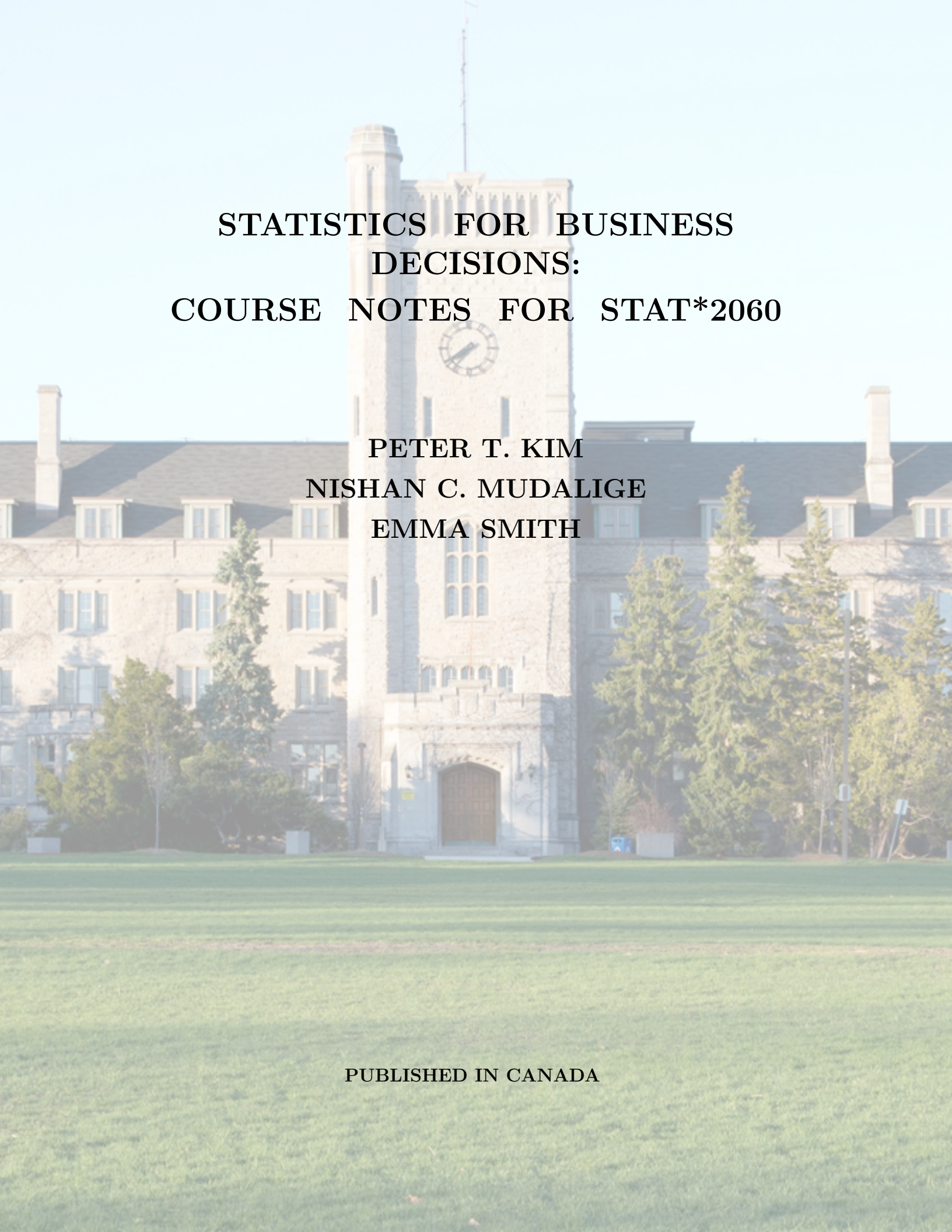


STATISTICS FOR BUSINESS DECISIONS: COURSE NOTES FOR STAT*2060

PETER T. KIM
NISHAN C. MUDALIGE
EMMA SMITH

PUBLISHED IN CANADA

Statistics for Business Decisions: Course Notes for STAT*2060

Peter T. Kim
Department of Mathematics and Statistics
University of Guelph

Nishan C. Mudalige
Department of Mathematics and Statistics
University of Guelph

Emma Smith
Department of Mathematics and Statistics
University of Guelph

First edition: August 2015

Kim, P.; Mudalige, N.; Smith, E.
University of Guelph Bookstore Press

University of Guelph Bookstore Press,
Guelph, Ontario, Canada

Contents

0 Overview **1**

1 Introduction **2**
- 1.1 Basics . . . 2
- 1.2 Types of Data . . . 4
- 1.3 Types of Studies . . . 6
- 1.4 Introduction to Inferential Statistics . . . 7
- 1.5 Issues with Sample Data . . . 9

2 Descriptive Statistics **12**
- 2.1 Representing Data . . . 12
- 2.2 Numerical Measures of Location . . . 13
  - 2.2.1 Mean . . . 13
  - 2.2.2 Median . . . 13
  - 2.2.3 Mode . . . 14
  - 2.2.4 Variance . . . 14
  - 2.2.5 Standard Deviation . . . 15
  - 2.2.6 Percentiles . . . 15
  - 2.2.7 Quartiles . . . 15
- 2.3 Graphical Techniques . . . 19
  - 2.3.1 Histograms . . . 21
  - 2.3.2 Boxplots . . . 24
  - 2.3.3 Stem and Leaf Plots . . . 28

3 Probability **30**
- 3.1 Introduction . . . 30
- 3.2 Terminology and Definitions . . . 30
- 3.3 Notation . . . 32
- 3.4 Properties . . . 33
- 3.5 Venn Diagrams . . . 36
- 3.6 Conditional Probability and Independence . . . 39
  - 3.6.1 Conditional Probability . . . 39
  - 3.6.2 Independence . . . 39
- 3.7 Bayes' Theorem . . . 40
- 3.8 Counting Techniques . . . 44
  - 3.8.1 Factorial . . . 44

3.8.2 Combinations . . . 44
3.9 Random Variables . . . 46
3.9.1 Discrete Random Variables . . . 46
3.9.1.1 Probability Mass Function . . . 47
3.9.1.2 Mean, Variance and Standard Deviation of a Discrete Random Variable . . . 47
3.9.2 Continuous Random Variable . . . 51
3.9.2.1 Probability Density Function . . . 51
3.9.2.2 Mean, Variance and Standard deviation of a Continuous Random Variable . . . 52

4 Distributions of Random Variables **54**
4.1 Distributions of Discrete Random Variables . . . 54
4.1.1 Bernoulli Distribution . . . 54
4.1.2 Binomial Distribution . . . 55
4.2 Distributions of Continuous Random Variables . . . 57
4.2.1 Continuous Uniform Distribution . . . 57
4.2.2 Normal Distribution . . . 58
4.2.2.1 The Standard Normal Distribution . . . 59
4.2.3 $t-$Distribution . . . 66

5 Foundations of Inference **70**
5.1 Sampling Distribution . . . 70
5.2 Central Limit Theorem . . . 70
5.3 Introduction to Inference . . . 72

6 Confidence Intervals **74**
6.1 Introduction . . . 74
6.2 Interpretation . . . 75
6.3 One Sample Confidence Intervals . . . 76
6.3.1 On the Mean . . . 76
6.3.1.1 When σ is Known . . . 76
6.3.1.2 When σ is Unknown . . . 78
6.3.2 On a Proportion . . . 79
6.3.3 Assumptions . . . 80
6.4 Two Sample Confidence Intervals . . . 81
6.4.1 On a Difference of Two Means . . . 81
6.4.1.1 When σ_1 and σ_2 are Known . . . 82
6.4.1.2 When σ_1 and σ_2 are Unknown . . . 84
6.4.1.2.1 When $\sigma_1 \neq \sigma_2$ . . . 84
6.4.1.2.2 When $\sigma_1 = \sigma_2$ . . . 87
6.4.1.3 Assumptions . . . 88
6.4.2 On a Difference of Two Proportions . . . 89
6.4.2.1 Assumptions . . . 91
6.5 On Paired Data . . . 91
6.5.1 Assumptions . . . 94

7 Hypothesis Tests **95**
7.1 One Sample Hypothesis Tests . . . 101
7.1.1 On the Mean . . . 101
7.1.1.1 When σ is Known . . . 101
7.1.1.2 When σ is Not Known . . . 102
7.1.2 On a Proportion . . . 104
7.1.3 Assumptions . . . 105
7.2 Two Sample Hypothesis Tests . . . 106
7.2.1 On a Difference of Two Means . . . 106
7.2.1.1 When σ_1 and σ_2 are Known . . . 106
7.2.1.2 When σ_1 and σ_2 are Not Known . . . 107
7.2.1.2.1 When $\sigma_1 \neq \sigma_2$ . . . 107
7.2.1.2.2 When $\sigma_1 = \sigma_2$ . . . 109
7.2.1.3 Assumptions . . . 110
7.2.2 On a Difference of Two Proportions . . . 111
7.2.2.1 Assumptions . . . 113
7.3 On Paired Data . . . 113
7.3.1 Assumptions . . . 114
7.4 Decision Errors . . . 115
7.5 Relationship Between Hypothesis Tests and Confidence Intervals . . . 116
7.6 Statistical Significance vs. Practical Significance . . . 117

8 Simple Linear Regression **119**
8.1 Introduction and Notation . . . 119
8.2 The Linear Regression Model . . . 123
8.2.1 Interpretation of the Slope and Intercept . . . 128
8.2.2 Interpolation and Extrapolation . . . 129
8.3 Residuals . . . 132
8.3.1 Residual Plots . . . 134
8.4 Model Assumptions . . . 139
8.5 Measuring Variability with a Regression Model . . . 140
8.5.1 Coefficient of Determination and Coefficient of Correlation . . . 142
8.6 Inference Procedures on the Slope . . . 144
8.6.1 Confidence Intervals on the Slope . . . 145
8.6.2 Hypothesis Tests on the Slope . . . 146

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

Introduction

1.1 Basics

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.*

When analyzing a set of data, our goal is to describe characteristics of a group of objects and make inferences about the group. Operationally this involves the collection of data through various means and the analysis and interpretation of data.

Statistics can be broken down into two broad categories: descriptive statistics and inferential statistics.

Definition 1.2 (Descriptive Statistics)**.** *Descriptive statistics are numerical and graphical methods used to analyze, interpret, and represent data.*

Definition 1.3 (Inferential Statistics)**.** *Inferential statistics use information from a sample to make generalizations about a larger population.*

Only a very cursory introduction to inferential statistics will be given in this chapter. In subsequent chapters, both descriptive statistics and inferential statistics will be discussed in greater detail.

It is important to familiarize yourself with the following definitions as they are used frequently in the field of statistics.

Definition 1.4 (Population)**.** *A (large) group of units that we are interested in studying.*

Definition 1.5 (Sample)**.** *A subset of the population.*

Definition 1.6 (Unit)**.** *The objects within a sample for which data are collected. This can be a person, a household, a rabbit, a plant, etc.*

Figure 1.1: Visualization of the a population, sample and a unit

Random sampling techniques allow inferences to be drawn on populations and are thus extremely important. The most basic random sample is known as a *simple random sample*.

Definition 1.7 (Simple Random Sample)**.** *A simple random sample selects n units from a population with equal probability such that every combination or sample of size n has an equal chance of selection. For a population of size N each unit has a $1/N$ chance of being selected.*

Note 1.1.
More complex sampling techniques include:

- *Stratified Sampling*
- *Cluster Sampling*
- *Multistage Sampling*

1.2 Types of Data

Data can be classified into two main categories: quantitative and qualitative data.

Definition 1.8 (Quantitative data)**.**
Data that can be measured numerically.

Some examples of quantitative data are:

- Number of hours you studied this week.
- Distance from your house to the university.
- Area (in ft^2) of the floor of a concourse.

Quantitative data can be further divided into discrete data and continuous data.

Definition 1.9 (Discrete data)**.**
Measurements can take only specific values.

Some examples of discrete data are:

- The number of heads you get when you toss a coin 5 times.
- The number of rooms in a residence.
- The number of 0.5 credit courses a student is currently enrolled.

Definition 1.10 (Continuous data)**.**
Measurements can take any value within a specified range.

Some examples of continuous data are:

- Height.
- Weight.
- The time taken to complete a task.

Note 1.2.
Quantitative data can also be categorized as:

- *Interval data*
- *Ratio data*

Definition 1.11 (Qualitative data)**.**
Data can not be measured numerically and instead falls into categories.

Some examples of qualitative data are:

- Favourite flavour of ice cream.
- Day of the week (Mon, Tue, ...) that an event occurred.
- City you live in.

Note 1.3.
Qualitative data can be further categorized as:

- *Nominal data*
- *Ordinal data*

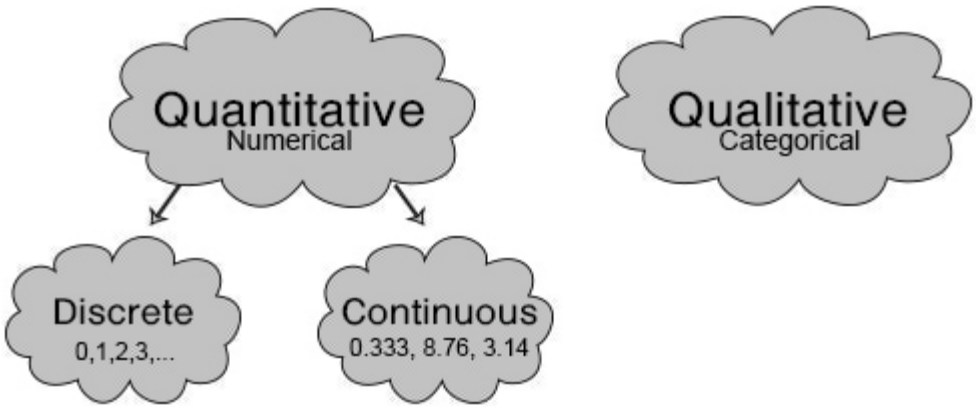


Figure 1.2: Simple breakdown of quantitative and qualitative data

1.3 Types of Studies

Definition 1.12 (Observational Studies)**.**
Experimenters observe units and take measurements without assigning treatments.

In an observational study, we can not decide which units get treatments and which do not. This may be due to ethical reasons or the nature of the study.

Examples of observational studies include:

- The effect of smoking on lung capacity.
- The effect of heroin on brain function.
- The difference in marks between students who take an in-class course and those who take an online course.

Definition 1.13 (Experimental Studies)**.**
Treatments are assigned to units and then the effects of the treatment are observed and measured.

In an experimental study we have control over which units do or do not receive a treatment. The group that does not receive a treatment is referred to as the *control group*. A control group is required in experimental studies in order to obtain a baseline for proper statistical comparison.

Examples of experimental studies include:

- Testing whether the packaging of a product is appealing to consumers before sending it out to the market.
- Providing one group with Windows computers, another (similar) group with Mac computers, and measuring the time taken to complete certain tasks.

Definition 1.14 (Sample surveys)**.**
Data is obtained from a selected part of the population using a survey instrument.

Sample surveys are one of the least expensive types of studies to conduct.

Some examples of sample surveys include:

- Satisfaction survey of guests at a resort.
- CSA O-week concert survey.

Note 1.4.
Other types of studies include:

- *Case control studies*
- *Cohort studies*
- *Cross sectional studies*

1.4 Introduction to Inferential Statistics

Before beginning to understand inferential statistics, there are several important definitions that one should become familiar with.

Definition 1.15 (Parameter)**.**
A numerical measure of a population.

The true value of a parameter is usually *unknown* as it is extremely difficult to take measurements on every unit in a population. The parameters of interest to this course are:

μ : Population mean

σ : Population standard deviation

p : Population proportion

The symbols μ (pronounced "mew"), σ (pronounced "sigma") and p are commonly used symbols to represent the population mean, population standard deviation and population proportion respectively.

Definition 1.16 (Statistic)**.**
A numerical measure of a sample.

Unlike parameters, statistics are *known* values as it is much easier to take measurements on all units in a sample. The statistics of interest to this course are:

- $\bar{x}$: Sample mean
- s : Sample standard deviation
- $\hat{p}$: Sample proportion

The symbols $\bar{x}$ (pronounced "*x bar*"), s, and $\hat{p}$ (which we call "*p hat*") are commonly used symbols to represent the sample mean, sample standard deviation and sample proportion respectively. A statistic is also often referred to as an *estimator*.

A conceptual picture of the relationship between population parameters and sample statistics is given in the figure below.

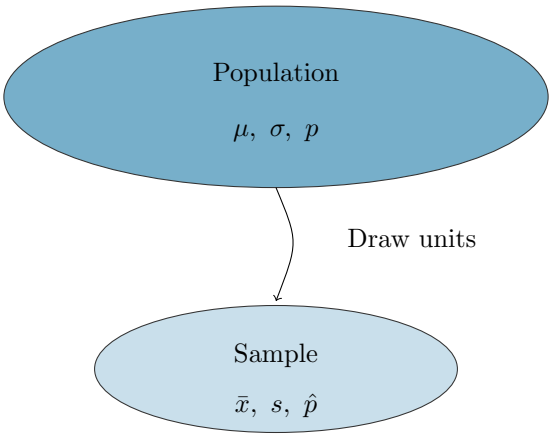


Figure 1.3: Conceptual diagram of population, sample, μ, σ, σ, $\bar{x}$, s and $\hat{p}$.

The aim of statistical inference is to produce estimators of the population parameters based on a smaller sample and to quantify the accuracy of these estimators in terms of a probability statement. In other words we are interested in the values of unknown parameters for a population thus we collect data from a sample and calculate statistics to estimate the parameters of interest. We then quantify our confidence that the statistic is representative of the parameter.

$$\bar{x} \xrightarrow{\text{Estimate}} \mu$$
$$s \xrightarrow{\text{Estimate}} \sigma$$
$$\hat{p} \xrightarrow{\text{Estimate}} p$$

In business settings, statistics allows us to make informed decisions which *(hopefully)* lead us to most "profitable" outcome. Examples of statistics in business include:

1. Data mining
2. E-commerce
3. Forecasting
4. Investment analysis
5. Marketing
6. Pricing strategies

Statistical inference is important in any good decision making process. No matter how one obtains data, data is typically expensive to acquire in terms of time, money and other resources. Sometimes obtaining good data can be very difficult. Furthermore once data is obtained it may need to be cleaned and filtered before it can be used for statistical inference. Once the data is obtained however, computation is *cheap* (i.e. not as resource intensive) if the infrastructure is in place. Your value as a statistician (to your employer) will be how you put together the statistical evidence that will ultimately translate into the profitable outcome.

1.5 Issues with Sample Data

It is usually very difficult (*or impossible*) to measure every unit in a population. It is more feasible and practical to draw a sample and take measurements on all sample units. It is important that the sample be *representative* of the population it is drawn from. This is because we would like to make generalizations about the population based on our sample. In some cases one can control for this using designed experiments or sample surveys. In other cases we do not have control over the sample which is usually the case in observational studies. As a result bias may occur in the sample. The standard forms of bias are as follows:

Definition 1.17 (Selection bias)**.** *The sample is not representative of population as a subset of the population has no chance of being selected for the sample.*

An example of selection bias is asking *only* listeners of a left-leaning radio show for their views on the Republican presidential candidate in order to gauge the candidate's nationwide popularity.

Definition 1.18 (Non-response bias)**.**
A respondent's refusal to participate may be related to the response variable.

An example of non-response bias is sampling employees working at a factory in order to determine the effects of a toxic chemical on cancer rates. If an employee has cancer, they would be unable to work and thus would not be included in the sample. Without accounting for this bias it would appear as though the chemicals have little effect.

Definition 1.19 (Measurement error bias)**.**
The response measured and recorded for an individual unit is not correct.

Measurement error bias may occur if study participants misrepresent themselves (e.g. giving a false age or weight) or if measurement tools are improperly calibrated.

Ideally we would ideally draw a random sample in order to eliminate (or at least minimize) any bias.

Example 1.1.
Company ABC is interested in determining the average job satisfaction of all its employees. In order to do so, it distributes a questionnaire to thirty employees working in its IT department. The questionnaire includes the following questions:

1. How many years have you worked for Company ABC?
2. Is your position at Company ABC full-time or part-time?
3. Rate your overall job satisfaction on a scale of 1-10.

(**a**) What is the population of interest?

All employees at Company ABC.

(**b**) What is the sample?

Thirty employees working in Company ABC's IT department.

(**c**) Is the data collected quantitative or qualitative?

Job satisfaction and length of employment are quantitative as they can be measured numerically. Position status is qualitative as employees can be classified as either full-time or part-time.

(**d**) Is the data discrete or continuous?

Job satisfaction is discrete as it can only take on the numbers $1, 2, 3, \ldots, 10$. Length of employment is continuous as it can take on any number such as $1, 2.5, 3.75, \ldots, 24.25$. Position status is qualitative and is therefore neither.

(**e**) What is the parameter of interest?

Average job satisfaction of all Company ABC employees.

(**f**) What is the statistic?

Average job satisfaction of thirty employees working in Company ABC's IT department.

(**g**) Is this an experimental study, observational study, or sample survey?

Sample survey.

(**h**) Is bias present in this sample?

Selection bias is present in the selected sample. Only employees in the IT department were surveyed thus employees working in other departments are not represented.

Chapter 2

Descriptive Statistics

2.1 Representing Data

Data comes to us in the form of *observations* (i.e. measurements) which we write symbolically as a lower case letter along with a subscript. For example, suppose we have taken a total of n observations. We have observation 1, observation 2, observation 3, ..., observation $n-1$ and observation n. By letting x represent an observation, all n observations can be represented as:

$$x_1,\ x_2,\ \ldots,\ x_n$$

where x_i is an individual observation, and the index $i = 1, \ldots, n$. Here n is referred to as the *sample size*. We can also represent this data in condensed form as:

$$x_i \ : \ \ i = 1, \ldots, n$$

or in tabular form as:

Observation	x_1	x_2	$\cdots$	x_n
Index	1	2	$\cdots$	n

Example 2.1.
The amount of time per week that a student spends studying for STAT*2060 was recorded over 5 weeks. The measurements recorded are 4, 3, 6, 20, 2. Represent this information in tabular form.

Solution:

Observation	4	3	6	20	2
Index	1	2	3	4	5

Here $n = 5,\ x_1 = 4,\ x_2 = 3,\ x_3 = 6,\ x_4 = 20$, and $x_5 = 2$.

2.2 Numerical Measures of Location

2.2.1 Mean

The *mean* (or more precisely the arithmetic mean) of a data set is obtained by adding up all of the observations and dividing by the sample size (number of observations). The mean is the typical average that we are all familiar with. sWe use the symbol $\bar{x}$ to represent the sample mean.

Definition 2.1 (Sample Mean)**.** *Let $x_1,\ x_2,\ x_3,\ \ldots,\ x_n$ represent a sample of n observations. The sample mean is given by:*

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n} \tag{2.2.1}$$

The symbol $\sum$ is referred to as *capital sigma* and symbolizes summation.

2.2.2 Median

Definition 2.2 (Median)**.** *The middle value of ordered data.*

The manner that the median is calculated depends on whether we have an odd or even number of observations.

Consider n observations $x_1, x_2,\ \cdots\ , x_n$.

If n is odd,

$$Median = observation\ \left(\frac{n+1}{2}\right)$$

If n is even,

$$Median = average\ of\ observation\ \left(\frac{n}{2}\right)\ and\ observation\ \left(\frac{n}{2}+1\right)$$

2.2.3 Mode

Definition 2.3 (Mode)**.**
The most frequent observation(s) relative to the rest of the data.

Note 2.1.
The mean, median and mode are referred to as measures of central tendency as they are indicators of how close the data is to the average and how spread out the data is.

2.2.4 Variance

The variance is a measure of the squared distance relative to the mean.

Definition 2.4 (Sample Variance)**.**
Let $x_1, x_2, x_3, \ldots, x_n$ represent a sample of n observations. Let $\bar{x}$ represent the sample mean of this data. The sample variance is given by:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1} \tag{2.2.2}$$

The variance is a non-negative value (i.e. the variance is always greater than or equal to zero). Notice in equation (2.2.2) that we are taking a difference, then squaring the value obtained before summing all the terms together. If the term $x_i - \bar{x}$ were not squared, it would be possible to have both positive and negative values. That is, $x_i - \bar{x}$ would be positive if $x_i > \bar{x}$ and negative if $x_i < \bar{x}$. By squaring the $x_i - \bar{x}$ term, the formula will provide us with non-negative values.

Note 2.2.
An alternative (and sometimes easier) way to calculate variance is:

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1} \tag{2.2.3}$$

2.2.5 Standard Deviation

In the calculation of the sample variance given by equation (2.2.3), the x_i terms are squared. This means that the units of measurement are also squared. In order to get back to the original units of measurement we take the square root of the variance. This value is referred to as the *standard deviation*.

Definition 2.5 (Sample Standard Deviation)**.** *Let $x_1,\ x_2,\ x_3,\ \ldots,\ x_n$ represent a sample of n observations. Let s^2 represent the sample variance of this data. The sample standard deviation is given by:*

$$s^2 = +\sqrt{s^2} \tag{2.2.4}$$

The standard deviation is a measure of the spread of the data relative to the mean. When we analyze data we typically describe the data in terms of the mean and standard deviation as the units are the same. However, the calculation of the sample variance is an important intermediate step.

2.2.6 Percentiles

Definition 2.6 (Percentile)**.** *For ordered data, the p^{th} percentile is the value such that p% of all observations lie below it.*

Suppose that a student obtained a mark of 68% on a test. Although 68% is not a very good grade, the test may have been difficult. Suppose you are told that this student who scored 68% is in the 90^{th} percentile. This means that they scored better than 90% of the rest of the class.

2.2.7 Quartiles

Definition 2.7 (Quartile)**.** *In an ordered data set, quartiles are the three values that divide the data into four groups such that each group consists of one fourth of the data.*

The three quartiles are the first quartile (Q_1), second quartile (Q_2) and third quartile (Q_3) respectively.

- First Quartile (Q_1) : A value such that 25% (i.e. a quarter) of all observations lie below it.

- Second Quartile (Q_2) : A value such that 50% (i.e. two quarters) of all observations lie below it.
- Third Quartile (Q_3) : A value such that 75% (i.e. three quarters) of all observations lie below it.

The quartiles are actual specific percentiles. The first quartile is the 25^{th} percentile, the second quartile is the 50^{th} percentile and the third quartile is the 75^{th} percentile. Notice that the second quartile is the median. We do not have a fourth quartile. By using the definition of percentiles, the fourth quartile would be a value such that 100% of all observations lie below it. This means that the fourth quartile is the largest value in the data set which is referred to as the *maximum*.

Definition 2.8 (Inter-Quartile Range)**.**
Let Q_1 and Q_3 represent the first and third quartiles of a data set respectively. The inter-quartile range (IQR) is

$$IQR = Q_3 - Q_1 \tag{2.2.5}$$

The inter-quartile range is a quick and simple measure of the dispersion of the data. The IQR is also referred to as the *mid-spread* since the difference between Q_3 and Q_1 contains 50% of the data.The inter-quartile range is also used to calculate whiskers in boxplots which we will cover in section 2.3.2.

Note 2.3.
The "inter-quartile range" should not be confused with the "range". The "range" of a data set is simply the difference between the largest value and the smallest value.

Example 2.2.
Given a sample with the following $n = 10$ observations,

$$9 \quad 4 \quad 1 \quad 10 \quad 8 \quad 11 \quad 15 \quad 6 \quad 6 \quad 16$$

calculate the following.

(**a**) The sample mean, $\bar{x}$.

$$\begin{aligned}\bar{x} &= \sum_{i=1}^{n} \frac{x_i}{n} = \sum_{i=1}^{10} \frac{x_i}{10} \\ &= \frac{9+4+1+10+\ldots+16}{10} = \frac{86}{10} = 8.6\end{aligned}$$

(**b**) The sample median. Rearranging the observations in ascending order,

$$1 \quad 4 \quad 6 \quad 6 \quad 8 \quad 9 \quad 10 \quad 11 \quad 15 \quad 16$$

Since $n = 10$ is even, the median is the average of ordered observation $\frac{n}{2} = \frac{10}{2} = 5$ and ordered observation $\frac{n}{2} + 1 = 6$. Therefore,

$$Median = \frac{8+9}{2} = \frac{17}{2} = 8.5$$

(**c**) The sample variance, s^2, and sample standard deviation, s.

$$\begin{aligned} s^2 &= \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1} = \sum_{i=1}^{10} \frac{(x_i - \bar{x})^2}{9} \\ &= \frac{(9-8.6)^2 + (4-8.6)^2 + \ldots + (16-8.6)^2}{9} \\ &= \frac{196.4}{9} = 21.82 \end{aligned}$$

Alternatively,

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

$$\sum_{i=1}^{n} x_i^2 = 9^2 + 4^2 + 1^2 + \ldots + 16^2 = 936$$

From part(a),

$$\sum_{i=1}^{n} x_i = 86$$

$$\Rightarrow s^2 = \frac{936 - \frac{(86)^2}{10}}{9} = \frac{196.4}{9} = 21.82$$

$$\Rightarrow s = \sqrt{s^2} = \sqrt{21.82} = 4.67$$

(**d**) The first and third quartiles, Q_1 and Q_3.

From part (b), the sample median is 8.5. Splitting the ordered observations into two sets yields

$$\begin{array}{ccccc} 1 & 4 & \underline{6} & 6 & 8 \\ 9 & 10 & \underline{10} & 15 & 16 \end{array}$$

Since both sets are of size $n = 5$ which is odd, Q_1 and Q_3 are simply the medians of each set. That is, $Q_1 = 6$ and $Q_3 = 11$.

(**e**) The inter-quartile range, IQR.

$$IQR = Q_3 - Q_1 = 11 - 6 = 5$$

Example 2.3.

Sir Shoes-A-Lot is releasing a new running shoe in 3 months. The company wants to ensure that the shoe is competitively priced and collects the prices of nine similar shoes,

20.75 49.50 33.25 23.50 39.00 49.75 22.50 25.25 48.75

(**a**) What is the sample mean price?

$$\begin{aligned}\bar{x} &= \sum_{i=1}^{9} \frac{x_i}{9} \\ &= \frac{20.75 + 49.50 + 33.25 + \ldots + 48.75}{9} \\ &= \frac{312.20}{9} = 34.69\end{aligned}$$

(**b**) What is the sample median price? Rearrange the prices in ascending order:

20.75 22.50 23.50 25.25 33.25 39.00 48.75 49.50 49.75

Since $n = 9$ is odd, the median price is equal to ordered observation number

$$\frac{9+1}{2} = \frac{10}{2} = 5$$

which is 33.25.

(**c**) What is the variance and standard deviation?

$$\begin{aligned}s^2 &= \sum_{i=1}^{9} \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \\ &= \frac{(20.75 - 34.69)^2 + (49.50 - 34.69)^2 + \ldots + (48.75 - 34.69)^2}{8} \\ &= \frac{1222.72}{8} = 152.72 \\ s &= \sqrt{s^2} = \sqrt{152.72} = 12.36\end{aligned}$$

(**d**) What are Q_1 and Q_3?

From part (b), the median price is 33.25. Splitting the data set into two sets yields

$$\begin{array}{ccccc} 20.75 & 22.50 & \underline{23.50} & 25.25 & 33.25 \\ 33.25 & 39.00 & \underline{48.75} & 49.50 & 49.75 \end{array}$$

Since both sets are of size $n = 5$ (odd), $Q_1 = 23.50$ and $Q_3 = 48.75$.

(**e**) What is the inter-quartile range?

$$IQR = 48.75 - 23.50 = 25.25$$

2.3 Graphical Techniques

Raw numbers on their own can be difficult to interpret. Pictures and plots can be a very useful for representing information. Graphical representations of data provide us with a more intuitive method to interpret information being analyzed.

Plots used for qualitative data include bar charts and pie charts. Plots used for quantitative data include histograms, box plots and stem and leaf plots. We can use plots to get a sense of the distribution our data follows as well as visualize features such as spread, central tendency and unusual data points.

Definition 2.9 (Skewness)**.**
Skewness is a measure of the symmetry of a distribution (i.e. how much a distribution leans towards a particular side).

We can describe data as being left skewed, right skewed or symmetric. If the data is described as *symmetric*, this implies that most observations are concentrated around the mean and tail off fairly evenly on both sides of the mean. If the data is described as *right skewed* or *positively skewed*, this implies that more observations are concentrated on smaller values and we observe a longer tail to the right side of the mean. If the data is described as *left skewed* or *negatively skewed*, this implies that more observations are concentrated on large values and we observe a longer tail to the left side of the mean.

Note 2.4.
Data that is right skewed may be referred to as being positively skewed and data that is left skewed may be referred to as being negatively skewed.

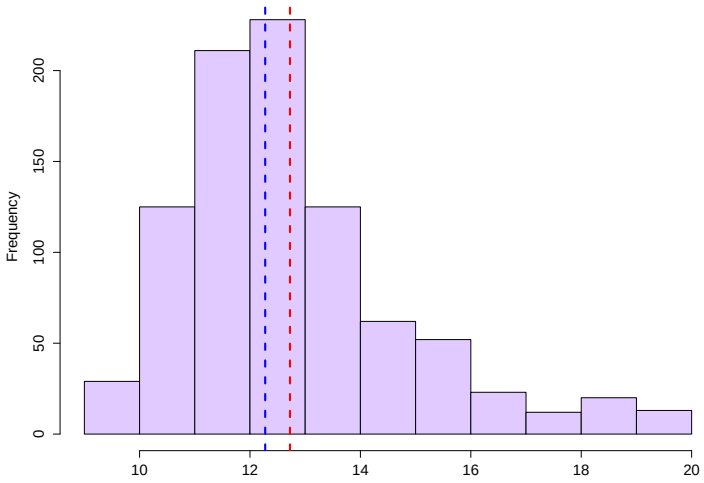


Figure 2.1: When data is skewed right the mean (red) is larger than the median (blue)

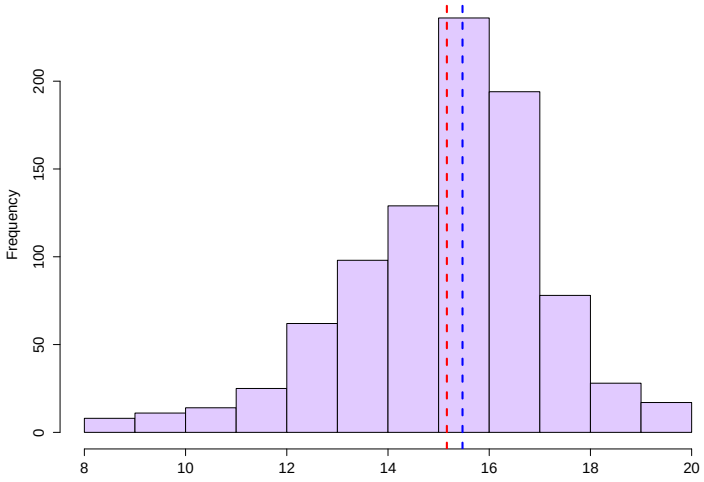


Figure 2.2: When data is skewed left the mean (red) is smaller than the median (blue)

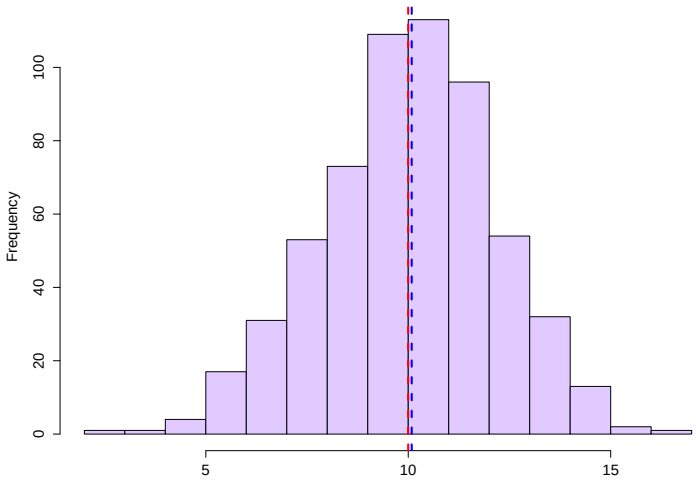


Figure 2.3: When data is symmetric the mean (red) is approximately equal the median (blue)

Definition 2.10 (Outlier)**.**
An outlier is an unusual data point which appears to lies outside the overall pattern of the rest of the data.

In other words, an outlier falls outside the range in which we would expect to see "typical" data values. Outliers may be present in data for a variety of reasons such as transcription error, measurement error, or we may have just measured a rare and unusual observation. In any case it is not a good practice to simply ignore outliers. All outliers should be investigated before deciding whether the observation should be included in data analysis or discarded.

2.3.1 Histograms

Histograms and boxplots are useful tools which can (*usually*) allow us to visually determine the skewness of a distribution. By noting skewness, we get even more information about our data. Note however that we may not always be able to determine skewness by visually observing a histogram. A histogram is similar to a bar chart with the distinction that histograms can only be created for *quantitative* data.

The width of the bars does not have any numerical meaning for bar charts, however bar width matters for histograms. They are an important factor in the creation of histograms. Interval width selection (or bin size selection) is an advanced topic that can be studied extensively and there are several rules available to select interval widths. However for the scope of this course we will keep things simple and allow the reader to intuitively choose bin size. After constructing a histogram, we can change the interval widths until we feel that we have a picture that is a good representation of our data.

In order to create a histogram, we must first

1. Construct class intervals (of equal or varying width) which will contain the data of interest.
2. Count the frequency at which data is observed in each of the class intervals.
3. Create a frequency table. A frequency tables contains both *frequencies* and *relative frequencies*.

Definition 2.11 (Frequency)**.** *Frequency, f_i, is a count of the number of observations in each class interval i, $i = 1, \ldots, m$.*

Definition 2.12 (Relative Frequency)**.** *Relative frequency, r_i, is the ratio of the frequency of class interval i, f_i, to the total frequency F where*

$$F = \sum_{i=1}^{m} f_i$$

That is,

$$r_i = \frac{f_i}{F} \quad i = 1, \ldots, m$$

Using our frequency table, we can create a *frequency histogram* or a *relative frequency histogram*. To create a frequency histogram, let the class intervals represent the width of the bars and the heights of these bars represent the frequencies of the data. To create a relative frequency histogram, let the class intervals represent the width of the bars and the heights of these bars represent the relative frequencies of the data.

Note 2.5. *When we use the term "histogram" on its own, we are usually referring to a frequency histogram.*

The general structure of a frequency table is given below.

Class Interval	Freq.	Relative Freq.	Cumulative Freq.	Cumulative Relative Freq.
$[a_1, b_1)$	f_1	$r_1 = f_1/F$	f_1	r_1
$[a_2, b_2)$	f_2	$r_2 = f_2/F$	$f_1 + f_2$	$r_1 + r_2$
$[a_3, b_3)$	f_3	$r_3 = f_3/F$	$f_1 + f_2 + f_3$	$r_1 + r_2 + r_3$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[a_m, b_m]$	f_m	$r_m = f_m/F$	$f_1 + \ldots + f_m = F$	$r_1 + \ldots + r_m = 1$
	$F = \sum_{i=1}^{m} f_i$	1		

This may appear overwhelming however it is not as bad as it looks. Lets do an example to illustrate.

Example 2.4.
Suppose we have the following set of 30 observations which represents manufacturing times (in days) for mining equipment:

53	51	92	53	77	78	77	76	53	40
45	60	99	64	44	93	64	45	53	26
58	114	35	64	58	118	74	37	48	39

It is hard to see any obvious patterns with just the raw data alone. Perhaps a picture will help. Construct a frequency histogram and a relative frequency histogram for this data.

Solution:

First we will begin by constructing a frequency table. An optional (but very useful) step is to sort the data.

26	35	37	39	40	44	45	45	48	51
53	53	53	53	58	58	60	64	64	64
74	76	77	77	78	92	93	99	114	118

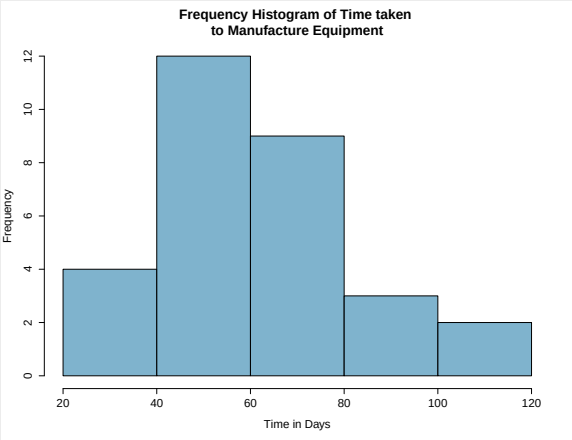
After sorting, the following class intervals appear intuitively "nice":

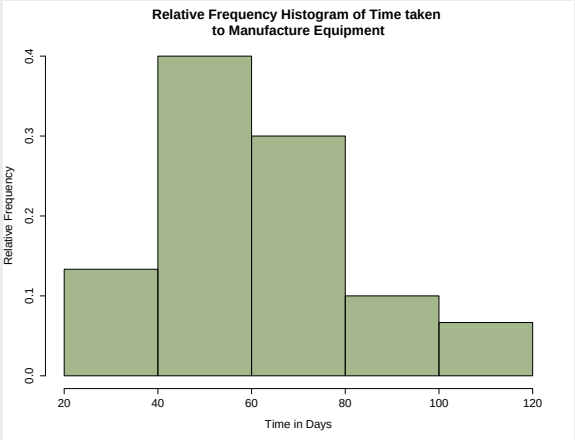
Interval	Mathematical representation
20 to 39	$[20, 40)$
40 to 59	$[40, 60)$
60 to 79	$[60, 80)$
80 to 99	$[80, 100)$
100 to 120	$[100, 120]$

We can now create our frequency table.

Class Interval	Frequency	Relative Frequency	Cumulative Frequency	Cumulative Relative Frequency
$[20, 40)$	4	0.13333333	4	0.13333333
$[40, 60)$	12	0.40	16	0.53333333
$[60, 80)$	9	0.30	25	0.83333333
$[80, 100)$	3	0.10	28	0.93333333
$[100, 120]$	2	0.06666667	30	1
	30	1		

The frequency table above can be used to create both a frequency histogram as well as a relative frequency histogram for our data.





Through visual inspection of the histograms, it appears as though the distribution of our data is right skewed. We can therefore deduce that the mean is less than the median.

2.3.2 Boxplots

Boxplots are another visual aid we can use to present and interpret data. They are one of the most simple graphical techniques to analyze visually. Boxplots are also known as *box-and-whisker plots* since they consists of figures resembling boxes along with a series of lines which we refer to as whiskers. The whiskers represent the quartiles as well as special values which we refer to as the *lower whisker* and the *upper whisker*.

Definition 2.13 (Upper & Lower Whiskers)**.** *Let Q_1, Q_3 and IQR represent the first quartile, third quartile and interquartile range respectively. Then,*

$$\begin{aligned} \textit{Lower Whisker} &= Q_1 - 1.5 \times IQR \\ \textit{Upper Whisker} &= Q_3 + 1.5 \times IQR \end{aligned}$$

Boxplots are very useful tool for identifying outliers as they are any values which fall outside of the lower and upper whiskers. We can also use boxplots to obtain information regarding the skewness of data. We achieve this by analyzing how close the quartiles are to each other graphically.

To construct a boxplot:

1. Draw whiskers at the values of the lower whisker and the upper whisker.
2. Draw the box around the values of the quartiles. The top of the box represents Q_3, the bottom of the box represents Q_1, and the line within the box represents the median or Q_2.
3. Any values that are greater than the upper whisker or less than the lower whisker are outliers and are represented by single points.

Note 2.6.
Boxplots can be oriented horizontally or vertically and the width of the boxplot does not have any significance.

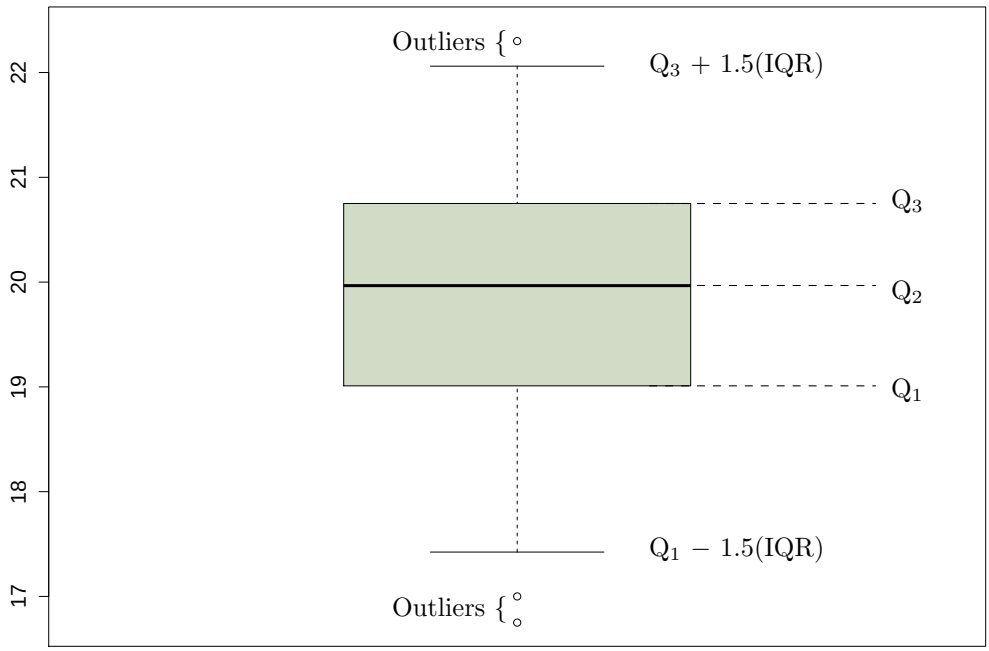


Figure 2.4: Interpreting a boxplot

We can also perform a side by side comparison of several boxplots in order to compare the skewness of different data sets, such as in figure 2.3.2.

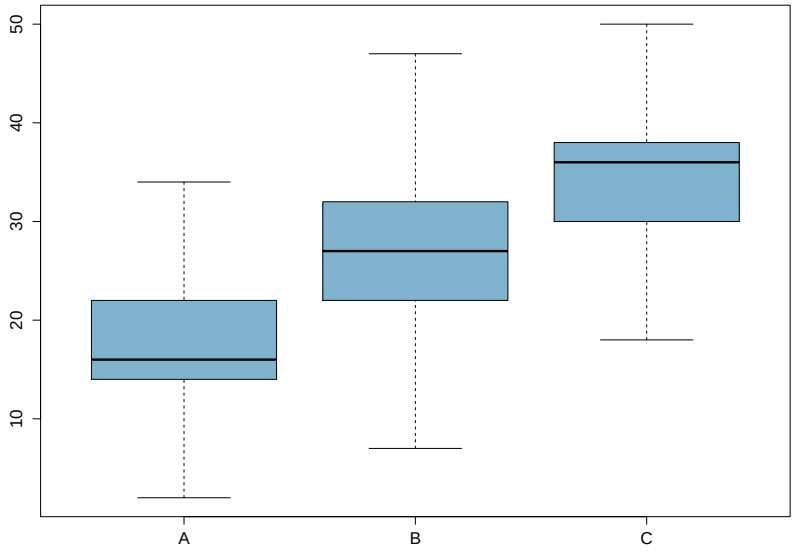


Figure 2.5: A positively skewed boxplot (A), a symmetric boxplot (B) and a negatively skewed boxplot (C)

Example 2.5.

Suppose we have the following $n = 10$ observations:

$$26 \quad 20 \quad 26 \quad 19 \quad 25 \quad 16 \quad 10 \quad 33 \quad 43 \quad 17$$

Construct a boxplot for this data.

Solution:

First, we find the median. Since $n = 10$ is even, the median is the average of the fifth and sixth ordered observations.

$$Median = \frac{20 + 25}{2} = 22.5$$

Splitting the data into two sets yields

$$10 \quad 16 \quad \underline{17} \quad 19 \quad 20$$
$$25 \quad 26 \quad \underline{26} \quad 33 \quad 43$$

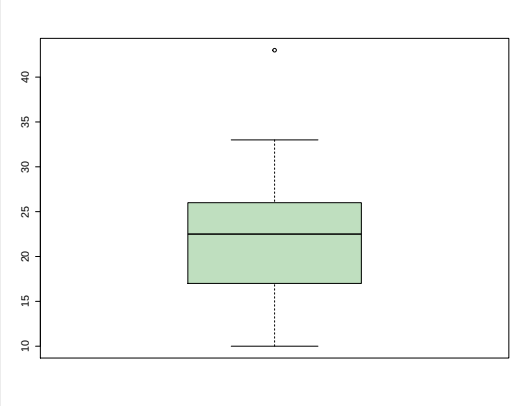
Therefore, $Q_1 = 17$ and $Q_3 = 26$. We can now calculate the IQR and the whiskers.

$$IQR = Q_3 - Q_1 = 27 - 17 = 10$$

$$\text{Lower whisker} = Q_1 - 1.5 \times IQR = 17 - 1.5(10) = 2$$

$$\text{Upper whisker} = Q_3 + 1.5 \times IQR = 26 + 1.5(10) = 41$$

The observation of 43 falls outside of the upper whisker and is therefore an outlier.



The upper whisker is drawn at the highest value below 41 which is 33. The lower whisker is drawn at the lowest value above 2 which is 10. The outlier of 43 is represented by a circle. The boxplot appears to suggest that the distribution may be slightly negatively (or left) skewed.

2.3.3 Stem and Leaf Plots

A *stem and leaf plot* (sometimes referred to as simply a *stem plot*) is a tabular method of displaying data. Each observation is split into a *stem* (the "larger" part of an observation) and a *leaf* (the smaller part of an observation).

One common partition is to let the stem represent a whole number and the leaves represent the decimal part of the number. Another common partition is to consider the stem to be the integer part of some power of 10 (ex. 10, 100, 1000 etc.) Stem and leaf plots are easier to understand with the aid of an example.

Example 2.6.

Consider the following set of (sorted) data points:

10.1	10.1	10.2	10.3	11.1	11.2	11.2	11.2	11.3	12.0
12.0	12.3	12.4	12.5	12.5	12.6	14.2	14.2	14.3	14.3
14.4	15.1	15.2	15.3	15.3	16.0	16.1	16.2	18.0	18.1

Create a stem and leaf plot for this data.

Solution:

Let's group terms and rewrite our data as follows:

10.1	10.1	10.2	10.3			
11.1	11.2	11.2	11.2	11.3		
12.0	12.0	12.3	12.4	12.5	12.5	12.6
14.2	14.2	14.3	14.3	14.4		
15.1	15.2	15.3	15.3			
16.0	16.1	16.2				
18.0	18.1					

We can now construct out stem and leaf plot. Let's choose our partition (|) to be at the decimal point. As such, the stem and leaf plot for this data is:

10	|	1123
11	|	12223
12	|	0034556
13	|	
14	|	22334
15	|	1233
16	|	012
17	|	
18	|	01

Chapter 3

Probability

3.1 Introduction

In a probability problem certain characteristics of the population (such as parameters) are known or at least assumed to be known. This is the key feature that distinguishes a probability problem from a statistics problem. We can then use this known information to answer questions regarding observing a specific outcome associated with a sample drawn from that population. The tools and techniques used to solve statistics problems (such as those covered in chapter 6 and 7) are derived and developed using probability as their foundation.

3.2 Terminology and Definitions

There are several terms that form the language of probability. It is important to become familiar with these terms.

Definition 3.1 (Experiment)**.**
An experiment is a random process in which the outcome is uncertain.

The meaning of the term "experiment" in statistics is not quite the same as that of regular English. In statistics, the following are examples of experiments:

1. Tossing a coin and observing the result.
2. Rolling a die and observing the result.
3. Choosing a marble at random from a jar of coloured marbles.

Definition 3.2 (Sample Space)**.**
A sample space is the set of all possible outcomes of an experiment.

Many texts use S to represent a sample space. In order to avoid confusion between a sample space and the standard deviation, we will use Ω (uppercase *Omega*) instead.

Definition 3.3 (Sample Point)**.**
A sample point is an element of a sample space.

Definition 3.4 (Event)**.**
An event is any collection of sample points.

Example 3.1.
Consider the experiment of flipping a coin twice.

(**a**) What is the sample space Ω for this experiment?

Let H represents heads and let T represents tails.

$$\Omega = \{HH,\ TT,\ HT,\ TH\}$$

The four sample points are then $HH,\ TT,\ HT,\ TH$.

(**b**) Let E represent the event that at least one H occurs. Explicitly write down E.

By examining Ω we observe that there are three sample points which contain H and thus

$$E = \{HH,\ HT,\ TH\}$$

Definition 3.5 (Probability)**.**
The probability of an event is the proportion of times the event would occur if we observed the random process an infinite number of times.

To understand probability better, consider the case where we toss a fair coin. If we toss it only twice, we will not always observe heads on one toss and tails on the other. Similarly, if we toss the coin 10 times, we will not always observe heads on 5 tosses and tails on 5 tosses. However if we toss the coin approximately 100 times, we will start to observe heads on almost 50 tosses and tails on almost 50 tosses. If we toss it 1000 times, we should observe heads on almost 500 tosses and tails on almost 500 tosses. As we increase the number of tosses, the proportion of heads we observe will start to approach 50% and the proportion of tails we observe will also start to approach 50%. If we keep tossing this coin forever (i.e. we keep tossing infinitely) then the proportion of heads we observe will be equal to 50% and the

proportion of tails we be equal to 50%. A simulation of coin tosses is shown in figure 3.2.

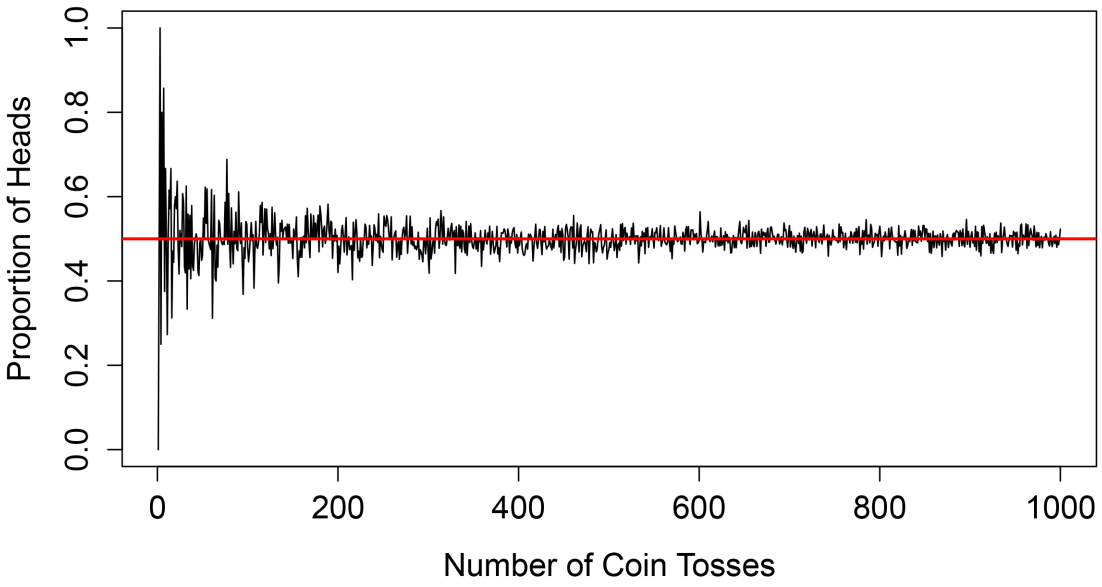


Figure 3.1: Simulation of a coin toss. Notice how the proportion of heads approaches 0.5 as the number of tosses becomes large.

Using this framework, we can attach probabilities to sample points and events occurring within a sample space in order to solve probability problems.

3.3 Notation

There are certain symbols and notation that are commonly used to represent probability problems. Probability statements are enclosed in an upper case P. By this we mean that a probability statement is written as $P(statement)$.

Let A and B be any 2 events in sample space Ω.

$P(A)$	:	Probability of event A occurring
$\cap$	:	Intersection $A \cap B$ means A intersection B. Common to <u>both</u> A and B.
$\cup$	:	Union $A \cup B$ means A union B. A or B or both. Everything in A as well as B.
$\square^{c}$	:	Complement A^{c} means the complement of A (or A complement) which is everything <u>except</u> A.

Table 3.1: Summary table of some of the common notation used for probability statements.

3.4 Properties

The foundation of all probability theory is built on *Kolmogorov's Axioms of Probability*. There are three axioms but only two are of interest to this course.

Property 3.1.
Let Ω be a sample space. Then

$$P(\Omega) = 1 \tag{3.4.1}$$

In other words Property 3.1 simply states that once we consider all possibilities, something must occur.

Property 3.2.
Let A be any event in sample space Ω. Then

$$0 \leq P(A) \leq 1 \tag{3.4.2}$$

Property 3.2 states that the probability that event A occurs must be between 0 and 1.

Events of particular interest to this course are *mutually exclusive events*.

Definition 3.6 (Mutually Exclusive Events)**.** *Two events are said to be mutually exclusive (or disjoint) if they can not occur at the same time. That is,*

$$P(A \cap B) = 0$$

Property 3.3. *Let A and B be any two events in sample space Ω. Then*

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{3.4.3}$$

Property 3.4. *Let A and B be mutually exclusive events in sample space Ω. Then*

$$P(A \cup B) = P(A) + P(B) \tag{3.4.4}$$

Property 3.5. *Let A be an event in sample space Ω. Then*

$$P(A^c) = 1 - P(A) \tag{3.4.5}$$

Example 3.2. A marketing team polled 500 random people and asked them for their opinion on two new radio jingles, jingle A and jingle B. 150 people liked jingle A but not B, 75 people liked jingle B but not A, and 50 people liked jingle A and B. Let A represent the event that the person liked jingle A. Let B represent the event that the person liked jingle B. Without using a venn diagram,

(**a**) Find $P(A)$ and $P(A^c)$.

150 liked *only* jingle A and 50 people liked *both* jingle A and B. Therefore, 200 people in total enjoyed jingle A.

$$\Rightarrow P(A) = \frac{200}{500} = 0.40$$

Using the property of complements,

$$P(A^c) = 1 - P(A) = 1 - 0.40 = 0.60$$

(**b**) Find $P(B)$ and $P(B^c)$.

75 people liked *only* jingle B and 50 people liked *both* jingle A and B. Therefore, 125 people in total enjoyed jingle B.

$$\Rightarrow P(B) = \frac{125}{500} = 0.25$$

$$P(B^c) = 1 - P(B) = 1 - 0.25 = 0.75$$

(**c**) Find $P(A \cap B)$.

50 people liked *both* jingle A and B.

$$\Rightarrow P(A \cap B) = \frac{50}{500} = 0.10$$

(**d**) Find $P(A \cup B)$.

Using the additive property,

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= 0.40 + 0.25 - 0.10 \\ &= 0.55 \end{aligned}$$

(**e**) How many people did not like either jingle?

From part **(d)** we know that $0.55 \times 500 = 275$ people liked jingle A, jingle B, or both. Therefore, $500 - 275 = 225$ did not like either jingle.

Alternatively, the probability that someone didn't like *both* A and B is the complement of the probability that someone likes jingle A or jingle B or both. That is,

$$\begin{aligned} P(A^c \cap B^c) &= 1 - P(A \cup B) \\ &= 1 - 0.55 \\ &= 0.45 \end{aligned}$$

which means $0.45 \times 500 = 225$ people did not like either jingle as above.

3.5 Venn Diagrams

Venn diagrams give a graphical representation of a sample space, events and sample points. They are an effective way of illustrating the concepts of set notation that we just learnt in the section 3.3 on common notation and are a very helpful tool to use when working with probability questions. Figures 3.2, 3.3, 3.4 and 3.5 are examples of Venn diagrams.

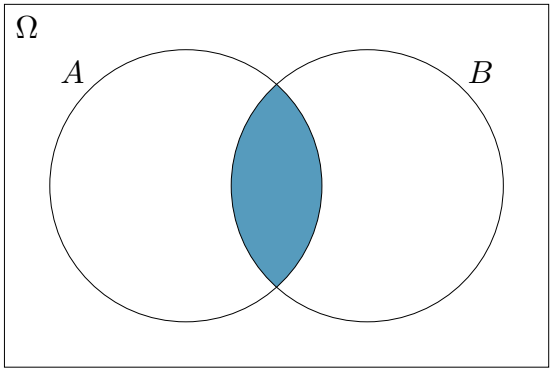


Figure 3.2: The shaded area represents $A \cap B$

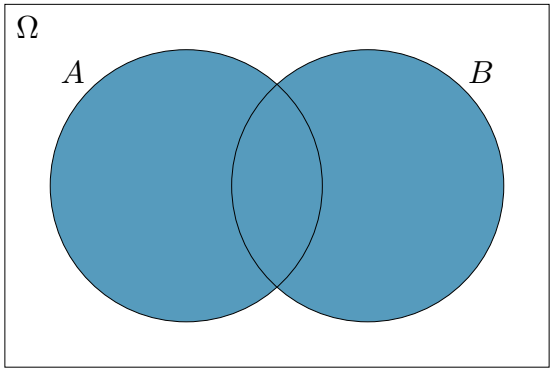


Figure 3.3: The shaded area represents $A \cup B$

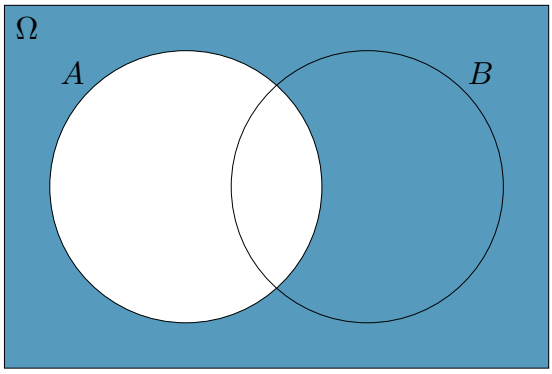


Figure 3.4: The shaded area represents A^c

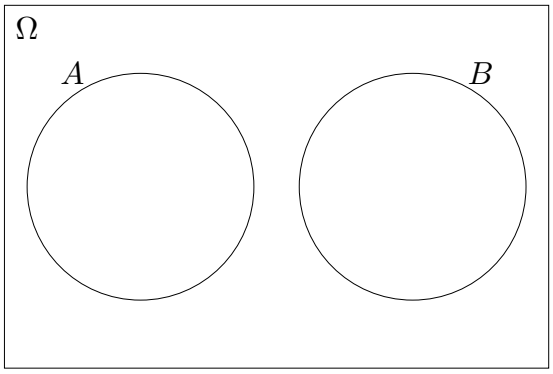
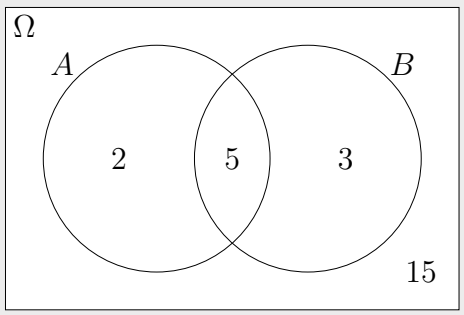


Figure 3.5: Mutually exclusive events A and B.

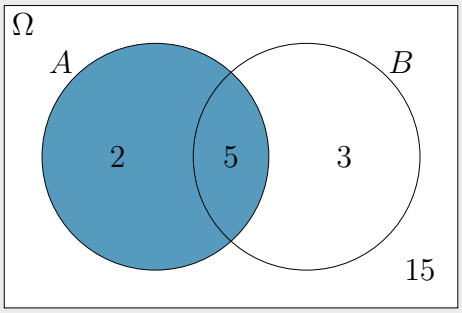
Example 3.3.
Consider the following Venn diagram. Each value represents the number of sample points in the region.



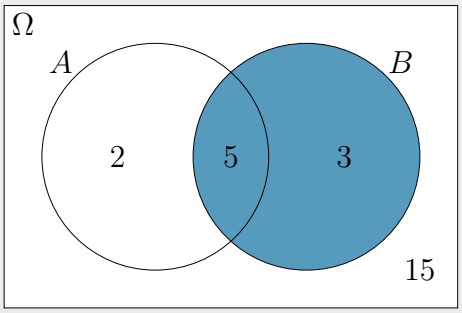
(**a**) How many sample points are in the sample space?

$$2 + 5 + 3 + 15 = 25$$

(**b**) What is $P(A)$? $P(B)$?

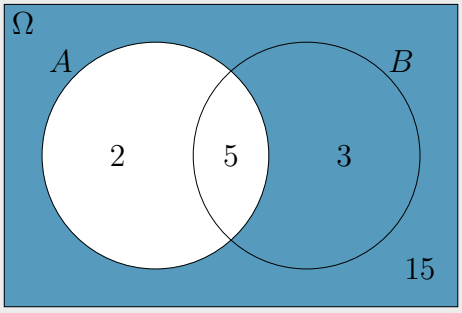


$P(A) = \frac{2+5}{25} = 0.28$

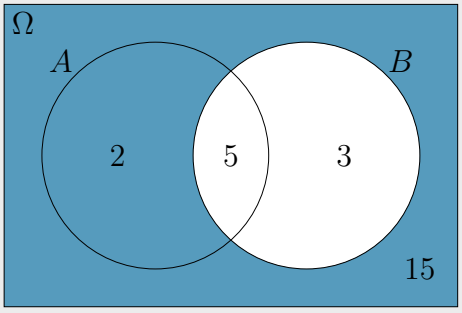


$P(B) = \frac{5+3}{25} = 0.32$

(**c**) What is $P(A^c)$? $P(B^c)$?

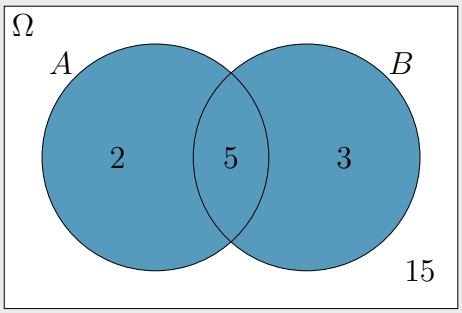


$P(A) = \frac{3+15}{25} = 0.72$

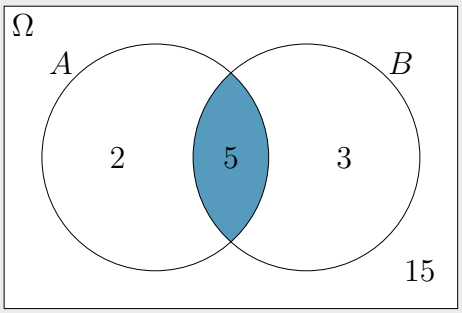


$P(B) = \frac{2+15}{25} = 0.68$

(**d**) What is $P(A \cup B)$? $P(A \cap B)$?

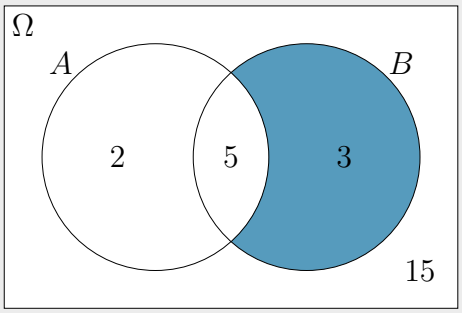


$P(A \cup B) = \frac{2+5+3}{25} = 0.40$

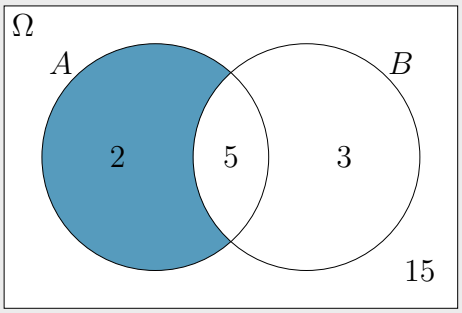


$P(A \cap B) = \frac{5}{25} = 0.20$

(**e**) What is $P(A^c \cap B)$? $P(A \cap B^c)$?

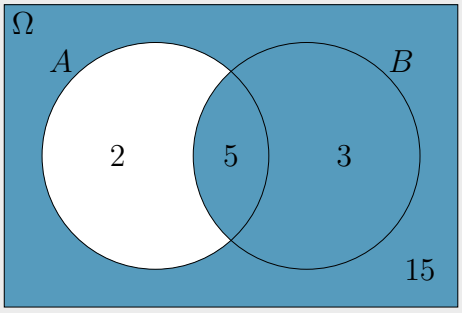


$P(A^c \cap B) = \frac{3}{25} = 0.12$

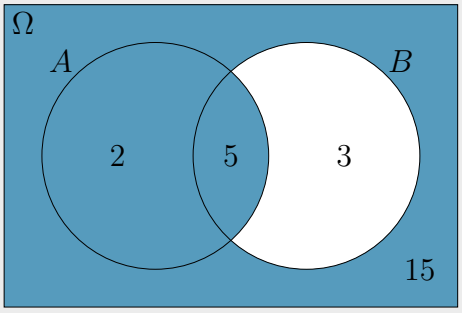


$P(A \cap B^c) = \frac{2}{25} = 0.08$

(**f**) What is $P(A^c \cup B)$? $P(A \cup B^c)$?

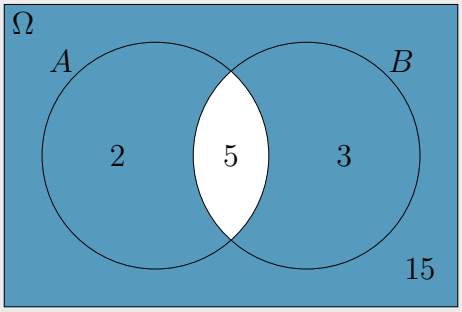


$P(A^c \cup B) = \frac{3+5+15}{25} = 0.92$

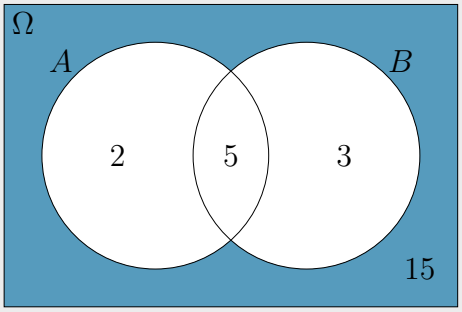


$P(A \cap B^c) = \frac{2+5+15}{25} = 0.88$

(**g**) What is $P(A^c \cup B^c)$? $P(A^c \cap B^c)$?



$P(A^c \cup B^c) = \frac{2+3+15}{25} = 0.8$



$P(A^c \cap B^c) = \frac{15}{25} = 0.6$

3.6 Conditional Probability and Independence

3.6.1 Conditional Probability

Conditional probability is the probability of an event occurring given that another event has already occurred. Consider two events A and B in the same sample space. A and B have certain probabilities of occurring on their own. However if we know that event B has occurred then we can use this information to calculate the probability that event A has also occurred.

The notation we use to express the conditional probability of event A occurring given that event B has occurred is $P(A|B)$.

Definition 3.7 (Conditional Probability)**.** *Let A and B be any two events in sample space Ω such that $P(B) \neq 0$. The conditional probability of A given B is*

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{3.6.6}$$

3.6.2 Independence

Independence in the statistical sense is slightly different to independence in regular English. Independence intuitively means that the occurrence of one event does not effect the occurrence of another event. Suppose we have events A and B that are independent. This means that if A occurs then it is neither more nor less likely that B occurs and vice versa. Mathematically the statistical independence is expressed as follows.

Definition 3.8 (Independence)**.** *Let A and B be any two events in sample space Ω. Events A and B are said to be independent if*

$$P(A|B) = P(A) \tag{3.6.7}$$

$$\iff P(B|A) = P(B) \tag{3.6.8}$$

$$\iff P(A \cap B) = P(A)P(B) \tag{3.6.9}$$

All three equations in definition 3.8 are equivalent. If one of the equalities holds true then the other two equalities will also hold true. If two events are not independent then they are said to be *dependent*.

Note 3.1.
In questions involving probability we should not automatically assume that events are independent unless explicitly stated.

3.7 Bayes' Theorem

Bayes' theorem is a very useful formula in statistics which is a result of manipulating conditional probabilities. Recall the definition of conditional probability from definition 3.8. Using conditional probability and some simple manipulation, we can show that:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A|B)P(B)}{P(A)} \tag{3.7.10}$$

This implies is that if we are provided information about $P(A|B)$, $P(B)$ and $P(A)$, then we can recover $P(B|A)$.

Theorem 3.1.
Let $B_1, B_2, \ldots, B_n$ be mutually exclusive events in sample space Ω such that $\sum_{i=1}^{n} P(B_i) = 1$. Let A be any event in Ω. Then:

$$P(B_i|A) = \frac{P(B_i) \cdot P(A|B_i)}{\sum_{i=1}^{n} P(B_i) \cdot P(A|B_i)} \tag{3.7.11}$$

Bayes' Theorem may look confusing but it is not so bad once we practice some questions and get comfortable with it. For the purposes of an introductory course such as ours it is sufficient to become familiar with the following forms of Bayes' Theorem:

$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A)} \tag{3.7.12}$$

and

$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A|B) \cdot P(B) + P(A|B^c) \cdot P(B^c)} \tag{3.7.13}$$

Example 3.4.

Banana releases a new high-end phone every 18 months. In order to gauge brand loyalty, *Bananaphone* polled 5000 shoppers nationwide. Of the 5000 shoppers, 500 recently bought the new *Bananaphone 5.0*. Of the 500 that bought the *Bananaphone 5.0*, 400 had previously purchased the *Bananaphone 4.0*. 2000 people purchased the *Bananaphone 4.0* in total.

Let $B5$ represent the event that the shopper owns the *Bananaphone 5.0*. Let $B4$ represent the event that the shoppers own the *Bananaphone 4.0*.

From the information provided, we can calculate $P(B5), P(B4)$ and $P(B4|B5)$. We know that 500 shoppers out of 5000 purchased the *Bananaphone 5.0*. The probability that a shopper owns the *Bananaphone 5.0* is therefore given by

$$P(B5) = \frac{500}{5000} = 0.1$$

Of the 500 that owned the *Bananaphone 5.0*, 400 had previously purchased the *Bananaphone 4.0*. The probability that a shopper previously purchased a *Bananaphone 4.0* given they currently own a *Bananaphone 5.0* is therefore given by

$$P(B4|B5) = \frac{400}{500} = 0.8$$

We also know that 2000 shoppers out of 5000 purchased the *Bananaphone 4.0*. The probability that a shopper purchased the *Bananaphone 4.0* is therefore given by

$$P(B4) = \frac{2000}{5000} = 0.40$$

Using Bayes' theorem, we can find the probability that a shopper owns the *Bananaphone 5.0* given they previously purchased the *Bananaphone 4.0*,

$$\begin{aligned} P(B5|B4) &= \frac{P(B4|B5)P(B5)}{P(B4)} \\ &= \frac{0.8 \times 0.1}{0.40} \\ &= 0.20 \end{aligned}$$

Only 20% of shoppers bought the *Bananaphone 5.0* after previously purchasing the *Bananaphone 4.0*. This may be due to the fact that the phones are high-end and therefore quite expensive. A new phone every 18 months allows only wealthier customers to keep up with releases. On the other hand, 80% of shoppers that bought the *Bananaphone 5.0* had previously purchased the *Bananaphone 4.0*. This indicates that customers are satisfied with the quality of the product and are relatively loyal to the *Bananaphone* brand.

Example 3.5.
Soylent Teal Manufacturing Co. held a professional development day and 600 employees were given the option to present their latest research at a seminar, renew their skill certifications through a workshop, or participate in a R & D brainstorming session. *Soylent Teal Manufacturing Co.* is interested in whether or not the participation in different activities led to higher promotion rates in the following 12 months. The collected data is presented in the table below.

•	Research Seminar (R)	Skills Workshop (S)	Brainstorming (B)
Promoted (Y)	22	20	6
Not Promoted (N)	120	340	92

We can create a tree diagram to help us visualize the data better. We start by creating branches for each of the activities: R, S, and B. Using the table we can easily find $P(R), P(S)$ and $P(B)$. We know 120 people participated in the research seminar and were not promoted and 22 people were promoted, meaning 142 people out of 600 in total participated in the research seminar.

$$P(R) = \frac{22 + 120}{600} = \frac{142}{600} \approx 0.24$$

Similarly,

$$P(S) = \frac{20 + 340}{600} = \frac{360}{600} = 0.60$$
$$P(B) = \frac{6 + 92}{600} = \frac{98}{600} \approx 0.16$$

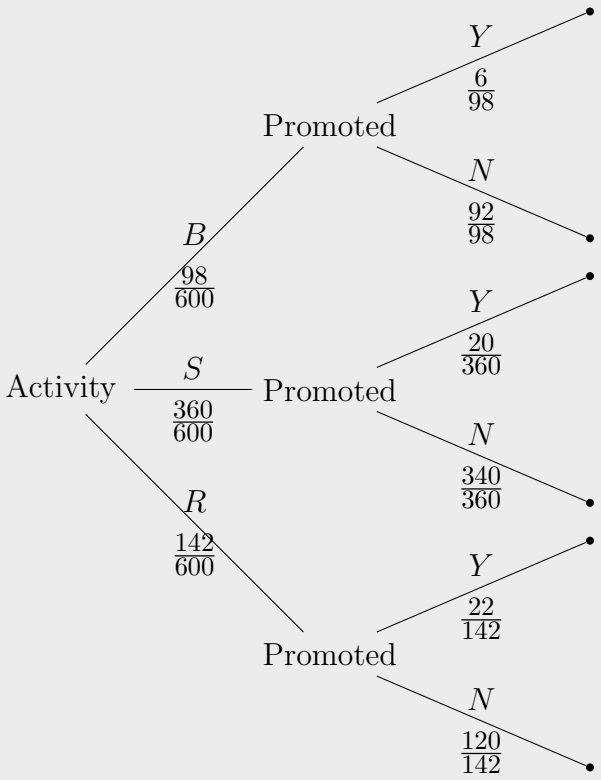
Now we need to create a two branches for each activity to represent promoted and not promoted. We now know that 142 people in total participated in the research seminar and of those 142, 22 got promoted. Therefore, the probability of an employee being promoted given they attended the research seminar is

$$P(Y|R) = \frac{22}{142} \approx 0.15$$

and thus

$$P(N|R) = \frac{120}{142} \approx 0.85$$

We can continue in this fashion to complete the tree diagram. The complete diagram is presented below.



(**a**) If an employee is selected at random, what is the probability that they attended the brainstorming session and were promoted?

We are interested in finding $P(B \cap Y)$. Recall Bayes' Theorem,

$$P(Y|B) = \frac{P(B \cap Y)}{P(B)}$$

Rearranging we find

$$P(B \cap Y) = P(Y|B) \cdot P(B)$$

From our tree we know that $P(B) = \frac{98}{600} \approx 0.16$ and $P(Y|B) = \frac{6}{98} \approx 0.06$. Plugging these into our equation for $P(B \cap Y)$ yields

$$P(B \cap Y) = P(Y|B) \cdot P(B) = 0.06 \cdot 0.16 \approx 0.01$$

(**b**) If an employee is selected at random, what is the probability that they attended the skills workshop and were promoted?

We are interested in $P(S \cap Y)$. From part (a) we know that

$$P(S \cap Y) = P(Y|S) \cdot P(S)$$

From our tree we know that $P(S) \approx 0.6$ and $P(Y|S) \approx 0.06$. Plugging these into our equation for $P(S \cap Y)$ yields

$$P(S \cap Y) = P(Y|S) \cdot P(S) = 0.6 \cdot 0.06 \approx 0.04$$

3.8 Counting Techniques

3.8.1 Factorial

The factorial is a a very useful function used in combinatorics. The factorial of a non-negative integer n is denoted as $n!$ and it is the product of all positive integers less than or equal to n.

Definition 3.9 (Factorial (!))**.**
Let n be any non-negative integer. The factorial of n is

$$n! = n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot 2 \cdot 1 \tag{3.8.14}$$

$$0! = 1 \tag{3.8.15}$$

The factorial is very useful in solving problems involving counting the number of ways we can arrange a number of items. The number of ways that we can arrange n items is $n!$ The reason that $0! = 1$ by definition is because there is only one way that we can arrange no items (and that is to not do anything). The reason for defining $0! = 1$ will also become more apparent when we study combinations and permutations.

3.8.2 Combinations

In counting problems we often would like to consider randomly drawing a sample of size k from a population of size n. The number of ways which we can achieve this is provided by binomial coefficient and this value is calculated as follows.

Definition 3.10 (Binomial Coefficient)**.**
The number of ways which we can choose k items out of a total of n items such that order does not matter is

$${}_nC_k = \binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{3.8.16}$$

Furthermore when we choose k items, they are not distinct items.

Example 3.6.
Suppose that the University of Guelph offers 23 courses that will satisfy business degree requirements but you can only choose 5 for the fall semester. How many different collections of 5 courses can you choose in the fall semester?

Solution:

There are ${}_{23}C_5$ different collections of 5 fall semester courses.

$$ {}_{23}C_5 = \binom{23}{5} = \frac{23!}{5!(23-5)!} = 33649 $$

In the following winter semester how many choices will you have, assuming that you passed all of your courses in the fall semester?

Since we already passed 5 out of the 23 possible courses, we only have 18 left to choose from. Therefore, there are ${}_{18}C_5$ ways of selecting 5 winter semester courses.

$$ {}_{18}C_5 = \binom{18}{5} = \frac{18!}{5!(23-5)!} = 8569 $$

Example 3.7.

Use factorials or the binomial coefficient to solve the following.

(**a**) How many ways can we arrange the numbers 0 through 9 if each number can only be selected once?
For the first number, we have 10 possibilities. Since we already selected the first number, we only have 9 possibilities for the second number. Since we already chose two numbers for the first and second number, we only have 8 possibilities for the third number, etc. Therefore, there are

$$ 10! = 10 \times 9 \times 8 \times 7 \times 6 \ldots \times 1 = 3,628,800 $$

ways to arrange the numbers 0 through 9.

(**b**) A standard license plate has 7 characters. If the characters can take on the values of 0-9, each number can only be used once, and each combination of numbers is unique, how many ways can we choose a license plate number?

Since we are only choosing 7 characters, we have 10 choices for the first number, 9 choices for the second number, ... , 4 choices for the seventh number. That is there are

$$ 10 \times 9 \times 8 \times \ldots \times 4 = 604,800 $$

ways to arrange seven numbers from 0 through 9. This is equivalent to

$$ \frac{10!}{3!} = \frac{10 \times 9 \times 8 \times \ldots \times 1}{3 \times 2 \times 1} $$

(**c**) If each combination of numbers on a license place is not unique, how many ways can we choose a license plate number?

From part (a) and via the factorial, we know that the 7 numbers we select can be arranged 7! ways. But since order doesn't matter, we assume that all 7! ways are the same. That is to say, 1 2 3 4 5 6 7 is equivalent to 7 6 5 4 3 2 1, etc. Therefore, assuming order doesn't matter we can choose a license plate with the numbers 0 through 9

$$\frac{10!}{3!7!} = 120$$

different ways. This is ${}_{10}C_7$.

3.9 Random Variables

A random variable is the realization of what we previously called an experiment with a numerical outcome. The term might be a little misleading since a random variable is actually a function.We usually denote a random variable with a capital letter (such as X) and the values it takes with a lowercase letter (such as x).

Definition 3.11 (Random Variable)**.** *Let Ω be a sample space. A random variable (RV) X is a function that maps events in Ω to a real number.*

$$X : \Omega \longrightarrow \mathbb{R} \tag{3.9.17}$$

Note 3.2. *Definition 3.11 above is not completely correct but it is sufficient for the purposes of an introductory course in statistics. We can consider a random variable to be a well defined map where every event in the sample space is mapped to some number. The proper definition of a random variable involves a mathematical set of events known as a sigma algebra (or sigma field) and is better suited for more advanced courses such as mathematical statistics, mathematical analysis or measure theory.*

3.9.1 Discrete Random Variables

Definition 3.12 (Discrete Random Variables)**.** *A discrete random variable is a random variable that can only take specific (i.e. discrete) values.*

Using a *probability distribution function*, we can assign probabilities to every possible value of a random variable.

3.9.1.1 Probability Mass Function

The *probability distribution function* of a discrete random variable is called a *probability mass function* (PMF) or mass function for short. If X is a discrete random variable with outcomes $x = x_1,\ x_2,\ \ldots$, the probability that X takes on a specific value x_i is $P(X = x_i)$ or simply $p(x_i)$.

Properties 3.1.
Let X be a discrete random variable with possible values $x = x_1,\ x_2,\ \ldots$ The probability mass function of X consists of individual probabilities $p(x_i)$ with the following properties:

1. $0 \leq p(x_i) \leq 1$

2. $\sum_{\forall i} p(x_i) = 1$

Note that we are being general in terms of the full range of X because it can be finite or infinite (This is why we have written $\forall i$ and not $i = 1, 2, \ldots, n$).

A probability mass function can be expressed in tabular form for convenience. For a discrete random variable X with possible values $x_1,\ x_2,\ \ldots\ ,\ x_n$,

$X = x$	x_1	x_2	$\ldots$	x_n
$P(X = x)$	$p(x_1)$	$p(x_2)$	$\ldots$	$p(x_n)$

Table 3.2: Probability mass function of discrete random variable X.

3.9.1.2 Mean, Variance and Standard Deviation of a Discrete Random Variable

Definition 3.13 (Mean of a Discrete Random Variable)**.**
Let X be a discrete random variable with outcomes $x_1, \ldots, x_n$ which have corresponding probabilities $P(X = x_1), \ldots, P(X = x_n)$. The mean of X is the sum of each outcome x_i

multiplied by their corresponding probability $P(X = x_i)$:

$$\mu = \mathbb{E}(X) = \sum_{\forall i} x_i \cdot p(x_i) \tag{3.9.18}$$

$$= x_1 \cdot p(x_1) + x_2 \cdot p(x_2) + \dots \tag{3.9.19}$$

This mean is the population mean and not the sample mean. The population mean is also referred to as the *expected value*, denoted by $\mathbb{E}(x)$.

Definition 3.14 (Variance & Standard Deviation of a Discrete RV)**.** *Let X be a discrete random variable with outcomes $x_1, \dots, x_n$ which have corresponding probabilities $P(X = x_1), \dots, P(X = x_n)$ and mean $\mu = \mathbb{E}(X)$. The variance of X is calculated as*

$$\sigma^2 = \mathbb{E}\big((X - \mu)^2\big) \tag{3.9.20}$$

$$= \sum_{\forall i} (x_i - \mu)^2 \cdot p(x_i) \tag{3.9.21}$$

$$= (x_1 - \mu)^2 \cdot p(x_1) + (x_2 - \mu)^2 \cdot p(x_2) + \dots \tag{3.9.22}$$

and the standard deviation of X is

$$\sigma = +\sqrt{\sigma^2} \tag{3.9.23}$$

Note 3.3.

An alternative (and sometimes easier) way to calculate the variance of a discrete random variable is:

$$\sigma^2 = \mathbb{E}(X^2) - \big(\mathbb{E}(X)\big)^2 \tag{3.9.24}$$

$$= \left(\sum_{\forall i} x^2 \cdot p(x_i) \right) - \mu^2 \tag{3.9.25}$$

$$= \left(x_1^2 \cdot p(x_1) + x_2^2 \cdot p(x_2) + \dots \right) - \mu^2 \tag{3.9.26}$$

Example 3.8.
Consider the following probability mass function for random variable X.

$X = x$	1	2	3	4	5	6	7	8
$P(X = x)$	0.05	0.02	0.07	0.36	0.33	0.06	0.05	0.06

(**a**) Is this a proper probability mass function?
In order to determine if this is a proper probability mass function, we need to make sure the following two properties hold.

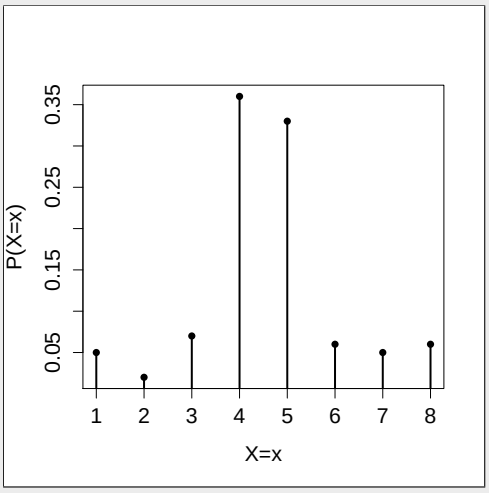
1. $0 \leq p(x_i) \leq 1$.
2. $\sum_{i=1}^{8} p(x_i) = 1$.

Through observation of the table, it is clear that $0 \leq p(x_1), \ldots, p(x_8) \leq 1$ and

$$\sum_{i=1}^{8} p(x_i) = 0.05 + 0.02 + \ldots + 0.06 = 1$$

Both properties hold and therefore this is a proper probability mass function.

(**b**) Sketch the mass function for X.



(**c**) Find the mean of X.

$$\begin{aligned}\mu = \mathbb{E}(X) &= \sum_{i=1}^{8} x_i \cdot p(x_i) \\ &= 0.05 \times 1 + 0.02 \times 2 + 0.07 \times 3 + \ldots + 0.05 \times 7 + 0.06 \times 8 \\ &= 4.58\end{aligned}$$

(**d**) Find the variance of X.

$$\begin{aligned}
\sigma^2 &= \mathbb{E}\Big((X-\mu)^2\Big) \\
&= \sum_{i=1}^{8}(x_i-\mu)^2 \cdot p(x_i) \\
&= (1-4.58)^2 \times 0.05 + (2-4.58)^2 \times 0.02 + (3-4.58)^2 \times 0.07 \\
&+ \ldots + (7-4.58)^2 \times 0.05 + (8-4.58)^2 \times 0.06 \\
&= 0.6408 + 0.1331 + 0.1747 + \ldots + 0.2928 + 0.7018 \\
&= 2.2436
\end{aligned}$$

Alternatively,

$$\sigma^2 = E(X^2) - \Big(E(X)\Big)^2$$

From part (b), we know $E(X) = 4.58$ and therefore $\Big(E(X)\Big)^2 = 4.58^2 = 20.9764$.

$$\begin{aligned}
E(X^2) &= \sum_{i=1}^{8} x_i^2 \cdot p(x_i) \\
&= 1^2 \times 0.05 + 2^2 \times 0.05 + 3^2 \times 0.05 + \ldots + 7^2 \times 0.05 + 8^2 \times 0.06 \\
&= 0.05 + 0.08 + 0.63 + 5.76 + 8.25 + 2.16 + 2.45 + 3.84 \\
&= 23.22 \\
\Rightarrow \sigma^2 &= 23.22 - 20.9764 = 2.2436 \quad \text{as before.}
\end{aligned}$$

(**e**) Find the standard deviation of X.

$$\sigma = +\sqrt{\sigma^2} = \sqrt{2.2436} = 1.498$$

Example 3.9.

Consider the following probability mass function.

X=x	-2	-1	0	1	2
P(X=x)	0.3	0.2	a	0.1	0.15

(**a**) What is a? In order to be a valid probability mass function,

$$\sum_{\forall i} p(x_i) = 1$$

Therefore we know that

$$0.3 + 0.2 + a + 0.1 + 0.15 = 1$$

Rearranging for a we find

$$a = 1 - 0.75 = 0.25$$

(**b**) Find μ.

$$\mu = \sum_{\forall i} x_i \cdot p(x_i) = (-2)(0.3) + (-1)(0.2) + 0 + (1)(0.1) + (2)(0.15) = -0.40$$

(**c**) Find σ^2.

$$\sigma^2 = E(X^2) - \Big(E(X)\Big)^2$$

$$E(X^2) = \sum_{\forall i} x_i^2 \cdot p(x_i) = (-2)^2(0.3) + (-1)^2(0.2) + \ldots + (2)^2(0.15) = 2.1$$

$$\sigma^2 = 2.1 - (-0.40)^2 = 1.94$$

3.9.2 Continuous Random Variable

A continuous random variable is a random variable that can take any value within a specified range. The support of a continuous variable has infinite precision meaning that we can make our measurement as fine as we would like.

3.9.2.1 Probability Density Function

Recall that in Section 3.9.1 a discrete random variable was associated with a probability mass function. We associate continuous random variables with a *probability density function* (PDF) or density function for short.

When we sketched the probability mass function, we got point masses at the values of the random variable. With continuous random variables, the density function is a continuous curve where the domain is the support of the random variable.

Properties 3.2.
Let X be a continuous random variable with support $[\alpha, \beta]$.
We say that $f(x)$ is the probability distribution function of X if it satisfies the following properties:

1. $f(x) \geq 0, \quad \forall x \in [\alpha, \beta]$

2. $P(a \leq x \leq b) = \int_a^b f(x)\,dx$

3. $\int_{-\infty}^{+\infty} f(x)\,dx = 1$

4. $P(x = c) = 0, \quad \forall c \in \mathbb{R}$

If the reader is familiar with calculus, recall that that the definite integral of a function $f(x)$ from limits a to b gives the area between the curve and the x-axis. With a density function the area under the curve gives us the probability. A deeper understanding of density functions requires a proper knowledge of integral calculus. For this introductory course a knowledge of calculus is not required however we will provide the definition of the mean and variance of a continuous random variable for completeness.

3.9.2.2 Mean, Variance and Standard deviation of a Continuous Random Variable

Definition 3.15 (Mean of a Continuous Random Variable)**.**
Let X be a continuous random variable with density function $f(x)$. The mean of X is given by:

$$\mu = \mathbb{E}(X) = \int_{-\infty}^{+\infty} x \cdot f(x)\,dx \tag{3.9.27}$$

Definition 3.16 (Variance and Standard Deviation of a Continuous RV)**.**
Let X be a continuous random variable with density function $f(x)$. The variance of X is

calculated as

$$Var(X) = \sigma^2 = \mathbb{E}[\big(X - \mathbb{E}(X)\big)^2] \tag{3.9.28}$$

$$= \int_{-\infty}^{+\infty} (x-\mu)^2 \cdot f(x)\, dx \tag{3.9.29}$$

and the standard deviation of X is

$$\sigma = +\sqrt{\sigma^2} \tag{3.9.30}$$

Example 3.10.

Suppose X is a continuous random variable with support $[0, 1]$.

(**a**) Is $f(x) = 3x^2$ a valid pdf for X?

$$\int_{-\infty}^{+\infty} f(x)dx = \int_0^1 3x^2 dx = \frac{3x^3}{3}\Big|_{x=0}^{x=1} = 1^3 - 0^3 = 1$$

This is a valid pdf.

(**b**) What is the mean of X?

$$\mu = \mathbb{E}(X) = \int_0^1 x \times 3x^2 dx = \int_0^1 3x^3 dx = \frac{3x^4}{4}\Big|_{x=0}^{x=1} = \frac{3}{4}(1^4) - \frac{3}{4}(0^4) = \frac{3}{4}$$

(**c**) What is the variance of X?

$$\mathbb{E}(X^2) = \int_0^1 x^2 \times 3x^2 dx = \frac{3x^5}{5}\Big|_{x=0}^{x=1} = \frac{3}{5}$$

$$\Rightarrow Var(X) = \mathbb{E}(X^2) - \Big(E(X)\Big)^2 = \frac{3}{5} - \left(\frac{3}{4}\right)^2 = \frac{3}{80}$$

(**d**) What is $P(X \geq 0.5)$?

$$P(X \geq 0.5) = \int_{0.5}^1 3x^2 dx = x^3\Big|_{x=0.5}^{x=1} = 1 - 0.5^3 = 0.875$$

Chapter 4

Distributions of Random Variables

4.1 Distributions of Discrete Random Variables

4.1.1 Bernoulli Distribution

A *Bernoulli* random variable is a discrete random variable that has exactly two possible outcomes: *success* or *failure*. The terms *success* and *failure* are relative to the problem being analyzed. For example, in a problem regarding quality control we may consider obtaining a faulty part as a success (while in real life this is typically not regarded as being a good thing).

The mass function of a Bernoulli random variable is given in mass function 4.1 below. If X is a Bernoulli random variable then X takes the value of 1 with probability of success p and the value of 0 with probability $1 - p$.

Probability Mass Function 4.1. *Let $X \sim$ Bernoulli(p). The mass function of X is*

$$P(X = x) = p^x(1-p)^{1-x}, \quad x = 0, 1 \tag{4.1.1}$$

where p represents the probability of success.

Another way to represent the Bernoulli Distribution is

$$P(X = x) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \end{cases} \tag{4.1.2}$$

This distribution can also be expressed in tabular form

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

The simplest example of a Bernoulli random variable is the outcome of tossing a fair coin once. If we consider a success to be getting tails and a failure to be getting heads, the probability of a success is $p = 0.5$ and the probability of a failure is $1 - p = 1 - 0.5 = 0.5$.

Definition 4.1 (Mean and Variance of a Bernoulli Random Variable)**.** *Let $X \sim$ Bernoulli (p). The mean of X is*

$$E(X) = \mu = p \tag{4.1.3}$$

and the variance of X is

$$Var(X) = \sigma^2 = p\ (1-p) \tag{4.1.4}$$

4.1.2 Binomial Distribution

The *binomial* distribution generalizes the Bernoulli distribution from a single trial to many trials. We use this distribution when we n independent Bernoulli trials occur and we are interested in x successes occurring in those n trials. The probability of success is denoted as p and it is the same each of the n trials. If X is a binomial random variable it is denoted as $X \sim \text{Bin}(n, p)$ where n is the number of trials and p is the probability of a success on any trial. The mass function of a binomial random variables involves the binomial coefficient from definition 3.10.

Probability Mass Function 4.2. *Let $X \sim Bin(n, p)$. The probability of observing x successes in these n independent trials is given by*

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \tag{4.1.5}$$

where

n represents the number of trials,

x represents the number of successes,

p represents the probability of success on any given trial,

$\binom{n}{x} = \dfrac{n!}{x!(n-x)!}$ *is the binomial coefficient.*

Let's examine mass function 4.2 closely. We have a total of n trials in which we are interested in x successes. As the n trials have only two possible outcomes, if x of these trials are successes then the remaining $n - x$ trials must be failures. Each success occurs with probability p and each failure occurs with probability $1 - p$. The expression $p^x(1-p)^{n-x}$ is the probability of obtaining x successes and $n - x$ failures. The binomial coefficient $\binom{n}{x}$ accounts for all possible combinations of x successes and $n - x$ failures in n trials.

A simple example of a binomial random variable is tossing a coin 10 times and calculating the probability of obtaining 6 heads on any of these trials.

Definition 4.2 (Mean and Variance of a Binomial Random Variable)**.** *Let $X \sim Bin(n, p)$. The mean of X is*

$$E(X) = \mu = np \tag{4.1.6}$$

and the variance of X is

$$Var(X) = \sigma^2 = np\ (1-p) \tag{4.1.7}$$

Example 4.1.
Suppose $X \sim Bin(10, 0.35)$.

(**a**) What is the mean of X?

$$\mu = \mathbb{E}(X) = np = 10 \times 0.35 = 3.50$$

(**b**) What is the variance of X?

$$\sigma^2 = np(1-p) = 10 \times 0.35(1-0.35) = 10 \times 0.2275 = 2.275$$

(**c**) What is the probability that $X = 5$?

The probability mass function of $X \sim Bin(10, 0.35)$ is

$$P(X = x) = \binom{10}{x}\ (0.35)^x\ (0.65)^{10-x}$$

Plugging in $x = 5$,

$$P(X = 5) = \binom{10}{5}\ (0.35)^5\ 0.65^{10-5} = 252 \times (0.35)^5 \times (0.65)^5 = 0.154$$

(**d**) What is the probability that $X \leq 5$?

$$\begin{aligned}
P(X \leq 5) &= P(X=5) + P(X=4) + P(X=3) + P(X=2) + P(X=1) + P(X=0) \\
&= \sum_{x=0}^{5} \binom{10}{x}\ 0.35^x\ 0.65^{10-x} \\
&= \binom{10}{0}\ 0.35^0\ 0.65^{10} + \binom{10}{1}\ 0.35^1\ 0.65^9 + \binom{10}{2}\ 0.35^2\ 0.65^8 \\
&+ \binom{10}{3}\ 0.35^3\ 0.65^7 + \binom{10}{4}\ 0.35^4\ 0.65^6 + \binom{10}{5}\ 0.35^5\ 0.65^5 \\
&= 0.013 + 0.072 + 0.176 + 0.252 + 0.238 + 0.154 \\
&= 0.905
\end{aligned}$$

(**e**) What is the probability that $X > 5$?

$P(X > 5)$ is the complement of $P(X \leq 5)$.

$$\Rightarrow P(X > 5) = 1 - P(X \leq 5) = 1 - 0.905 = 0.095$$

(**f**) What is the probability that $X \geq 5$?

$$P(X \geq 5) = P(X > 5) + P(X = 5) = 0.095 + 0.154 = 0.249$$

(**g**) What is the probability that $3 \leq X \leq 6$?

$$P(3 \leq X \leq 6) = P(X = 3) + P(X = 4) + P(X = 5) + P(X = 6)$$

From part (d), we know

$$P(X = 5) = 0.154, \quad P(X = 4) = 0.238, \quad P(X = 3) = 0.252$$

We only need to find $P(X = 6)$.

$$P(X = 6) = \binom{10}{6} 0.35^6 \, 0.65^4 = 0.069$$
$$\Rightarrow P(3 \leq X \leq 6) = 0.252 + 0.238 + 0.154 + 0.069 = 0.713$$

4.2 Distributions of Continuous Random Variables

4.2.1 Continuous Uniform Distribution

A *continuous random variable* follows a *uniform* probability distribution if any value within a defined interval is equally likely. If a random variable X is distributed uniformly on interval $[a, b]$ we denote this as $X \sim \mathrm{U}(a, b)$. The distribution function of a uniform random variable is very simple.

Probability Mass Function 4.1. ———————————————————

Let $X \sim U(a, b)$. The density function of X is

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

Definition 4.3 (Mean and Variance of a Uniform Random Variable)**.** *Let $X \sim U(a, b)$. The mean of X is*

$$\mathbb{E}(X) = \mu = \frac{a+b}{2} \tag{4.2.8}$$

and the variance of X is

$$Var(X) = \sigma^2 = \frac{(b-a)^2}{12} \tag{4.2.9}$$

4.2.2 Normal Distribution

The *Normal* distribution (also referred to as the *Gaussian* distribution) is an extremely important distribution in statistics and many other areas including mathematics, econometrics, finance, physics, astronomy etc. The normal distribution models many real life processes and events. This distribution is the bell curve that we are used to seeing. If X is a random variable which follows a normal distribution we denote this as $X \sim N(\mu, \sigma^2)$ where μ and σ^2 are the mean and standard deviation respectively

Probability Mass Function 4.2.
Let $X \sim N(\mu, \sigma^2)$. The density function of X is

$$f(x) = \frac{1}{(\sqrt{2\pi})\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad -\infty < x < +\infty \tag{4.2.10}$$

The normal distribution is completely described by the mean μ and standard deviation σ. Note that the support of the normal distribution is the entire real number line (i.e. the support is infinite).

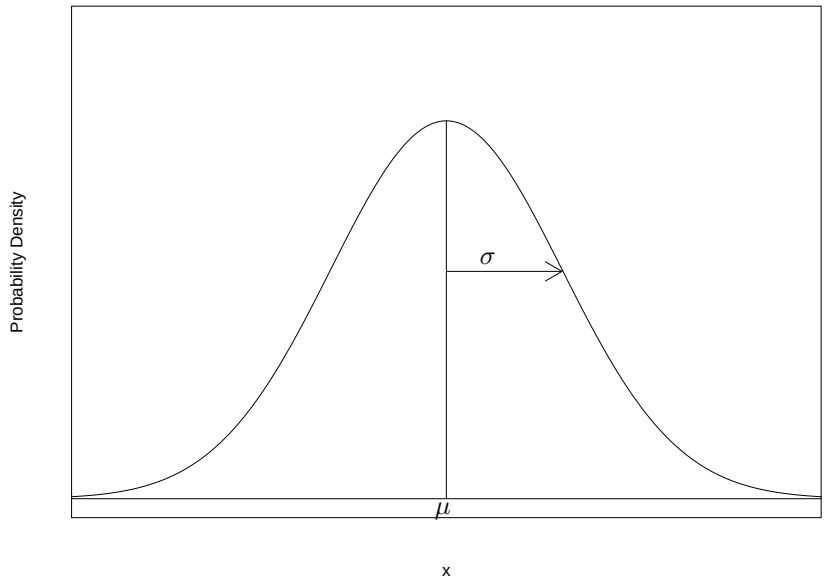


Figure 4.1: The normal distribution

The normal distribution is symmetric about μ (i.e. 0.5 probability below μ and 0.5 probability above μ). The mean, median and mode are all the same value of μ. Some examples of normal distributions with different means and standard deviations are given in Figure 4.2.2.

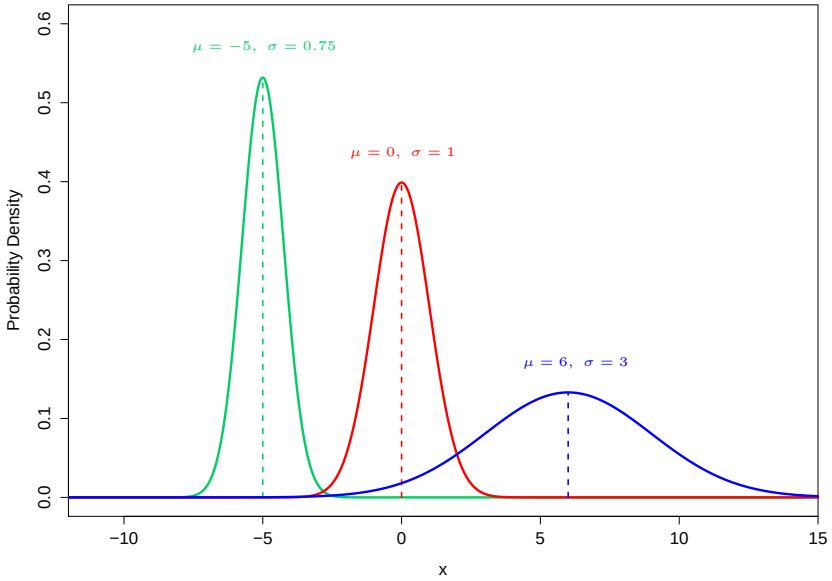


Figure 4.2: Examples of normal distributions with different means and standard deviations

4.2.2.1 The Standard Normal Distribution

The *standard normal* distribution is a normal distribution with $\mu = 0$ and $\sigma = 1$.

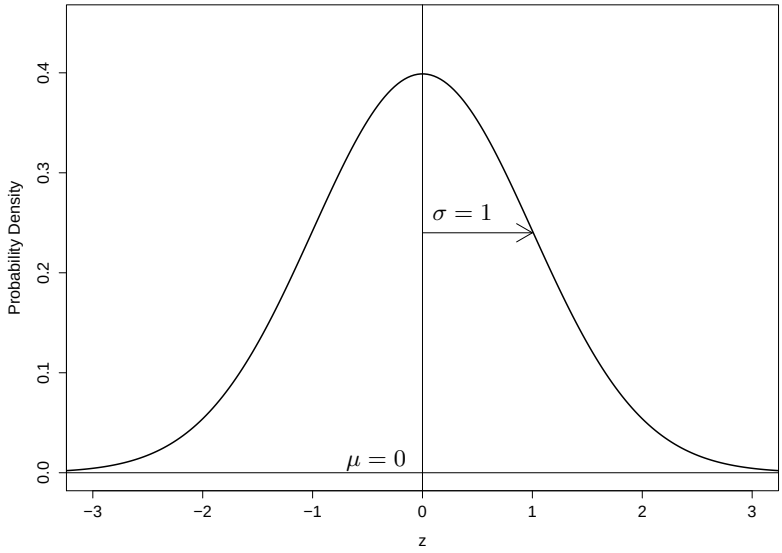


Figure 4.3: The standard normal distribution.

A random variable that follows a standard normal distribution is usually associated with the letter Z and is denoted as $Z \sim N(0,\ 1)$. Probabilities under the standard normal curve are in a table at the back of this text (Appendix ???: Table ???. An equivalent copy of this table is also on courselink). It is extremely important to familiarize yourself with this table.

A normal distribution can have any mean and any standard deviation, so there are infinitely many of them. We transform a probability question about a random variable that follows any normal distribution in to an equivalent problem in terms of the standard normal distribution. This way we have a single standard tool (i.e. the standard normal table) which we can use to solve problems involving a normal distribution.

Definition 4.4 (Z score)**.** *The Z score of an observation is the number of standard deviations by which it falls above or below the mean.*

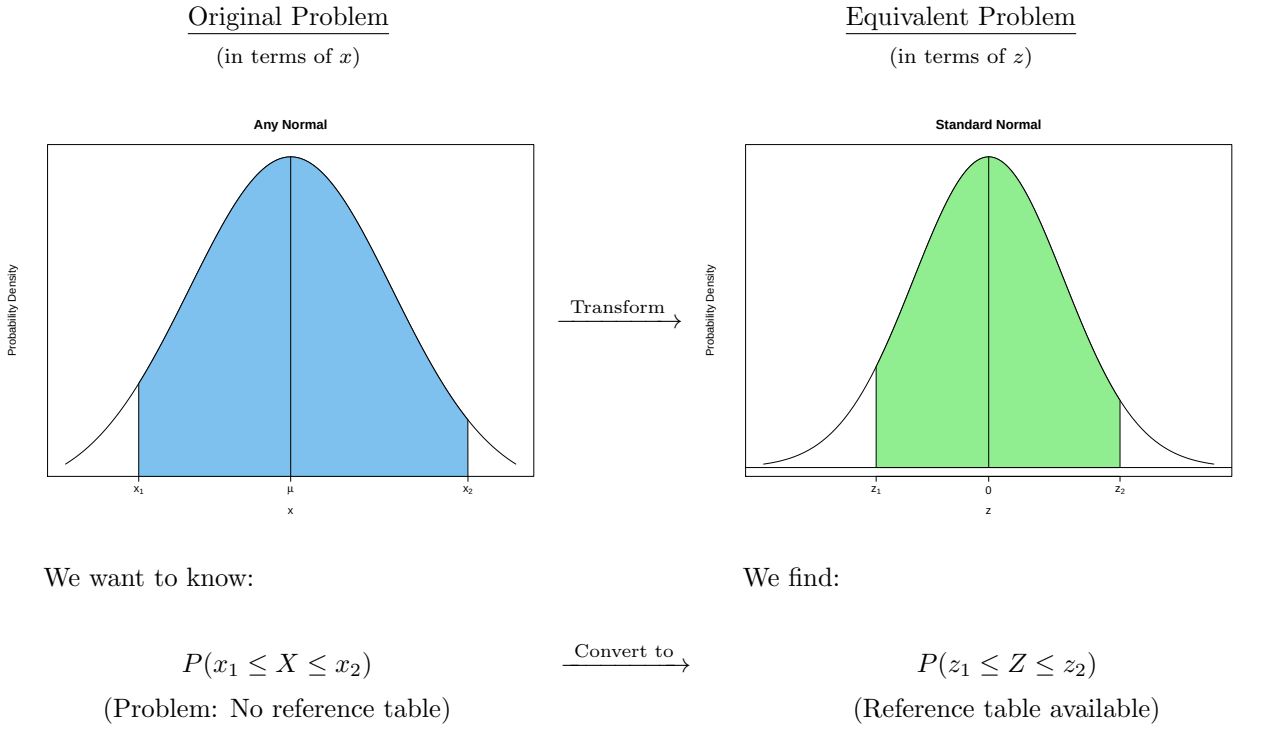


Figure 4.4: Transforming a problem for random variable $X \sim N(\mu, \sigma^2)$ into an equivalent problem in terms of $Z \sim N(0, 1)$.

In order to transform a normal random variable, $X \sim N(\mu, \sigma^2)$ into a standard normal variable, $Z \sim N(0, 1)$, we compute the Z score for an observation x using transformations 4.1 and 4.2.

Transformation 4.1.

Let $X \sim N(\mu, \sigma^2)$. The Z score for a single observation x is

$$z = \frac{x - \mu}{\sigma} \tag{4.2.11}$$

Transformation 4.2.

The Z score for an average is Let $X \sim N(\mu, \sigma^2)$.

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \tag{4.2.12}$$

In Transformation 4.2 the denominator $(\sigma/\sqrt{n})$ is referred to as the standard error of the

mean. This transformation will be used to describe sampling distributions which will be discussed in the following chapter.

The *Empirical Rule*, also referred to as the "68-95-99/7" rule or the three-sigma rule, is a quick rule of thumb for determining the probability of falling within 1, 2, and 3 standard deviations of the mean of the normal distribution. This rule also works for well for any distribution that is approximately bell shaped. It is useful in settings where would like to make quick estimates and do not have access to a calculator or normal tables.

Rule 4.1 (Empirical rule)**.**

Let X be a random variable with a probability distribution that is approximately bell-shaped. Then

Approximately 68% of values lie within 1 standard deviation of the mean
Approximately 95% of values lie within 2 standard deviations of the mean
Approximately 99.7% of values lie within 3 standard deviations of the mean

Formally,

$$P(\mu - \sigma \leq x \leq \mu + \sigma) \approx 0.68 \tag{4.2.13}$$
$$P(\mu - 2\sigma \leq x \leq \mu + 2\sigma) \approx 0.95 \tag{4.2.14}$$
$$P(\mu - 3\sigma \leq x \leq \mu + 3\sigma) \approx 0.997 \tag{4.2.15}$$

Example 4.2.

If $X \sim N(5, 4)$,

(**a**) Find the Z-score of X.

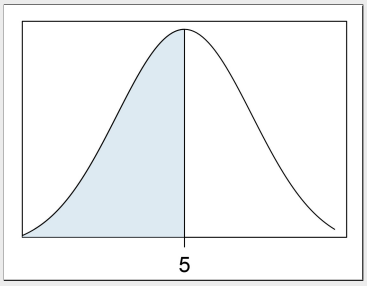
$$z = \frac{x - \mu}{\sigma} = \frac{x - 5}{2}$$

(**b**) Find $P(X = 5)$.
Since X is a continuous random variable, the probability of a single point is zero.

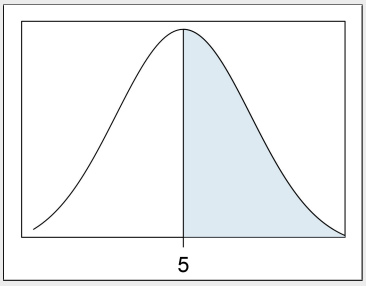
$$\Rightarrow P(X = 5) = 0$$

(**c**) Find $P(X \leq 5)$ and $P(X \geq 5)$.
Since the mean of X is $\mu = 5$ and the normal distribution is symmetric,

$$P(X \leq 5) = P(X \geq 5) = 0.5$$



$P(X \leq 5)$



$P(X \geq 5)$

(**d**) Find $P(X \leq 3)$.

First transform X into the Z statistic. When $X = 3$,

$$z = \frac{x-5}{2} = \frac{3-5}{2} = -1$$

Therefore,

$$P(X \leq 3) = P(Z \leq -1)$$

Using our standard normal table,

$$P(Z \leq -1) = 0.159 \Rightarrow P(X \leq 3) = 0.159$$

(**e**) Find $P(X \leq 7)$.

When $X = 7$,

$$z = \frac{x-5}{2} = \frac{7-5}{2} = 1$$

Therefore,

$$P(X \leq 7) = P(Z \leq 1)$$

Using our standard normal table,

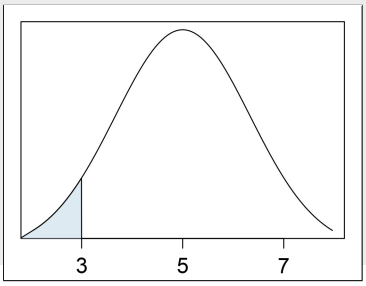
$$P(Z \leq 1) = 0.841 \Rightarrow P(X \leq 7) = 0.841$$

(**f**) Find $P(3 \leq X \leq 7)$.

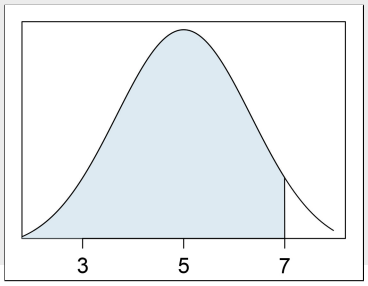
We want to find the probability that X is greater than 3 but less than 7. From parts (b) and (c) we know

$$P(X \leq 3) = 0.159, \quad P(X \leq 7) = 0.841$$

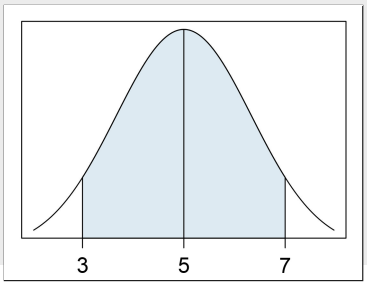
$P(X \leq 3)$ gives us the probability of X being any number between $-\infty$ and 3. $P(X \leq 7)$ gives us the probability of X being any number between $-\infty$ and 7. If we subtract $P(X \leq 3)$ from $P(X \leq 7)$, we will be left with the probability of X being between 3 and 7. Graphically,



$P(X \leq 3)$



$P(X \leq 7)$



$P(3 \leq X \leq 7)$

$$\Rightarrow P(3 \leq X \leq 7) = P(X \leq 7) - P(X \leq 3)$$
$$= 0.841 - 0.159 = 0.682$$

Alternatively, we could recognize that $\mu - \sigma = 5 - 2 = 3$ and $\mu + \sigma = 5 + 2 = 7$. By the *three-sigma rule* we know that $P(\mu - \sigma \leq x \leq \mu + \sigma) \approx 0.68$.

(**g**) If $n = 25$, find the Z-score of $\bar{X}$.

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{x} - 5}{2/\sqrt{25}} = \frac{\bar{x} - 5}{2/5} = \frac{\bar{x} - 5}{0.40}$$

(**h**) Find $P(\bar{X} \leq 4 \cup \bar{X} \geq 6)$.
Recall the *additive property*,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Since it is impossible for $\bar{X}$ to be less than 4 and greater than 6 at the same time,

$$P(\bar{X} \leq 4 \cap \bar{X} \geq 6) = 0$$

Therefore,

$$P(\bar{X} \leq 4 \cup \bar{X} \geq 6) = P(\bar{X} \leq 4) + P(\bar{X} \geq 6)$$

Plugging in $\bar{X} = 4$ and $\bar{X} = 6$ into our Z-score equation from part (g) yields

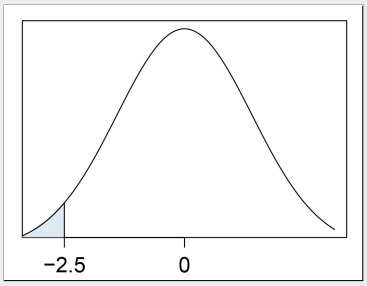
$$z = \frac{\bar{x} - 5}{0.40} = \frac{4 - 5}{0.40} = -2.5$$
$$\Rightarrow P(\bar{X} \leq 4) = P(Z \leq -2.5)$$

and

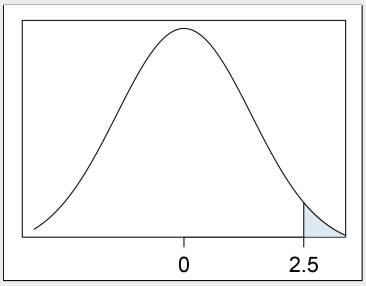
$$z = \frac{\bar{x} - 5}{0.40} = \frac{6 - 5}{0.40} = 2.5$$
$$\Rightarrow P(\bar{X} \geq 6) = P(Z \geq 2.5)$$

Due to the symmetry of the normal distribution,

$$P(Z \geq 2.5) = P(Z \leq -2.5)$$



$P(Z \leq -2.5)$



$P(Z \geq 2.5)$

Using our standard normal table,

$$\begin{aligned} P(\bar{X} \leq 4 \cup \bar{X} \geq 6) &= P(\bar{X} \leq 4) + P(\bar{X} \geq 6) \\ &= P(Z \leq -2.5) + P(Z \leq -2.5) \\ &= 0.00621 + 0.00621 \\ &= 0.01242 \end{aligned}$$

Example 4.3.

The corn ethanol industry is worried that unfavourable weather will lead to low crop yields in the next year. Researchers determine that yearly North American corn yield under similar weather conditions is normally distributed with a mean of 17.5 billion bushels and a standard deviation of 10 billion bushels.

(**a**) The industry puts out a statement saying that crop yield needs to remain at or above 15 billion bushels to support current demand. What is the probability that crop yield drops below 15 billion bushels?

First we find our Z statistic.

$$z = \frac{x - \mu}{\sigma} = \frac{15 - 17.5}{10} = -0.25$$

Using our standard normal table,

$$P(X < 15) = P(Z < -0.25) = 0.4013$$

The probability that crop yield drops below 15 billion bushels is 0.4013.

(**b**) If corn yields do fall below levels needed to support current demand, the industry would like to know by how much. What is the probability that corn yields are greater than 10 billion bushels given they are less than 15 billion bushels?

We are interested in finding $P(X > 10|X < 15)$. Via Bayes' Theorem,

$$P(X > 10|X < 15) = \frac{P(X > 10 \cap X < 15)}{P(X < 15)} = \frac{P(10 < X < 15)}{P(X < 15)}$$

From part (a) we know that $P(X < 15) = 0.4013$. Then

$$\begin{aligned} P(10 < X < 15) &= P(X < 15) - P(X < 10) = 0.4013 - P\left(Z < \frac{10 - 17.5}{10}\right) \\ &= 0.4013 - P(Z < -0.75) = 0.4013 - 0.2266 = 0.1747 \end{aligned}$$

Plugging this into Bayes' Theorem,

$$P(X > 10|X < 15) = \frac{P(10 < X < 15)}{P(X < 15)} = \frac{0.1747}{0.4013} = 0.4353$$

(**c**) If 3 years are randomly sampled, what is the probability mean corn yield of these 3 years is between 10 billion and 20 billion bushels?

We need to find $P(10 \leq \bar{X} \leq 20)$.

$$P(10 \leq \bar{X} \leq 20) = P(\bar{X} \leq 20) - P(\bar{X} \leq 10) \tag{4.2.16}$$

For $\bar{x} = 20$,

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{20 - 17.5}{10/\sqrt{3}} \approx 0.43$$

Using our standard normal table,

$$P(Z < 0.43) = 0.6664$$

For $\bar{x} = 10$,

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{10 - 17.5}{10/\sqrt{3}} \approx -1.30$$

Using our standard normal table,

$$P(Z < -1.30) = 0.0968$$

Putting it all together,

$$P(10 \leq \bar{X} \leq 20) = P(Z < 0.433) - P(Z < -1.30) = 0.6664 - 0.0968 = 0.5698$$

4.2.3 $t-$Distribution

The Student's $t-$distribution or t distribution for short, is a continuous family of distributions that are commonly used in statistical inference techniques, particularly when the population standard deviation σ is not known. Generally, the $t-$distribution has a bell shape similar to the normal distribution, however the thickness of the tails of the $t-$distribution varies.

Note 4.1.
The $t-$distribution was developed by William Sealy Gosset while working on quality control for the Guinness Brewery in Dublin, Ireland. He was not allowed to publish his results as Guinness did not want competitors to obtain information about the research conducted at Guinness. Gosset secretly published his work under the pseudonym student.

The $t-$distribution is centered at zero and depends on the *degrees of freedom*. The degrees of freedom, *df*, describe the bell shape form of the $t-$distribution and the thickness of its tails.

Definition 4.5 (Degrees of Freedom)**.**
The degrees of freedom refer to the number of free values that we can vary in the calculation of an estimate.

Example 4.4.
Recall the sample standard deviation is calculated using

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \tag{4.2.17}$$

In order to calculate s we first calculate the mean of the sample, $\bar{x}$, then calculate the sum of the squared deviations of each observations from $\bar{x}$. There are n squared deviations but only $(n-1)$ of them are free to assume any value. This is because one $(x_i - \bar{x})^2$ must include a value of x_i such that $\frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$ that was calculated. All of the other $(n-1)$ values of $(x_i - \bar{x})^2$ can take any value (in theory). As such s is said to have $n-1$ degrees of freedom. For example, if it is known that $n = 3$, $\bar{x} = 5$, $x_1 = 6$, and $x_2 = 5$, x_3 must be fixed at 4 in order for our calculation of $\bar{x} = 5$ to hold.

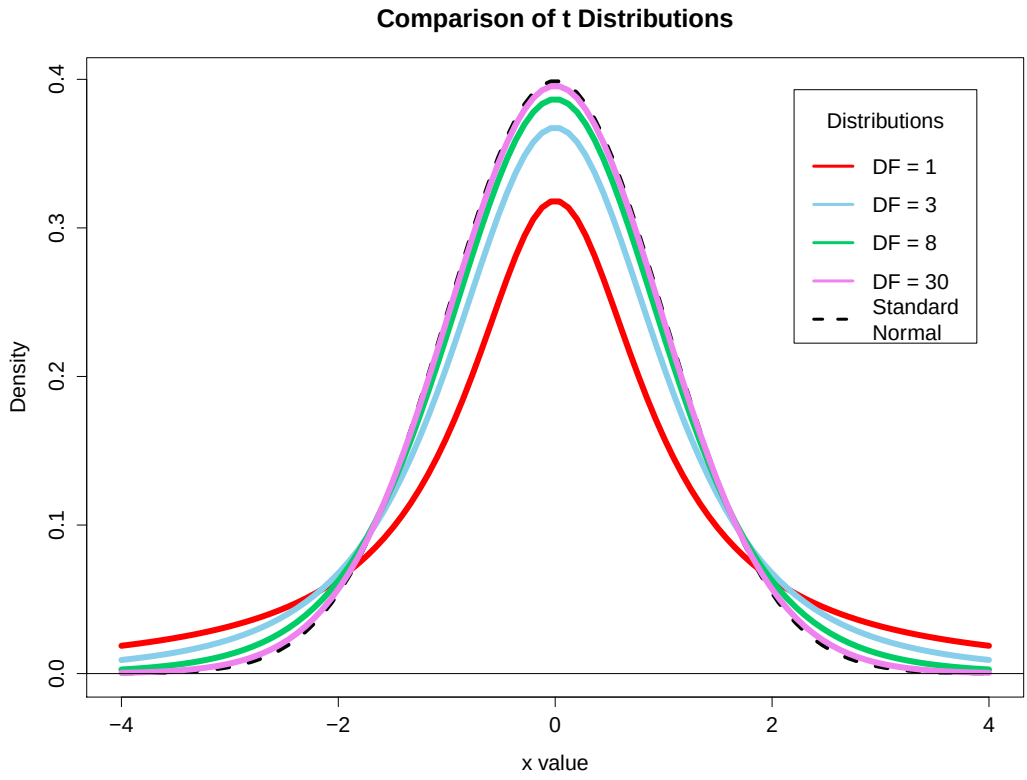


Figure 4.5: Several $t-$distributions for 1,3,8 and 30 degrees of freedom along with the standard normal distribution.

Notice that the $t-$distribution has thicker tails than the normal distribution in general. As the degrees of freedom approach infinity, the t-distribution approaches the standard normal distribution.

Note 4.2.

The degrees of freedom of a distribution can be any positive real value. They do not need to by positive integers.

We use the $t-$table in Appendix ??? to read values in the tails of the $t-$distribution.

Example 4.5.

Assuming the sample is of size $n = 10$, use the t-distribution table to determine the following.

(**a**) Find $P(t \leq 0)$ and $P(t \geq 0)$.
The mean of the t-distribution is 0 and therefore $P(t \leq 0) = P(t \geq 0) = 0.5$.

(**b**) Find $P(t \leq 2.262)$ and $P(t \geq 2.262)$.
Since $n = 10$,

$$df = n - 1 = 10 - 1 = 9$$

To find $P(t \leq 2.262)$, we find the row that matches our degrees of freedom and move right until we find the column containing 2.262. Once we find 2.262, we move up the column until we hit the probability. Using our table yields $P(t \leq 2.262) = 0.975$. Therefore, $P(t \geq 2.262) = 1 - 0.975 = 0.025$.

(**c**) Find $P(t \leq 2)$.
We notice that the value of 2 is not present in the $df = 9$ row. In order to estimate $P(t \leq 2)$, we determine the closest number above and below 2. In this case it would be 1.833 and 2.262. We find the $P(t \leq 1.833)$ and $P(t \leq 2.262)$ and average them to estimate $P(t \leq 2)$.

$$P(t \leq 2) = \frac{0.95 + 0.975}{2} = 0.9625$$

(**d**) Find $P(t \geq -0.883)$.

$$P(t \geq -0.883) = P(t \leq 0.883)$$

Using our table,

$$P(t \leq -0.883) = P(t \geq -0.883) = 0.80$$

(**e**) Find the value q such that $P(t \leq q) = 0.95$.
We find the column for the 95% percentile and match it with the row representing $df = 9$. This value is $q = 1.833$.

(**f**) Find the value q such that $P(t \geq q) = 0.90$.
Due to the symmtery of the t-distribution, we know $P(t \geq q) = P(t \leq -q) = 0.90$. We find the column for the 90% percentile and match it with the row representing $df = 9$. This value is $q = 1.383$. Since $P(t \geq q) = P(t \leq -q)$, $P(t \geq -1.383) = 0.90$.

(**g**) Find the value q such that $P(t \leq q) = 0.825$.
We notice that there is no 82.5% percentile column in our table. Therefore we go to the nearest percentile above and below the one of interest. That is, 80% and 85%. We then take the average of the two numbers corresponding to these percentiles in the $df = 9$ row to approximate q,

$$q = \frac{0.883 + 1.100}{2} = 0.9915$$

Chapter 5

Foundations of Inference

In section 4.2.2, we learnt how to solve problems involving normally distributed populations. In this chapter we will discuss the tools we have for conducting statistical inference on populations that are not normally distributed.

5.1 Sampling Distribution

The concept of sampling distributions is very important for many statistical inference techniques.

Definition 5.1 (Point Estimate)**.** *A point estimate is a single value (i.e. a single point on the real number line) that estimates the value of a parameter.*

In other words, a point estimate is a single value that can be regarded as the best guess of a parameter. A point estimate of μ is $\bar{x}$ and a point estimate of σ is s.

Definition 5.2 (Sampling Distribution)**.** *The distribution of a statistic is referred to as the sampling distribution of the statistic.*

We are particularly interested in the sampling distribution of the sample mean. By this we mean that we are interested in the distribution of $\bar{x}$.

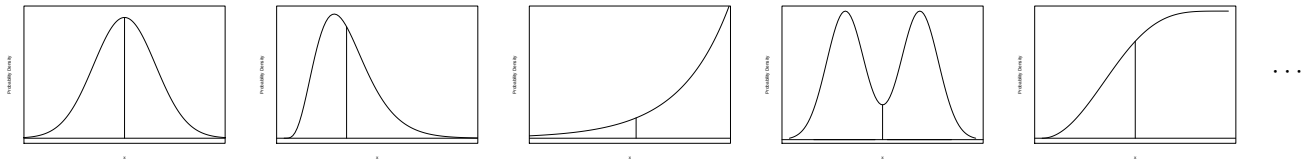
5.2 Central Limit Theorem

The *Central Limit Theorem* is an extremely important theorem related to sampling distributions and is of fundamental importance to inferential techniques on the population mean.

Theorem 5.1.

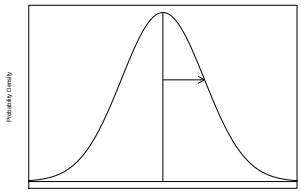
Consider m random samples, each of fixed size n, drawn from a population (following any distribution) with mean μ and standard deviation σ. When m, the number of samples drawn, is sufficiently large the sampling distribution of the sample mean $\bar{x}$ is approximately normally distributed with mean $\mu_{\bar{x}} = \mu$ and standard deviation $\sigma_{\bar{x}} = \sigma/\sqrt{n}$. That is, $\bar{x} \sim N(\mu,\ \sigma^2/n)$.

Population Distribution:



Sample 1	Sample 2	Sample 3	⋯	⋯	⋯
-	-	-	⋯	⋯	⋯
-	-	-	⋯	⋯	⋯
⋮	⋮	⋮	⋯	⋯	⋯
-	-	-	⋯	⋯	⋯
$\bar{x}_1$	$\bar{x}_2$	$\bar{x}_3$	⋯	⋯	⋯

Distribution of $\bar{x}$'s:



Mean : $\mu_{\bar{x}} = \mu$

Dev : $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$

Figure 5.1: Implementation of the Central Limit Theorem for creating the sampling distribution of $\bar{x}$.

$\sigma/\sqrt{n}$ is referred to as the *standard deviation of the sample mean*. This is the standard deviation of $\bar{x}$'s estimate of μ. If we did not know σ then we would calculate the standard error of the sample mean using s (the sample standard deviation). We will learn more about these terms in section 5.3. For a better understanding of the manner in which sampling distributions are created it is recommended that the reader should visit an application that can be found at *http://onlinestatbook.com/stat_sim/sampling_dist/index.html*. This java applet can be used to visualize various aspects of sampling distributions and how they are affected by the sample size n.

Note 5.1.
In order to use the URL above, Java should be installed and enabled in your browser. This application may not work with some browsers (in particular Google Chrome). There are some stability issues and this application may cause your browser to crash, however your computer will not be harmed.

Example 5.1.
Assume $X \sim N(500, 90)$. Suppose we take 1000 samples of size $n = 25$ from this population. Use the central limit theorem to find the distribution of $\bar{x}$.

Solution:

The mean of $\bar{x}$ is

$$\mu_{\bar{x}} = \mu = 500$$

The standard deviation of $\bar{x}$ is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{90}}{\sqrt{25}} = 6$$

$$\Rightarrow \bar{x} \sim N(500, 36)$$

5.3 Introduction to Inference

Statistical inference is the study of estimating parameters and quantifying the degree of certainty of each estimate. One of the most important assumptions we make in many inference techniques is the following.

Assumption 5.1 (Important Assumption for Inference)**.**
The value of a parameter is fixed.

By assuming the value of a parameter to be fixed, we are assuming that the value of the parameter of interest does not change (at least for the time frame that we are interested in). For instance suppose we are interested in the average salary of all Canadians for the current year. By assumption 5.1 it does not matter whether we draw a sample on January 1^{st} of this year, or on December 31^{st} of this year or any day in between, the average salary for all Canadians for the current year will not change.

Recall that a point estimate is our best single guess for the estimate of a parameter. However due to the nature of randomness point estimates will likely not be exactly equal to the parameter that they are estimating. Furthermore the value of a point estimate will vary from sample to sample. Thus it is important we are able to quantify the accuracy of point estimates.

The standard deviation of the sampling distribution of an estimate (or statistic) indicates how much an estimate deviates from the actual population parameter. In other words, this standard deviation describes the typical error associated with point estimates. As a result, we usually refer to this standard deviation as the standard error (SE) of an estimate.

Definition 5.3 (Standard Error of an Estimate)**.**
The standard error of a point estimate is its associated standard deviation.

The standard error is used in constructing confidence intervals in chapter 6 and conducting hypothesis tests in chapter 7.

Note 5.2.
Although estimates usually are not exactly equal to the parameter they are estimating, they become more accurate as more data becomes available (i.e. as the sample size increases). As such we get a smaller standard error as the sample size increases.

Chapter 6

Confidence Intervals

6.1 Introduction

We learnt in chapter 5 that a point estimate is the best single guess for the numeric value of a parameter and that due to the nature of randomness, a point estimate will likely not be exactly equal to the parameter that it is estimating. Using properties of sampling distributions, we can create an interval around our point estimate which we believe will capture our parameter with a certain level of confidence. This interval is referred to as a *confidence interval* for the parameter of interest.

Definition 6.1 (Confidence Interval)**.** *A confidence interval is a plausible range of values that captures a parameter with a quantified degree of confidence.*

Suppose we are interested in the average mark for STAT*2060 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:

$$\textit{estimator} \pm \underbrace{\left(\begin{array}{c}\textit{value from reference}\\ \textit{distribution}\end{array}\right) \times \left(\begin{array}{c}\textit{standard error}\\ \textit{of estimate}\end{array}\right)}_{\textit{margin of error}} \tag{6.1.1}$$

The value from the reference distribution in the skeleton above will be either a value from the standard normal distribution discussed in section 4.2.2 or the Student t-distribution discussed in section 4.2.3. The margin of error (MOE) can be considered as the distance around our estimator in which the true value of the parameter of interest will be found, with a specified level of confidence.

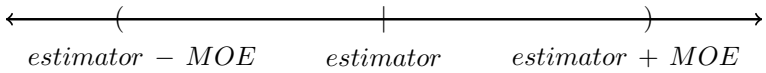


Figure 6.1: Visualization of a confidence interval on the real number line. The margin of error is abbreviated as MOE. The estimator is the centre of the interval. The confidence interval consists of all values between the estimator$-MOE$ and the estimator$+MOE$.

The exact form of a confidence interval depends on the information we have and the parameter we are estimating.

Note 6.1.

The confidence intervals that we will be constructing are referred to as "two-sided confidence intervals" as they provide both a lower and upper bound for a plausible range of values that the parameter of interest can take.

One-sided confidence intervals can also be constructed however these types of confidence intervals are not constructed as often as two-sided confidence intervals. One-sided confidence intervals are constructed on rare occasions when a researcher is interested in just a lower or an upper bound for the value of a parameter.

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.

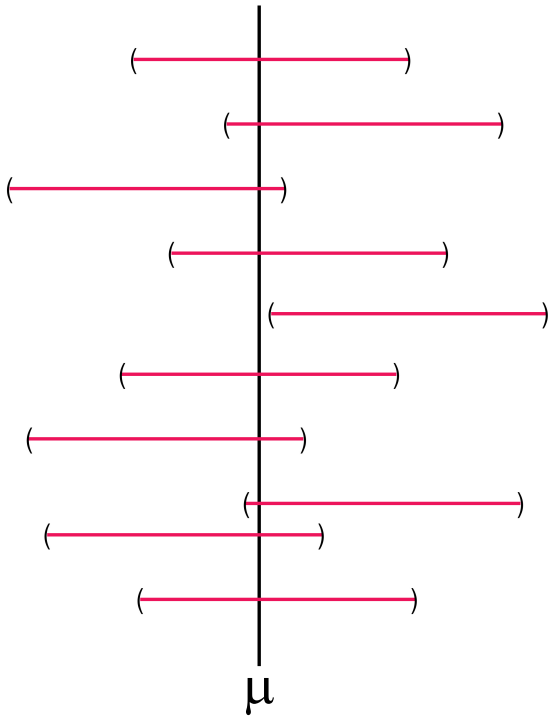


Figure 6.2: Graphical interpretation of a 90% confidence interval for μ. Note that 9 of the 10 constructed intervals capture the true value of μ.

Note 6.2.
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. Recall from assumption 5.1 that 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 One Sample Confidence Intervals

6.3.1 On the Mean

6.3.1.1 When σ is Known

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*

$$\bar{x} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \tag{6.3.2}$$

The $z_{\alpha/2}$ value is obtained from standard normal tables. The standard error in 7.1.4 is $\frac{\sigma}{\sqrt{n}}$ and the margin of error is $z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$.

Example 6.1.

Corporation A's closing stock prices over the past year are assumed to be approximately normally distributed. 10 closing stock prices were sampled from the past year:

49.43 56.47 52.76 56.34 47.67 40.73 49.55 56.12 37.32 53.68

The corporation is interested in estimating μ, the true mean closing stock price over the past year. If the true standard deviation of Corporation A stock prices is known to be $\sigma = 10$, construct a 95% confidence interval for μ.

Solution:

First, we find $\bar{x}$.

$$\begin{aligned}\bar{x} &= \sum_{i=1}^{n}\frac{x_i}{n} = \sum_{i=1}^{10}\frac{x_i}{10}\\ &= \frac{49.43+56.47+52.76+56.34+\ldots+37.32+53.68}{10}\\ &= 50\end{aligned}$$

Second, we find the margin of error.

$$\text{Margin of Error } = z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

For a 95% confidence interval, $\alpha = 0.05$ and $z_{\alpha/2} = z_{0.025} = 1.96$.

$$\Rightarrow \text{Margin of Error} = 1.96\left(\frac{10}{\sqrt{10}}\right) = 1.96 \times 3.16 = 6.19$$

The 95% confidence interval for μ when $\sigma = 10$ is

$$\bar{x} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) = 50 \pm 6.19 = (43.81,\ 56.19)$$

6.3.1.2 When σ is Unknown

When we do not know the population standard deviation σ, we estimate it with the sample standard deviation s and construct a confidence interval for μ in the following manner.

Confidence Interval 6.2 (Confidence Interval on μ when σ is Unknown)**.**
A $(100-\alpha)\%$ confidence interval on μ when σ is unknown is

$$\bar{x} \pm t_{(\alpha/2,\ n-1)}\left(\frac{s}{\sqrt{n}}\right) \tag{6.3.3}$$

The $t_{(\alpha/2,\ n-1)}$ value in 6.3.3 is obtained from the $t-$distribution by referring to the $t-$tables with $n-1$ degrees of freedom. The standard error in 6.3.3 is $s/\sqrt{n}$ and the margin of error is $t_{(\alpha/2,\ n-1)}(s/\sqrt{n})$.

Example 6.2.
Corporation A's closing stock prices over the past year are assumed to be approximately normally distributed. 10 closing stock prices were sampled from the past year:

49.43 56.47 52.76 56.34 47.67 40.73 49.55 56.12 37.32 53.68

The corporation is interested in estimating μ, the true mean closing stock price over the past year. If the true standard deviation of Corporation A stock prices is unknown, construct a 95% confidence interval for μ.

Solution:

Since we do not know σ, we must estimate it using s.

$$\begin{aligned}
s^2 &= \sum_{i=1}^{n}\frac{(x_i-\bar{x})^2}{n-1} = \sum_{i=1}^{10}\frac{(x_i-50)^2}{9} \\
&= \frac{(49.43-50)^2}{9}+\frac{(56.47-50)^2}{9}+\frac{(52.76-50)^2}{9}+\ldots+\frac{(53.68-50)^2}{9} \\
&= \frac{0.32+41.86+7.62+40.20+5.43+85.93+0.20+37.45+160.78+13.54}{9} \\
&= 43.704 \\
\Rightarrow s &= \sqrt{s^2} = \sqrt{43.704} = 6.61
\end{aligned}$$

To calculate the margin of error, we can no longer use $z_{\alpha/2}$ as σ is unknown. Therefore we use $t_{\alpha/2,\ n-1}$.

$$\text{Margin of Error} = t_{(\alpha/2,\ n-1)}\left(\frac{s}{\sqrt{n}}\right)$$

$t_{\alpha/2,\ n-1}$ is the probability that t is greater than $t_{\alpha/2}$ or less than $-t_{\alpha/2}$ for a t-distribution with $n-1$ degrees of freedom. Using our t-distribution table, we find that $t_{0.025,9}$ is 2.26.

$$\Rightarrow \text{Margin of Error} = 2.26\left(\frac{6.61}{\sqrt{10}}\right) = 2.26 \times 2.09 = 4.72$$

The 95% confidence interval for μ when σ is unknown is

$$\bar{x} \pm t_{(\alpha/2,\ n-1)}\left(\frac{s}{\sqrt{n}}\right) = 50 \pm 4.72 = (45.28,\ 54.72)$$

6.3.2 On a Proportion

When we are interested in constructing a confidence interval for a proportion, we use the sample proportion $\hat{p}$.

Confidence Interval 6.3 (Confidence Interval on a Proportion)**.** *A* $(100-\alpha)\%$ *confidence interval on* p *is*

$$\hat{p} \ \pm \ z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{6.3.4}$$

The standard error in 6.3.4 is $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ and the margin of error is $z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

Example 6.3.

Corporation A's closing stock prices over the past year are assumed to be approximately normally distributed. 10 closing stock prices were sampled from the past year:

49.43 56.47 52.76 56.34 47.67 40.73 49.55 56.12 37.32 53.68

Suppose Corporation A is interested in the proportion of closing stock prices during the past year that were greater than fifty dollars. Using the information provided, construct a 95% confidence interval for p.

Solution:

First we need to find $\hat{p}$.

$$\hat{p} = \frac{\text{number of closing stock prices} > \ \$50 \text{ in sample}}{n} = \frac{5}{10} = 0.5$$

From Example 6.1, we know that $z_{\alpha/2}$ is $z_{0.025} = 1.96$ for a 95% confidence interval. The margin of error is then

$$\text{Margin of Error} = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 1.96\sqrt{\frac{0.5(1-0.5)}{10}} = 1.96 \times 0.158 = 0.31$$

The 95% confidence interval for p is

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.50 \pm 0.31 = (0.19, 0.81)$$

6.3.3 Assumptions

Assumptions 6.1 (Assumptions for One-Sample Confidence Intervals for μ)**.** ———————
In order to construct confidence intervals on the population mean μ, the following assumptions must be met in order for their construction to be valid.

1. *Data is from a random sample of a large population.*
2. *Observations in the sample are independent of each other.*
3. *If the sample size is small, the population distribution must be approximately normal. For large sample sizes, the population does not need not be approximately normal due the effect of the Central Limit Theorem (refer to section 5.2).*

Assumptions 6.2 (Assumptions for One-Sample Confidence Intervals for p)**.** ———————

In order to construct confidence intervals on the population proportion p, the following assumptions must be met in order for their construction to be valid.

1. *Data is from a random sample of a large population.*
2. *Observations in the sample are independent of each other.*
3. $np \geq 10$ *and* $n(1-p) \geq 10$.

Note 6.3.
Assumption 3 in 6.2 can be tested by verifying whether $\hat{p} \geq 10$ *and* $n(1-\hat{p}) \geq 10$

Example 6.4.
Are the assumptions for the confidence intervals constructed in examples 6.1, 6.2 and 6.3 met?

Solution:

In order to construct confidence intervals for μ and p observations in the sample must be independent of each other. Our sample contains daily closing stock prices from the past year. It is unlikely that these closing prices are independent. If a stock performs poorly today, it is likely that the price will continue to drop tomorrow. Alternatively, if a stock performs well today, it is likely that the price will continue to rise tomorrow.

In order to construct a confidence interval for p, the following must hold:

1. $n\hat{p} \geq 10$
2. $n(1-\hat{p}) \geq 10$

However, for $n = 10$ and $\hat{p} = 0.5$,

$$n\hat{p} = 10 \times 0.5 = 5 < 10$$
$$n(1-\hat{p}) = 10(1-0.5) = 10 \times 0.5 = 5 < 10$$

The assumptions are not met and therefore the confidence intervals are not valid.

6.4 Two Sample Confidence Intervals

We are often interested in determining whether there is a significant difference in the parameters of two populations. In such cases we draw independent samples from each population and construct confidence intervals for the difference between the parameters of interest.

6.4.1 On a Difference of Two Means

Consider two populations. One population has a mean μ_1 and standard deviation σ_1 and the other population has mean μ_2 and standard deviation σ_1. We draw a sample of size n_1 from the first population and calculate the sample mean, standard deviation and size of the sample drawn from this population. We draw an independent sample of size n_2 from the second population and calculate the same statistics.

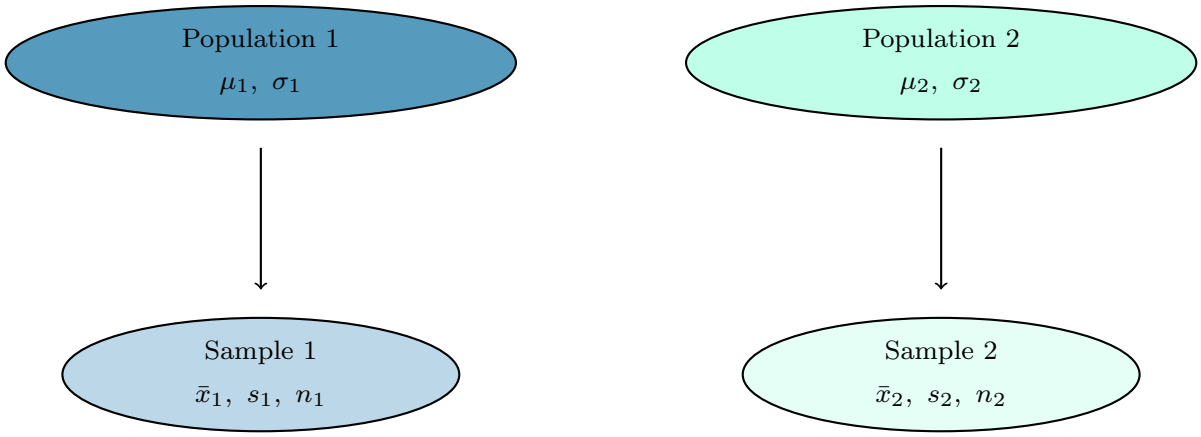


Figure 6.3: Sampling from two populations to construct a confidence interval on a difference of means.

Suppose we are interested in the difference in average salary between males and females at the management level in a certain region. The two populations of interest all males and all females in at the management level in this region. Let μ_1 represent the average salary of all males at the management level and μ_2 represent the average salary of all females at the management level. The parameter of interest is $\mu_1 - \mu_2$. If both the lower and upper bounds of interval $\mu_1 - \mu_2$ are greater than 0, the interval provides evidence that $\mu_1 - \mu_2 > 0$ which therefore suggests that $\mu_1 > \mu_2$. In other words, it provides evidence that the average salary of males is greater than the average salary of females at the management level in a certain region. If both the lower and upper bounds of interval $\mu_1 - \mu_2$ are less than 0, the interval provides evidence that $\mu_1 < \mu_2$. If the interval $\mu_1 - \mu_2$ contains 0 then it is plausible that $\mu_1 - \mu_2 = 0$ which suggests that $\mu_1 = \mu_2$. In other words, it provides evidence that the average salary of males is equal to the average salary of females at the management level in the region.

We could have also let μ_1 represent the average salary of all females at the management level and μ_2 represent the average salary of all males and also constructed this confidence interval. We just have to be careful about the manner in which we interpret the confidence interval we constructed. In the end however our conclusion would be the same.

6.4.1.1 When σ_1 and σ_2 are Known

Confidence Interval 6.4 (Confidence Interval on $\mu_1 - \mu_2$ when σ_1 and σ_2 are Known)**.** – *A* $(100 - \alpha)\%$ *confidence interval on* $\mu_1 - \mu_2$ *when* σ_1 *and* σ_2 *are both known is*

$$\left(\bar{x}_1 - \bar{x}_2\right) \ \pm \ z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \tag{6.4.5}$$

Example 6.5.
Company A is worried about their stock market performance during different periods of the business cycle. It wants to determine whether there is a significant difference in the daily return of their stock in recessionary years vs. boom years. Recessionary and boom periods were identified over the past ten years and a sample of ten daily returns was collected for both periods.

Period	Daily Return (%)									
Boom	26.87	30.92	25.82	37.98	31.65	25.90	32.44	33.69	32.88	28.47
Recession	22.68	5.85	-9.32	-33.22	16.87	-0.67	-0.24	14.16	12.32	8.91

If it is known that the standard deviation of daily returns in boom years is $\sigma_B = 40\%$ and the standard deviation of deviations in recessionary years is $\sigma_R = 30\%$, construct a 95% CI for the difference in mean daily returns, $\mu_B - \mu_R$.

Solution:

We start by finding $\bar{x}_B$ and $\bar{x}_R$.

$$\begin{aligned}\bar{x}_B &= \sum_{i=1}^{n_B} \frac{x_i}{n_B} = \sum_{i=1}^{10} \frac{x_i}{10} \\ &= \frac{26.87 + 30.92 + 25.82 + \ldots + 32.88 + 28.47}{10} \\ &= \frac{306.62}{10} = 30.66\end{aligned}$$

$$\begin{aligned}\bar{x}_R &= \sum_{i=1}^{n_R} \frac{x_i}{n_R} = \sum_{i=1}^{10} \frac{x_i}{10} \\ &= \frac{22.68 + 5.85 - 9.32 + \ldots + 12.32 + 8.91}{10} \\ &= \frac{37.34}{10} = 3.73\end{aligned}$$

$$\Rightarrow \bar{x}_B - \bar{x}_R = 30.66 - 3.73 = 26.93$$

$z_{\alpha/2} = 1.96$ for a confidence level of 95%. $n_B = n_R = 10$ as we sampled 10 daily returns for each period. The margin of error is then

$$\begin{aligned}\text{Margin of Error} &= z_{\alpha/2}\sqrt{\frac{\sigma_B^2}{n_B} + \frac{\sigma_R^2}{n_R}} \\ &= 1.96\sqrt{\frac{40^2}{10} + \frac{30^2}{10}} \\ &= 1.96\sqrt{160 + 90} = 1.96 \times 15.81 = 30.99\end{aligned}$$

Putting everything together, a 95% confidence interval for $\mu_B - \mu_R$ is

$$(\bar{x}_B - \bar{x}_R) \pm z_{\alpha/2}\sqrt{\frac{\sigma_B^2}{n_B} + \frac{\sigma_R^2}{n_R}} = 26.93 \pm 30.99 = (-4.06, 57.92)$$

The 95% confidence interval contains 0 and therefore suggests that there is evidence that the mean daily return during boom periods does not differ from the mean daily return during recessionary periods.

6.4.1.2 When σ_1 and σ_2 are Unknown

We now consider the more realistic case where both population standard deviations are both unknown. This case is broken down into two sub-cases which are when σ_1 and σ_2 are assumed or known to be unequal and when σ_1 and σ_2 are assumed or known to be unequal. There are statistical tests available to test whether the population standard deviations are equal or not but these tests are beyond the scope of this course. However since this is an introductory text, we will consider that we either know or at least we can assume that the population standard deviations are equal or different based on additional information available from the study or expert knowledge in the area of interest. If the population standard deviations are in fact equal, the sample standard deviations will typically be reflective of this (i.e. if $\sigma_1 = \sigma_2$, we would expect $s_1 \approx s_2$).

Note 6.4.
In general, do not automatically assume that $\sigma_1 = \sigma_2$.

Rule 6.1 (Rule of Thumb for Testing Equality of Variances)**.**
There is a crude rule of thumb that can be implemented quickly to test whether we can assume equality of variances or not. Divide the larger sample standard deviation by the smaller sample standard deviation. If the resulting value is greater than or equal to two, we should not assume $\sigma_1 = \sigma_2$. Formally,

$$\frac{\max(s_1, s_2)}{\min(s_1, s_2)} \geq 2 \quad \longrightarrow \quad \textit{Do } \underline{\textit{not}} \textit{ assume } \sigma_1 = \sigma_2 \tag{6.4.6}$$

6.4.1.2.1 When $\sigma_1 \neq \sigma_2$

If the population variances are not equal we construct a confidence interval in the following manner.

Confidence Interval 6.5 (Confidence Interval on $\mu_1 - \mu_2$ when $\sigma_1 \neq \sigma_2$)**.** *A $(100-\alpha)\%$ confidence interval on $\mu_1 - \mu_2$ when $\sigma_1 \neq \sigma_2$ is*

$$\left(\bar{x}_1 - \bar{x}_2\right) \pm t_{(\alpha/2,\, d)}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{6.4.7}$$

where a conservative estimate of d is given by $d = \min(n_1 - 1,\ n_2 - 1)$.

Confidence interval 6.6 is also known as the *Welch-Satterthwaite* method or *Welch's* method.

Note 6.5.
We stated a conservative estimate of the degrees of freedom, d, for confidence interval 6.6. A more accurate calculation of the degrees of freedom is given by:

$$d = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_2^2/n_2\right)^2}{n_2 - 1}} \tag{6.4.8}$$

With this calculation of d, we will typically not get a whole number and we round down to the nearest integer.

Example 6.6.
Company A is worried about their stock market performance during different periods of the business cycle. It wants to determine whether there is a significant difference in the daily return of their stock in recessionary years vs. boom years. Recessionary and boom periods were identified over the past ten years and a sample of ten daily returns was collected for both periods.

Period	Daily Return (%)									
Boom	26.87	30.92	25.82	37.98	31.65	25.90	32.44	33.69	32.88	28.47
Recession	22.68	5.85	-9.32	-33.22	16.87	-0.67	-0.24	14.16	12.32	8.91

If the standard deviation of daily returns is unknown for both periods and they are assumed to be different, construct a 95% CI for the difference in mean daily returns, $\mu_B - \mu_R$.

Solution:

It is known that $\bar{x}_B = 30.66$ and $\bar{x}_R = 3.73$ from previous examples. In order to construct the 95% confidence interval when σ_B and σ_R are unknown, we first find s_B^2 and s_R^2.

$$
\begin{aligned}
s_B^2 &= \sum_{i=1}^{n_B} \frac{(x_i - \bar{x}_B)^2}{n_B - 1} \\
&= \sum_{i=1}^{10} \frac{(26.87 - 30.66)^2 + (30.92 - 30.66)^2 + \ldots + (28.47 - 30.66)^2}{9} \\
&= \frac{14.38 + 0.07 + \ldots + 4.80}{9} \\
&= 15.24
\end{aligned}
$$

$$
\begin{aligned}
s_R^2 &= \sum_{i=1}^{n_R} \frac{(x_i - \bar{x}_R)^2}{n_R - 1} \\
&= \sum_{i=1}^{10} \frac{(22.68 - 3.73)^2 + (5.85 - 3.73)^2 + \ldots + (8.91 - 3.73)^2}{9} \\
&= \frac{358.95 + 4.48 + \ldots + 26.79}{9} \\
&= 257.38
\end{aligned}
$$

We use the conservative estimate of d,

$$d = min(n_B - 1, n_R - 1) = 9$$

Using our t-distribution table we find $t_{(0.025,9)} = 2.26$. Our margin of error is then

$$
\begin{aligned}
\text{Margin of Error} &= t_{(0.025,9)} \sqrt{\frac{s_B^2}{n_B} + \frac{s_R^2}{n_R}} \\
&= 2.26 \sqrt{\frac{15.24}{10} + \frac{257.38}{10}} = 2.26 \sqrt{1.52 + 25.74} = 2.26 \times 5.22 \\
&= 11.80
\end{aligned}
$$

Putting it all together, a 95% confidence interval for $\mu_B - \mu_R$ is

$$(\bar{x}_B - \bar{x}_R) \pm t_{(\alpha/2,d)} \sqrt{\frac{s_B^2}{n_B} + \frac{s_R^2}{n_R}} = 26.93 \pm 11.80 = (15.13, 38.73)$$

The 95% confidence interval does not contain 0 and therefore suggests that there is evidence that the mean daily return during boom periods differs from the mean daily return during recessionary periods. Furthermore, since the interval was constructed as $\bar{x}_B - \bar{x}_R$ and both the lower and upper bounds are positive, this suggests that the mean daily return is greater during boom periods than recessionary periods.

6.4.1.2.2 When $\sigma_1 = \sigma_2$ If the population variances are equal we construct a confidence interval in the following manner.

Confidence Interval 6.6 (Confidence Interval on $\mu_1 - \mu_2$ when $\sigma_1 = \sigma_2$)**.**
A $(100 - \alpha)\%$ confidence interval on $\mu_1 - \mu_2$ when $\sigma_1 = \sigma_2$ is

$$(\bar{x}_1 - \bar{x}_2) \pm t_{(\alpha/2,\, n_1+n_2-2)} \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \tag{6.4.9}$$

where

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \tag{6.4.10}$$

Note 6.6.
The value of s_p^2 in confidence interval 6.6 is called the pooled sample variance. It is an average of the variances of both samples that takes into account the size of each sample. If we take the square root of s_p^2 we get s_p which is called the pooled sample standard deviation.

Confidence interval 6.6 is also referred to as the *pooled method*.

Example 6.7.
Company A is worried about their stock market performance during different periods of the business cycle. It wants to determine whether there is a significant difference in the daily return of their stock in recessionary years vs. boom years. Recessionary and boom periods were identified over the past ten years and a sample of ten daily returns was collected for both periods.

Period	Daily Return (%)									
Boom	26.87	30.92	25.82	37.98	31.65	25.90	32.44	33.69	32.88	28.47
Recession	22.68	5.85	-9.32	-33.22	16.87	-0.67	-0.24	14.16	12.32	8.91

If the standard deviation of daily returns is unknown for both periods and they are assumed to be equal, construct a 95% CI for the difference in mean daily returns, $\mu_B - \mu_R$.

Solution:

It is known that $\bar{x}_B = 30.66$, $\bar{x}_R = 3.73$, $s_B^2 = 15.24$, and $s_R^2 = 257.38$ from previous examples. Since we have made the assumption that the variances are equal, we must find the pooled sample variance s_p^2.

$$\begin{aligned} s_p^2 &= \frac{(n_B-1)s_B^2 + (n_R-1)s_R^2}{n_B+n_R-2} \\ &= \frac{(10-1)(15.24)+(10-1)(257.38)}{10+10-2} \\ &= \frac{2453.58}{18} = 136.31 \end{aligned}$$

Using our t-distribution table we find $t_{(0.025,18)} = 2.101$. The margin of error is then

$$\begin{aligned} \text{Margin of Error} &= t_{(0.025,18)}\sqrt{s_p^2\left(\frac{1}{n_B}+\frac{1}{n_R}\right)} \\ &= 2.101\sqrt{136.31 \times 0.2} \\ &= 2.101 \times 5.22 = 10.97 \end{aligned}$$

Putting it all together, a 95% confidence interval for $\mu_B - \mu_R$ is

$$(\bar{x}_B - \bar{x}_R) \pm t_{(0.025,18)}\sqrt{s_p^2\left(\frac{1}{n_B}+\frac{1}{n_R}\right)} = 26.93 \pm 10.97 = (15.96, 37.90)$$

The 95% confidence interval does not contain 0 and therefore suggests that there is evidence that the mean daily return during boom periods differs from the mean daily return during recessionary periods. Furthermore, since the interval was constructed as $\bar{x}_B - \bar{x}_R$ and both the lower and upper bounds are positive, this suggests that the mean daily return is greater during boom periods than recessionary periods.

Note: It is not appropriate to assume equal variances in this instance as

$$\frac{max(s_B, s_R)}{min(s_B, s_R)} = \frac{max(3.90, 16.04)}{min(3.90, 16.04)} = \frac{16.04}{3.90} = 4.11 > 2$$

Unequal variances should be assumed.

6.4.1.3 Assumptions

Assumptions 6.3 (Assumptions for Two-Sample Confidence Intervals for $\mu_1 - \mu_2$)**.** ——

1. *Both samples are taken randomly from large populations.*
2. *For a given sample, all observations within the sample are independent.*
3. *Both samples are independent of each other.*

4. *Both populations are approximately normally distributed.*
5. *For the pooled method* : *The two populations have the same variance. This assumption is known as the assumption of homogeneity of variance.*

   For Welch's method : *The two populations do not have the same variance. This assumption is known as the assumption of heterogeneity of variance.*

6.4.2 On a Difference of Two Proportions

Consider two populations. One population has a population proportion p_1 and the other population has population proportion p_2. Using a sample of size n_1 drawn from the first population and sample of size n_2 drawn from the second population, we calculate their respective sample proportions.

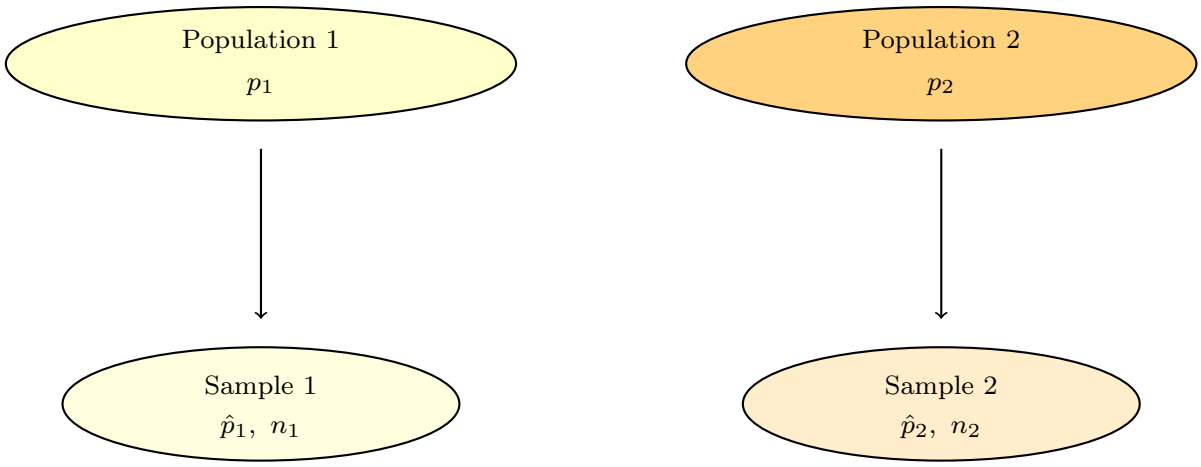


Figure 6.4: Sampling from two populations to construct a confidence interval on a difference of proportions.

A common example where we are interested in a difference in proportions is when two political candidates are running for the same position in an election in a certain region. We are interested to determine whether they are likely to receive an equal proportion of the votes or whether one politician has a lead.

Confidence Interval 6.7 (Confidence Interval on $p_1 - p_2$)**.**
A $(100-\alpha)\%$ *confidence interval on* $p_1 - p_2$ *is constructed using*

$$\left(\hat{p}_1 - \hat{p}_2\right) \ \pm \ z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \tag{6.4.11}$$

In 6.4.11 the standard error of proportion consists of all terms under the square root and this value multiplied by $z_{\alpha/2}$ is the margin of error.

Example 6.8.

Sahara, an up-and-coming shopping website, recently introduced free shipping to consumers that spend above a threshold amount of \$50. *Sahara* is interested in whether there is a difference in spending habits before and after adding this new feature. They sample 200 pre-tax checkout totals before the new feature and 100 pre-tax checkout totals after the new feature. Before the new feature, 55 consumers spent 50 or more dollars on the website. Following the new feature, 72 consumers spent 50 or more dollars on the website. That is,

$$\hat{p}_1 = \frac{55}{200} = 0.275 \quad \hat{p}_2 = \frac{72}{100} = 0.72$$

where $\hat{p}_1$ represents the sample proportion of consumers that spent 50 or more dollars on *Sahara* before the new free shipping feature was implemented and $\hat{p}_2$ represents the sample proportion of consumers that spent 50 or more dollars on *Sahara* after the new free shipping feature was implemented. Construct a 97.5% confidence interval for the difference in proportions, $p_1 - p_2$.

Solution:

Since we are given $\hat{p}_1 = 0.275$ and $\hat{p}_2 = 0.72$ it is easy to calculate $\hat{p}_1 - \hat{p}_2$,

$$\hat{p}_1 - \hat{p}_2 = 0.275 - 0.72 = -0.445$$

For a 97.5% confidence interval, $z_{\alpha/2} = z_{0.0125} = 2.24$. Our margin of error is then,

$$\begin{aligned}\text{Margin of Error} &= z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \\ &= 2.25\sqrt{\frac{0.275(1-0.275)}{200} + \frac{0.72(1-0.72)}{100}} \\ &= 2.25 \times 0.0549 = 0.124\end{aligned}$$

Putting this all together, a 97.5% confidence interval for $p_1 - p_2$ is

$$\left(\hat{p}_1 - \hat{p}_2\right) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} = -0.445 \pm 0.124 = (-0.569, -0.321)$$

The 97.5% confidence interval does not contain 0 and therefore suggests that there is evidence that the proportion of consumers spreading \$50 or more on *Sahara* before the free shipping feature differs from the proportion of consumers spreading \$50 or more on *Sahara* after the free shipping feature was implemented. Furthermore, the interval was constructed as $\hat{p}_1 - \hat{p}_2$ and both the lower and upper bounds are negative, suggesting that the proportion of consumers spending \$50 or more following the new feature is greater than the proportion of consumers spending \$50 or more before the new feature.

6.4.2.1 Assumptions

Assumptions 6.4 (Assumptions for Confidence Intervals on a Difference of Two Proportions)**.**

1. *Both samples are drawn from large populations.*
2. *For a given sample, each observation is independent.*
3. *Both samples are independent.*
4. $n_1p_1 \geq 10$ *and* $n_1(1-p_1) \geq 10$*;*

   $n_2p_2 \geq 10$ *and* $n_2(1-p_2) \geq 10$*.*

Note 6.7.
Assumption 4 in 6.4 can be tested by verifying $n_1\hat{p}_1 \geq 10$*,* $n_1(1-\hat{p}_1) \geq 10$*,* $n_2\hat{p}_2 \geq 10$*, and* $n_2(1-\hat{p}_2) \geq 10$*.*

6.5 On Paired Data

A matched pairs test (also known as a paired t-test) is performed when we have two measurements dependent on a single unit. The population of interest for us is the population of differences.

Definition 6.2 (Paired Data)**.**
Paired data consists of two sets of observations where each observation in one set has a dependence (or some other type of similar connection) with exactly one observation in the other set.

For instance we could be interested in the effect of a drug that is supposed to lower cholesterol. The blood cholesterol of each unit in the study is measured and recorded prior to administering the drug (Measurement 1). The drug is then given to each unit and after the latent period has elapsed, a measurement is taken for each unit again and recorded (Measurement 2). Since we would like to analyze the affect of the drug, we record the difference for each unit (Difference = Measurement 1 − Measurement 2).

Unit	Measurement 1	Measurement 2	Difference
1	x_{11}	x_{12}	$x_{11} - x_{21}$
2	x_{21}	x_{22}	$x_{21} - x_{22}$
3	x_{31}	x_{32}	$x_{31} - x_{32}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	x_{n1}	x_{n2}	$x_{n1} - x_{n2}$
		mean =	$\bar{x}_d$
		st. dev =	s_d

Table 6.1: General layout of table used to construct confidence intervals on paired data.

Once we have our collection of differences, we can find the mean ($\bar{x}_d$) and standard deviation (s_d) of the differences.

Note 6.8.

We can calculate

$$\textit{Difference} = \textit{Measurement 1 - Measurement 2} \tag{6.5.12}$$

or

$$\textit{Difference} = \textit{Measurement 2 - Measurement 1} \tag{6.5.13}$$

Our final conclusion will be the same. We just have to be careful about the manner in which we interpret the results.

Confidence Interval 6.8 (Confidence Interval on Paired Data)**.**

A $(100 - \alpha)\%$ *confidence interval on paired data is*

$$\bar{x}_d \pm t_{(\alpha/2,\ n-1)}\left(\frac{s_d}{\sqrt{n}}\right) \tag{6.5.14}$$

where $\bar{x}_d$ *is the sample mean of all of the differences and* s_d *is the sample standard deviation of all of the differences.*

Note 6.9.

Notice the similarity between 6.5.14 in confidence interval 6.8 and 6.3.3 in confidence interval 6.2. When we have paired data, the confidence interval we use has essentially the same structure as a one-sample confidence interval when the variance is unknown.

Example 6.9.

Banque de Monet is opening a new branch in Guelph and wants to recruit the best talent in the area. In order to ensure that its salaries are competitive, it finds job postings by its competitor, *Picasso's Pennies*, and pairs each job with an identical one being offered at the new branch. This is done for 20 positions in total. The differences in the salaries offered by the two banks are presented in the following table.

Position	*Banque de Monet* Salary	*Picasso's Pennies* Salary	Difference
Service Rep.	35,000	32,000	3,000
Marketing Dir.	75,000	80,000	-5,000
Branch Manager	92,500	92,500	0
IT Team Member	45,000	42,500	2,500
⋮	⋮	⋮	⋮
Sr. Stat Analyst	225,000	175,000	-50,000
		Mean $\bar{x}_d$	3,500
		Standard Deviation s_d	1,750

Construct a 95% confidence interval for the paired data, that is the mean difference in salary.

Solution:

From the table and the information provided we know $\bar{x}_d = 3,500$, $s_d = 1,750$, and $n = 20$. Our margin of error is

$$\begin{aligned}\text{Margin of Error} &= t_{(\alpha/2,n-1)}\left(\frac{s_d}{\sqrt{n}}\right)\\ &= t_{0.025,19}\left(\frac{1750}{\sqrt{20}}\right)\\ &= 2.093 \times 391.312 \approx 819\end{aligned}$$

Putting it all together, the 95% confidence interval for the mean difference in salary is

$$\bar{x}_d \pm t_{(\alpha/2,n-1)}\left(\frac{s_d}{\sqrt{n}}\right) = 3500 \pm 819 = (2681, 4319)$$

Under repeated sampling, we are 95% confident that the true value of the mean difference in salary will fall within the constructed interval 95% of the time.

6.5.1 Assumptions

Assumptions 6.5 (Assumptions for Confidence Intervals on Paired Data)**.**

1. *Each observation has two measurements dependent on the unit.*
2. *The sample size should be less than 10% of the population.*
3. *Measurements on each unit are independent of measurements on other units.*
4. *The population of differences is normally distributed.*

Chapter 7

Hypothesis Tests

Hypothesis testing is a very important topic in statistical inference. When carrying out a hypothesis test, a claim about the value of a parameter is tested using statistical methods. In Chapter 6, we discussed how to use confidence intervals to draw conclusions about the value of the parameter based on the bounds of a constructed interval. By using hypothesis testing, we are able to take this a step further by quantifying the strength of our conclusion with a probability statement.

The mathematical background behind a hypothesis test begins with a decision rule. A *decision rule* is a rule that is used to decide on a conclusion to make in a hypothesis test based on based on a calculated value obtained from data. In the instance of hypothesis testing, the value of a test statistic determines whether we accept our hypothesis or reject it. A test statistic is a value calculated using sample data. The mathematical formulae underlying test statistics are derived using decision rules and the famous likelihood ratio test, which are covered in greater detail in advanced undergraduate or graduate courses in mathematical statistics.

We will skip the mathematical derivation of a hypothesis test and get right into the procedure of conducting them. This makes the process of conducting a hypothesis test somewhat algorithmic in nature.

Steps involved in conducting a hypothesis test:

1. *State the null and alternative hypothesis.*
2. *Find the appropriate test statistic.*
3. *Find the p-value*
4. *Compare p-value to a level of significance (α).*
5. *Make a conclusion.*

---

1. State the null and alternative hypothesis

The first step in hypothesis testing is to state the null and alternative hypothesis. The null hypothesis is the conservative or skeptical belief and the alternative hypothesis is the claim we are testing. The alternative is usually a researchers' belief. The null hypothesis is represented by H_0 and the alternative hypothesis is represented by H_a

Definition 7.1 (Null and Alternative Hypothesis)**.**

H_0 : *Null hypothesis*

Represents either a skeptical or conservative perspective on a claim to be tested.

H_a : *Alternative hypothesis*

Represents an alternative claim or alternative belief under consideration. Usually considers a range of possible values for a parameter.

Note 7.1.

Some textbooks use the notation H_1 to refer to the alternative hypothesis. We will be using H_a consistently throughout this text.

When conducting a hypothesis test, we operate under the assumption that H_0 is true. Our goal is to find evidence in support of H_0 or against H_0.

Definition 7.2 (Types of Hypothesis Tests)**.**

Suppose γ is the parameter we are interested in and γ_0 is the hypothesized value of γ under the null hypothesis. The 3 possible hypothesis tests we can perform are:

1. $H_0 : \gamma = \gamma_0$ *vs.* $H_a : \gamma > \gamma_0$
2. $H_0 : \gamma = \gamma_0$ *vs.* $H_a : \gamma < \gamma_0$
3. $H_0 : \gamma = \gamma_0$ *vs.* $H_a : \gamma \neq \gamma_0$

Hypothesis tests 1 and 2 are referred to as one-sided or one-tailed tests. Hypothesis test 3 is referred to as a two-sided or two tailed test. This test may also be referred to as a non-directional test.

Note 7.2.
We will always write the null hypothesis as an equality ("=") and the alternative hypothesis as a strict inequality (either "<" or ">".

2. Find the Appropriate Test Statistic

Definition 7.3 (Test Statistic)**.**
A statistic calculated from sample data that is used to conduct a hypothesis test.

The test statistic we calculate depends on the information available as well as the assumed value of the parameter under the null hypothesis.

General form of a test statistic:

$$\textit{Test Statistic} = \frac{\textit{(a statistic)} - \textit{(hypothesized value of parameter)}}{\textit{(standard error of statistic)}} \tag{7.0.1}$$

The calculated test statistic follows a certain reference distribution. In this course the reference distributions we will be using are either the standard normal distribution or the $t-$distribution.

Note 7.3.
A statistic as basic as the sample mean $\bar{x}$ on its own can also be used as a test statistic, however this is not common practice and many test statistics are in the form of 7.0.1.

3. Find the p-value

P-values are used to quantify the strength of the evidence supporting (or against) the null hypothesis via a probability statement. The p-value is calculated using the test statistic and is dependent on the form of the alternative hypothesis. We formally define a p-value in definition 7.4 but caution the reader that it is difficult to understand the definition of a p-value immediately and a proper comprehension of p-values may take time.

Definition 7.4 (P-value)**.** *Assuming that H_0 is true, the p-value is the probability of calculating a test statistic at least as extreme as the one computed using sample data solely by chance.*

Recall that a test statistic is calculated using sample data and the assumed value of the parameter under the null hypothesis. A p-value is the probability of calculating the same test statistic or a test statistic that is even more unlikely than the one just calculated by pure chance and without any other external factors. If the p-value is extremely small, it is unlikely that the assumed value of the parameter under the null hypothesis is correct.

Figures 7.1, 7.2, 7.3, 7.4 and 7.5. provide some examples of p-values when the reference distibution is either the normal distribution of the $t-$distribution. The shaded area represents the p-value.

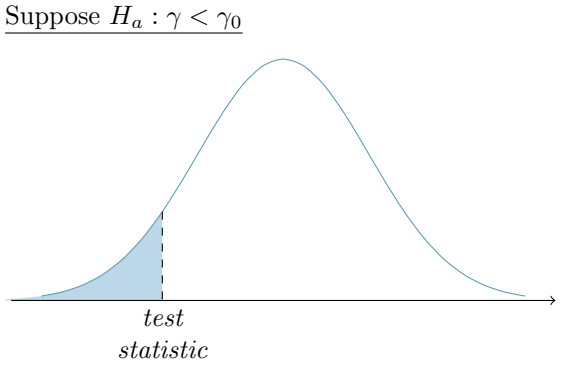


Figure 7.1: A p-value in the small tail of a reference distribution when the alternative is that the parameter is less than a hypothesized value.

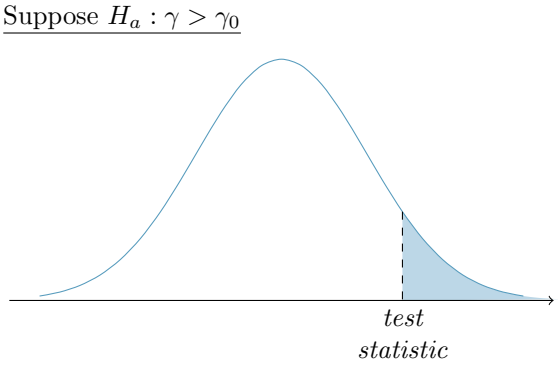


Figure 7.2: A p-value in the small tail of a reference distribution when the alternative is that the parameter is greater than a hypothesized value.

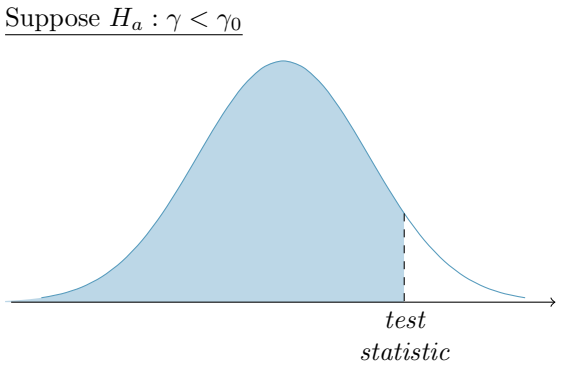


Figure 7.3: A p-value in the small tail of a reference distribution when the alternative is that the parameter is less than a hypothesized value.

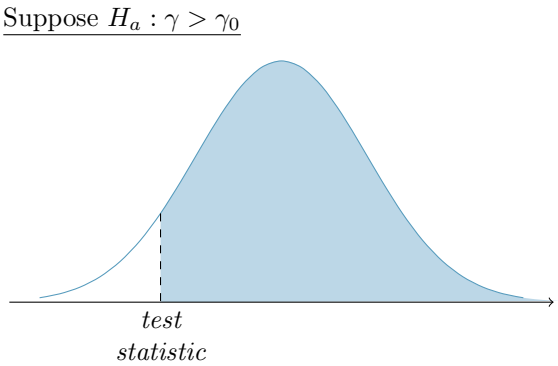


Figure 7.4: A p-value in the large tail of a reference distribution when the alternative is that the parameter is greater than a hypothesized value.

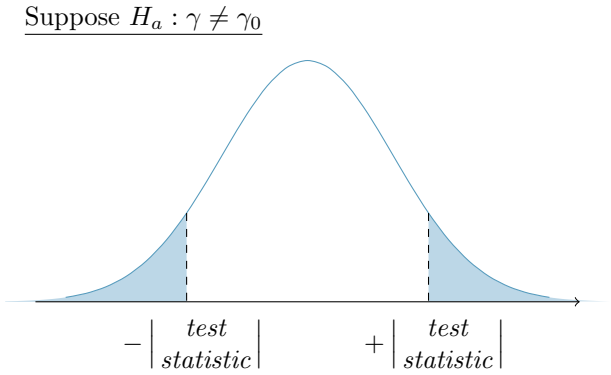


Figure 7.5: The p-value in the area in the extreme tails of a reference distribution when the alternative is that the parameter is not equal to a hypothesized value.

Note 7.4.
Since the p-value is a probability it must be a value between 0 and 1.

4. Compare with a Level of Significance (α)

We compare the p-value with a level of significance (or significance level), (α). The level of significance can be considered as a tolerance level used to determine whether we have sufficient evidence to reject the null hypothesis or not. Formally,

Definition 7.5 (Level of Significance (α))**.**
The probability of rejecting the null hypothesis when it is actually true.

The level of significance is usually pre-determined before a statistical test is performed. If multiple tests are performed on the same data set, we typically use the same value of α for all tests in order to obtain consistent results.

Note 7.5.
If α is not specified, we usually take the default value of $\alpha = 0.05$.

We can also decide on a level of significance after we calculate the p-value, however this is not usually done as it is considered bad practice. Choosing a level of significance after calculating a p-value can result in the abuse and manipulation of statistical tests.

5. Make a Conclusion

The conclusion we make depends on the p-value calculated and the level of significance.

Rule 7.1 (Making a Conclusion on a Hypothesis Test)**.**

$$\begin{aligned} \textit{p-value} > \alpha \implies & \textit{Evidence supports } H_0. \\ & \textit{Do not reject } H_0\textit{, conclude } H_0. \end{aligned}$$

$$\begin{aligned} \textit{p-value} < \alpha \implies & \textit{Evidence against } H_0. \\ & \textit{Reject } H_0\textit{, conclude } H_a. \end{aligned}$$

The manner in which we make a conclusion for hypothesis tests is always the same.

Note 7.6.
A common mistake that students make is to refer back to the alternative hypothesis when they decide the manner in which to make a conclusion. Once we have the p-value we follow rule 7.1. There is no need to refer back to check the sign of the alternative hypothesis.

In Chapter 6, we stressed the importance of using careful language when drawing conclusions on confidence intervals. This also extends to hypothesis testing. We cannot conclude tests by stating that the null hypothesis is *definitely* correct or that the null hypothesis is *definitly* wrong as this would imply that we are completely certain of our result. When concluding hypothesis tests, the statement we make is based on probability so there is always a chance that we have made an error (see section 7.4). Furthermore, due to the nature of random sampling, we may have obtained a sample that is not representative of the true population. In this case, a misleading test statistic would be calculated and lead us to draw the wrong conclusions about the population. There may also be hidden covariates that are not evident (i.e. there may be a hidden relationship between the data and other factors that we are not aware of). As such we make conclusions by stating that the evidence *suggests* we should either support or reject the null hypothesis. If we reject the null hypothesis, we conclude the alternative hypothesis since as it is the only other choice available to us.

Note 7.7.
Hypothesis tests use an evidence based approach to make conclusions, hence we make conclusions by stating that either the evidence supports the null or the evidence is against the null hypothesis.

If we have a very small p-value, we can state that we have strong evidence against the null hypothesis and if we have a large p-value we can state that we have strong evidence supporting the null hypothesis.

7.1 One Sample Hypothesis Tests

7.1.1 On the Mean

7.1.1.1 When σ is Known

When we know the population standard deviation σ we perform the following hypothesis test on μ.

Hypothesis Test 7.1 (Hypthesis Test on μ when σ is Known)**.**
Suppose we are interested in any one of the following hypothesis tests on the population mean:

- $H_0 : \mu = \mu_0$ *vs.* $H_a : \mu > \mu_0$
- $H_0 : \mu = \mu_0$ *vs.* $H_a : \mu < \mu_0$
- $H_0 : \mu = \mu_0$ *vs.* $H_a : \mu \neq \mu_0$

The test statistic for a hypthesis test on μ when σ is known is

$$z^* = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \tag{7.1.2}$$

Reference distribution: the standard normal distribution.

Alternative Hypothesis	*P-value*
$H_a : \mu > \mu_0$	*Area to the right of z^**
$H_a : \mu < \mu_0$	*Area to the left of z^**
$H_a : \mu \neq \mu_0$	*Sum of the areas in the tails of z^* and $-z^*$*

Example 7.1.
Moolah Marketing claims that they can boost web traffic for its customers by an average of 100 visitors per day using its advanced advertising techniques. However, there have been recent complaints from customers that the average visitor boost is *less than* 100. In order to support their claim, *Moolah Marketing* randomly samples the number of visitors per day on 20 of its customer's websites and determines the number of visits above baseline levels. From the sample it is determined that $\bar{x} = 97.5$. If it is known that $\sigma = 10$, conduct the following hypothesis test at the $\alpha = 0.05$ and $\alpha = 0.10$ level.

Solution:

$$H_0 : \mu = 100 \;\; vs. \;\; H_a : \mu < 100$$

First we calculate the test statistic.

$$z^* = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{97 - 100}{10/\sqrt{20}} = -1.342$$

Since $H_a : \mu < \mu_0 = 100$ is a one-sided hypothesis, the p-value is the area to the left of z^*,

$$p - value = P(Z < z^*) = P(Z < -1.342)$$

Using our normal distribution table, we find that the p-value is 0.0898.

When $\alpha = 0.05$, $p - value = 0.0898 > \alpha = 0.05$. Via Rule 7.1, this implies that there is evidence to support the null hypothesis and thus we accept the null hypothesis and conclude that *Moolah Marketing* can boost website traffic by an average of 100 visitors per day using its advertising techniques.

When $\alpha = 0.10$, $p - value = 0.0898 < \alpha = 0.10$. Via Rule 7.1, this implies that there is evidence against the null hypothesis and thus we reject the null hypothesis in favour of the alternative and conclude that *Moolah Marketing* boosts website by an average of less than 100 visitors per day using its advertising techniques.

7.1.1.2 When σ is Not Known

When we do not know the population standard deviation σ we perform the following hypothesis test on μ.

Hypothesis Test 7.2 (Hypthesis Test on μ when σ is Not Known)**.** ———————
Suppose we are interested in any one of the following hypothesis tests on the population mean:

- $H_0 : \mu = \mu_0$ *vs.* $H_a : \mu > \mu_0$
- $H_0 : \mu = \mu_0$ *vs.* $H_a : \mu < \mu_0$
- $H_0 : \mu = \mu_0$ *vs.* $H_a : \mu \neq \mu_0$

The test statistic for a hypthesis test on μ when σ is not known is

$$t^* = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \tag{7.1.3}$$

Reference distribution: the $t-$distribution at $n-1$ degrees of freedom.

Alternative Hypothesis	*P-value*
$H_a : \mu > \mu_0$	*Area to the right of* t^*
$H_a : \mu < \mu_0$	*Area to the left of* t^*
$H_a : \mu \neq \mu_0$	*Sum of the areas in the tails of* t^* *and* $-t^*$

Note 7.8.

We typically can not use the exact p-value when we use the t−tables. We can however provide a range of possible p-values which will usually be good enough to make a conclusion. In the t−table we go down the degrees of freedom (df) column until we reach $n-1$ degrees of freedom. We then go across the row and see which values our t^* test statistic lies between. This will indicate the tail probability that are enclosed by the values in the t−table which are closest to the t^* test statistic we calculated. As a result we can now provide a range of possible values for the p-value.

Example 7.2.

Moolah Marketing claims that they can boost web traffic for its customers by an average of 100 visitors per day using its advanced advertising techniques. However, there have been recent complaints from customers that the average visitor boost is not equal to 100. In order to support their claim, *Moolah Marketing* randomly samples the number of visitors per day on 20 of its customer's websites and determines the number of visits above baseline levels. From the sample it is determined that $\bar{x} = 97.5$ and $s = 15$. If σ is unknown, conduct the following hypothesis test at the $\alpha = 0.05$ level.

Solution:

$$H_0 : \mu = 100 \;\; vs. \;\; H_a : \mu \neq 100$$

First we calculate the test statistic. Since we do not know σ, we use the t-test statistic instead of the z-test statistic.

$$t^* = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{85 - 100}{15/\sqrt{20}} = -4.472$$

$$df = n - 1 = 20 - 1 = 19$$

Since $H_a : \mu \neq \mu_0 = 100$ is a two-sided hypothesis, the p-value is the sum of the area to the left of $-|t^*|$ and to the right of $+|t^*|$.

$$p - value = P(t < -|t^*|) + P(t > +|t^*|) = P(t < -4.472) + P(t > 4.472)$$

Due to the symmetry of the t-distribution,

$$p - value = 2 \times P(t < -4.472) \approx 2 \times 0.0001 = 0.0002$$

At the $\alpha = 0.05$ level, $p - value = 0.0002 < \alpha = 0.05$. Via Rule 7.1, this implies that there is evidence against the null hypothesis and thus we reject the null hypothesis in favour of the alternative and conclude that *Moolah Marketing* does not boost traffic by an average of 100 visitors per day using its advertising techniques.

7.1.2 On a Proportion

Hypothesis Test 7.3 (Hypthesis Test on p)**.**
Suppose we are interested in any one of the following hypothesis tests on the population proportion:

- $H_0 : p = p_0$ *vs.* $H_a : p > p_0$
- $H_0 : p = p_0$ *vs.* $H_a : p < p_0$
- $H_0 : p = p_0$ *vs.* $H_a : p \neq p_0$

The test statistic is:

$$z^* = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \tag{7.1.4}$$

Reference distribution: the standard normal distribution.

Alternative Hypothesis	*P-value*
$H_a : p > p_0$	*Area to the right of* z^*
$H_a : p < p_0$	*Area to the left of* z^*
$H_a : p \neq p_0$	*Sum of the areas in the tails of* z^* *and* $-z^*$

Note 7.9.
It is important to note in hypothesis test 7.3 that the calculation of the denominator of test statistic depends on the assumed value of p under the null (i.e. p_0).

Example 7.3.
Moolah Marketing is interested in the number of repeat visitors on its client's websites. It believes that greater than 30% of visitors return on a regular basis. It samples 10,000 visitor's habits and determines that the sample proportion of return visitors is 33.7%. Conduct the following hypothesis test at the $\alpha = 0.01$ level.

Solution:

$$H_0 : p = 0.30 \quad vs. \quad H_a : p > 0.30$$

First we calculate the test statistic.

$$z^* = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.337 - 0.30}{\sqrt{\frac{0.337(1-0.337)}{1000}}} = 2.478$$

Since $H_a : p > 0.30$ is a one-sided hypothesis, the p-value is the area to the right of z^*.

$$p - value = P(Z > z^*) = P(Z > 2.478)$$

Using our normal distribution table, we find the p-value is 0.007.

At the $\alpha = 0.01$ level, $p - value = 0.007 < \alpha = 0.01$. Via Rule 7.1, this implies that there is evidence against the null hypothesis and thus we reject the null hypothesis in favour of the alternative and conclude that more than 30% of visitors return to client websites on a regular basis.

7.1.3 Assumptions

Assumptions 7.1 (Assumptions for One-Sample Hypothesis Tests on μ)**.** *In order to construct confidence intervals on the population mean μ, the following assumptions are necessary in order for their construction to be valid.*

1. *Data is from a random sample from a large population.*
2. *Observations in the sample must be independent of each other.*
3. *If the sample size is small, the population distribution must be approximately normal.*
4. *If the sample size is large, population does not need not be approximately normal (Recall the effect of the central limit theorem from section 5.2).*

Assumptions 7.2 (Assumptions for One-Sample Hypothesis Tests on p)**.** *In order to construct confidence intervals on the population proportion p, the following assumptions are necessary in order for their construction to be valid.*

1. *Data is from a random sample from a large population.*
2. *Observations in the sample must be independent of each other.*
3. *$np \geq 10$ and $n(1-p) \geq 10$.*

Note 7.10.
Assumption 3 in 7.2 can be tested by verifying whether $\hat{p} \geq 10$ *and* $n(1-\hat{p}) \geq 10$.

7.2 Two Sample Hypothesis Tests

7.2.1 On a Difference of Two Means

In this section we will discuss hypothesis tests on a difference of two means.

Note 7.11.
We will use the notation D_0 *to represent the hypothesized difference between the means of two populations.*

7.2.1.1 When σ_1 and σ_2 are Known

Hypothesis Test 7.4 (Hypthesis Test on $\mu_1 - \mu_2$ when σ_1 and σ_2 are Known)**.**
Suppose we are interested in any one of the following hypothesis tests on a difference between population means when the standard deviation of both populations are known:

- $H_0 : \mu_1 - \mu_2 = D_0$ *vs.* $H_a : \mu_1 - \mu_2 > D_0$
- $H_0 : \mu_1 - \mu_2 = D_0$ *vs.* $H_a : \mu_1 - \mu_2 < D_0$
- $H_0 : \mu_1 - \mu_2 = D_0$ *vs.* $H_a : \mu_1 - \mu_2 \neq D_0$

The test statistic is:

$$z^* = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{7.2.5}$$

Reference distribution: the standard normal distribution.

Alternative Hypothesis	*P-value*
$H_a : \mu_1 - \mu_2 > D_0$	*Area to the right of* z^*
$H_a : \mu_1 - \mu_2 < D_0$	*Area to the left of* z^*
$H_a : \mu_1 - \mu_2 \neq D_0$	*Sum of the areas in the tails of* z^* *and* $-z^*$

Example 7.4.
Arrow is closing one of its department stores in Ravenholm and must decide between two stores. It believes that Store A has greater mean monthly sales than Store B. It samples 15 monthly sales totals from each store over the past 10 years. It is determined that $\bar{x}_A = 120$ and $\bar{x}_B = 105$ in thousands of dollars. If it is known that $\sigma_A = 20$ and $\sigma_B = 10$, conduct a hypothesis test to test whether the the average monthly sales of store A are greater than that of store B. Use $\alpha = 0.05$ level.

Solution:

$$H_0 : \mu_A = \mu_B \;\; vs. \;\; H_a : \mu_A > \mu_B$$

$H_0 : \mu_A = \mu_B$ is equivalent to $H_0 : \mu_A - \mu_B = 0$, implying that $D_0 = 0$. The appropriate test statistic is

$$z^* = \frac{(\bar{x}_A - \bar{x}_B) - D_0}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}} = \frac{(120 - 105) - 0}{\sqrt{\frac{20^2}{15} + \frac{10^2}{15}}} = 2.598$$

Since $H_a : \mu_A > \mu_B$ is a one-sided hypothesis, the p-value is the area to the right of z^*.

$$p - value = P(Z > z^*) = P(Z > 2.598) = P(Z < -2.598) = 0.005$$

At the $\alpha = 0.05$ level, $p - value = 0.005 < \alpha = 0.05$. Via Rule 7.1, this implies that there is evidence against the null hypothesis and thus we reject the null hypothesis in favour of the alternative and conclude that Store A's mean monthly sales are greater than that of Store B's mean monthly sales.

7.2.1.2 When σ_1 and σ_2 are Not Known

7.2.1.2.1 When $\sigma_1 \neq \sigma_2$

Hypothesis Test 7.5 (Hypthesis Test on $\mu_1 - \mu_2$ when σ_1 and σ_2 are Unknown and Different)**.**
Suppose we are interested in any one of the following hypothesis tests on a difference between population means when the standard deviation of both populations are not known:

- $H_0 : \mu_1 - \mu_2 = D_0 \;\; vs. \;\; H_a : \mu_1 - \mu_2 > D_0$
- $H_0 : \mu_1 - \mu_2 = D_0 \;\; vs. \;\; H_a : \mu_1 - \mu_2 < D_0$
- $H_0 : \mu_1 - \mu_2 = D_0 \;\; vs. \;\; H_a : \mu_1 - \mu_2 \neq D_0$

The test statistic is:

$$t^* = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{7.2.6}$$

Reference distribution: the $t-$distribution where a conservative estimate of the degrees is the smaller of $n_1 - 1$ and $n_2 - 1$.

Alternative Hypothesis	*P-value*
$H_a : \mu_1 - \mu_2 > D_0$	*Area to the right of* t^*
$H_a : \mu_1 - \mu_2 < D_0$	*Area to the left of* t^*
$H_a : \mu_1 - \mu_2 \neq D_0$	*Sum of the areas in the tails of* t^* *and* $-t^*$

Note 7.12.

We stated a conservative estimate of the degrees of freedom for the reference distribution in 7.5. A more accurate calculation of the appropriate degrees of freedom is given by:

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_2^2/n_2\right)^2}{n_2 - 1}} \tag{7.2.7}$$

Example 7.5.

Arrow is closing one of its department stores in Ravenholm and must decide between two stores. It believes that Store A has greater mean monthly sales than Store B. It samples 15 monthly sales totals from each store over the past 10 years. It is determined that $\bar{x}_A = 120$, $\bar{x}_B = 105$, $s_A = 10$, and $s_B = 25$ in thousands of dollars. If σ_A and σ_B are unknown and assumed to be different, conduct the following test at the $\alpha = 0.01$ level.

$$H_0 : \mu_A = \mu_B \quad vs. \quad H_a : \mu_A > \mu_B$$

Solution:

Our test statistic is

$$t^* = \frac{(\bar{x}_A - \bar{x}_B) - 0}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}} = \frac{(120 - 105) - 0}{\sqrt{\frac{10^2}{15} + \frac{25^2}{15}}} = 2.158$$

with $df = min(n_A - 1, n_B - 1) = min(14, 14) = 14$ degrees of freedom.

Since $H_a : \mu_A > \mu_B$ is a one-sided hypothesis, the p-value is the area to the right of z^*.

$$p - value = P(Z > z^*) = P(Z > 2.158) = P(Z < -2.158) = 0.02$$

At the $\alpha = 0.01$ level, $p - value = 0.02 > \alpha = 0.01$. Via Rule 7.1, this implies that there is evidence to support the null hypothesis and thus we accept the null hypothesis and conclude that Store A's mean monthly sales are equal to Store B's mean monthly sales.

7.2.1.2.2 When $\sigma_1 = \sigma_2$

Hypothesis Test 7.6 (Hypthesis Test on $\mu_1 - \mu_2$ when σ_1 and σ_2 are Unknown and Equal)**.**

Suppose we are interested in any one of the following hypothesis tests on a difference between population means when the standard deviation of both populations are not known:

- $H_0 : \mu_1 - \mu_2 = D_0$ *vs.* $H_a : \mu_1 - \mu_2 > D_0$
- $H_0 : \mu_1 - \mu_2 = D_0$ *vs.* $H_a : \mu_1 - \mu_2 < D_0$
- $H_0 : \mu_1 - \mu_2 = D_0$ *vs.* $H_a : \mu_1 - \mu_2 \neq D_0$

The test statistic is:

$$t^* = \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \tag{7.2.8}$$

where

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \tag{7.2.9}$$

Reference distribution: the $t-$distribution at $n_1 + n_2 - 2$ degrees of freedom.

Alternative Hypothesis	*P-value*
$H_a : \mu_1 - \mu_2 > D_0$	*Area to the right of* t^*
$H_a : \mu_1 - \mu_2 < D_0$	*Area to the left of* t^*
$H_a : \mu_1 - \mu_2 \neq D_0$	*Sum of the areas in the tails of* t^* *and* $-t^*$

Example 7.6.
Arrow is closing one of its department stores in Ravenholm and must decide between two stores. It believes that Store A and Store B have different mean monthly sales. It samples 15 monthly sales totals from each store over the past 10 years. It is determined that $\bar{x}_A = 120$, $\bar{x}_B = 105$, $s_A = 10$, and $s_B = 25$ in thousands of dollars. If σ_A and σ_B are unknown and assumed to be equal, conduct the following test at the $\alpha = 0.05$ level.

$$H_0 : \mu_A = \mu_B \;\; vs. \;\; H_a : \mu_A \neq \mu_B$$

Solution:

We must first find the pooled sample variance,

$$s_p^2 = \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2} = \frac{14 \times 12^2 + 14 \times 17^2}{15 + 15 - 2} = 216.50$$

Our test statistic is

$$t^* = \frac{(\bar{x}_A - \bar{x}_B) - 0}{\sqrt{s_p^2\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}} = \frac{(120 - 105) - 0}{\sqrt{216.50\left(\frac{1}{15} + \frac{1}{15}\right)}} = 2.79$$

with $df = n_A + n_B - 2 = 15 + 15 - 2 = 28$ degrees of freedom.
Since $H_a : \mu_A \neq \mu_B$ is a two-sided hypothesis, the p-value is the sum of the area to the left of $-|t^*|$ and right of $+|t^*|$.

$$p - value = P(t < -|t^*|) + P(t > +|t^*|) = 2 \times P(t < -2.79) = 0.01$$

At the $\alpha = 0.05$ level, $p - value = 0.01 < \alpha = 0.05$. Via Rule 7.1, this implies that there is evidence against the null hypothesis and thus we reject the null hypothesis in favour of the alternative hypothesis and conclude that Store A's mean monthly sales are not equal to Store B's mean monthly sales.

7.2.1.3 Assumptions

Assumptions 7.3 (Assumptions for Two-Sample Hypothesis Tests on $\mu_1 - \mu_2$)**.**

1. *Data from both samples are taken from random samples from large populations.*
2. *Observations in a sample must be independent of each observation from the same sample.*
3. *Observations in a sample must be independent of each observation from the other sample.*

4. *Both populations are approximately normally distributed.*
5. *When* $\sigma_1 \neq \sigma_2$: *The two populations have the same variance. This assumption is called the assumption of homogeneity of variance.*

   When $\sigma_1 = \sigma_2$: *The two populations do not have the same variance. This assumption is called the assumption of heterogeneity of variance.*

7.2.2 On a Difference of Two Proportions

Hypothesis Test 7.7 (Hypthesis Test on $p_1 - p_2$)**.** *Suppose we are interested in any one of the following hypothesis tests on a difference between population proportions:*

- $H_0 : p_1 - p_2 = 0$ *vs.* $H_a : \mu_1 - \mu_2 > 0$
- $H_0 : p_1 - p_2 = 0$ *vs.* $H_a : \mu_1 - \mu_2 < 0$
- $H_0 : p_1 - p_2 = 0$ *vs.* $H_a : \mu_1 - \mu_2 \neq 0$

The test statistic is:

$$z^* = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \tag{7.2.10}$$

where

$$\hat{p}_1 = \frac{x_1}{n_1}\ , \qquad \hat{p}_2 = \frac{x_2}{n_2} \tag{7.2.11}$$

and

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} \tag{7.2.12}$$

Reference distribution: the standard normal distribution.

Alternative Hypothesis	*P-value*
$H_a : p_1 - p_2 > 0$	*Area to the right of* z^*
$H_a : p_1 - p_2 < 0$	*Area to the left of* z^*
$H_a : p_1 - p_2 \neq 0$	*Sum of the areas in the tails of* t^* *and* $-t^*$

Example 7.7.

Arrow is closing one of its department stores in Ravenholm and must decide between two stores. It is interested in potential customer base growth. While sales are believed to be higher at Store A, the neighbourhoods around Store B are becoming more popular. *Arrow* wants to know the proportion of customers shopping at each store for the first time in the past month. 2500 customers in Store A were surveyed and 2000 customers in Store B were surveyed. It was determined that $\hat{p}_1 = 0.05$ and $\hat{p}_2 = 0.075$, where $\hat{p}_1$ and $\hat{p}_2$ represent the proportion of customers shopping at Stores A and B for the first time, respectively. Conduct the following test at the $\alpha = 0.05$ level.

$$H_0 : p_1 = p_2 \; vs. \; H_a : p_1 < p_2$$

Solution:

We know $\hat{p}_1 = 0.05$, $\hat{p}_2 = 0.075$, $n_1 = 2500$ and $n_2 = 2000$. We need to find x_1 and x_2 so that we can find $\hat{p}$.

$$\hat{p}_1 = \frac{x_1}{n_1} \;\Rightarrow x_1 = \hat{p}_1 n_1 = 0.05(2500) = 125$$

$$\hat{p}_2 = \frac{x_2}{n_2} \;\Rightarrow x_2 = \hat{p}_2 n_2 = 0.075(2000) = 150$$

$$\Rightarrow \hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{125 + 150}{2500 + 2000} = 0.06$$

After finding $\hat{p}$, we can calculate the test statistic.

$$z^* = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.05 - 0.075}{\sqrt{0.06(1-0.06)\left(\frac{1}{2500} + \frac{1}{2000}\right)}} = -3.482$$

Since $H_a : p_1 < p_2$ is a one-sided hypothesis, the p-value is the area to the left of z^*.

$$p - value = P(Z < z^*) = P(Z < -3.482) = 0.0002$$

At the $\alpha = 0.05$ level, $p - value = 0.0002 < \alpha = 0.05$. Via Rule 7.1, this implies that there is evidence against the null hypothesis and thus we reject the null hypothesis in favour of the alternative hypothesis and conclude that the proportion of customers shopping at Store B for the first time is greater than the proportion of customers shopping at Store A for the first time.

7.2.2.1 Assumptions

Assumptions 7.4 (Assumptions for Confidence Intervals on a Difference of Two Proportions)**.**

1. *Data from both samples are taken from random samples from large populations.*
2. *Observations in a sample must be independent of each observation from the same sample.*
3. *Observations in a sample must be independent of each observation from the other sample.*
4. $n_1p_1 \geq 10$ *and* $n_1(1-p_1) \geq 10$*;*

   $n_2p_2 \geq 10$ *and* $n_2(1-p_2) \geq 10$*.*

Note 7.13.
Assumption 4 in 7.4 can be tested by verifying whether $n_1\hat{p}_1 \geq 10$ *and* $n_1(1-\hat{p}_1) \geq 10$ *and also whether* $n_2\hat{p}_2 \geq 10$ *and* $n_2(1-\hat{p}_2) \geq 10$*.*

7.3 On Paired Data

Hypothesis Test 7.8 (Hypthesis Test on Paired Data)**.**
Suppose we are interested in any one of the following hypothesis tests on paired data:

- $H_0 : \mu_D = D_0$ *vs.* $H_a : \mu_D > D_0$
- $H_0 : \mu_D = D_0$ *vs.* $H_a : \mu_D < D_0$
- $H_0 : \mu_D = D_0$ *vs.* $H_a : \mu_D \neq D_0$

The test statistic is:

$$t^* = \frac{\bar{x}_d - D_0}{s_d/\sqrt{n}} \tag{7.3.13}$$

Reference distribution: the $t-$*distribution at* $n-2$ *degrees of freedom.*

Alternative Hypothesis	*P-value*
$H_a : \mu_D > D_0$	*Area to the right of* t^*
$H_a : \mu_D < D_0$	*Area to the left of* t^*
$H_a : \mu_D \neq D_0$	*Sum of the areas in the tails of* t^* *and* $-t^*$

7.3.1 Assumptions

Assumptions 7.5 (Assumptions for Hypothesis Tests on Paired Data)**.**

1. *Two measurements from each observation are dependent on the unit from which they were measured.*
2. *The sample size should be less than 10% of the population.*
3. *Measurements on each unit are independent of each measurement on other units.*
4. *The population of differences is normally distributed.*

Example 7.8.

Banque de Monet is opening a new branch in Guelph and wants to recruit the best talent in the area. In order to ensure that its salaries are competitive, it finds job postings by its competitor, *Picasso's Pennies*, and pairs each job with an identical one being offered at the new branch. This is done for 20 positions in total. The differences in the salaries offered by the two banks are presented in the following table.

Position	*Banque de Monet* Salary	*Picasso's Pennies* Salary	Difference
Service Rep.	35,000	32,000	3,000
Marketing Dir.	75,000	80,000	-5,000
Branch Manager	92,500	92,500	0
IT Team Member	45,000	42,500	2,500
⋮	⋮	⋮	⋮
Sr. Stat Analyst	225,000	175,000	-50,000
	•	Mean $\bar{x}_d$	3,500
	•	Standard Deviation s_d	1,750

Banque de Monet is interested in determining whether or not the true mean difference in salaries is equal to \$3000. Conduct the appropriate hypothesis test at the $\alpha = 0.05$ level.

Solution:

Since *Banque de Monet* is interested in whether or not there is a difference with no specified direction, we should use a two-sided test.

$$H_0 : \mu_d = 3000 \; vs. \; H_a : \mu_d \neq 3000$$

We start by calculating the test statistic using the information provided by the question.

$$t^* = \frac{\bar{x}_d - D_0}{s_d/\sqrt{n}} = \frac{3500 - 3000}{1750/\sqrt{20}} = 1.278$$

with $df = n - 2 = 20 - 2 = 18$ degrees of freedom.
Since $H_a : \mu_d \neq 3000$ is a two-sided hypothesis, the p-value is the sum of the area to the left of $-|t^*|$ and to the right of $+|t^*|$.

$$p - value = P(t < -|t^*|) + P(t > +|t^*|) = 2 \times P(t < -1.278) = 2 \times 0.1097 = 0.218$$

At the $\alpha = 0.05$ level, $p - value = 0.218 > \alpha = 0.05$. Via Rule 7.1, this implies that there is evidence to support the null hypothesis and thus we accept the null hypothesis and conclude that the true mean difference in salaries is \$3000.
Note: In Section 6 we determined a 95% confidence interval for μ_d,

$$(2681, 4319)$$

$\mu_d = 3000$ falls well within this interval and supports the results of our hypothesis test.

7.4 Decision Errors

The conclusions drawn using a hypotheses test may not always be correct. Due to the nature of random sampling, the data used in our analysis may be unrepresentative of the population being studied. As a result the data would produce statistics that would result in the incorrect conclusion for a hypothesis test. There are 2 types of errors that can occur when we conduct a hypothesis test:

1. Type I error
2. Type II error

Definition 7.6 (Type I Error)**.**
A type I error occurs if the null hypothesis is rejected when it is true.

Type I errors are also known as false positives. The probability of making a type I error is equal to the level or significance level α. We have some control over the probability of

making a type I error since we can control the level of α.

Definition 7.7 (Type II Error)**.**
A type II error occurs if the null hypothesis is not rejected when it is false.

Type II errors are also known as a false negatives. The probability of making a type II error is denoted by β. The calculation of β depends on the power of a test (not in the scope of this course). The value of β depends on the actual value of the parameter that we are testing. We can not control the probability of a type II error.

Note 7.14.
The mainstream belief is that type I errors are considered worse than type II errors. This is because the null hypothesis is considered the conservative belief or the current belief and if we make a type I error we are erroneously concluding that the "safe" belief is incorrect.

Note 7.15.
When designing a hypothesis test, we can either control the probability of a type I error or a type II error but not both. Since type I errors are considered worse (see Note 7.14), hypothesis tests are designed to control for type I errors.

Table 7.4 provides a summary of type I and type II errors.

		Reality	
		H_0 true	H_0 false
Conclusion from hypothesis test	Reject H_0	Type I error (α)	No error
	Do not reject H_0	No error	Type II error (β)

Table 7.1: Summary table of type I and type II errors. All 4 possible scenarios are presented for conclusions of hypothesis tests.

7.5 Relationship Between Hypothesis Tests and Confidence Intervals

There is a very intuitive relationship between confidence intervals and hypothesis tests. Recall from note 6.1 that the confidence intervals that we constructed are referred to as "two-sided confidence intervals". If a 100(1 - α)% confidence interval does not contain

a hypothesized value of a parameter then a two-sided hypothesis test will reject the null hypothesis that the parameter is equal to that value.

Example 7.9.
Suppose we conduct the following test:

$H_0 : \mu = 12.0$
$H_a : \mu \neq 12.0$

If we reject H_0 at the 5% significance level then the corresponding 95% confidence interval created using the same data will not contain the value of 12.0. If we do not reject H_0 at the 5% significance level, then the corresponding 95% confidence interval created using the same data will contains 12.0.

7.6 Statistical Significance vs. Practical Significance

As statisticians the term statistically significant has a very particular meaning. When we perform a hypothesis test and reject a null hypothesis, we say that our result is *statistically significant* (at a specific level of significance α). However a meaningful question is whether the result is *practically significant*. A result can be considered practically significant if it will affect a real world decision. Statistical significance is a mathematical conclusion and practical significance is subjective conclusion based on other factors.

Example 7.10.
A company would like to know whether men and women at the management level get paid the same annual salary or whether there is a difference. The hypothesis test conducted is:

$H_0 : \mu_{males} - \mu_{females} = 0$
$H_a : \mu_{males} - \mu_{females} \neq 0$

After conducting the hypothesis test, we conclude that we reject the null and accept the alternative. How useful is this result? All we know is that there is a difference, but not how big the extent of this difference actually is. Even small differences between means might lead to a rejection of the null hypothesis if sample sizes are very large.

Suppose after further investigation the the following test is conducted:

$H_0 : \mu_{males} - \mu_{females} = 10$
$H_a : \mu_{males} - \mu_{females} \neq 10$

After conducting the hypothesis test, we conclude that we do not reject the null hypothesis. In other words, we conclude that men get paid an extra \$10 per year on average than their

female counterparts. In the grand scheme of things, does a $10 difference between the genders really matter?

Chapter 8

Simple Linear Regression

8.1 Introduction and Notation

Regression analysis is the study of fitting a curve through a set of data points. A mathematical framework is used to create a model that will have the best possible fit for the points. The quality of the fit is assessed using some goodness of fit criteria. Regression analysis is very useful in analyzing relationships between variables and also for making predictions. Simple linear regression is fitting a model in the form of a straight line.

Suppose we would like to explore the relationship between a single numeric response variable using just one predictor variable. We can achieve this using a linear model that we can construct using data and we can also use this model to also make predictions.

Definition 8.1 (Independent Variable)**.** *An independent variable is a variable whose values dfoes not depend on changes in the values of other variables.*

Definition 8.2 (Dependent Variable)**.** *A dependent variable is variable whose value depends on that of another variable. It is the variable being tested in a scientific experiment.*

We use y to represent the dependent variable and we use x to represent the variable that we believe is good at predicting y. These variables are referred to by other names. A summary of the terminology used for x and y is given in table 8.1.

y : Dependant variable \ Response
x : Independent variable \ Predictor \ Explanatory variable

Table 8.1: Common notation used for variables in regression analysis analysis

Once we have chosen the variables we would like to study we go out into the real world and collect data on units. We take take measurements on both x and y for each unit. We can tabulate the results as in table 8.1.

x	y
x_1	y_1
x_2	y_2
x_3	y_3
$\vdots$	$\vdots$
x_n	y_n

Table 8.2: Table of measurements for n units.

Note 8.1.

We will assume that we have complete data (i.e. for each unit we are able to record both and x and a y value). There are more advanced courses in statistics that study model construction when there is missing data.

We can also plot the data points in order to visually analyze the relationship between x and y. The x variable is plotted on the horizontal axis and the y variable is plotted on the vertical axis. This will allow us to determine whether we have a linear relationship between x and y. We would like to see whether y increases as x increases or if y decreases as x increases. Figures 8.3 and 8.4 are examples of data that appear to show a linear relationship between variables.

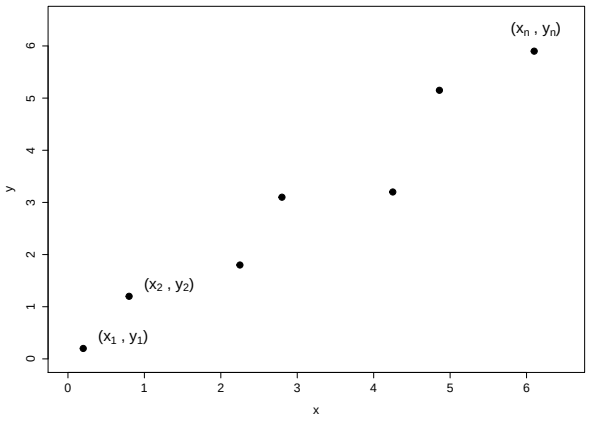


Figure 8.1: Increasing linear relationship between variables x and y.

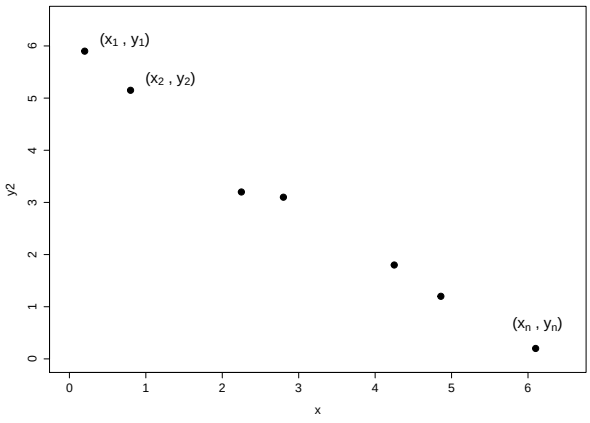


Figure 8.2: Decreasing linear relationship between variables x and y.

If there appears to be a relationship between x and y we would like to use the pairs of data points (x, y) to create a linear model that will fit the data points.

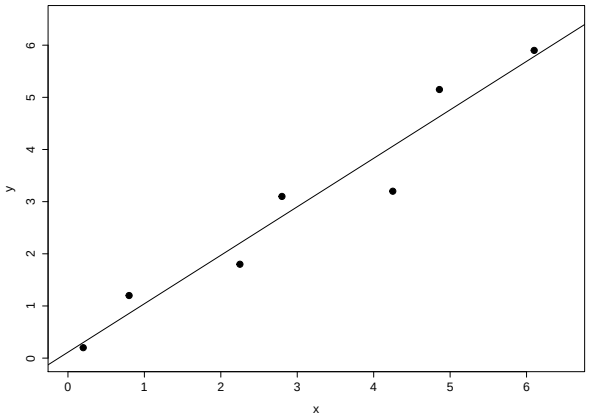


Figure 8.3: Increasing linear relationship between variables x and y with regression line superimposed.

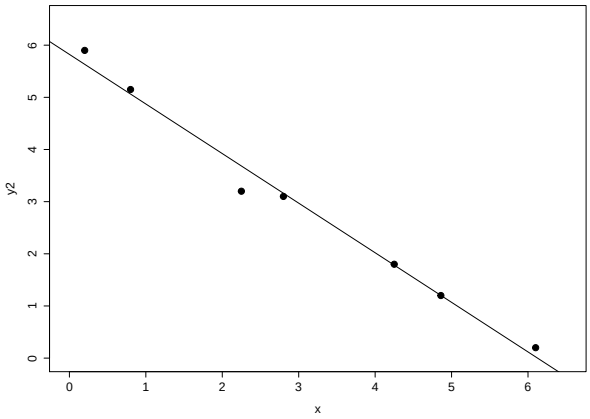


Figure 8.4: Decreasing linear relationship between variables x and y with regression line superimposed.

Recall that a straight line takes the form:

$$y = mx + b \tag{8.1.1}$$

where m is the slope and b is the $y-$intercept. The linear model we create takes the same form as 8.1.1. However the notation we use is different. The model we are interested in is represented as:

$$y = \beta_1 x + \beta_0 + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2) \tag{8.1.2}$$

Model 8.1.2 is called the population regression model (or population regression line) and the terms in this mode are explained in table 8.1 below.

β_1	:	Population slope
β_0	:	Population intercept
ε	:	Error terms

Table 8.3: Summary of terms in model 8.1.2

Note the similarity between 8.1.2 and 8.1.1. Model 8.1.2 is the target model of the population hence β_1 and β_0 are parameters. This is the model that we could construct if we were able to take measurements on every single unit in the population. The error terms ε are in the model because even if we were able to collect measurements on all possible units in the population there will still be some difference between the value predicted by our model and the value observed in real life. The ε terms are very important in the model and certain conditions are required on the ε terms in order for the model to be valid By $\varepsilon \sim N(0, \sigma^2)$ we mean that the error terms are normally distributed with a mean of 0 and some constant variance σ^2. We will learn more about model assumptions in section 8.4).

Since we are unable to take measurements on every single unit in a population we create a model that uses estimates of the parameters in 8.1.2. Using the data we collect we can construct:

$$\hat{y} = \hat{\beta}_1 x + \hat{\beta}_0 \tag{8.1.3}$$

The terms $\hat{\beta}_1$ and $\hat{\beta}_0$ are predicted values of their corresponding parameters and $\hat{y}$ refers to the predicted (or estimated) value of y. The terms in model 8.1.3 are explained in table 8.1.

$\hat{y}$	:	Predicted value of the response y
$\hat{\beta}_1$	:	Population slope
$\hat{\beta}_0$	:	Population intercept

Table 8.4: Summary of terms in 8.1.3

The sign of $\hat{\beta}_1$ determines the direction of the slope. If $\hat{\beta}_1$ is positive we get an increasing slope (such as in figure 8.3) and if $\hat{\beta}_1$ is negative we get an decreasing slope (such as in figure 8.4). We interpret $\hat{y}$ as the predicted value of the response for a particular value of x on average. Model 8.1.3 is called the **least squares regression model** and this term will be explained more in section 8.3 when we cover residuals.

Note 8.2.

The notation of using β_1, β_0 and $\hat{\beta}_1$, $\hat{\beta}_0$ used may appear strange at first, however using notation in this is manner becomes elegant if we make more advanced models which have mode predictors. For instance, suppose we have three possible predictors x_1, x_2 and x_3 for response y. One possible model that we might want to consider is a model that has x_1 as a

quadratic term, x_2 and x_3 as linear terms as well as an interaction term between x_1 and x_2. This model can be neatly expressed as:

$$\hat{y} = \hat{\beta}_1 x_1^2 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_{12} x_1 x_2 \tag{8.1.4}$$

Note 8.3.

Some texts will use b_1 instead of β_1 and b_0 instead of β_0 to represent the population intercept and slope as well as to use $\hat{b}_1$ instead of $\hat{\beta}_1$ and $\hat{b}_0$ instead of $\hat{\beta}_0$ to represent estimates of the slope and intercept. Other textbooks might use a and b to represent the intercept and slope. We will use notation involving β's since this notation is more elegant and meaningful when we construct more complicated models and it is also the notation used in the vast majority journal articles and scientific papers across all fields.

Note 8.4.

We reiterate that the interpretation of $\hat{y}$ is the predicted value of the response for a given value of x on average. It is important to use the word "average" in the interpretation. This is because $\hat{y} = \mathbb{E}(y|x)$ and when we have an expectation in a statement, this refers to the expected value or average. The derivation of $\hat{y} = \mathbb{E}(y|x)$ is theoretical result from regression analysis.

8.2 The Linear Regression Model

Once we have collected our data (i.e. our pairs of data points (x, y)) we can proceed to construct the linear regression model. As mentioned in note 8.1, we will assume that we do not have any missing data. There are several values that are required in order to calculate $\hat{\beta}_1$ and $\hat{\beta}_0$. We will need the mean of all the x values $\bar{x}$ as well as the mean of all the y values $\bar{y}$. These are obtained in the same manner as the sample mean in section 2. We will also need some new values which are

---

$$SS_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{8.2.5}$$

$$SS_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{8.2.6}$$

$$SS_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \tag{8.2.7}$$

SS_{xx} is the sum of the squares of x, SS_{yy} is the sum of the squares of y and SS_{xy} is the cross sum for x and y. We will not require SS_{yy} to construct the regression model but we will be using it later for inference techniques as well as assessing the goodness of fit of the regression model. Table 8.5 summarizes the manner in which we calculate these values.

x	y	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})\ (y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
x_1	y_1	$(x_1 - \bar{x})$	$(y_1 - \bar{y})$	$(x_1 - \bar{x})\ (y_1 - \bar{y})$	$(x_1 - \bar{x})^2$	$(y_1 - \bar{y})^2$
x_2	y_2	$(x_2 - \bar{x})$	$(y_2 - \bar{y})$	$(x_2 - \bar{x})\ (y_2 - \bar{y})$	$(x_2 - \bar{x})^2$	$(y_2 - \bar{y})^2$
x_3	y_3	$(x_3 - \bar{x})$	$(y_3 - \bar{y})$	$(x_3 - \bar{x})\ (y_3 - \bar{y})$	$(x_3 - \bar{x})^2$	$(y_3 - \bar{y})^2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_n	y_n	$(x_n - \bar{x})$	$(y_n - \bar{y})$	$(x_n - \bar{x})\ (y_n - \bar{y})$	$(x_n - \bar{x})^2$	$(y_n - \bar{y})^2$
$\sum_{i=1}^{n} x_i$	$\sum_{i=1}^{n} y_i$			$\underbrace{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}_{SS_{xy}}$	$\underbrace{\sum_{i=1}^{n}(x_i - \bar{x})^2}_{SS_{xx}}$	$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{SS_{yy}}$

Table 8.5: Summary table for the values required to obtain SS_{xx}, SS_{yy} and SS_{xy}

Once we have $\bar{x}$, $\bar{y}$, SS_{xx} and SS_{xy} we can calculate $\hat{\beta}_1$ using equation 8.2.8:

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} \tag{8.2.8}$$

and we can calculate $\hat{\beta}_0$ using equation 8.2.9

$$\bar{y} = \hat{\beta}_1 \bar{x} + \hat{\beta}_0 \tag{8.2.9}$$

Note 8.5.

Equation 8.2.9 can easily be manipulated to find $\hat{\beta}_0$. The terms can be rearranged to express $\hat{\beta}_0$ as:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{8.2.10}$$

Note 8.6.

We can use equation 8.2.9 to solve for $\hat{\beta}_0$ since the regression line 8.1.3 will always pass through the point $(\bar{x}, \bar{y})$.

Note 8.7.

x	y	xy	x^2	y^2
x_1	y_1	$x_1 y_1$	x_1^2	y_1^2
x_2	y_2	$x_2 y_2$	x_2^2	y_2^2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_n	y_n	$x_n y_n$	x_n^2	y_n^2
$\sum_{i=1}^{n} y_i$	$\sum_{i=1}^{n} y_i$	$\sum_{i=1}^{n} x_i y_i$	$\sum_{i=1}^{n} x_i^2$	$\sum_{i=1}^{n} y_i^2$

Table 8.6: Summary table of values for an alternate way to calculate SS_{xx}, SS_{yy} and SS_{xy}

An alternate way to calculate SS_{xx} and SS_{yy} is to use

$$SS_{xx} = \left(\sum_{i=1}^{n} x_i^2 \right) - \frac{\left(\sum_{i=1}^{n} x_i \right)^2}{n} \tag{8.2.11}$$

$$SS_{yy} = \left(\sum_{i=1}^{n} y_i^2\right) - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} \tag{8.2.12}$$

and an alternate way to calculate SS_{xy} *is to use*

$$SS_{xy} = \left(\sum_{i=1}^{n} x_i y_i\right) - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n} \tag{8.2.13}$$

Using equations 8.2.13 and 8.2.13 will provide solutions to SS_{xy} *and* SS_{xx} *faster than using equations 8.2.7 and 8.2.5.*

Note 8.8.

An alternate but longer way to calculate $\hat{\beta}_1$ *is to use:*

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{SS_{xx}} \tag{8.2.14}$$

Example 8.1.

company A is interested in the relationship between the length of employment and current salary of its employees. It collects the following data from a sample of five employees.

Years Employed	1	2	5	7	10
Current Salary (in thousands)	40	53	78	95	121

(**a**) Find SS_{xx}, SS_{yy}, and SS_{xy}.

In this example, our independent variable x is length of employment in years and our dependent variable y is current salary. In order to find SS_{xx}, SS_{yy}, and SS_{xy} we need to complete the following summary table.

x	y	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})\ (y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
x_1	y_1	$(x_1 - \bar{x})$	$(y_1 - \bar{y})$	$(x_1 - \bar{x})\ (y_1 - \bar{y})$	$(x_1 - \bar{x})^2$	$(y_1 - \bar{y})^2$
x_2	y_2	$(x_2 - \bar{x})$	$(y_2 - \bar{y})$	$(x_2 - \bar{x})\ (y_2 - \bar{y})$	$(x_2 - \bar{x})^2$	$(y_2 - \bar{y})^2$
x_3	y_3	$(x_3 - \bar{x})$	$(y_3 - \bar{y})$	$(x_3 - \bar{x})\ (y_3 - \bar{y})$	$(x_3 - \bar{x})^2$	$(y_3 - \bar{y})^2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_n	y_n	$(x_n - \bar{x})$	$(y_n - \bar{y})$	$(x_n - \bar{x})\ (y_n - \bar{y})$	$(x_n - \bar{x})^2$	$(y_n - \bar{y})^2$
$\sum_{i=1}^{n} x_i$	$\sum_{i=1}^{n} y_i$			$\underbrace{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}_{SS_{xy}}$	$\underbrace{\sum_{i=1}^{n}(x_i - \bar{x})^2}_{SS_{xx}}$	$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{SS_{yy}}$

The first step is to find $\bar{x}$ and $\bar{y}$. Once we have found those two values, we can easily calculate the rest of the table and find the necessary sums of squares.

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{1+2+5+7+10}{5} = \frac{25}{5} = 5$$
$$\bar{y} = \sum_{i=1}^{n} \frac{y_i}{n} = \frac{40+53+78+95+121}{5} = \frac{387}{5} = 77.4$$

By following the formula at the top of each column, the completed summary table is

x	y	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
1	40	-4	-37.4	149.6	16	1398.8
2	53	3	24.4	73.2	9	595.4
5	78	0	0.6	0	0	0.36
7	95	2	17.6	35.2	4	309.8
10	121	5	43.6	218	25	1900.9
25	387	•	•	476	54	4205.3

The calculations for the first row of the table are as follows:

$$(x_1 - \bar{x}) = 1 - 5 = -4$$
$$(y_1 - \bar{y}) = 40 - 77.4 = -37.4$$
$$(x_1 - \bar{x})(y_1 - \bar{y}) = (-4)(-37.4) = 149.6$$
$$(x_1 - \bar{x})^2 = (-4)^2 = 16$$
$$(y_1 - \bar{y})^2 = (-37.4)^2 = 1398.8$$

From our table,

$$SS_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = 476$$
$$SS_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = 54$$
$$SS_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = 4205.3$$

(**b**) Find $\hat{\beta}_1$.

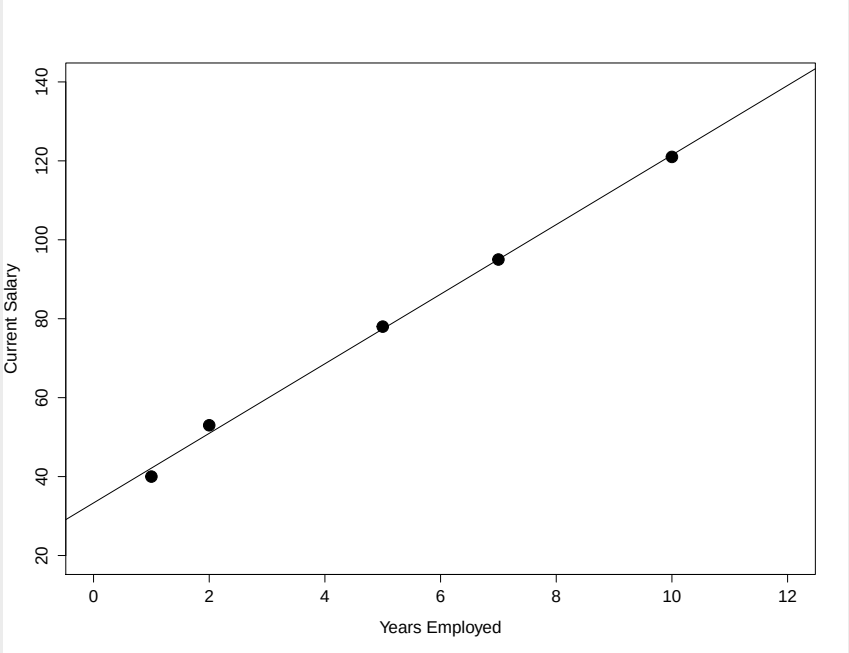
$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{476}{54} = 8.815$$

(**c**) Find $\hat{\beta}_0$.

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 77.4 - 8.815(5) = 33.325$$

(**d**) What is the least squares regression model?

$$\hat{y} = \hat{\beta}_1 x + \hat{\beta}_0 = 8.815\ x + 33.325$$



8.2.1 Interpretation of the Slope and Intercept

The manner in which we interpret the estimate of the slope $\hat{\beta}_1$ is given in definition 8.3

Definition 8.3 (Interpretation of Estimate of the Slope)**.** *An increase in the independent variable x by 1 unit will result in an increase/decrease in the response y by $\hat{\beta}_1$ units on average.*

In definition 8.3 we state that an increase in x will result in an increase or decrease in y since the direction of the change depends on the sign of $\hat{\beta}_1$. If $\hat{\beta}_1$ is positive then an increase in x will result in an increase in y and If $\hat{\beta}_0$ is negative then an increase in x will result in a decrease in y.

An interpretation of the and intercept $\hat{\beta}_0$ is given in definition 8.4 below.

Definition 8.4 (Interpretation of the Estimate of the Intercept)**.** *$\hat{\beta}_0$ is the predicted value of the response y when the value of the independent variable x is 0 on average.*

The intercept may not always have a meaningful practical interpretation and is sometimes difficult to explain.

Example 8.2.
In Example 8.1, Company A determined that the relationship between the length of employment (in years) and current salary (in thousands of dollars) of its employees could be modelled using the following least squares regression model.

$$\hat{y} = 8.815\ x + 33.325$$

An increase in an employee's length of employment by 1 year will result in an increase in current salary of \$8,815 on average.

\$33,325 is the predicted average current salary of an employee whose length of employment is 0 years.

8.2.2 Interpolation and Extrapolation

Definition 8.5 (Interpolation)**.** *Interpolation is calculating predicted values of the response using our model while working within the range of x in which data was available to construct our model.*

Definition 8.6 (Extrapolation)**.** *Extrapolation is calculating predicted values of y using our 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 preformed carefully since extrapolations that are performed without any foresight can be very inaccurate.

Example 8.3.
In Example 8.1, Company A determined that the relationship between the length of employment (in years) and current salary (in thousands of dollars) of its employees could be modelled using the following least squares regression model.

$$\hat{y} = 8.815\ x + 33.325$$

(**a**) What is the average current salary of an employee that has worked at Company A for 6 years?

This is an interpolation as our data contains lengths of employment ranging between 1 and 10 years and 6 years falls within this range. We can use our model to interpolate this by plugging in $x = 6$,

$$\hat{y} = 8.815\ x + 33.325 = 8.815(6) + 33.325 = 86.215$$

The estimated average current salary of an employee that has worked at Company A for 6 years is \$86,215.

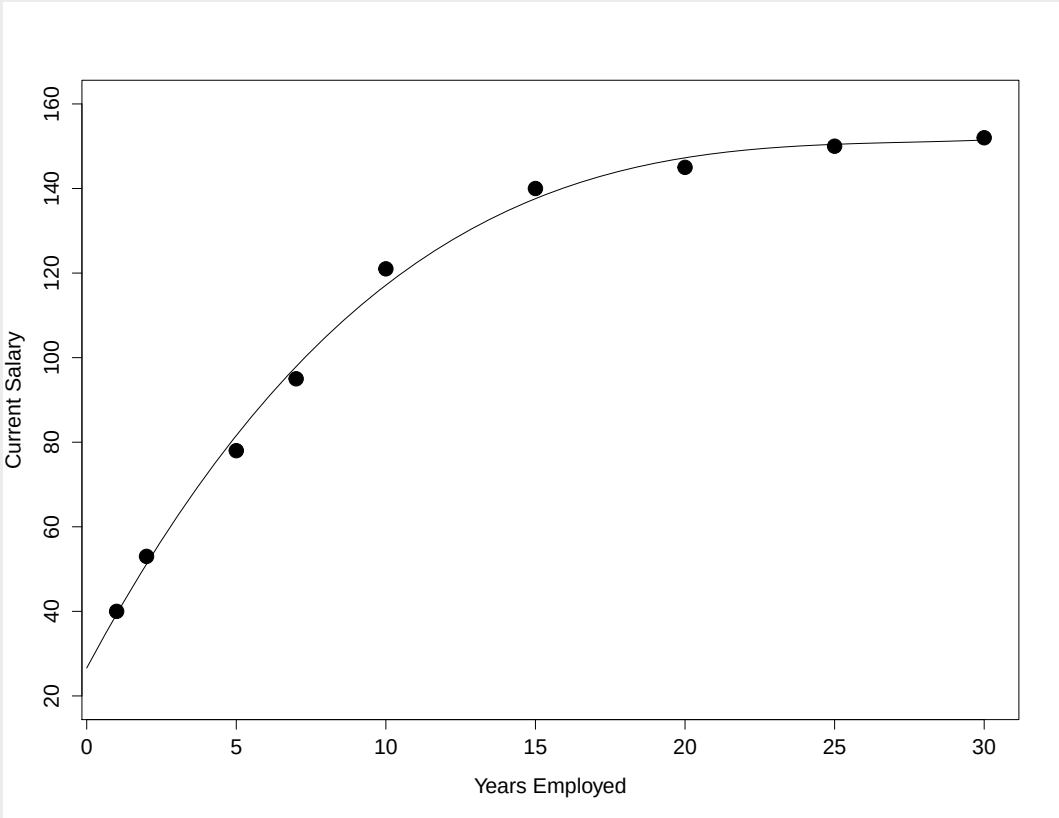
(**b**) What is the average current salary of an employee that has worked at Company A for 20 years?

This is an extrapolation as our data contains lengths of employment ranging between 1 and 10 years and 20 years falls outside this range. We can use our model to extrapolate this by plugging in $x = 20$,

$$\hat{y} = 8.815\ x + 33.325 = 8.815(20) + 33.325 = 209.625$$

The estimated average current salary of an employee that has worked at Company A for 6 years is \$209,625.

In actuality, the relationship between current salary and length of employment at Company A resembles the following figure. This means that our extrapolation of current salary at 20 years of employment is a gross overestimation! By examination of the figure below, the current salary of an employee who has worked at Company A for 20 years should be closer to \$140,000.



If Company A had sampled more than five data points, they may have been able to detect the non-linearity in the model.

8.3 Residuals

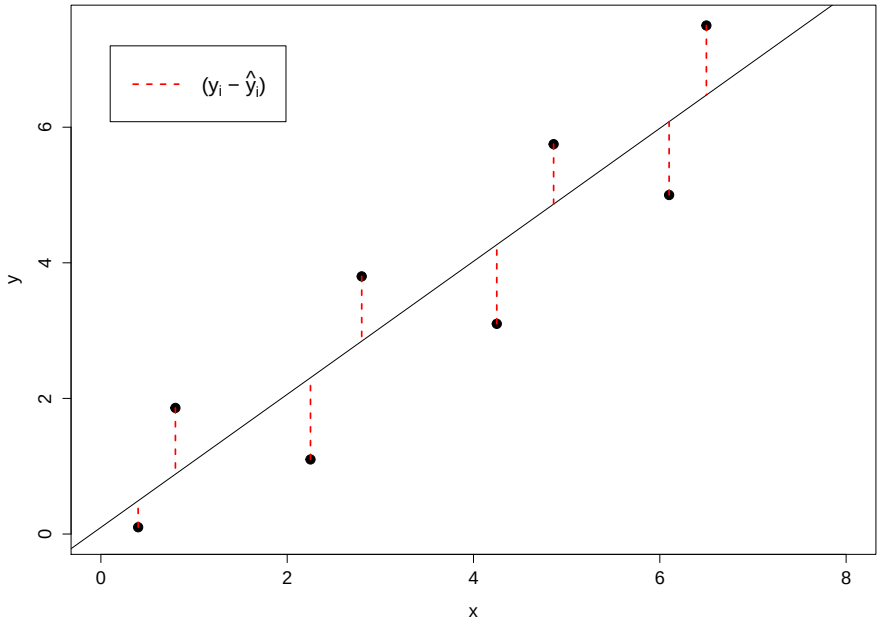


Figure 8.5: The red dashed lines represent the residuals which are the distance between an observed value of a response and the corresponding predicted value obtained from the model $(y_i - \hat{y}_i)$.

A residual is the vertical distance between an actual observed value and the fitted value from the model. Figure 8.3 shows residuals for all the data points used to create the particular regression line seen in the figure.

Definition 8.7 (Residual)**.** *A residual e_i is the difference between the observed value of the dependent variable y_i and the corresponding predicted value $\hat{y}_i$ for each x_i in the data.*

$$\textit{residual} = \textit{observed value} - \textit{fitted value} \tag{8.3.15}$$

$$e_i = y_i - \hat{y}_i \tag{8.3.16}$$

Residuals are very important in regression analysis. They assist us in determining the validity of model assumptions as well as in the calculation of statistics that are used in inference procedures on the slope. If we have n data points we have n residuals. Residuals can be positive or negative. A residual is positive if the observed value of the response is above the regression line and a residual is negative if the observed response is below the regression line.

Properties 8.1.

1. *The sum of the residuals is 0.*

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} y_i - \hat{y}_i = 0 \tag{8.3.17}$$

2. *By finding $\hat{\beta}_1$ and $\hat{\beta}_0$ in the manner that we have, the sum of the residuals squared is as small as possible.*

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} \quad \textit{and} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad \Longrightarrow \quad \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \ \textit{ is a } \min \tag{8.3.18}$$

The reason for the sum of the residuals to be 0 as mentioned in property 1 in 8.1 is because some of the observed values will be above the regression line and the rest will be below the regression line. As mentioned earlier in this section points about the regression line will result in positive residuals and points below the regression line will result in negative residuals. The positive residuals and the negative residuals will add up to 0.

The value of the sum of the squared residuals is called the sum of squared error (SSE). It is an important value as it will be used in inference procedures on the slope in section 8.6.

Definition 8.8 (Sum of Squared Error (SSE))**.**
The sum of squared error (SSE) is the sum of the squared residuals.

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{8.3.19}$$

Property 2 in Properties 8.1 is very important. The regression model 8.1.3 is the only regression model where the sum of the squared residuals (i.e. the SSE) is as small as possible. This is the reason that we call our regression model the **least squares regression model**. We will learn more about the SSE in section 8.5.

Note 8.9.
Property 1 in 8.1 is not unique to our regression line. Many possible lines exist in which this

property is true. Property 2 however is unique to our regression line.

Definition 8.9 (Least Squares Regression Line)**.** *The least squares regression line is the regression line with the smallest possible value for the sum of the squares of the residuals. (i.e. a regression line such that* $SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ *is as small as possible).*

Note 8.10.
Alternate ways to calculate the SSE are:

$$SSE = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_1 x_i - \hat{\beta}_0\right)^2 \tag{8.3.20}$$

$$= SS_{yy} - \hat{\beta}_1 SS_{xy} \tag{8.3.21}$$

Note 8.11.
Some textbooks may refer to the SSE as the residual sum of squares (RSS).

8.3.1 Residual Plots

Residual plots are constructed in order to assess the validity of the model assumptions listed in 8.1. We plot the residuals e_i on the vertical axis against their corresponding fitted values $haty_i$ on the horizontal axis. In order for our assumptions to be valid we look for the following conditions to be satisfied in our regression plot.

1. Random scattering of the points (i.e. no obvious ordering or pattern).
2. Approximately half the points above 0 and half below 0.
3. The majority of points appear to be within a symmetric band about the horizontal.

A random scattering of the points in a residual plot suggests that the residuals (and hence the measurements) are independent of each other. If we notice a pattern in when we plot the residuals in time order then our assumption of independent measurements is violated and therefore assumption 2 in 8.1 is violated. By having approximately half of the residuals

above zero and approximately half of the residuals zero as well as having the majority of points appearing within a symmetric band about the horizontal suggest that assumptions 4 and 5 in 8.1 are satisfied.

Another analysis we can perform on residuals is to plot them in the order measurements were taken (i.e. in time order). This is because the order in which measurements are taken may effect the residuals. For example a person may be using a specific instrument to take measurements on certain units. This individual may be using this particular type of instrument for the first time. It may be the case that the individual would not be very used to the new equipment they are using so there might be substantial error in the initial measurements taken. However as this individual took more and mode measurements they became accustomed to the equipment they were using and consequently took more accurate and/or precise measurements. If this were the case initial measurements would consist of a lot of error in terms of accuracy and/or precision (resulting in large residual values) and measurements taken after the individual became more accustomed to the equipment would be more accurate and/or precise to the actual measurement. (resulting in smaller residual values).

Example 8.4.

As gas prices climb, the Canadian government considers giving auto manufacturers subsidies to produce electric cars. However, it wants to quantify the relationship between the number of electric cars purchased and gas prices before proceeding. It collects the following data from the past ten years. Gas prices are measured in cents per litre and the number of electric cars purchased is measured in thousands of units.

Year	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Average Gas Price	110	114	118	119	120	121	124	127	133	135
Electric Cars Purchased	45	75	80	99	101	104	111	128	132	175

It is determined that $\sum x_i = 1221$, $\sum x_i^2 = 149641$, $\sum y_i = 1050$, $\sum y_i^2 = 121622$, and $\sum x_i y_i = 130626$.

(**a**) Find the least squares regression model.

Using the information provided, we use the appropriate formulae to calculate the sums of squares needed to find $\hat{\beta}_1$.

$$SS_{xy} = \left(\sum_{i=1}^{n} x_i y_i\right) - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n} = 130626 - \frac{1221 \times 1050}{10} = 2421$$

$$SS_{xx} = \left(\sum_{i=1}^{n} x_i^2\right) - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} = 149641 - \frac{1221^2}{10} = 556.9$$

Plugging this into our formula for $\hat{\beta}_1$,

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{2421}{556.9} = 4.347$$

In order to find $\hat{\beta}_0$, we need to find $\bar{y}$ and $\bar{x}$. Since we are provided with $\sum y_i$ and $\sum x_i$,

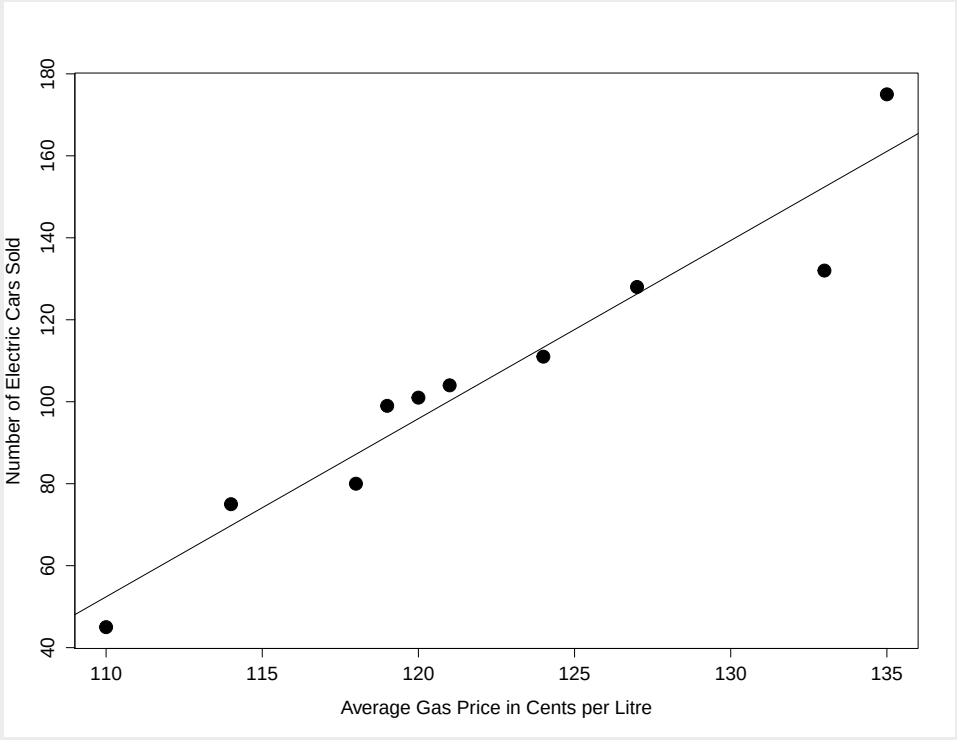
$$\bar{y} = \frac{\sum y_i}{10} = \frac{1050}{10} = 105 \qquad \bar{x} = \frac{\sum x_i}{10} = \frac{1221}{10} = 122.1$$

Plugging this into our formula for $\hat{\beta}_0$,

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 105 - 4.347 \times 122.1 = -425.769$$

Therefore, the least squares regression model is

$$\hat{y} = 4.347\ x - 425.769$$



(**b**) Interpret $\hat{\beta}_1$ and $\hat{\beta}_0$.

An increase in average gas prices by 1 cent per litre will result in an increase in the number of electric cars sold per year by 4,347 units on average.

-425,769 electric cars are sold per year on average when the average gas price is 0 cents per litre. This doesn't make sense but that's okay! As discussed in section 8.2.1, the intercept will not always be meaningful. In this instance, we can take this to mean that Canadian consumers have no incentive to buy electric cars when the average gas price is 0 cents per litre.

(**c**) How many electric cars are sold per year on average when the average gas price is 130 cents per litre?

We can use our model to interpolate this by plugging in $x = 130$,

$$\hat{y} = 4.347 \times 130 - 425.769 = 139.341$$

The estimated average number of electric cars sold per year is 139,341 when the average gas price is 130 cents per litre.

(**d**) Calculate the residuals.

From Definition 8.7,

$$e_i = y_i - \hat{y}_i$$

Using the least squares regression model we found in part (a),

$$\begin{aligned}
\hat{y}_1 &= 4.347 \times 110 \text{ - } 425.769 &&= 52.401 \\
\hat{y}_2 &= 4.347 \times 114 \text{ - } 425.769 &&= 69.789 \\
\hat{y}_3 &= 4.347 \times 118 \text{ - } 425.769 &&= 87.177 \\
&\vdots \\
\hat{y}_{10} &= 4.347 \times 135 \text{ - } 425.769 &&= 161.076
\end{aligned}$$

Therefore

$$\begin{aligned}
e_1 &= 45\text{-}52.401 &&= \text{-}7.401 \\
e_2 &= 75\text{-}69.789 &&= 5.211 \\
e_3 &= 80\text{-}87.177 &&= \text{-}7.177 \\
&\vdots \\
e_{10} &= 175 \text{ - } 161.076 &&= 13.924
\end{aligned}$$

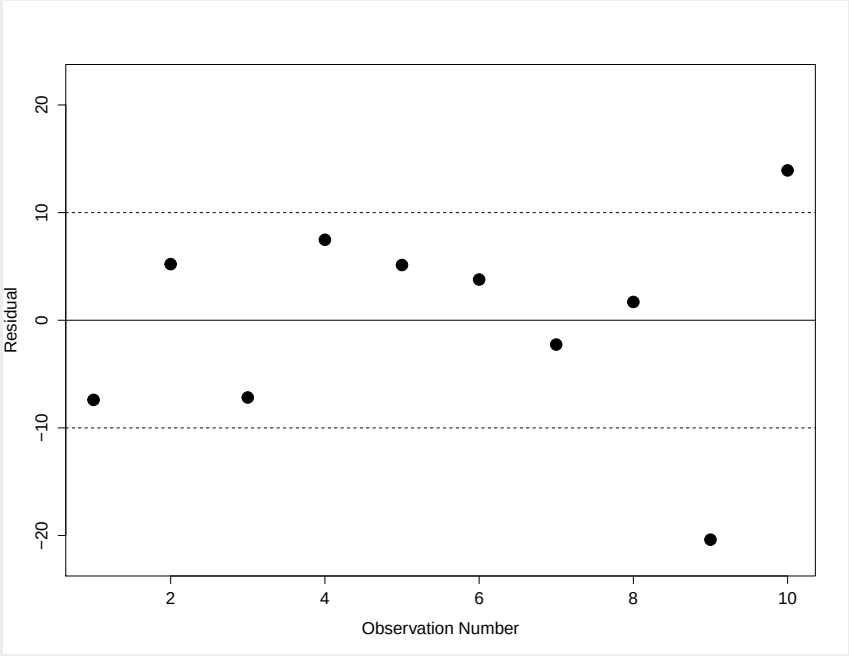
In summary

Year	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Observed	45	75	80	99	101	104	111	128	132	175
Predicted	52.401	69.789	87.177	91.523	95.871	100.218	113.260	126.302	152.385	161.080
Residual	-7.401	5.211	-7.177	7.477	5.129	3.782	-2.260	1.698	-20.385	13.920

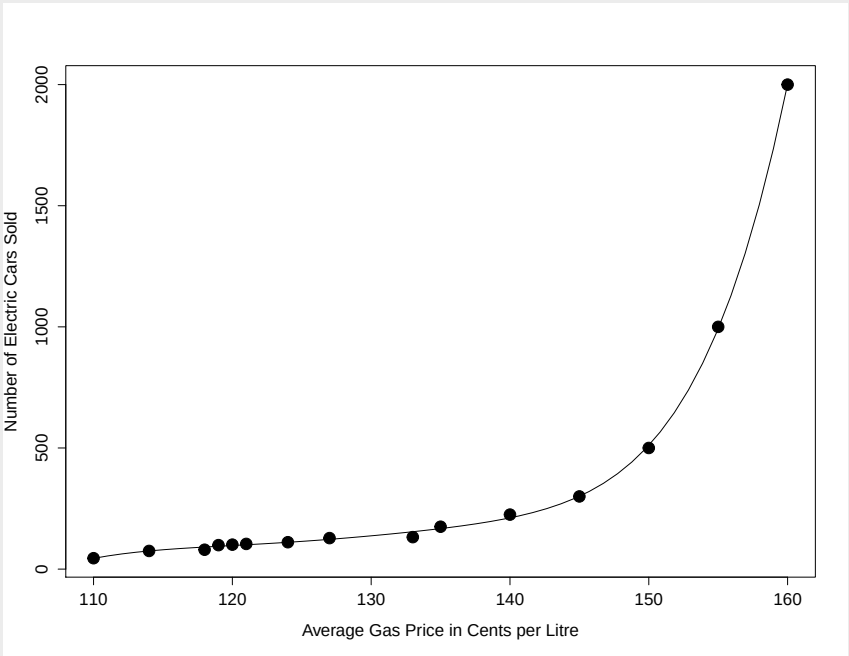
(**e**) Find the sum of squared errors (SSE).

$$SSE = \sum_{i=1}^{10} e_i^2 = (-7.401)^2 + (5.211)^2 + \ldots + (-20.385)^2 + (13.920)^2 = 847.236$$

(**f**) Comment on the residual plot.



There does not appear to be any obvious pattern in the residuals. Approximately half of the values fall above and below zero. Observations 1 through 8 fall within a reasonable distance from zero but observations 9 and 10 may raise concerns about heteroscedasticity or a non-constant mean. In the context of this problem, this is not unlikely. As gas prices continue to increase, the number of electric cars purchased may change exponentially due to a snowball effect as they gain popularity.



8.4 Model Assumptions

Assumptions 8.1 (Assumptions for Regression)**.**

1. *The deterministic component y is a linear function of the predictor x.*
2. *The ε terms are independent for each observation.*
3. *The ε terms are normally distributed.*
4. *The ε terms have a mean of 0.*
5. *The ε terms have constant variance σ^2 for all values x.*

Definition 8.10 (Homoscedacity)**.**
The variance of the residuals around the regression line is the same for all values of the predictor.

Definition 8.11 (Heteroscedasticity)**.**
The variance of the residuals around the regression line is not the same for all values of the predictor.

A more formal way of expressing assumption 5 in 8.1 is to state that we assume that the error terms are homoscedastic. This assumption is violated if the residuals are heteroscedastic.

Note 8.12.
A consequence of assumption 2 in 8.1 is that the ε are also independent of each other and that each of responses y_i are also independent of each other.

Note 8.13.
A consequence of assumption 3 in 8.1 is that the responses y_i are also normally distributed.

Note 8.14.
A consequence of assumption 5 in 8.1 is that the responses y_i also have some constant variance.

8.5 Measuring Variability with a Regression Model

In this section we will learn more about variability in the dependent variable y. If it appears that our model is able to explain a lot of the variability in the response, then this suggests that we have a good model.

Let's start by revisiting the SSE. Recall that the SSE is the sum of the squared difference between an actual observed response value and the predicted value of the response obtained from the model. (We can refer back to figure 8.3 to visualize the residuals).

A new value of interest is the sum of squares regression (SSR). This is the sum of the squared distance between a predicted response and the mean of all the responses.

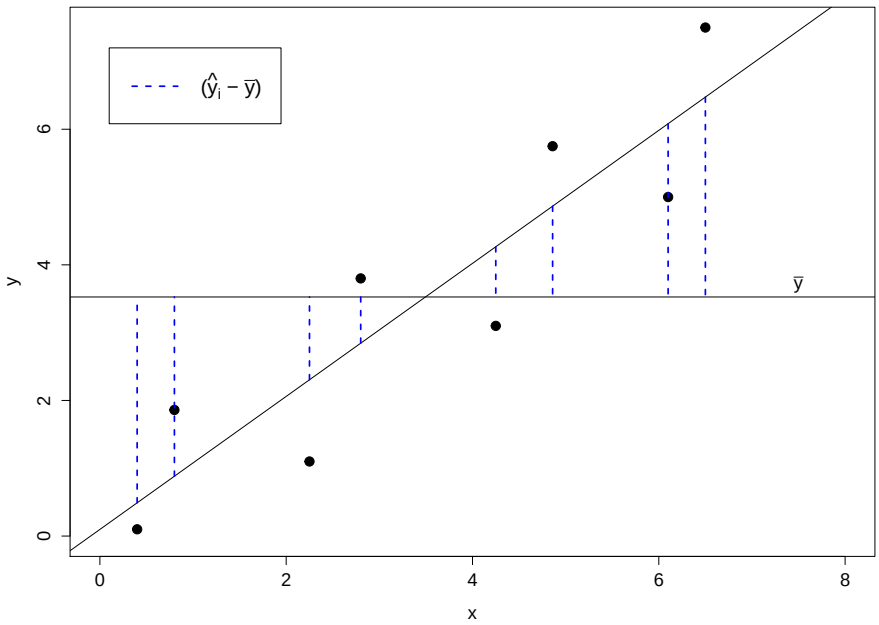


Figure 8.6: The blue dashed lines represent the distance between a predicted value of a response and the mean of all responses in the data set.

Definition 8.12 (Sum of Squares of Regression (SSR))**.**
The sum of squares of regression (SSR) is the sum of the squared differences between the

predicted value of the response for each observation and the mean of all recorded responses in the data set.

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \tag{8.5.22}$$

Figure 8.5 provides an example of the distances between predicted values of y and mean of all the measured responses $\bar{y}$. If we square these distances and add them up then we get the sum of squares of regression.

The final value of interest for measuring variability is the total sum of squares $SSTotal$.

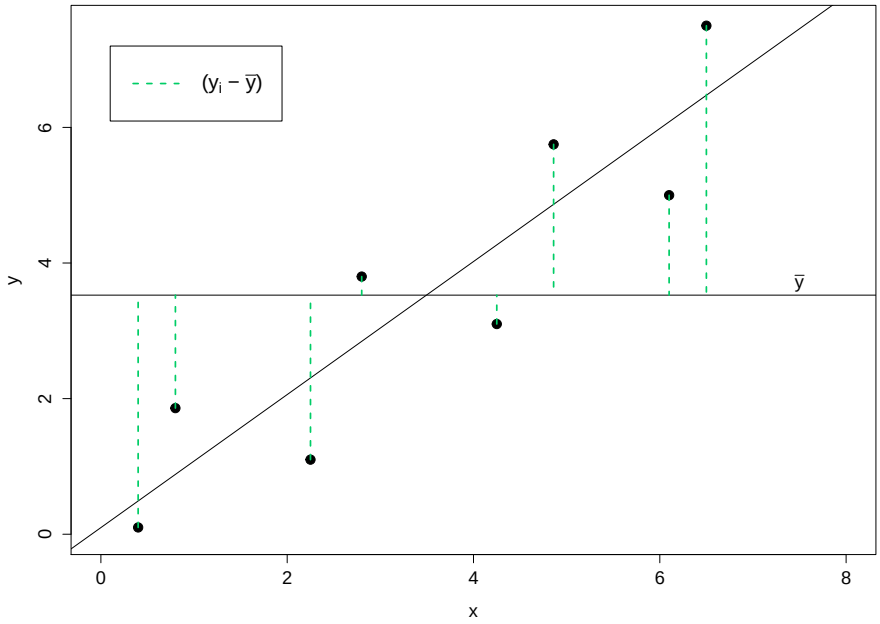


Figure 8.7: The blue dashed lines represent the distance between a predicted value of a response and the mean of all responses in the data set.

Definition 8.13 (Total Sum of Squares ($SSTotal$))**.** *The total sum of squares (SSTotal) is the sum of the squared differences of the response each observation from the mean of all recorded responses in the data set.*

$$SSTotal = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{8.5.23}$$

Note 8.15.
Some textbooks may may annotate SSTotal as SST or TSS.

Note 8.16.
The SSTotal is the same as SS_{yy}.

$$SSTotal = SS_{yy} \tag{8.5.24}$$

We have two labels for the same value since it is more appropriate to refer to one over the other depending on the context.

The total sum of squares is a measure of the amount of variability in the data. For linear regression models the total sum of squares is also obtained by adding the SSE and the SSR.

$$SSTotal = SSE + SSR \tag{8.5.25}$$

8.5.1 Coefficient of Determination and Coefficient of Correlation

The coefficient of determination (r^2) and the coefficient of correlation (r) are values that we can use to assess the quality of our model. Recall that we can calculate the SS_{yy} using 8.2.6 or 8.2.12 and by note 8.16. We can use SS_{yy} along with the SSE or the SSR to calculate r^2 and r.

Definition 8.14 (Coefficient of Determination (r^2))**.**
The coefficient of determination (r^2) is a measure of the total variability in the dependent variable (y) that is explained by the independent variable (x) through a regression model.

$$r^2 = \frac{explained\ variation}{total\ variation} \tag{8.5.26}$$

$$= \frac{SSR}{SSTotal} \tag{8.5.27}$$

$$= \frac{SS_{yy} - SSE}{SS_{yy}} \tag{8.5.28}$$

Note 8.17.
Using some simple manipulation we can easily show that 8.5.28 can be expressed as

$$r^2 = 1 - \frac{SSE}{SS_{yy}} \tag{8.5.29}$$

The coefficient of determination is a value that is between 0 and 1 (i.e. $0 \leq r^2 \leq 1$). High values of r^2 (i.e. values of r^2 close to 1) suggest that a lot of variability in the response is explained by the predictor and low values of r^2 (i.e. values of r^2 close to 0) suggest that very little variablity in the response is explained by the predictor. High r^2 values suggest that we have a good regression model and low r^2 values suggest that we do not have a good regression model for the particular data being analyzed.

We now move on to the coefficient of correlation which is another value we use to determine whether we have a good regression model or not.

Definition 8.15 (Coefficient of Correlation (r))**.**
The coefficient of correlation is measure of the strength of the relationship between the response y and the predictor x.

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}} \tag{8.5.30}$$

The coefficient of correlation is a value that is between -1 and $+1$ (i.e. $-1 \leq r^2 \leq 1$). Values of r that are close to $+1$ suggest that we have strong positive correlation and values of r that are close to -1 suggest that we have strong positive correlation. Both of these cases suggest that there is a strong linear relationship between the response and the predictor. In simple linear regression, a value of r that is close to 0 suggests that a linear relationship does not exist between the response and the predictor.

We can also obtain r from r^2 and vice versa. If we already know r then we simply have to square this value to obtain r^2. However if we know r^2 we have to be careful when we use it to obtain r since r can be either positive or negative. The sign of r depends on the sign of the slope.

$$r = \pm\sqrt{r^2} \tag{8.5.31}$$

The sign of r depends on the sign of β_1 in the linear model $\hat{y} = \hat{\beta}_1 x + \hat{\beta}_0$.

$$\begin{cases} \hat{\beta}_1 > 0 \implies r > 0 \\ \hat{\beta}_1 < 0 \implies r < 0 \end{cases} \tag{8.5.32}$$

Note 8.18.
We can also use r^2 to calculate $\hat{\beta}_1$using :

$$\hat{\beta}_1 = r\frac{s_{yy}}{s_{xx}} \tag{8.5.33}$$

Note 8.19.
r^2 and r are both unit-less values.

8.6 Inference Procedures on the Slope

In regression analysis we are mainly interested in inferences procedures on the slope. In the case of simple linear regression this means that we are interested in $\hat{\beta}_1$. We will use a value called the standard error of regression s for our inference procedure.

Recall we mentioned in assumptions 8.4 that for a linear model of the form $y = \beta_1 x + \beta_0 + \varepsilon$ one of the assumptions is that $\varepsilon \sim N(0, \sigma^2)$ (i.e. the ε terms are normally distributed with mean 0 and some constant variance σ^2). We can use s^2 to estimate σ^2 which is the estimate of the variance of the error terms.

Definition 8.16 (s^2)**.**
In a linear model of the form $y = \beta_1 x + \beta_0 + \varepsilon$ where $\varepsilon \sim N(0, \sigma^2)$, we can estimate σ^2 with s^2 which is

$$s^2 = \frac{SSE}{n-2} \tag{8.6.34}$$

If we take the square root of s^2 we get s which is referred to as the standard error of the regression.

Definition 8.17 (Standard Error of Regression)**.**
The standard error of regression is the average distance between an observed values and their corresponding predicted value on the regression line.

$$s = \sqrt{s^2} \tag{8.6.35}$$

The standard error of regression measures how far away the predicted values on the regression line are from actual values of the response variable on average. A smaller s suggests that we have a better fit on average.

We are typically interested in whether our model should include a slope of not and we can achieve this by constructing confidence intervals on the slope and hypothesis tests on the slope.

8.6.1 Confidence Intervals on the Slope

Confidence Interval 8.1 (Confidence Interval on β_1)**.**
A $(100-\alpha)\%$ confidence interval on β_1 is constructed using

$$\hat{\beta}_1 \ \pm \ t_{(\alpha/2,\ n-2)}\left(\frac{s}{\sqrt{SS_{xx}}}\right) \tag{8.6.36}$$

Note 8.20.
It's important to note that the value from the $t-$distribution is taken at $n-2$ degrees of freedom. We use $n-2$ degrees of freedom since we do not know β_1 or β_0 (i.e. two unknown values).

The value $\frac{s}{\sqrt{SS_{xx}}}$ is called the standard error of the slope. This value multiplied by $t_{(\alpha/2,\ n-2)}$ is the margin of error of the slope.

When we construct a confidence interval on β_1 we typically interested in whether the confidence interval contains 0 or not. This occurs when the bounds of the confidence interval

are of opposite sign (i.e. the lower bound is negative and the upper bound is positive), If the confidence interval contains 0, this suggests that 0 is a plausible value for β_1. This in turn suggests that our model should not include a slope and there does not exist a relationship between x and y. This is because if the true value of β_1 really is 0, then it does not matter which value of x we enter into our model since multiplying the entered x value by 0 will not have an effect on the response.

8.6.2 Hypothesis Tests on the Slope

Hypothesis Test 8.1 (Hypthesis Test on β_1)**.**
Suppose we are interested in any one of the following hypothesis tests on the population mean:

- $H_0 : \beta_1 = \beta_{hyp}$ *vs.* $H_a : \beta_1 > \beta_{hyp}$
- $H_0 : \beta_1 = \beta_{hyp}$ *vs.* $H_a : \beta_1 < \beta_{hyp}$
- $H_0 : \beta_1 = \beta_{hyp}$ *vs.* $H_a : \beta_1 \neq \beta_{hyp}$

The test statistic for a hypthesis test on β_1 is

$$t^* = \frac{\hat{\beta}_1 - \beta_{hyp}}{s/\sqrt{SS_{xx}}} \tag{8.6.37}$$

Reference distribution: the $t-$distribution at $n-2$ degrees of freedom.

Alternative Hypothesis	*P-value*
$H_a : \beta_1 > \beta_{hyp}$	*Area to the right of t^**
$H_a : \beta_1 < \beta_{hyp}$	*Area to the left of t^**
$H_a : \beta_1 \neq \beta_{hyp}$	*Sum of the areas in the tails of t^* and $-t^*$*

Note 8.21.
We use βhyp to represent the hypothesized value of the slope in 8.1.

Example 8.5.
Casinollama would like to quantify the benefits of its customer loyalty program. When a player frequents the casino, they can use their casino player card to gain points that can be

redeemed for free parking service, hotel rooms, refreshments, etc. This of course comes at a cost to the casino but allows them to keep players in the casino for longer. *Casinollama* samples the opening and closing balances of 500 player accounts over the course of the day as well as the number of hours spent by each player in the casino. 4 observations from the data are shown in the table below. When working with losses in statistics, it is often convention to represent gains as negative losses. For example, Player 03144 caused the casino a loss of −\$9000 which is in fact a gain of \$9000.

Player ID	Starting Time	Ending Time	Total Time (in hrs)	Opening Balance	Closing Balance	Loss to Casino
11378	11:00	13:30	2.5	\$1000	\$1050	\$50
12794	18:30	00:30	6	\$50	\$500	\$450
03144	21:30	3:30	6	\$10000	\$1000	-\$9000
⋮	⋮	⋮	⋮	⋮	⋮	⋮
14183	17:30	20:00	2.5	\$500	\$400	-\$100

Casinollama calculates the following using the sample data.

$$\sum x_i = 2602 \quad \sum y_i = -142536$$
$$\sum x_i^2 = 17449 \quad \sum y_i^2 = 51619463$$
$$\sum x_i y_i = -887526$$

where y represents the loss incurred by the casino and x represents time spent in the casino. Using the information provided,

(**a**) Estimate β_0 and β_1.

Start by finding SS_{xx}, SS_{yy}, and SS_{xy}.

$$SS_{xy} = \left(\sum x_i y_i\right) - \frac{(\sum x_i)(\sum y_i)}{n} = -887526 - \frac{(2602)(-142536)}{500} \approx -145769$$
$$SS_{xx} = (\sum x_i^2) - \frac{(\sum x_i)^2}{n} = 17449 - \frac{(2602)^2}{500} \approx 3908$$
$$SS_{yy} = (\sum y_i^2) - \frac{(\sum y_i)^2}{n} = 51619463 - \frac{(-142536)^2}{500} \approx 10986440$$

Using our formula for $\hat{\beta}_1$,

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{-145769}{3908} = -37.30$$

In order to estimate β_0, we need to find $\bar{x}$ and $\bar{y}$.

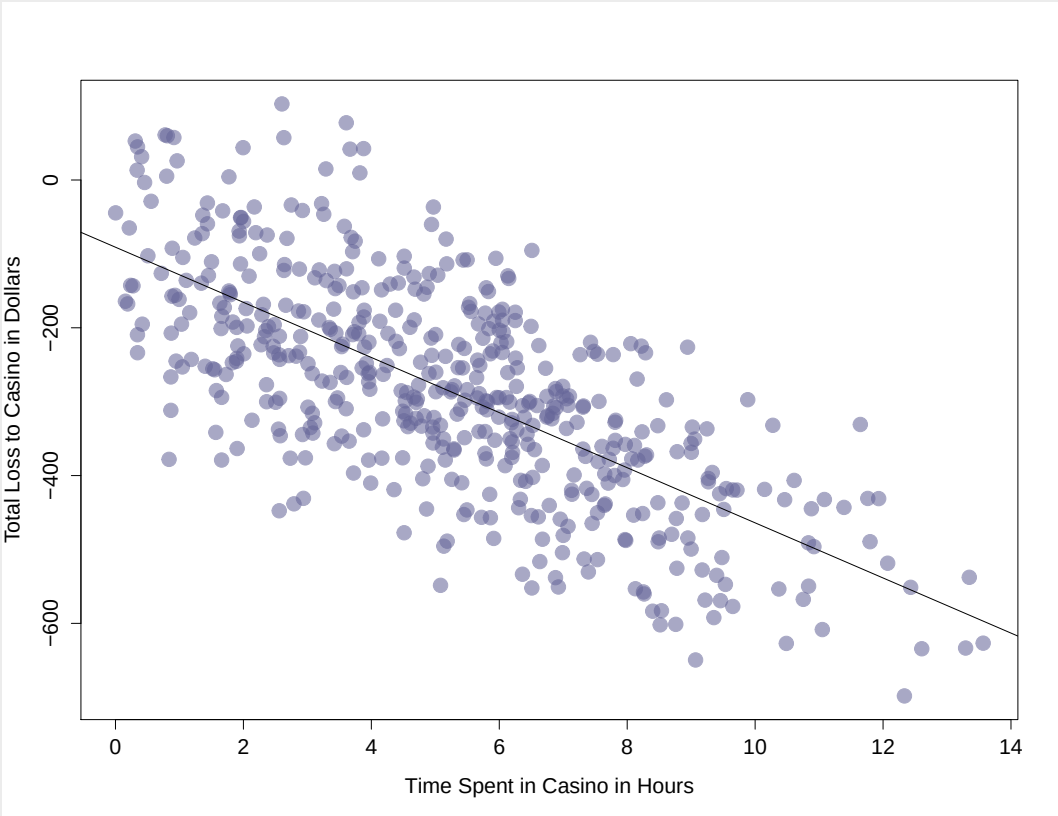
$$\bar{y} = \frac{\sum y_i}{n} = \frac{-142536}{500} = -285.07 \quad \bar{x} = \frac{\sum x_i}{n} = \frac{2602}{500} = 5.20$$

Using our formula for $\hat{\beta}_0$,

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = -285.07 - (-37.30)(5.20) = -91.11$$

Our least squares regression model is

$$\hat{y} = -37.30\ x - 91.11$$



(**b**) Interpret $\hat{\beta}_1$ and $\hat{\beta}_0$.

For every one hour increase in time spent at the casino, the total loss to the casino decreases by \$37.30 dollars (or increases by -\$37.30) on average.

A player that spends 0 hours at the casino results in a total loss to the casino of -\$91.11 (that is, a gain of \$91.11 by the casino) on average.

(**c**) Find SSE.

Using the values we calculated in part (a),

$$SSE = SS_{yy} - \hat{\beta}_1 SS_{xy} = 10986440 - (-37.30)(-145769) \approx 5549256$$

(**d**) Find SSR.

Recall that $SSTotal = SS_{yy}$. Therefore,

$$SSTotal = SSE + SSR \Rightarrow SSR = SSTotal - SSE = 10986440 - 5549256 = 5437184$$

(**e**) Find the coefficient of determination, r^2.

Using our calculations from part (a) and (b),

$$r^2 = 1 - \frac{SSE}{SS_{yy}} = 1 - \frac{5549256}{10986440} = 0.49$$

Alternatively, using our calculations from part (b) and (c),

$$r^2 = \frac{SSR}{SSTotal} = \frac{5437184}{10986440} = 0.49$$

(**f**) Find the coefficient of correlation, r.

Since $\hat{\beta}_1 = -37.30 < 0$,

$$r = -\sqrt{r^2} = -\sqrt{0.49} = -0.7$$

(**g**) Find s.

$$s^2 = \frac{SSE}{n-2} = \frac{5549256}{498} = 11143.08 \quad \Rightarrow s = +\sqrt{s^2} = 105.56$$

(**h**) Find a 95% confidence interval for β_1.

A 95% confidence for β_1 is of the form

$$\hat{\beta}_1 \pm t_{(\alpha/2,n-2)} \left(\frac{s}{\sqrt{SS_{xx}}} \right)$$

We have all of the components we need to build the interval from parts (a)-(g) except for $t_{(\alpha/2),n-2)}$. Using our t-distribution table,

$$t_{(0.025,498)} = 1.96$$

Notice that this is approximately the same as $z_0.025$! This is because our sample size is relatively large.

Plugging in all of our values, a 95% confidence interval for β_1 is

$$\hat{\beta}_1 \pm t_{(\alpha/2,n-2)} \left(\frac{s}{\sqrt{SS_{xx}}} \right) = -37.30 \pm 1.96 \left(\frac{105.56}{\sqrt{3908}} \right) = -37.30 \pm 3.31 = (-40.61, -33.99)$$

(**i**) Conduct the following hypothesis test at the $\alpha = 0.05$ level.

$$H_0 : \beta_1 = 0 \; vs. \; H_a : \beta_1 \neq 0$$

We start by finding our test statistic.

$$t^* = \frac{\hat{\beta}_1 - \beta_{hyp}}{s/\sqrt{SS_x x}} = \frac{-37.30 - 0}{105.56/\sqrt{3908}} = -22.09$$

Recall that the mean of the t-distribution is zero. Since our test statistic t^* is so extreme (i.e very far from zero), our p-value will be extremely small.

Since $H_a : \beta_1 \neq 0$ is a two-sided hypothesis, the p-value is the sum of the area to the left of $-|t^*|$ and $+|t^*|$.

$$p - value = P(t < -|t^*|) + P(t > +|t^*|) = 2 \times P(t < -22.09)$$

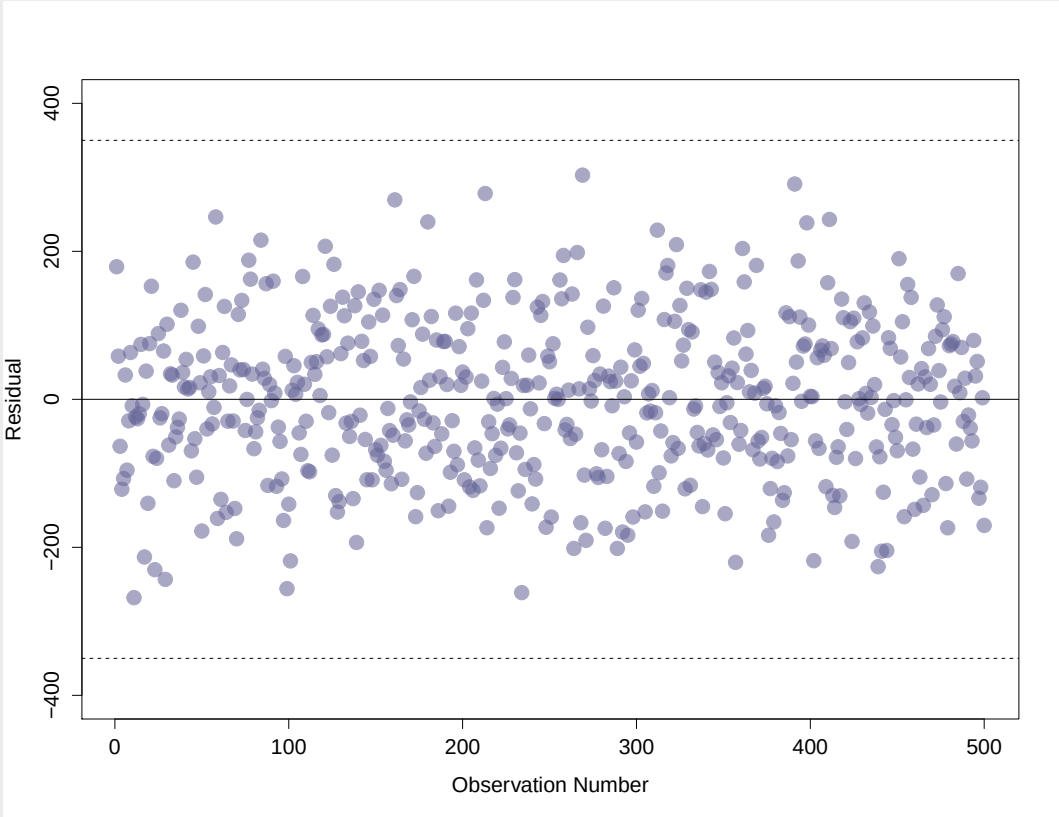
Using software, it is determined that $p - value < 2.2 \times 10^{-16}$.

At the $\alpha = 0.05$ level, $p - value < 2.2 \times 10^{-16} < \alpha = 0.05$. Via Rule 7.1 this implies that there is strong evidence against the null hypothesis and thus we reject the null in favour of the alternative hypothesis that $\beta_1 \neq 0$.

(**j**) Comment on the results of the 95% confidence interval and hypothesis test.

Since 0 was not contained within the 95% confidence interval for β_1 and we rejected the null hypothesis that $\beta_1 = 0$, there is strong evidence of a relationship between the length of time spent in the casino and the losses incurred by the casino. With a negative sample slope coefficient $\hat{\beta}_1 = -37.30$, we believe that the more time that a player spends within the casino, the less money the casino loses. Based on these results, it is within the casino's best interest to continue their player loyalty program.

(**k**) The residual plot is presented in the figure below. Does the residual plot raise any concerns regarding the validity of our results?



There does not appear to be any obvious pattern in the residuals. Through observation, it appears that approximately half of the values fall above and below zero. All of the values seem to fall within a band around zero. In conclusion, the residual plot does not raise any concerns regarding the validity of our results.

Index

Additive property, 34
    Of mutually exclusive events, 34

Bayes' Theorem, 40
Bias
    Measurement error bias, 10
    Non-response bias, 10
    Selection bias, 9
Binomial coefficient, 44
Boxplots, 24

Central limit theorem, 70
Coefficient of correlation, 143
Coefficient of determination, 142
Combinations, 44
Complement, 33
    Property of complements, 34
Confidence intervals, 74
    Assumptions, 80, 88, 91, 94
    Definition, 74
    On a difference of two means, 81
    On a difference of two proportions, 89
    On a proportion, 79
    On paired data, 91
    On the mean, 76
    On the slope, 145
    One sample, 76
    Two sample, 81

Data
    Continuous data, 4
    Discrete data, 4
    Paired data, 91
    Qualitative data, 5
    Quantitative data, 4
    Types of data, 4
Decision error, 115
    Type I error, 115
    Type II error, 116
Decision rule, 95
Degrees of freedom, 67
Distribution
    Bernoulli, 54
    Binomial, 55
    Continuous uniform, 57
    Normal, 58
    Sampling distribution, 70
    t-distribution, 66

Empirical rule, 62
Event, 31
    Mutually exclusive events, 34
Experiment, 30
Extrapolation, 129

Factorial, 44
Frequency, 22
    Relative frequency, 22

Histograms, 21
Hypothesis
    Alternative hypothesis, 96
    Null hypothesis, 96
Hypothesis tests, 95
    Assumptions, 105, 110, 113, 114
    On a difference of two means, 106
    On a difference of two proportions, 111
    On a proportion, 104
    On paired data, 113
    On the mean, 101
    On the slope, 146
    One sample, 101
    Two sample, 106
    Types of, 96

Inference, 72
Introduction to inferential statistics, 7
On the slope, 144
Inter-quartile range, 16
Intercept, 122
Interpretation, 129
Interpolation, 129
Intersection, 33
Introduction, 2

Level of significance, 99

Mean, 13
Sample mean, 13
Median, 13
Mode, 14

Outlier, 21
Overview, 1

P-value, 98
Parameter
Definition, 7
Percentile, 15
Plots
Bosplots, 24
Histograms, 21
Residual plots, 134
Stem and leaf plots, 28
Point estimate, 70
Pooled method, 87
Population
Definition, 3
Probability, 30
Conditional probability, 39
Density function, 51
Independence, 39
Mass function, 47

Quartile, 15

Random variables, 46
Continuous, 51
Definition, 46
Discrete, 46
Regression
Assumptions, 139
Least squares regression line, 134
Simple linear regression, 119
Residuals, 132
Heteroscedasticity, 139
Homoscedacity, 139
Residual plots, 134

s^2, 144
Sample
Definition, 3
Pooled sample standard deviation, 87
Sample point, 31
Sample space, 30
Sample standard deviation, 15
Sample surveys, 7
Sample variance, 14
Significance, 117
Practical significance, 117
Statistical significance, 117
Skewness, 19
Slope, 122
Interpretation, 128
Standard deviation, 15
Sample standard deviation, 15
Standard error, 73
Standard error of the regression, 145
Statistic
Definition, 8
Test statistic, 97
Statistics
Definition, 2
Descriptive statistics, 2, 12
Inferential statistics, 2
Introduction, 1
Stem and leaf plots, 28
Studies
Experimental studies, 6
Observational studies, 6
Sum of squared error, 133
sum of squares of regression, 140

Total sum of squares, 141

Union, 33

Variable
Dependent variable, 119
Independent variable, 119
Variance, 14

Pooled sample variance, 87
Sample variance, 14
Venn diagrams, 36
Welch-Satterthwaite method, 85
Z score, 60