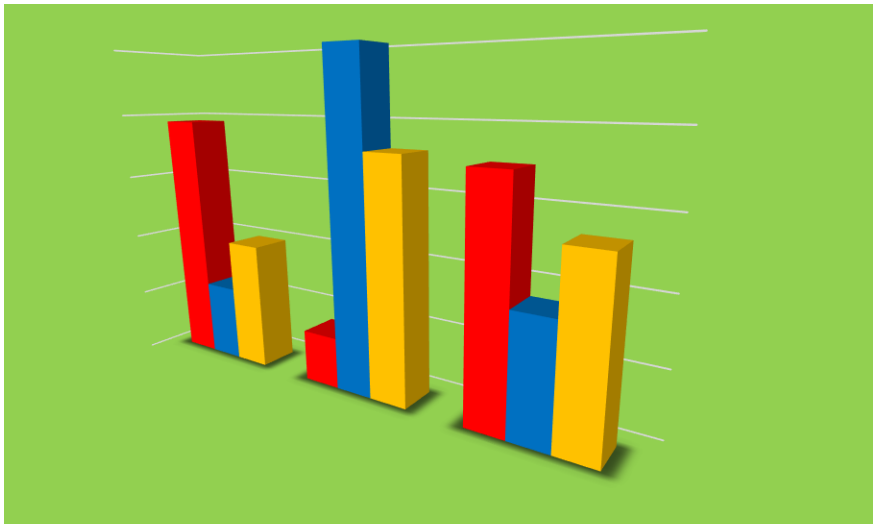
STATISTICS WITH TECHNOLOGY FOR HIGH SCHOOL



Maria Emilia Alfaro & John Hicks

East Tennessee State University

Statistics with Technology for High School

Maria Emilia Alfaro and Dr. John Hicks

2023

Statistics with Technology for High School

Contents

About the Authors

Maria Emilia Alfaro was born in 1999 in Azul, Buenos Aires, Argentina. She completed her secondary education in Colegio Inmaculada Concepción de Azul and studied English as a second language and the CEI English Institute in Azul. In 2018, she decided to pursue a Bachelor's degree in the United States while playing Division 1 college tennis. She played college tennis for five years and graduated from East Tennessee State University (ETSU) with a Bachelor's in Mathematics concentrated in Statistics on May 2022. She is currently pursuing a Master's degree in Applied Data Science at ETSU. She has also worked as a mathematics, statistics and Spanish tutor and as a mathematics and statistics summer instructor for the ETSU Upward Bound Summer Program while being a student.

Dr. John Hicks has lived in East Tennessee his entire life. John received his Master's and Bachelor's Degrees from East Tennessee State University in Mathematics with a Concentration in Statistics. He received his PhD. in Applied Mathematics from the University of Tennessee.

Introduction

This booklet was developed as a part of the STEM L.D project at East Tennessee State University, founded by the Niswonger Foundation. For its elaboration, the authors of this work have drawn upon their extensive knowledge and experience to curate a collection of original explanations, examples and exercises. These materials have been meticulously crafted to illustrate key concepts and provide readers with valuable opportunities for practice and skill development, built upon the fundamental framework of the Scientific Method, which is at the heart of an introductory course in Probability and Statistics. The Scientific Method can be summarized as follows:

1. Research Question
2. Data Collection
3. Analysis
4. Conclusion

By embracing this systematic approach, readers will gain a solid foundation in Probability and Statistics, and the practical application of R will aid in more complex computations and data visualization.

The following is a work that is not intended to be a comprehensive work on Probability and Statistics, but just an introduction to some of the topics as well as an introduction to using R. R is an open sourced statistical software program that at its core, is a versatile tool functioning as a calculator, in which the inputting of code results in valuable output that aim to aid in complex computations and data visualization. A brief overview of R is found at the end of this booklet, and code excerpts are provided to demonstrate the concepts applications in R.

This booklet serves as just an introduction, by not going into great detail on some topics, but when appropriate, further elaboration is given to gain a better understanding.

1 What is Statistics

We will present here a list of terms that will be used throughout this booklet.

1.1 Terminology

Statistics: A science that studies the collection, analysis and interpretation of data.

Individual: Who the data is measured upon. That is the subjects we collect information on. This could be people, places, or objects.

Data: Any kind of information with meaning or context that can be collected and analyzed from an individual. Examples of data are the number of inhabitants from a city, the number of female and male students in a school, the blood pressure of patients in a hospital.

- **Individual**: Who the data is measured upon. That is the subjects we collect information on. This could be people, places, or objects.
- **Variables**: Different attributes or aspects that a set of data can have. For example, suppose we construct a data set of dogs from an animal shelter and we take note of the age, gender, breed and weight of each dog. Age, gender, breed and weight are the variables.

Data can be classified into two types based on its nature: categorical or numerical data.

- **Categorical Data**: Data that can be classified into categories. It is also known as qualitative data. Some examples are gender, religion or education level. It can be further divided into ordinal or nominal data.
  - **Ordinal Data:** Categorical data with an order of hierarchy. For example, the education level or rating scales.
  - **Nominal data:** Categorical data without any numerical order. For example, gender, religion, or movie genre.
- **Numerical data:** Data that can be quantified and measured using numbers (it makes sense to find an average for, and that average has an interpretation). It is also known as quantitative data. Some examples are height, blood pressure, or ACT score. It can be further divided into continuous or discrete data.
  - **Continuous data:** Numerical data that can take any real value within an interval or range of values. For example, height or blood pressure.
  - **Discrete data:** Numerical data that can only be measured using whole numbers. That is, a variable that takes on only a finite number of values. For example, number of speeding tickets, or ACT score.

Distribution: The distribution of a variable represents the specific values the variable takes and the frequency the variable takes these values.

Statistical Analysis: The process of using statistical methods to collect, analyze and draw conclusions about data. Typically, we utilize sample statistics to come up with estimates for a population parameter.

Population: It is the entire group or set of individuals that we are interested in describing. For example, if we wanted to study the life expectancy of blue whales, then our population is all blue whales in the world. Because it is impossible to gather together all the blue whales in the world, we tend to work with samples.

Parameter: A numerical value describes the population of interest. The values of the population parameters are typically unknown and will be estimated by a sample statistic. The notation for some common parameters and their notation are as follows:

- Mean (μ)
- Standard Deviation (σ)
- Proportion (p)

Sample: A representative subset of the entire population. For example, following the above example, a possible sample for the study would be a group of 30 blue whales. This in not all existing blue whales in the world but a small subset of them that we will use to gather and analyze data.

Statistic: A numerical that describes the sample. The statistics will be discussed in greater detail as we proceed in this booklet. The value of a statistic will act as an estimate to the unknown population parameter. Some common examples of statistics and their notation are as follows:

- Mean ($\bar{x}$)
- Standard Deviation (s)
- Proportion ($\hat{p}$)

Data collection: The process of gathering data.

Sampling Designs: Methods of collecting data. Some of them are:

- **Simple Random Sampling (SRS)**: A sampling technique in which every individual in the population has an equal chance of being selected. In another words, as the name indicates, it is random. For example, randomly selecting 100 individuals from a directory with 1245 people.
- **Systematic Sampling**: A sampling technique in which every nth individual is selected. For example, selecting all individuals located in a multiple of 20 positions from a phone directory.
- **Cluster Sampling**: A sampling technique in which the population is divided into smaller groups with certain characteristics in common and randomly selecting some of these subgroups. For example, we can divide the student population from a university into freshmen, sophomores, juniors and seniors and then randomly select two of these categories.
- **Stratified Sampling**: A sampling technique in which the population is divided into smaller groups with certain characteristics in common, called **strata**, and a certain number of individuals from each subgroup is randomly chosen from each. For example, dividing the student population from a university into freshmen, sophomores, juniors and seniors and then randomly selecting 20 individuals from each subgroup or stratum.
- **Convenience Sampling**: It consists of taking a sample based on availability or ease. For example, selecting your neighbors to create a sample of citizens from a town because they are close to your location. This would be an example of a sample design to avoid since it will lack randomness.
- **Census**: Selecting all the individuals from the population.

Research Studies: Studies involving data analysis to try to discover or validate knowledge. Research studies classify their variables into predictor variables and response variable.

- **Predictor Variables**: These are variables that are hypothesized to have an effect on the response variable, so they are used to predict or explain the values of the response. They are also known as an **independent** variables or **explanatory** variables. For example, in a study investigating the impact of study time on exam scores, the amount of study time would be the predictor variable.

- **Response Variable**: It is the single variable that is being studied or analyzed to observe how it changes in response to changes in the predictor variable. It is also known as a **dependent** variable or **outcome** variable. In the previous example, the exam scores would be the response variable.

There are two types of studies:

- **Observational Study**: Type of study in which the variables are collected through the recording of data from individuals in their "natural state", without manipulating any conditions about the individuals involved. With observational studies we are not going to be able to discuss cause-and-effect relationships since we are not controlling any outside influences (**lurking variables**) on the individuals. We simply ask a question and record the response given by the individual of interest.

- **Experimental Study**: Type of study in which the investigator manipulated different treatments to try to compare outcomes of the response variable. For example, conducting a study about pesticide performance. The data collected will be based on plants growing naturally (no pesticide) or about plants in which the pesticide is applied. These are the treatments that are being manipulated based on the factor (explanatory variable) of type of application to plants.

Experimental Designs: The structure of an experimental study involving how variables are manipulated, subjects are assigned to different treatment groups, and data is collected and analyzed. Through proper design, researchers can draw meaningful conclusions about cause-and-effect relationships. Some different types of experimental designs are as follows:

- **Randomized Comparative Experiment**: An experiment that utilizes random allocation of subjects into at least two treatment groups. Comparisons are made on the response variable after the treatments have been applied.

- **Matched Pairs Experiment**: A design that can be thought of as a before and after situation. Two treatments are applied to subjects as similar as possible if not the same, and analysis is conducted on the differences within the treatments on the individual. For example, the weights of individuals prior to starting a diet and the weights of individuals after the diet has been implemented. The analysis is conducted on the differences with the pre and post weights on the individual.

- **Block Design**: A design in which the subjects are separated into similar disjoint groups referred to as blocks. The experiment is conducted within each of the blocks. (Similar to the logic behind a stratified random sample).

- **Blind Experiment**: A design in which the treatment group the individuals are participating in is kept secret from them. For example, administering a drug and a placebo to different patients without them knowing whether they took the real drug or the placebo.

- **Double-Blind Experiment**: A design in which neither the researcher nor the subjects know which treatment the individuals are receiving. For example, administering a drug and a placebo to different patients without them nor the investigator knowing who took the real drug and who took the placebo.

Exercises

1. Classify the following variables as either a categorical variable or quantitative variable.

Variable	**Type**
Age of students	
Brand of vehicle	
Flavor of Ice-Cream	
Monthly salary	
State of residency	
Level of satisfaction (Low, Medium, High)	
Temperature	

2. Classify the following categorical variables into ordinal or nominal.

Categorical Variable	**Type**
Plant species	
T-shirt size	
Race/Ethnicity	
Type of soil	
Are you fluent in another language?	
Sporting weight class	
Hair color	

3. Classify the following variables into discrete or continuous.

Quantitative Variable	**Type**
Number of days of pregnancy	
Body Mass Index (BMI)	
Distance traveled	
Car mileage	
Points scored	
Credit hours	
Number of citizens in a city	

4. Identify the population and the sample in each of the following examples.

   (a) Suppose we are interested in finding out what is the median expenditures in groceries per household in a small city in Tennessee. For this, we interviewed 48 households.

   (b) We are interested in studying the growth rate of new palm trees planted in Miami. To do this we recorded several variables (such as location area, temperature, etcetera) of 112 new palm trees planted in Miami for 75 days.

5. Identify the sampling techniques used in each of the following scenarios.

   (a) For a satisfaction survey conducted by a ground transportation survey, a questionnaire was handed to every 5th individual entering a bus from New York to Boston.

   (b) To know the population total in Uruguay, a set of volunteers knocked at every house in Uruguay to ask how many people lived in each of them.

(c) In a laboratory, mice are separated by species. We randomly selected 10 mice from each species to conduct a study.

(d) Aiming to study the average number of hours people watch TV, we interviewed the people who live in the same building as us.

(e) To study the average weight of Oregon-grew Russet potatoes, we randomly selected 30 potatoes and recorded their weights.

(f) Aiming to study transportation habits in cities with less than 50,000 inhabitants in Romania, we created a list of cities with less than 50,000 inhabitants and randomly selected three of those cities.

2 Categorical Data

The most common way to visualize categorical data is by utilizing **frequency tables**, **bar charts** or **pie charts**.

2.1 One-Way Tables

One-way tables (also known as **frequency tables**) are simple representations of categorical data that display the number of observations that fall into each specific category within the variable. We can display them vertically or horizontally. The vertical display consists of two columns and as many rows as specific values within the categorical variable. The horizontal display consists of two rows and as many columns as different categories within the variable.

Example:

Suppose a company collected data about how their employees' most frequent transportation choice to get to their offices. 42 workers stated they most frequently drive their cars, 13 workers bike, 26 take a bus, and 4 most frequently walk.

Using this information, we can construct a frequency table in both vertical and horizontal display.

Vertical display:

Trasnportation Choice	**Number of Workers**
Car	42
Bike	13
Bus	26
Foot	4

Horizontal Display:

Transportation Choice	Car	Bike	Bus	Foot
Number of Workers	42	13	26	4

2.2 Univariate Graphs

In this section we will examine the approriate graphs for examining one categorical variable at a time.

2.2.1 Bar Charts

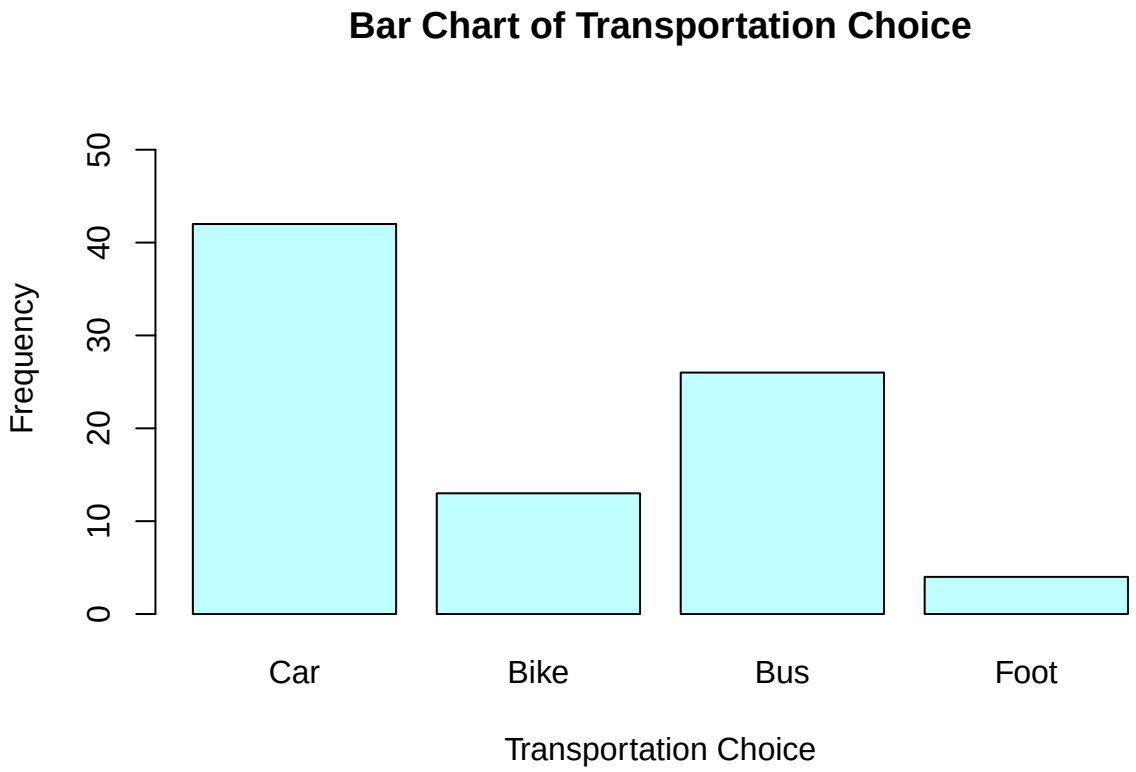
A **bar chart** is a graphical representation of data that uses bars to compare the specific categories within the variable. The height of each bar represents the frequency or relative frequency (proportion or percentage out of the total sample size) of occurrence of a particular category.

Characteristics of a bar chart:

- The x-axis displays the categories
- The y-axis displays the frequencies or relative frequencies
- Rectangles need to be the same width and are separated by gaps
- The order of the rectangles in not important

Example:

Suppose we are working with the same data about employees' transportation. A bar chart for these data would look like this:



An alternative representation of a bar chart is given in terms of a **relative frequency**. A relative frequency is the frequency within each specific value of the categorical variable to the total number observed, that is, a relative frequency is the proportion that is relative to the total number observed.

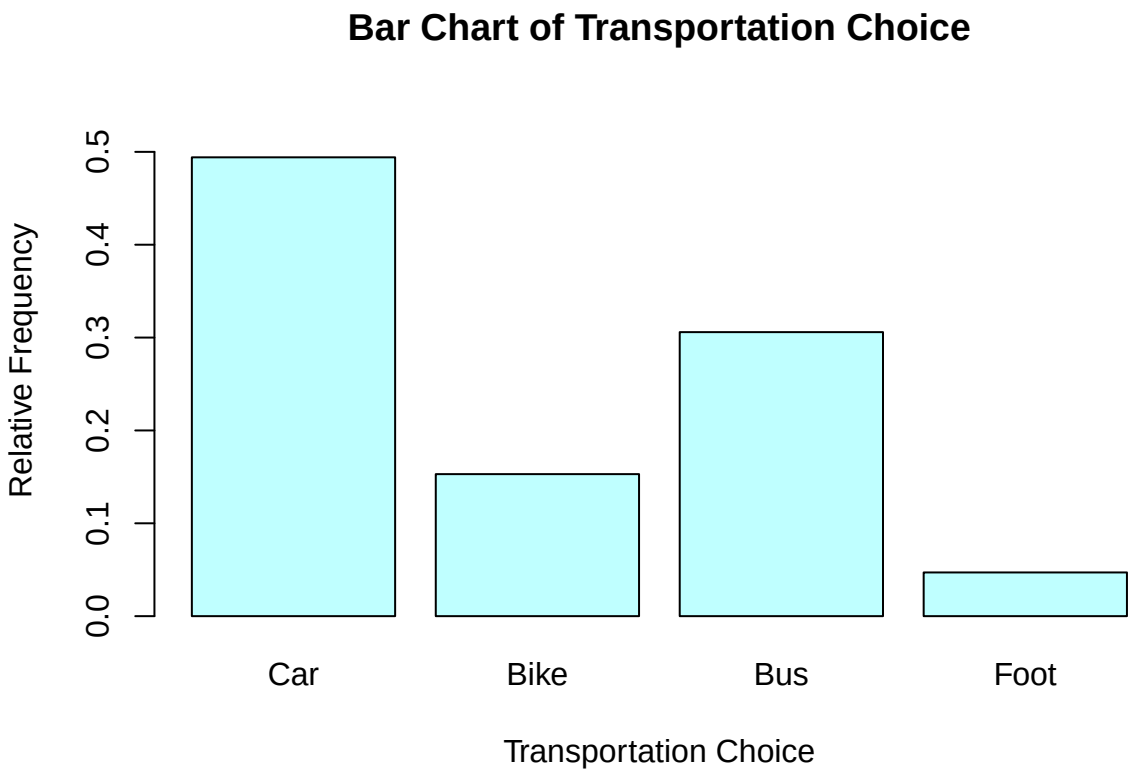
Example:

From our frequency table above, we will construct the relative frequencies by dividing the number of workers belonging to each category by the overall total of workers considered.

In this case the overall total is 85.

Trasnportation Choice	**Number of Workers**	**Relative Frequence**
Car	42	$42 \div 85 = 0.494$
Bike	13	$13 \div 85 = 0.153$
Bus	26	$26 \div 85 = 0.306$
Foot	4	$4 \div 85 = 0.047$

The relative frequency bar chart looks like this:



2.2.2 Pie Charts

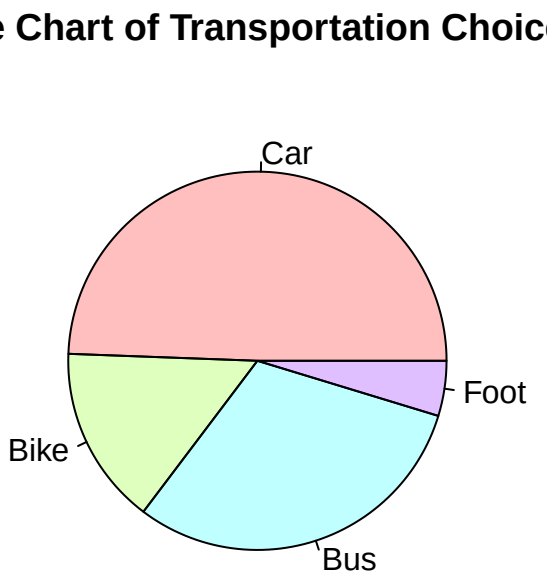
A **pie chart** is a type of data visualization that represents data as a circular shape divided into slices. Each slice represents the percentage of observations that belongs to each specific category within the variable.

Characteristics of a pie chart:

- There are no axes
- The entire circle represents the sum of all the values in the data set
- Slices represent the percentage of observations that belong to each category
- Usually, slices are identified with different colors
- There are no gaps between the slices

Example:

A pie chart for the above data would look like the following:



2.3 Two-Way Tables

Two-way tables are representations of two categorical variables that allow us to examine the relationship between the two variables. One categorical variable is represented as the column variable and the other variable as the row variable in the table. The intersection between the specific values of the column variable and row variable represents the number of individuals that belong to both specific values of the variable simultaneously. The totals are referred to as **marginal distributions** and represent the distribution of the individual row or column variable.

Example:

Suppose we work with the same data about employees' most frequent transportation choice, but this time we will also categorize them based on gender. Out of the 42 workers stated they most frequently drive their cars, 21 were females and 21 were male. Out of the 13 workers that bike, 3 were females and 10 were males. Out of the 26 that take a bus, 15 were females and 11 were males. Lastly, out of the 4 employers who walk, 3 were females and 1 was a male. We can represent this data in the following way using a two-way table.

	Female	**Male**	**Total**
Car	21	21	42
Bike	3	10	13
Bus	15	11	26
Walk	3	1	4
Total	42	43	85

Notice two-way tables are very rich in information. For example, notice that some of the observations that we can make are as follows:

- There are a total of 43 males
- 13 people bike to work
- There are 15 females who ride a bus
- The overall total is 85

Can you come up with more observations?

2.4 Multivariate Graphs

In this section, we will examine the appropriate graphs to examine the relationship between two categorical variables.

2.4.1 Clustered Bar Charts

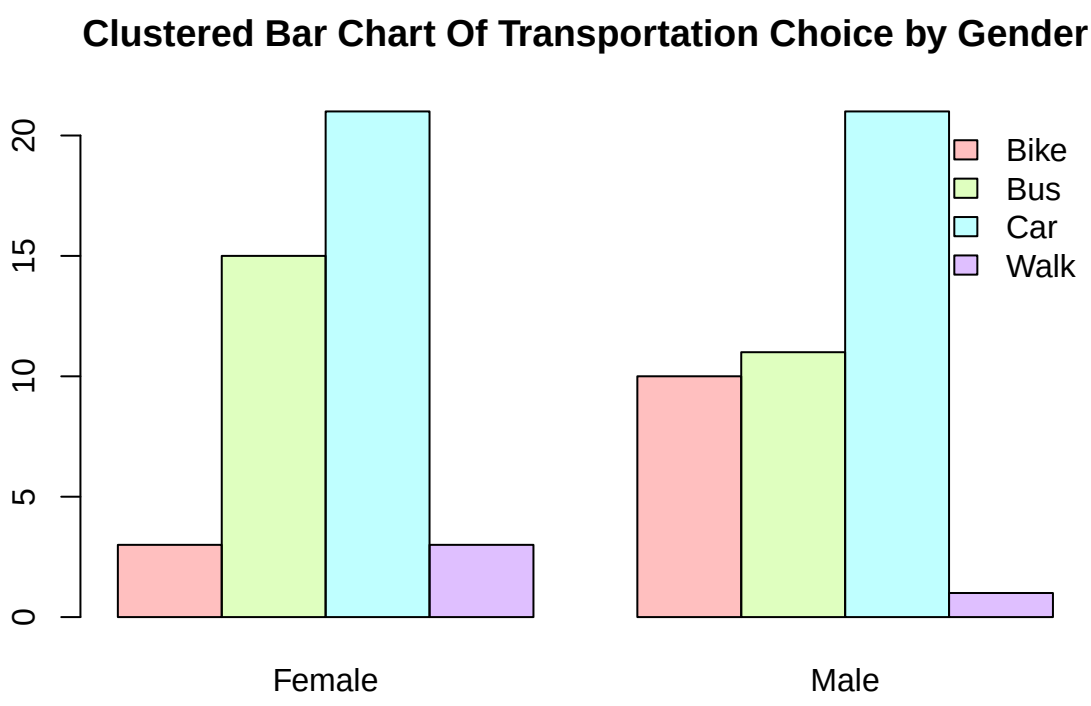
Clustered bar charts can be used to represent the relationship between two categorical variables. This plot will represent the intersection, or cells, of the two-way table by using either the counts or the relative frequencies.

Characteristics of a bar chart:

- The x-axis displays the specific categories within variable 1
- The categories within variable 2 are further subdivided within the specific values of variable 1 to represent the intersection
- The y-axis displays the frequency or relative frequency
- Rectangles that belong to the same category for variable 1 are not separated by gaps
- The order of the categories for variable 1 is not important

Example:

Suppose we are working with the same example about employers' preferred transportation and gender. A clustered bar chart for this example would look like this:


Clustered Bar Chart Of Transportation Choice by Gender
0
5
10
15
20
Bike
Bus
Car
Walk
Female
Male

Categorical Data in R

The following section will introduce some R code to help with the information graphing categorical variables and examining their distributions.

Character Strings and Visualization

Consider a simple random sample of 11 students and their classification. Suppose we have 2 freshmen, 3 sophomores, 4 juniors and 2 seniors.

To enter this information in R, we begin by creating a variable and filling it with a **character string**. To do this, we need to choose a name for our variable (in this case we will call it *Classification*) and we enter the observations using the **c()** function, separating them by commas, and writing each of them between quotation marks, because they are characters (categories).

```
# Creating a Character String
Classification<-c("Freshman","Freshman","Sophomore","Sophomore","Sophomore","Junior",
                  "Junior","Junior","Junior","Senior","Senior")
Classification
```

```
 [1] "Freshman"  "Freshman"  "Sophomore" "Sophomore" "Sophomore" "Junior"
 [7] "Junior"    "Junior"    "Junior"    "Senior"    "Senior"
```

We can also repeat certain values automatically by using the **rep()** function in R. This function takes in the values you wish to repeat and the number of times you wish to repeat it.These are two different ways of entering the same data.

```
# Creating a Larger Character String Using Repetition
Classification<-c(rep("Freshman",2),rep("Sophomore",3),
                  rep("Junior",4),rep("Senior",2))
Classification
```

```
 [1] "Freshman"  "Freshman"  "Sophomore" "Sophomore" "Sophomore" "Junior"
 [7] "Junior"    "Junior"    "Junior"    "Senior"    "Senior"
```

To visualize this information table formats, we can create frequency tables or relative frequency tables, using the following code.

```
# Frequency Table
tab<-table(Classification)
tab
```

```
Classification
 Freshman    Junior    Senior Sophomore
        2         4         2         3
```

```
# Relative Frequency Table
prop.table(tab)
```

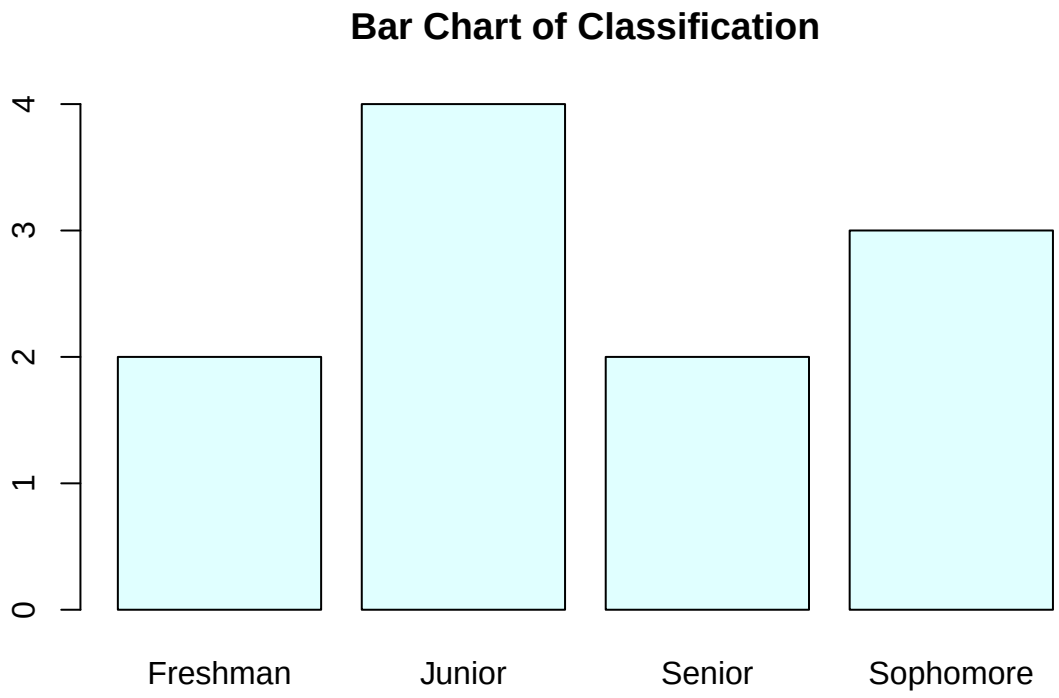
```
Classification
 Freshman    Junior    Senior Sophomore
0.1818182 0.3636364 0.1818182 0.2727273
```

Now, we can use these tables to create bar charts or pie charts of the data.

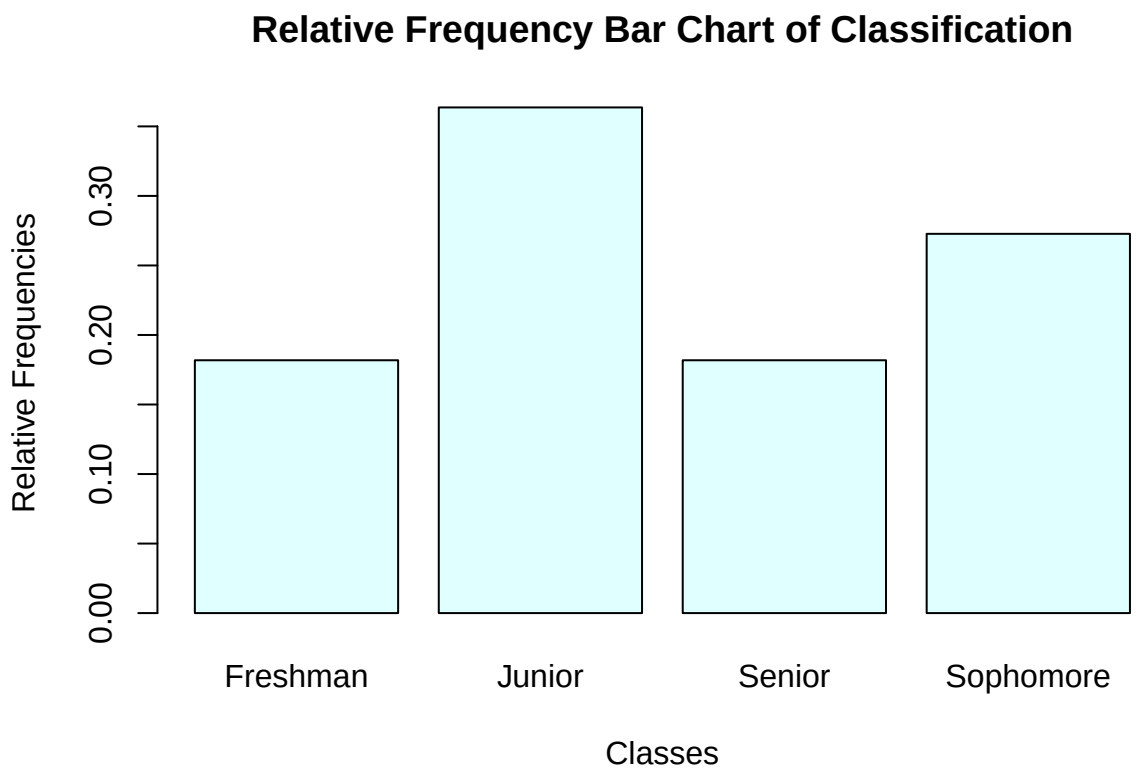
Note: The *col="..."* argument you will see below is optional and it is used to pick the color of the charts. Colors can be referenced to with color names, color codes or color palettes as we will see through the examples in this and the following chapters.

For bar charts,

```
# Frequency Bar Chart
barplot(tab,main="Bar Chart of Classification",col="lightcyan1")
```

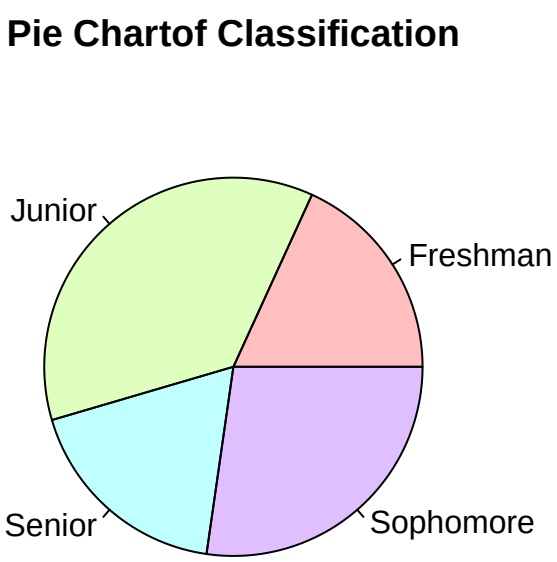



```
# Relative Frequency Bar Chart
barplot(prop.table(tab),main="Relative Frequency Bar Chart of Classification",
        col="lightcyan1",ylab="Relative Frequencies",xlab="Classes")
```



For pie charts we use,

```
# Pie Chart
pie(tab,main="Pie Chartof Classification", col=rainbow(4,alpha=0.25))
```



Two-Way Tables

Suppose we recorded our sample performance in a pass-fail exam and obtained the following information:

	Passed	Failed
Freshmen	1	1
Sophomore	2	1
Junior	2	2
Senior	2	0

To create this table in R we first need to manually add the grades keeping in mind the must match the order of the *Classification* variable that we created before. Once this is done, we can proceed to create the table.

```
# See the Variable
Classification
```

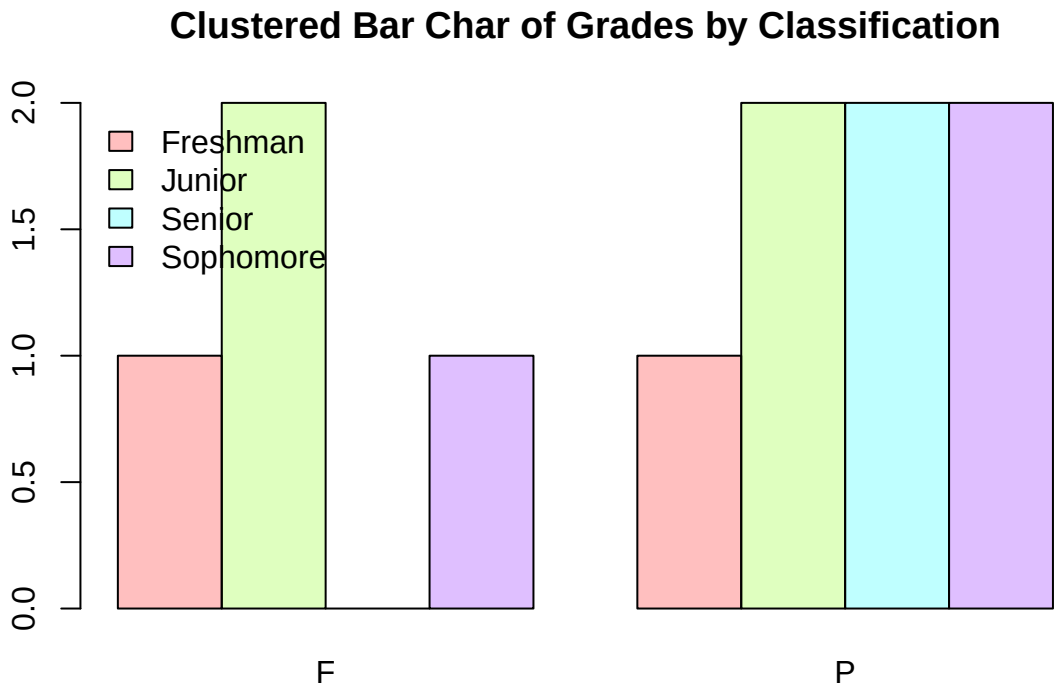
```
 [1] "Freshman"  "Freshman"  "Sophomore" "Sophomore" "Sophomore" "Junior"
 [7] "Junior"    "Junior"    "Junior"    "Senior"    "Senior"
```

```
# Create the Pass/Fail Variable
Grade<-c("P","F","P","P","F","P","P","F","F","P","P")

# Create the Two-Way Table
table(Classification,Grade)
```

```
               Grade
Classification F P
     Freshman  1 1
     Junior    2 2
     Senior    0 2
     Sophomore 1 2
```

And we create a clustered bar chart using,

```
# Create the Bar Chart
barplot(table(Classification,Grade),beside=TRUE,
        col=rainbow(4,alpha=0.25), main="Clustered Bar Char of Grades by Classification")

# Add a Legend
legend("topleft",legend=rownames(table(Classification,Grade))
       ,fil=rainbow(4,alpha=0.25),bty="n")
```



Exercises

1. Create a bar chart of the following table that summarizes favorite ice-cream flavor of 12 5-year-old kids.

Flavor	**Frequency**
Strawberry	2
Vanilla	3
Chocolate	5
Other	2

2. The following clustered bar chart summarizes the audience of a play presented in a local theater. Based on the graph, answer the following questions.

   (a) How many female adults came to the play?

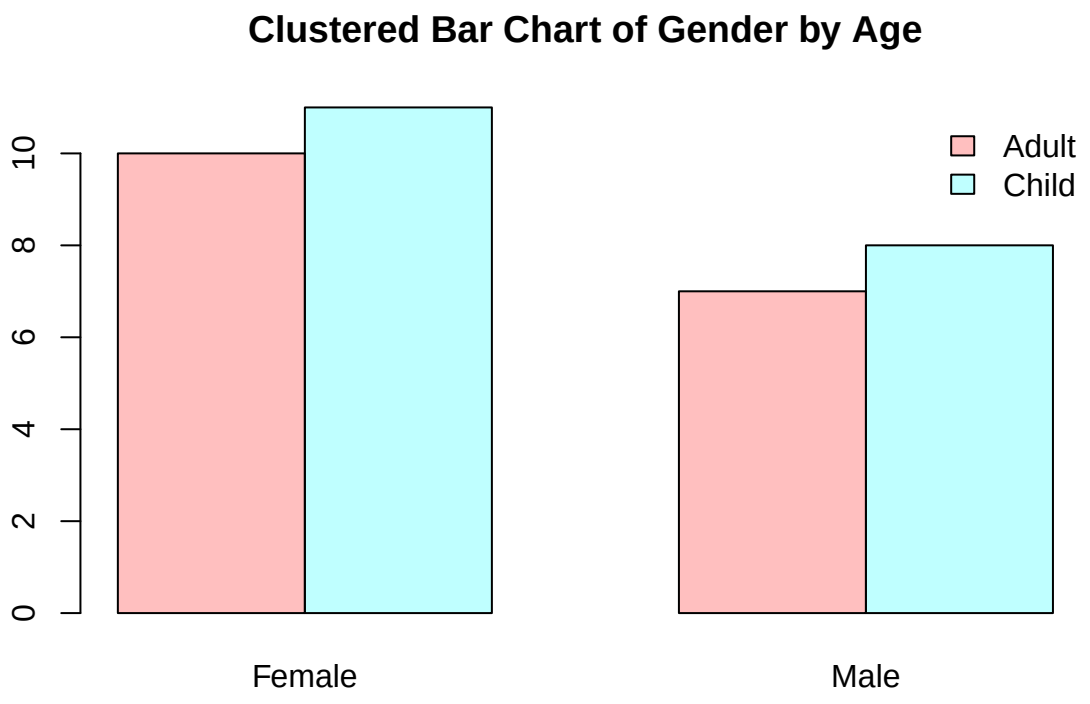
   (b) How many children came to the play?

   (c) How many males came to the play?

   (d) How many male or female children came to the play?

   (e) Complete the following two-way table with the information provided in the clustered bar chart.

	Adult	**Child**	**Total**
Female			
Male			
Total			



3 Numerical Data

Recall that a **parameter** is a numerical description of a population and a **statistic** is a numerical description of a sample. We will not consider the calculation of population parameters in this series of notes, but what follows is a description of some common sample statistics.

Below is a list of the statistical measures that we will explore along with their population and sample symbols respectively, when appropriate:

- Totals (N, n)
- Mean $(\mu, \bar{X})$
- Median
- Mode
- Range
- Interquartile Range (IQR)
- Standard Deviation (σ, s)
- Variance (σ^2, s^2)

3.1 Totals

Population total (N): The total number of observations in a population.

Sample total (n): The total number of observations in a sample.

3.2 Measures of Center

The following are typical measures of central tendency.

3.2.1 Mean

Population Mean (μ): The average value for the entire population. This is our parameter value and the formula will not be discussed here.

Sample Mean ($\bar{x}$): The average values of the sample.

Sample mean formula:

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

where: - x is a generic variable name
- x_i is the value of an observation
- i is a subscript used for indexing, (it runs from the first observation to the last one, $i = 1, 2, \ldots, n$)
- $\sum_{i=1}^{n}$ is shorthand for sum up all the values of the observations
- n is the sample size

Properties of the mean:

- It is a measure of center for a symmetric distribution
- It is a non-resistant measure. That is it is easily influenced by the values of the observations
- It cannot be less than the minimum value or more than the maximum value
- The sum of the deviations away from the mean will always add to zero: $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$. (A deviation is the distance an observations value is away from the mean)

Example:

Suppose we collect a sample of five residents of an elderly shelter and measure record their age. The ages were 82, 74, 79, 79 and 90. Compute the sample mean.

Solution:

In this case we have

x_1	x_2	x_3	x_4	x_5
83	74	79	79	90

Therefore, the sample mean $\bar{x}$ is

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{83+74+79+79+90}{5} = 81$$

3.2.2 Median

The **median** is, from lowest to highest, the value that lies exactly in the middle of the data set, splitting the data into two groups with the same number of observations. In another words, it is the 50% mark of the data. That is, half the observations live below the median and half the observation live above the median. This is often referred to as the second quartile (Q_2).

Steps to calculating the median by hand:

1. Order the observations from least to greatest.
2. Apply the locator formula $\frac{n+1}{2}$ to find the location of the middle observation. The median is found at this location.
3. Find the value corresponding to this location. This is the median.

Using the locator formula, we are only able to find the position of the median and not its value. This will require us to look at cases based on the sample size n being even or odd.

- Case 1: The sample size n is even. In this case our locator tool will take us to the middle of two observations and we must average the adjoining two.
- Case 2: The sample size n is odd. In this case our locator tool will take us to an exact observation.

Properties of Median:

- The median is a measure of center for a skewed distribution or when outliers are present
- The median is a resistant measure, (it is based on the number of observations and not their values)

Example:

Consider the above example.

1.First, we order it from lowest to highest. 74, 79, 79, 83, 90
2.Then, we apply the locator formula $\frac{5+1}{2} = 3$. The median is located in the position number 3.
3.The median is 79.

3.2.3 Mode

The **mode** is the value that is repeated the most times in the data. It is possible for a data set to have more than one mode if there are more than one observation that are repeated the same number of times. It is also possible for a data set to not have a mode, if each value is unique.

Example:

Consider the same data set as above. The mode is 79.

3.3 Measures of Dispersion

A measure of dispersion is a number used to describe the spread of the observations, usually about a measure of central tendency, in the data set.

3.3.1 Range

The **range** is the difference between the highest and the lowest values from the data set.

Formula:

$$\text{Range} = max - min$$

3.3.2 Five Number Summary

The numbers composing the **5-number summary** of a data set are as follow, for an ordered list of data:

1. Minimum Value
2. Quartile 1 (Q_1)
3. Median (Q_2)
4. Quartile 3 (Q_3)
5. Maximum Value

Quartile: A statistical value that separates the data into four equal pieces, with each part representing one quarter of the data. There are three quartiles.

- Quartile 1 (Q_1) represents the 25^{th} percentile of the data.
- Quartile 2 (Q_2) represents the 50^{th} percentile of the data (median).
- Quartile 3 (Q_3) represents the 75^{th} percentile of the data.

The procedure to calculate the quartiles is the same as for the median, in which we find the three values that break the data set into four pieces instead of two.

Steps to calculating the quartiles by hand:

1. Order the observations from least to greatest.
2. Apply the locator formulas to find $Q1$, $Q2$, and $Q3$.
    - $Q1$: Use $\frac{n+1}{4}$ to find the location of $Q1$, and look for the value that corresponding to this location.

- $Q2$: Use $\frac{n+1}{2}$ to find the location of the median and look for the value that corresponds to this location.
- $Q3$: Use $\frac{3(n+1)}{4}$ to find the location of $Q3$, and look for the value that corresponding to this location.

Note: If the position for a quartile lies between two values, we average them.

Example:

Obtain the 5-number summary for the following data.

3	3	5	7	8	11	11	12	15	16	20

Solution:

1. Minimum $= 3$
2. $Q_1 = 5$
3. $Q_2 = 11$
4. $Q_3 = 15$
5. Maximum $= 20$

Note: Some software programs may compute this summary differently than hand computation. This is due to different techniques that may be used to calculate the quartiles.

3.3.3 Interquartile Range

The **Interquartile Range (IQR)** is the difference between Q_3 and Q_1 of the data.

Interquartile range formula:

$$IQR = Q_3 - Q_1$$

Example:

For the above example, the IQR is $15 - 5 = 10$.

3.3.4 Variance and Standard Deviation

Population variance σ^2: It is the square of the average distance of all observations in the population away from the population mean.

Sample variance (s^2): It is the square of the average distance away from the sample mean.

Sample variance formula:

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

where: - x_i is an observation
- i is a subscript used for indexing
- $\bar{x}$ is the sample mean
- n is the sample total

Population Standard Deviation (σ): It is the average distance of all observations in the population away from the population mean.

Sample Standard Deviation (s): It is the average distance away from the sample mean.

Note: In order to calculate the standard deviation we must first find the sample mean.

Sample standard deviation formula:

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

where: - x_i is an observation
- i is a subscript used for indexing
- $\bar{x}$ is the sample mean
- n is the sample total

Note: Notice that we square the deviations away from the mean because the sum of the deviations away from the mean will always add to zero. We square them to get rid of the negative deviations, values that live below the mean generate a negative deviation, so that we have a sum to actually take the average. Then we take the square root to remedy for the squares added in the formula.

Properties of the Standard Deviation:

- It is the appropriate measure of spread when the mean is the appropriate measure of center
- It is a non-resistant measure
- It is always greater than or equal to 0

Steps to calculate the sample standard deviation by hand:

1. Identify each x_i
2. Identify the sample size (n)
3. Identify the sample mean ($\bar{x}$)
4. Expand the formula and plug in the values to solve

Example:

Suppose we had a sample of three hotel prices for a standard room for two people in a certain city area. The observed prices were \$50, \$55, and \$48. Compute the standard deviation.

Solution:

1.

x_1	x_2	x_3
50	55	60

2. $n = 3$
3. $\bar{x} = 55$
4. $s = \sqrt{\frac{1}{3-1}[(50-55)^2 + (55-55)^2 + (60-55)^2]} = 5$

3.3.5 Outliers

An **outlier** is an observation that lies considerably far from other observations in the data. Outliers can be low or high. We can find out which observations from a set of data are outliers by using the 1.5 IQR rule. This rule consists of finding an upper and a lower bound for the data, to classify any observation that lies outside of them as an outlier. We typically represent outliers in a type of graph called **box plots**, that will be explored shortly.

Lower Bound $= Q_1 - 1.5 \times IQR$

Upper Bound $= Q_3 + 1.5 \times IQR$

Any observation that is lower than the lower bound or higher than the upper bound will be considered an **outlier**.

Example:

Consider the following data about foxes' weight (expressed in kilos):

5.2	6.1	7.3	6.9	5.5	6.7	5.8	7.2	6.4	9.6	6.0

Solution

$IQR = 7.2 - 5.8 = 1.4$

$1.5 \times IQR = 1.5 \times 1.4 = 2.1$

Upper Bound: $Q_3 + 1.5 \times IQR = 7.2 + 2.1 = 9.3$

Lower Bound: $Q_1 - 1.5 \times IQR = 5.8 - 2.1 = 3.7$

Now, we need to look if there are any values lower than 3.7 or higher than 9.3.

Because 9.6 is greater than 9.3, we consider this observation an outlier.

3.4 Plots of Quantitative Variables

We will examine three plots for a single quantitative variable. These are:

- Stem and Leaf Plots
- Histograms
- Box Plots

Plots for numerical variables are helpful to visualize their **distributions**.

Shape of a distribution: It refers to the pattern in which the data is spread out across different values. We will classify the shape of the distribution for a quantitative variable as one of the following descriptions

1. Symmetric: The pattern is about a mirror image about the middle. Considering the number of modes of a distribution we can further classify symmetric distribution into:
    - Uniform – if the distribution has no peak
    - Unimodal – if the distribution is single peaked

- Bimodal – if the distribution has two clusters of data
- Multimodal – if the distribution has multiple clusters of data

2. Skewed: The pattern is asymmetrical, off the center, and has a direction. Depending on where most of the observations are concentrated, we can classify skewed distributions into:
   - Skewed left – if the majority of the observations occur in the larger values and less frequent observations occur in the smaller values (the tail drifts off to the left)
   - Skewed right – if the majority of the observations occur in the smaller values and less frequent observations occur in the larger values (the tail drifts off to the right)

3.4.1 Stem and Leaf Plots

Steam and leaf plots are graphs used to organize data. They are recommended only if the sample size is small.

Characteristics:

- Each number is separated into two parts, the stem and the leaf
- The stem represents the first digit or digits of the number
- The leaf represents the last digit of the number
- Stems are displayed vertically in an increasing order
- Leaves are arranged horizontally in increasing order, and are separated by a comma, for the stem they belong to
- The stems are represented in one column and the leaves in another one, separated by a " | " or in a table format

Example:

Consider the following sample of 11 students' average commuting time to school (in minutes):

3	8	12	15	17	20	22	23	31	35	44

A stem and leaf plot of the data would look like the following.

```
The decimal point is 1 digit(s) to the right of the |

0 | 38
1 | 257
2 | 023
3 | 15
4 | 4
```

We can also do a process known as splitting the stems. This process will provide more intervals by subdividing each stem into two intervals.

```
The decimal point is 1 digit(s) to the right of the |

0 | 3
0 | 8
1 | 2
```

```
1 | 57
2 | 023
2 |
3 | 1
3 | 5
4 | 4
```

3.4.2 Box Plots

Box plots are a visual representation of the 5-number summary.

Characteristics:

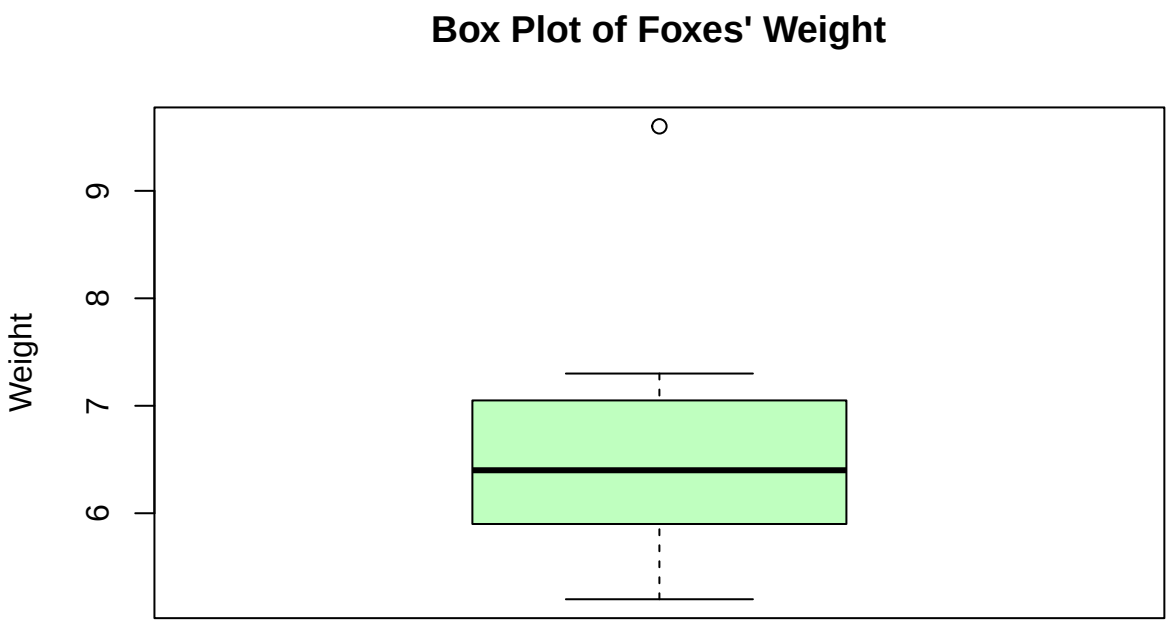
- They can be vertical or horizontal
- There is only one axis representing the data values
- The box goes from Q_1 to Q_3
- Whiskers extend from the Q_1 to the lower fence created by the $1.5IQR$ Rule and from the Q_3 to the upper fence created by the $1.5IQR$ Rule
- The line inside of the box represents the median
- Outliers are represented with dots or asterisks

Example:

Consider the same data about foxes' weight (expressed in kilos) used earlier.

5.2	6.1	7.3	6.9	5.5	6.7	5.8	7.2	6.4	9.6	6.0

The 5-number summary was 5.2, 5.9, 6.4, 7.05, 9.6 with the value of 9.6 being classified as an outlier. The box plot is:



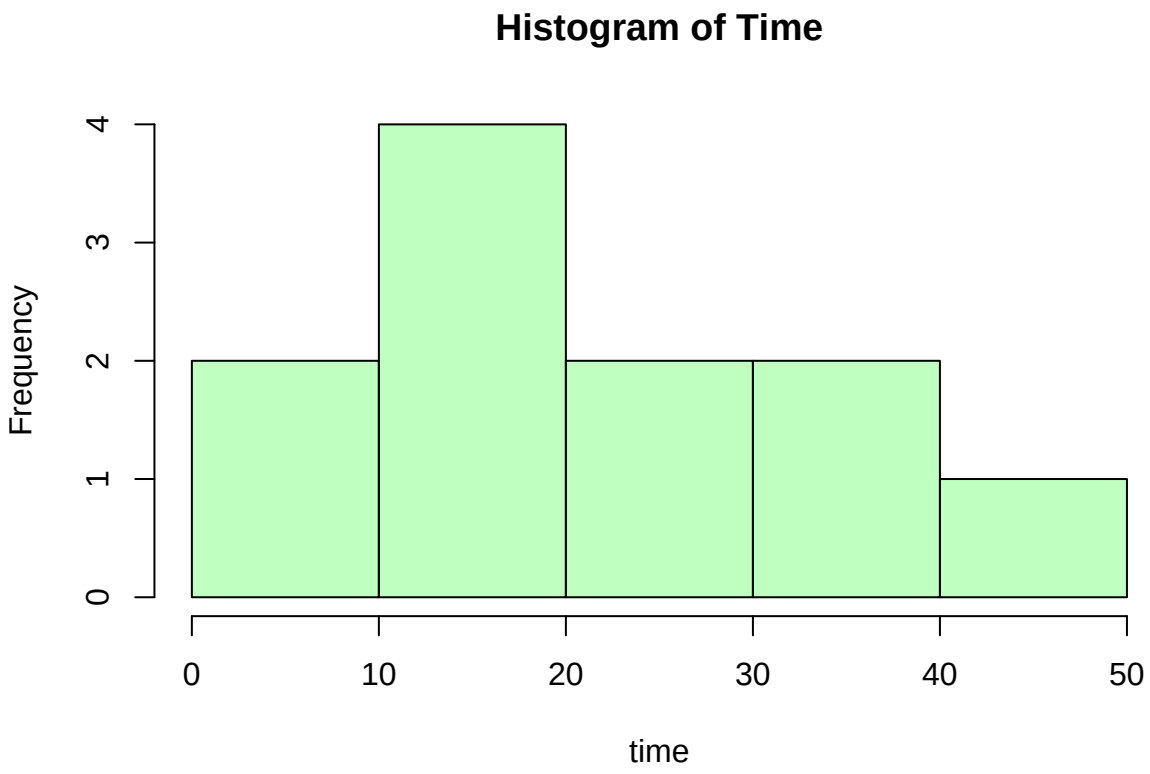
3.4.3 Histograms

A **histogram** is a type of graph in which the data is divided into intervals that are represented in the x-axis, and vertical bars or rectangles are drawn up to the number of observations belonging to that interval, represented by the y-axis.

Characteristics:

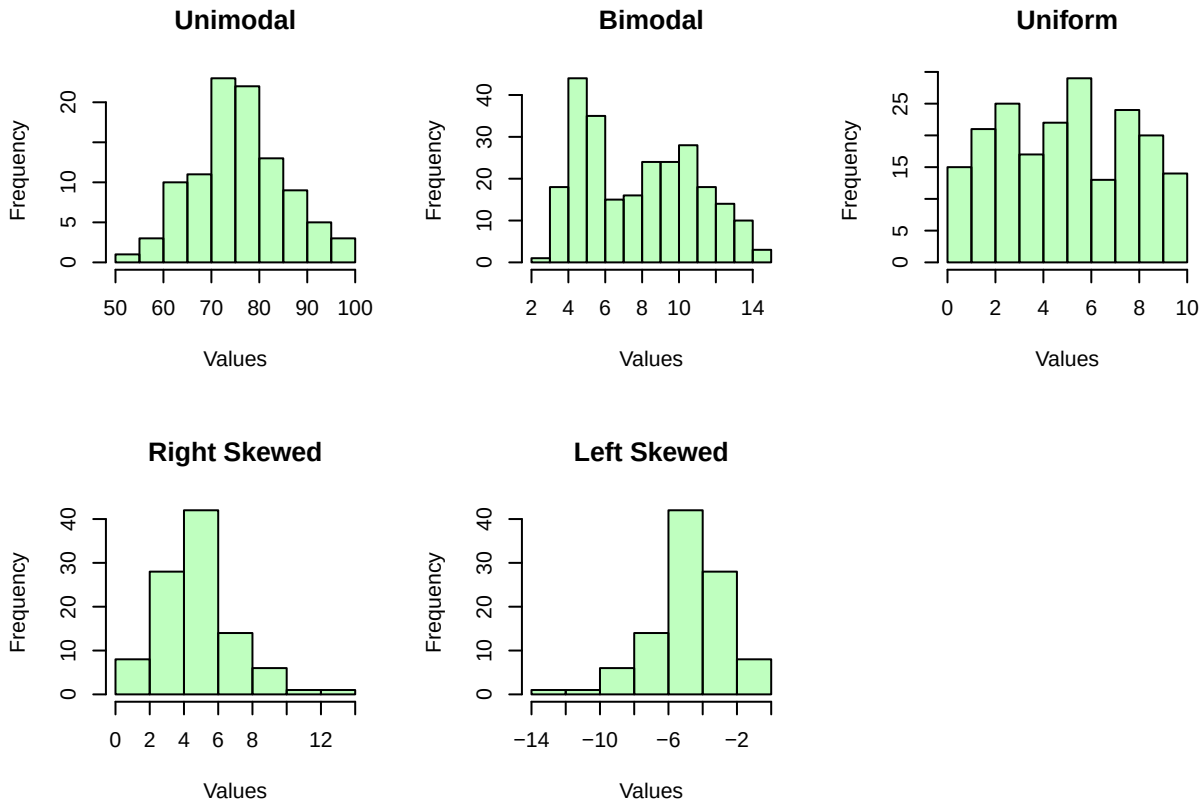
- The x-axis represents disjoint intervals of values
- The y-axis represents the frequency
- All intervals are the same width
- They are useful to see the distribution of the data
- There is no space between the vertical bars for contiguous intervals with a frequency of at least 1

Using the example introduced with our stem and leaf plot, a histogram for the same data is:



Here the disjoint (mutually exclusive) intervals are 0, 10, 20, 30, 40, 50 and the counts within each interval are 2, 4, 2, 2, 1 respectively.

Below are some other examples along with descriptions of the distributions.



3.4.4 Comparing Measures of Center to Determine Shapes

Understanding the notion of resistance and non-resistance allows us to gain a better understand of the shape of the distribution when we compare measures of center. That is,

1. Symmetric implies the mean and median are approximately equal.
2. Skewed to the right implies the mean is larger than the median.
3. Skewed to the left implies the mean is less than the median.

Numerical Data in R

In this section we will introduce the R code to describe the distribution of a quantitative variable.

Example:

A survey was conducted on 16 students asking a variety of questions. One question asked: "On a typical weeknight, how many minutes do you spend watching videos on your phone or other electronic device?" The responses are summarized in the table below.

300	300	180	100	250	245	90	190	200	150	180	180	360	240	300	350

We will examine the distribution in R using the following code.

```
# Define a Variable and Fill in the Values of the Variable
videotime<-c(300,310,180,110,250,245,90,190,260,150,280,280,360,250,300,350)
videotime
```

```
 [1] 300 310 180 110 250 245  90 190 260 150 280 280 360 250 300 350
```

```
# Mean
mean(videotime)
```

```
[1] 244.0625
```

```
# Median
sort(videotime)
```

```
 [1]  90 110 150 180 190 245 250 250 260 280 280 300 300 310 350 360
```

```
median(videotime)
```

```
[1] 255
```

```
# Range
max(videotime)-min(videotime)
```

```
[1] 270
```

```
# Five Number Summary (Approach 1)
fivenum(videotime)
```

```
[1]  90 185 255 300 360
```

```
# Five Number Summary (Approach 2)
quantile(videotime)
```

```
   0%   25%   50%   75%  100%
 90.0 187.5 255.0 300.0 360.0
```

```r
# Five Number Summary Including the Mean (Approach 3)
summary(videotime)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   90.0   187.5   255.0   244.1   300.0   360.0
```

```r
# IQR
IQR(videotime)
```

```
[1] 112.5
```

```r
# Variance
var(videotime)
```

```
[1] 6384.062
```

```r
# Standard Deviation
sd(videotime)
```

```
[1] 79.90033
```

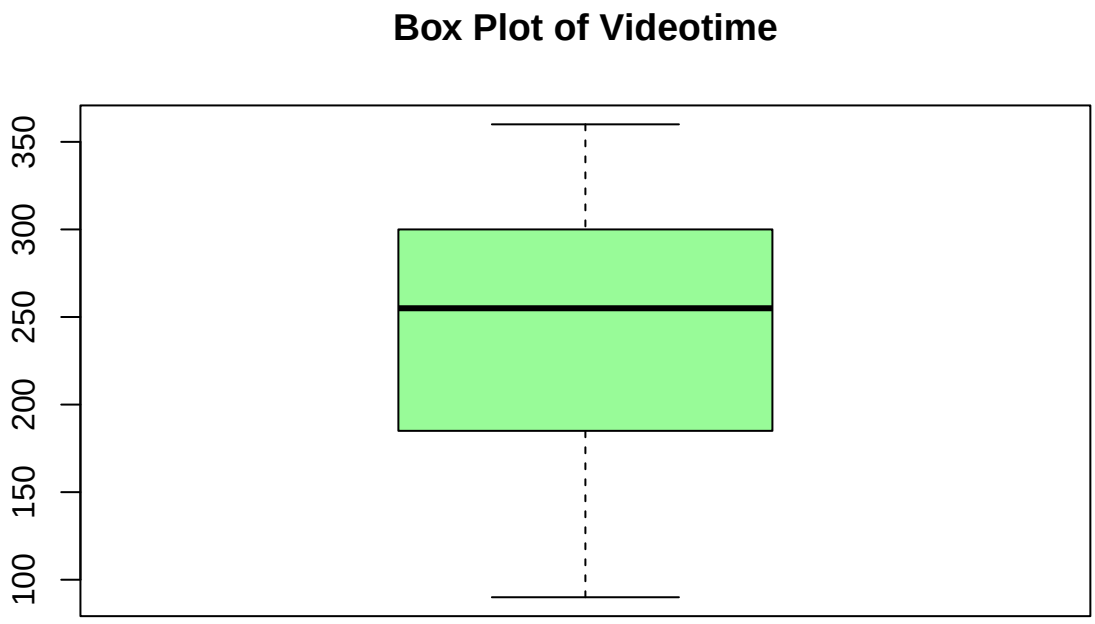
```r
# Stem and Leaf Plot
stem(videotime)
```

```
  The decimal point is 2 digit(s) to the right of the |

  0 | 9
  1 | 1589
  2 | 555688
  3 | 00156
```

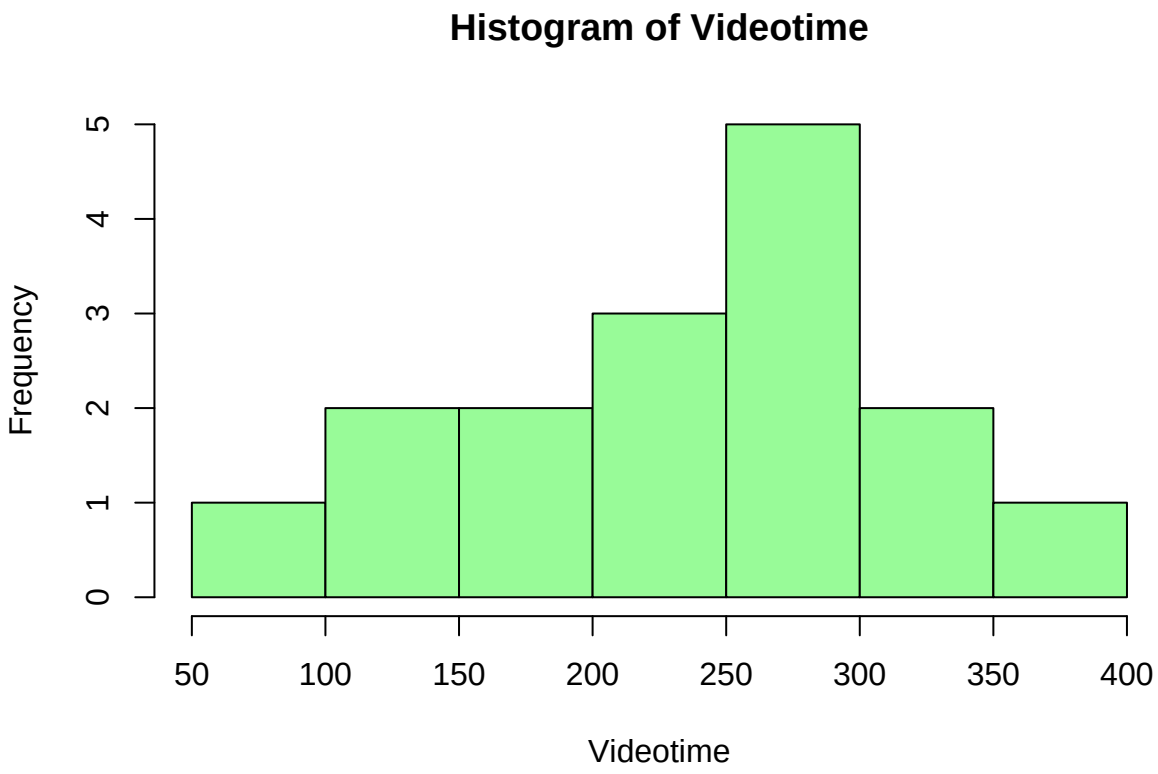
```r
# Stem and Leaf Plot Splitting the Stems
stem(videotime,2)
```

```
  The decimal point is 2 digit(s) to the right of the |

  0 | 9
  1 | 1
  1 | 589
  2 |
  2 | 555688
  3 | 001
  3 | 56
```

```r
# Box Plot
boxplot(videotime, main="Box Plot of Videotime", col="palegreen")
```




```
# Histogram
hist(videotime, main="Histogram of Videotime", xlab="Videotime", ylab="Frequency",col="palegreen")
```


Histogram of Videotime
Frequency
0
1
2
3
4
5
50
100
150
200
250
300
350
400
Videotime

Exercises

1. A small company asked their employees to state the average number of times per week that they drink coffee. The data is summarized below in the variable named coffee.

7	0	8	5	5	6	12	7	10	9	7	6

   In R, type

   coffee<-c(7,0,8,5,5,6,12,7,10,9,7,6)

   (a) Find the mean.

   (b) Calculate the standard deviation.

   (c) Calculate the five number summary.

   (d) Calculate the range.

   (e) Construct a box plot of this distribution.

   (f) Are there any outliers?

   (g) Construct a stem and leaf plot.

   (h) Construct a histogram.

   (i) How would you describe the shape, center, and spread of this distribution?

4 Variables Relationships

In this chapter we will examine the relationships between two quantitative variables. This will be accomplished through regression and the **Method of Least Squares** to find the line of best fit.

4.1 Scatter plots

Scatter Plots: They are graphical representations of two quantitative variables. The independent variable (explanatory) is represented on the horizontal axis and the dependent variable (response) is represented on the vertical axis. Each variable's position in the grid is the intersection between its x-value and its y-value. That is each observations is plotted as a point within the xy-plane.

Recall that the **explanatory variable** is the variable used to explain a measured outcome, while the **response variable** is the measured outcome. Their relationship will be identified based on the research question.

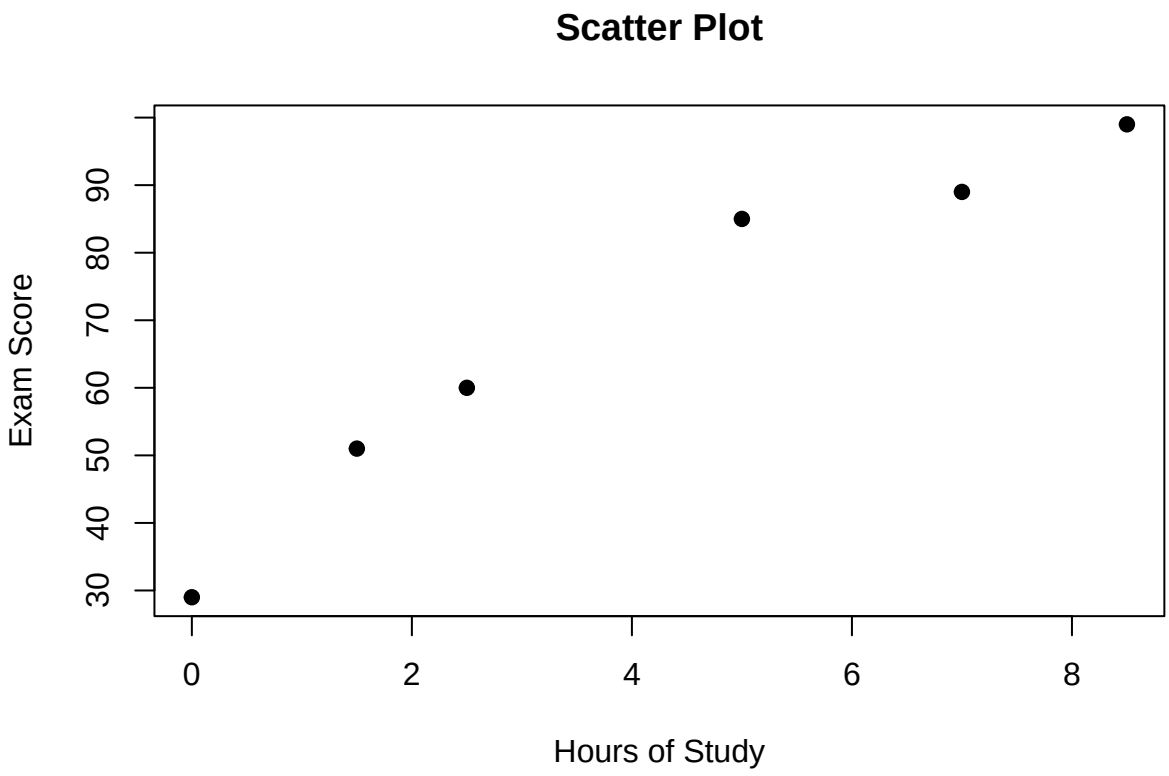
Example:

Consider that we are interested in examining the relationship between the number of hours a student studies for an exam and the actual score on the exam. We record the hours of study and the exam score (out of 100) for six students. Here the explanatory variable is the hours of study and the response variable is exam score. The observations are as follows.

Hours of Study	0	8.5	1.5	7	5	2.5
Exam Score	29	99	51	89	85	60

A scatter plot for the data and the R code is given by,

```
# Creating the Variables
Hours <- c(0,8.5,1.5,7,5,2.5)
Score <- c(29,99,51,89,85,60)

# Creating the Scatter Plot
plot(Hours,Score, xlab = "Hours of Study", main="Scatter Plot",
     ylab = "Exam Score", pch=19)
```



4.1.1 Relationships

The **relationship** between two variables is determined by the behavior of the dependent variable as the independent variable changes.

Some common variable relationships are linear, quadratic, exponential and logarithmic.

We will primarily examine and model the linear relationships, but in the end of the chapter we will demonstrate some of the other relationship types.

4.2 Linear Relationships

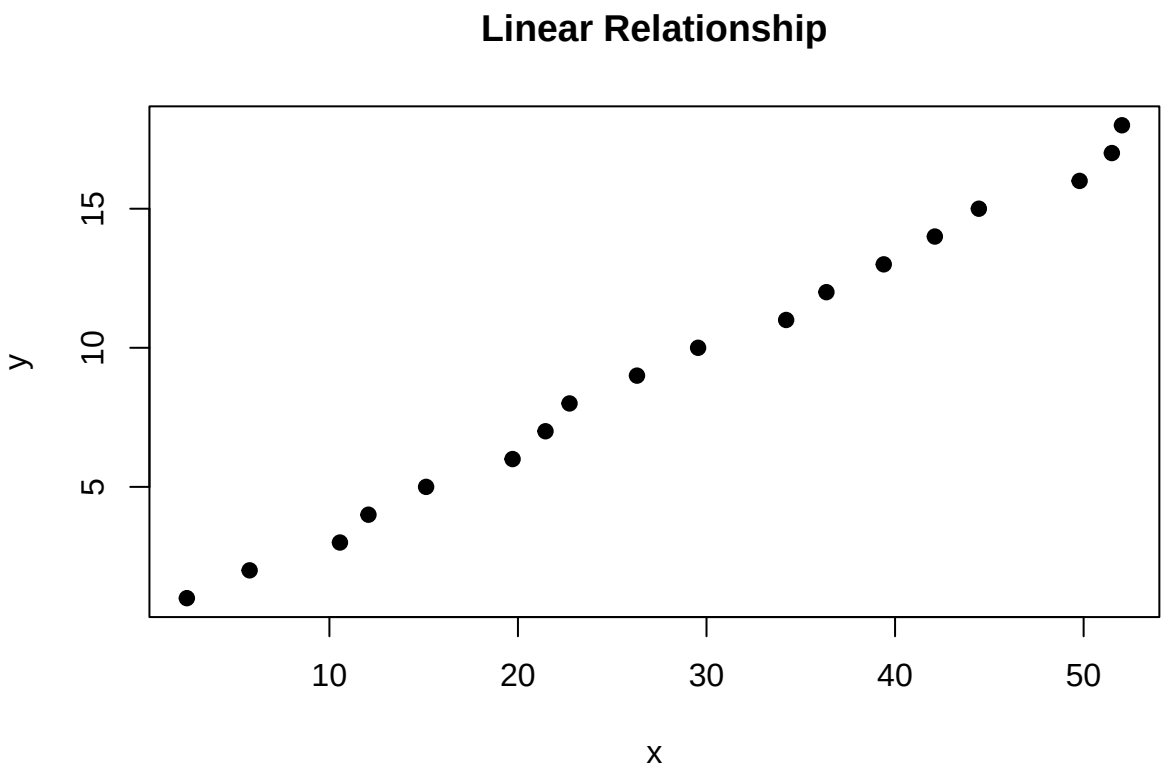
A **linear relationship** is one of which it appears that a straight line model would describe the scatter plot appropriately. This can be further explained by saying that, as the explanatory variable increases, the response variable seems to change by a fixed amount.

Characteristics:

- They can be represented using a linear function: $\hat{y} = b + ax$, where a is the slope of the relationship and b is the y-intercept.
  - Slope: (rise/run). The fixed amount by which the dependent variable increases for each unit increase in x. This can be interpreted as "For every unit increase in x, the slope is the expected change in y."
  - y-intercept: The value where we cross the y-axis. This will mainly be needed for scaling and interpretation will only be relevant when the data passes through zero for the explanatory variable.

Note: With real world data linear relationships do not have to be perfect. In reality it is very hard to find variables that are strictly linear. Normally, we encounter some noise in the data. However, if the overall pattern is linear, we can describe the relationship as such.

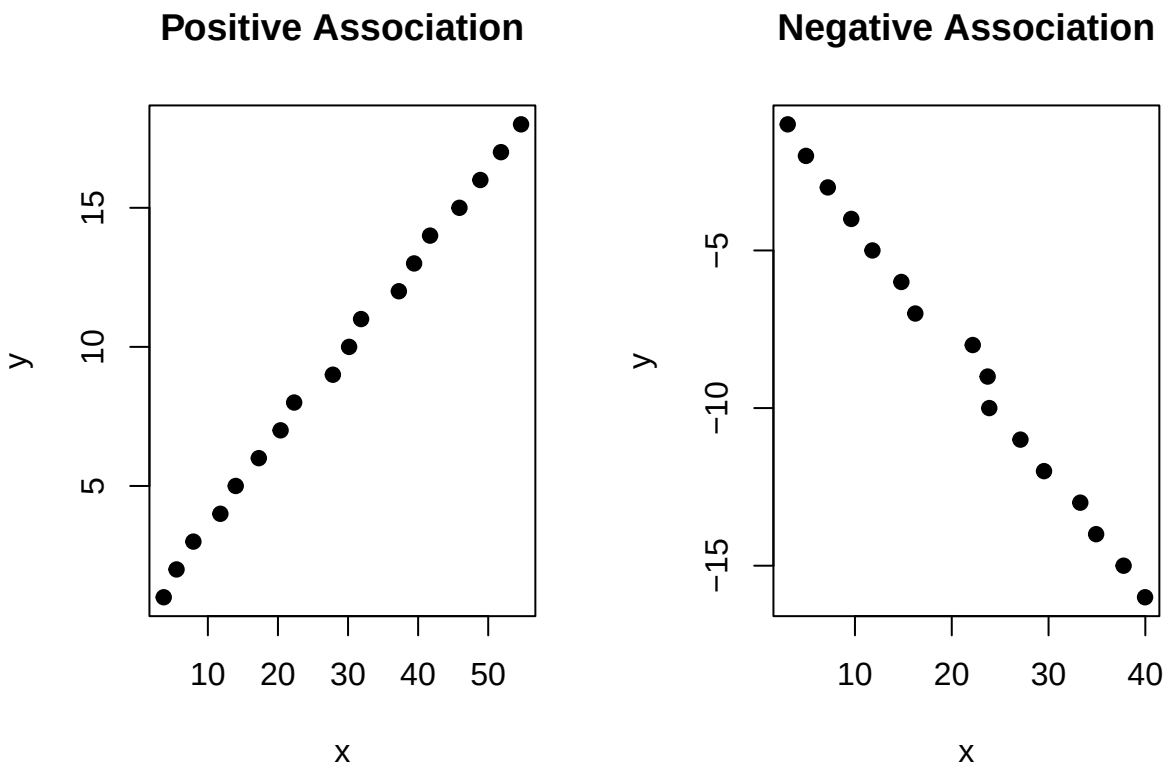
Example:



When we have a linear relationship in our scatter plot, it will make sense to discuss the direction of the relationship. The **direction** of the linear relationship will be broken into one of two types:

1. **Positive Association:** As the values of the explanatory variable increase, the values of the response Variable also increase
2. **Negative Association:** As the values of the explanatory variable increase, the values of the response variable decrease

Examples:



4.3 Correlation

Correlation: A statistical measure that describes the strength of the linear association between two variables.

Correlation coefficient (r): It is our measure of linear strength. That is, how well our data will fit to a straight line pattern. It is the average of the product of standardized values.

Correlation coefficient formula:

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

where: - x_i is an observation from the independent variable
- y_i is an observation from the dependent variable
- $\bar{X}$) is the sample mean for the independent variable
- $\bar{y}$) is the sample mean for the dependent variable
- s_x is the sample standard deviation for the independent variable
- s_y is the sample standard deviation for the dependent variable
- n is the sample total

Example:

Consider the example above with study hours and score on exam. The correlation coefficient is 0.9732958, calculated using,

```r
cor(Hours,Score)
```

```
[1] 0.9732958
```

Note: We will not concentrate in the by-hand calculation of the correlation coefficient.

Properties of the Correlation Coefficient:

- It only measures linear strength
- It ranges between -1 and 1
- It is a unitless number
- The sign specifies the direction, (this will also be the sign of the slope for the method of least squares)
- It can only be found between two quantitative variables
- A change in units does not influence its value
- It does not discriminate between dependent and independent variable
- It is a non-resistant measure (if observations change, r changes)

Types of correlation:

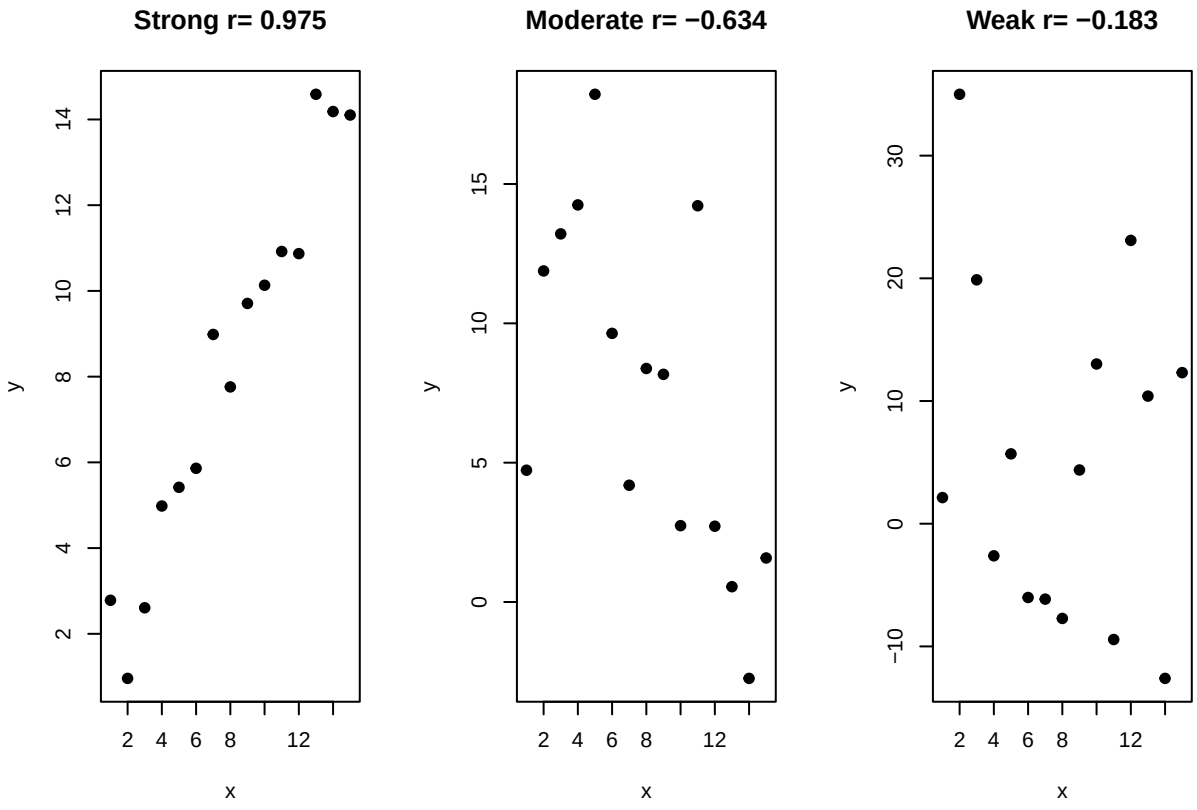
Based on its Direction: Positive or negative. Recall this is determined by the sign of the correlation coefficient.

Based on its Strength: Very strong, strong, moderate, weak, very weak. This is determined by the absolute value of the correlation coefficient. (Recall that the closer to |1|, the stronger the correlation is).

The table below summarizes the description of correlation based on both direction and strength.

Strong, Neg.	Moderate, Neg.	Weak, Neg.	None	Weak, Pos.	Moderate, Pos.	Strong, Pos.
-1 to -0.80	-0.79 to -0.50	-0.49 to -0.10	-0.09 to 0.09	0.10 to 0.49	0.5 to 0.79	0.80 to 1

Examples including scatter plots:



Interpretation:

The interpretation is based on the correlation types and the variables involved. For example, suppose the hours of study and the exam score gave us a correlation coefficient of 0.97. Then, our interpretation would be something like: "The hours of study and the exam score have a very strong positive correlation".

Note: Correlation does not imply causation.

The fact that two variables may be correlated does not imply that one is causing the other to behave in a certain manner. For example, we the number of firefighters sent to a fire scene is positively correlated with property damage. However, sending more firefighter does not cause the property damage. Instead, there is a **lurking variable**: the size of the fire. As the size of the fire and the resulting need for additional resources might lead to both increased property damage and a higher number of firefighters being deployed.

Lurking Variable: A variable that is not considered in the study but might have an impact on the variables that were included in the study.

4.4 Linear Models

Model: A **statistical model** is a mathematical function that describes the relationship between a dependent variable and one or more independent variables.

Linear Model: A **simple linear model** is a mathematical linear function that describes the linear relationship between two variables. (Recall a linear function has the form $\hat{y} = b + ax$).

4.4.1 Method of Least Squares

The Method of Least Squares is a method to obtain the **least squares regression equation** $\hat{y} = b + ax$ that will minimize the sum of the squared residuals. Using this method, we find $a = r\frac{s_y}{s_x}$ and $b = \bar{y} - a\bar{x}$.

Residual: The vertical distance from the observed response y to the predicted response $\hat{y}$. We calculate the residual by $y - \bar{y}$.

Fitted values: It is a synonym for the predicted response. The fitted values are the $\hat{y}$ values obtained by plugging the x values into the regression equation.

Example:

Consider the example above dealing with study hours and exam score. We will exemplify this providing the R code, as R utilizes the method of least squares to automatically come up with the line of best fit.

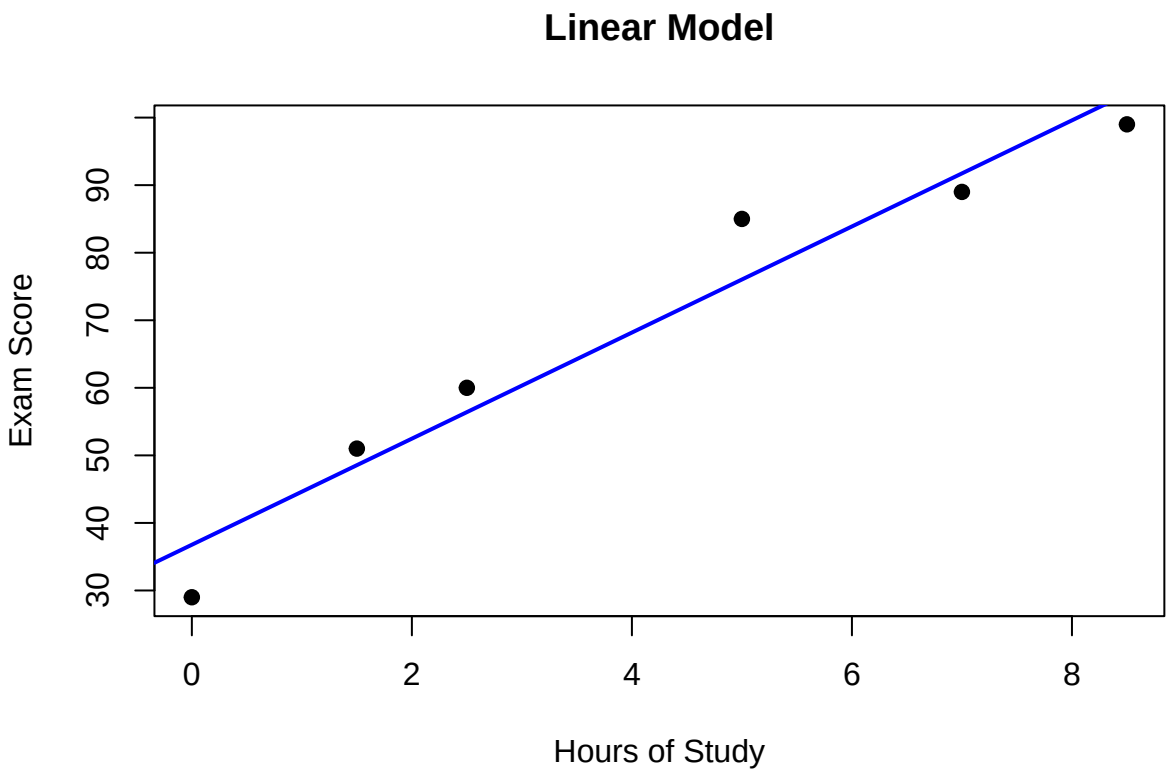
Using R to find the linear equation,

```
# Creating the Variables
Hours <- c(0,8.5,1.5,7,5,2.5)
Score <- c(29,99,51,89,85,60)

# Fitting the Linear Model
fit<-lm(Score~Hours)
fit
```

```
Call:
lm(formula = Score ~ Hours)

Coefficients:
(Intercept)        Hours
     36.770        7.852
```

Looking at the output, we find the lest squares regression equation to be $\hat{y} = 36.77 + 7.85(x)$.

```
# Creating a Scatter Plot
plot(Hours,Score, xlab = "Hours of Study", main="Linear Model",
     ylab = "Exam Score", pch=19)

# Adding the Line of Best Fit
abline(fit,col="blue", lwd=2)
```



To find the fitted values, the residuals and visualize the distance between them,

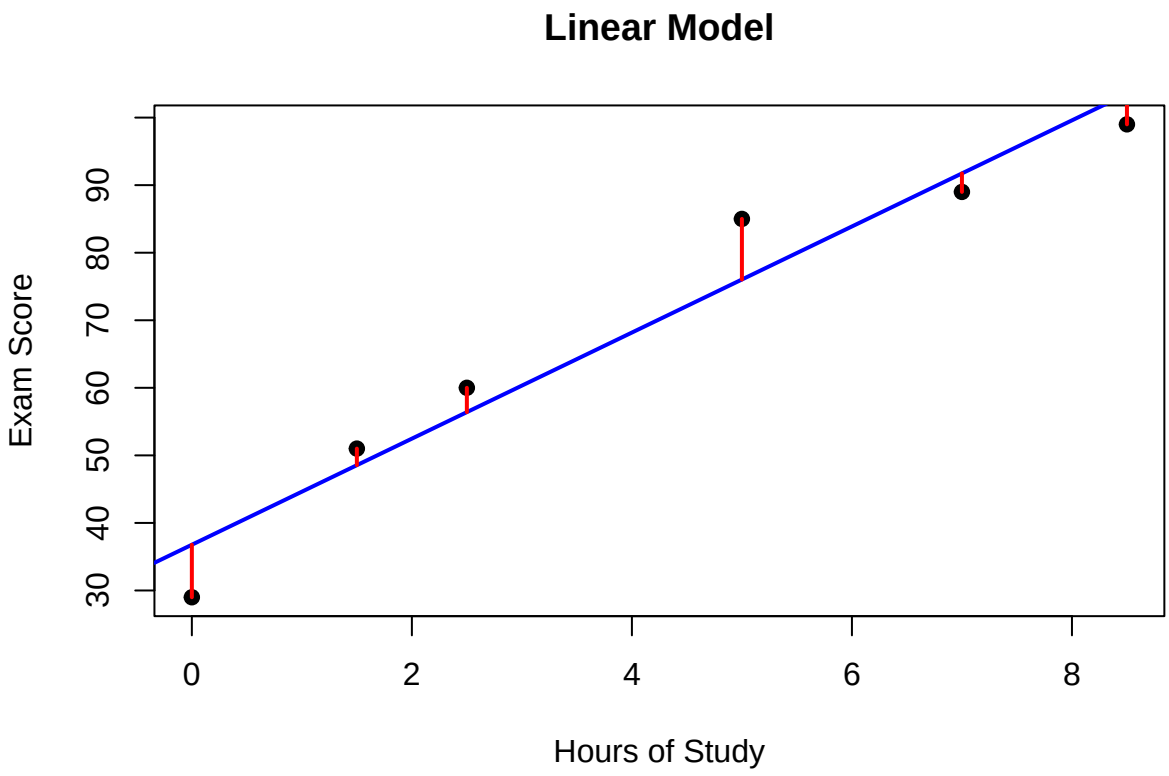
```
# Finding the Fitted Values
fit$fitted.values
```

```
       1        2        3        4        5        6
36.76999 103.51409  48.54836  91.73572  76.03123  56.40061
```

```
# Finding the Residuals
fit$residuals
```

```
       1        2        3        4        5        6
-7.769992 -4.514090  2.451637 -2.735720  8.968774  3.599391
```

```
# Plotting the Line of Best Fit and the Residuals
plot(Hours,Score, xlab = "Hours of Study",main="Linear Model",
     ylab = "Exam Score", pch=19)
abline(fit,col="blue", lwd=2)

# Adding Segments from the Residual Values to the Line of Best Fit
segments(Hours[1],Score[1],Hours[1],fit$fitted.values[1],col="red",lwd=2)
segments(Hours[2],Score[2],Hours[2],fit$fitted.values[2],col="red",lwd=2)
segments(Hours[3],Score[3],Hours[3],fit$fitted.values[3],col="red",lwd=2)
segments(Hours[4],Score[4],Hours[4],fit$fitted.values[4],col="red",lwd=2)
segments(Hours[5],Score[5],Hours[5],fit$fitted.values[5],col="red",lwd=2)
segments(Hours[6],Score[6],Hours[6],fit$fitted.values[6],col="red",lwd=2)
```



As displayed in the plot, a residual is essentially a measure of how much we missed our mark, meaning how far off our predicted response is from the actual is observed response.

Additionally, we could also solve this problem in R by using the formulas for the slope (a) and the y-intercept (b) using the following code.

```
# Computing the Mean and Standard Deviation of Each Variable
xbar<-mean(Hours)
sx<-sd(Hours)
ybar<-mean(Score)
sy<-sd(Score)

#Finding the Correlation Coefficient
r<-cor(Score,Hours)

# Finding the Slope
a=r*sy/sx
a
```

```
[1] 7.852247
```

```
#Finding the y-intercept
b<-ybar-a*xbar
b
```

```
[1] 36.76999
```

```r
# Comparing the Results to the Previous Approach
fit$coefficients
```

```
(Intercept)       Hours
  36.769992    7.852247
```

The obtained values are the same as the ones given by the least squares method.

Again, the least squares regression equation for this example is $\hat{y} = 36.77 + 7.85(x)$ or $\hat{score} = 36.77 + 7.85 \times (hours)$.

4.4.2 Model Interpretations

Interpretation of the slope:

The slope represents the rate of change for the response variable as the independent variable increases. We have a general interpretation of the slope. That is: **For every unit increase in x, the slope is the expected change in y.**. We will merely tailor this interpretation to our example at hand.

Example:

For every hour increase in the number of hours a student studies, we expect that the score on the exam will increase by 7.85 points.

4.4.3 Making Predictions Based on the Model

We can use our constructed models to make predictions about new observations. The way to do this is by plugging in the value of the new observation in the constructed function.

Example:

For a student that studies for 5 hours, what is the expected score on the exam?

Solution:

$\hat{Score} = 36.77 + 7.85 \times (hours)$
$\hat{Score} = 36.77 + 7.85 \times 5$
$\hat{Score} = 76.02$

Answer: The expected exam score for a student that studies for 5 hours is 76.02.

4.5 Do We Have a Good Fit?

Good fit: It refers to the model being a good representation of the behavior of the data.

We will explore two options for measuring the goodness of fit, the **coefficient of determination** and **residual plots**.

4.5.1 The Coefficient of Determination

The coefficient of determination (R^2) is a value used to determine if we have a good fit. It is the square of the correlation coefficient and it represents the fraction of variation in the response variable that is explained by the least squares regression of the response on the independent variable. Values above 0.50 imply that we have explained over half of the variation using our explanatory variable, which is considered good. The remaining variation that is not explained by the least squares regression of the response on the predictors is due to the residual values.

In R,

```
# Finding the Coefficient of Determination
summary(fit)$r.squared
```

```
[1] 0.9473048
```

```
# or
cor(Hours,Score)^2
```

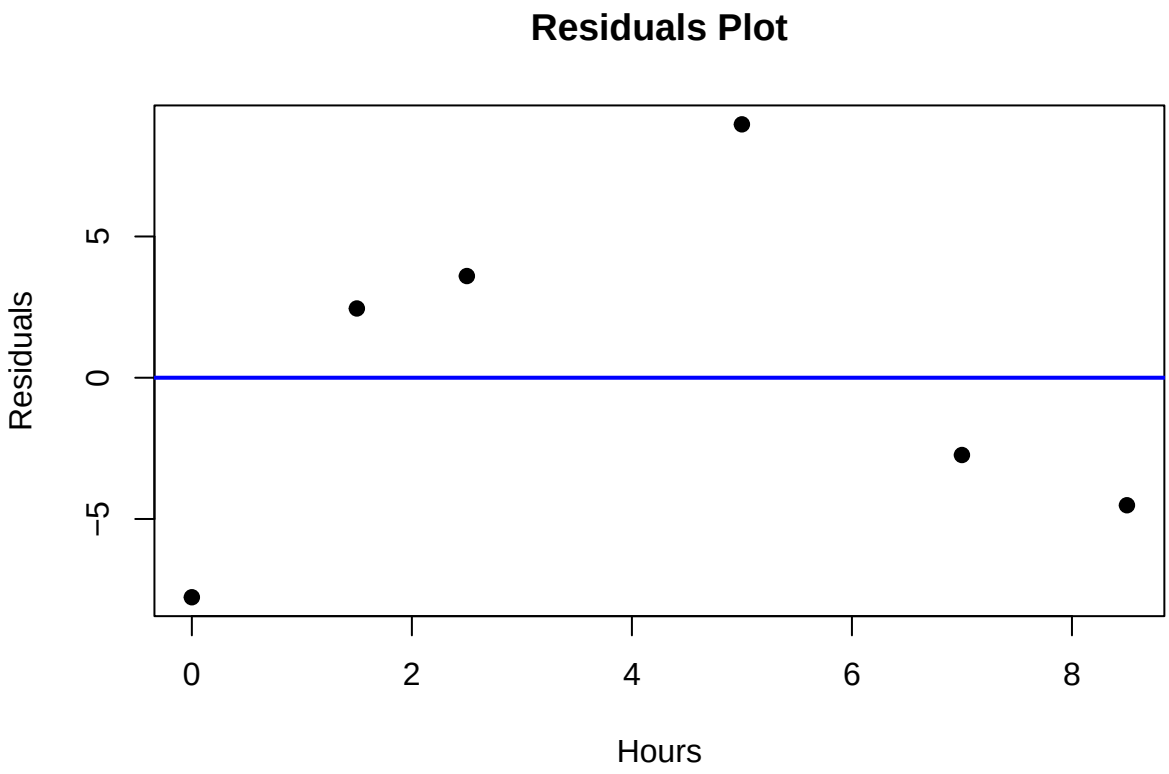
```
[1] 0.9473048
```

4.5.2 Residual Plots

A **residual plot** is a scatter plot of the independent variable versus the residuals. It is helpful to examine the how far away our observed responses (represented by the horizontal line at $x = 0$) are from our predicted responses. Non-major fluctuations around 0 will indicate a good fit. A specific distance around 0 will be dependent on the scale of the response variable.

In R,

```
# Creating the Residuals Plot
plot(Hours,fit$residuals,pch=19,main="Residuals Plot", ylab="Residuals")

# Adding a Horizontal Line at y=0
abline(h=0,col="blue",lwd=2)
```



4.6 Model Considerations

Extrapolation: It refers to using your regression equation to make prediction far outside the range or scope of your explanatory variable. This prediction may be untrustworthy since we have no data to describe the behavior far outside our explanatory variable's scope.

For example, trying to predict the exam score for an individual who studied for 24 hours using our model, because this observation is way bigger than the data that we used to construct the model.

Lurking variables: Recall that a lurking variable is a variable that is not considered in the study but might be affecting the response.

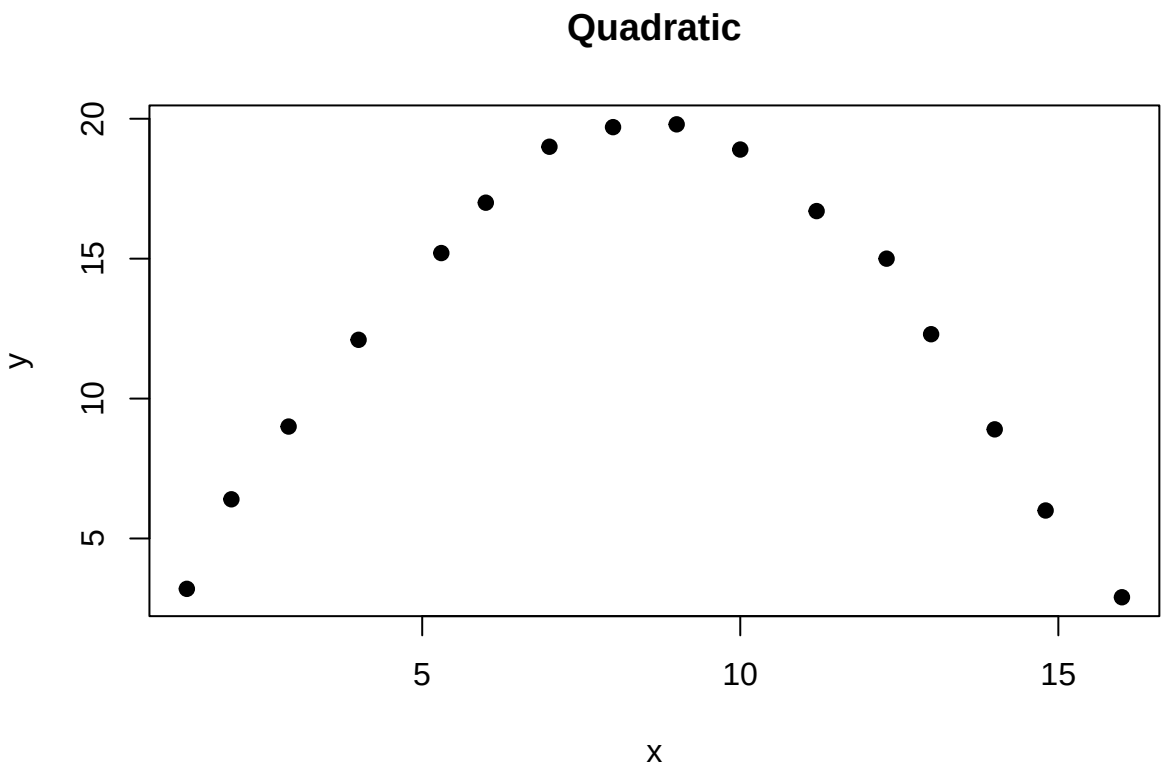
For example, exam scores might also be related to student's IQ. Individuals with higher IQ scores may have a natural aptitude for learning, grasping concepts more quickly and easily. As a result, they may require less time to study effectively and still achieve higher exam scores.

4.7 Other Types of Relationships

4.7.1 Quadratic Relationships

Quadratic relationships are those in which one variable is related to the square of another variable in a way that they present a parabolic shape (U-shape or inverted U-shape).

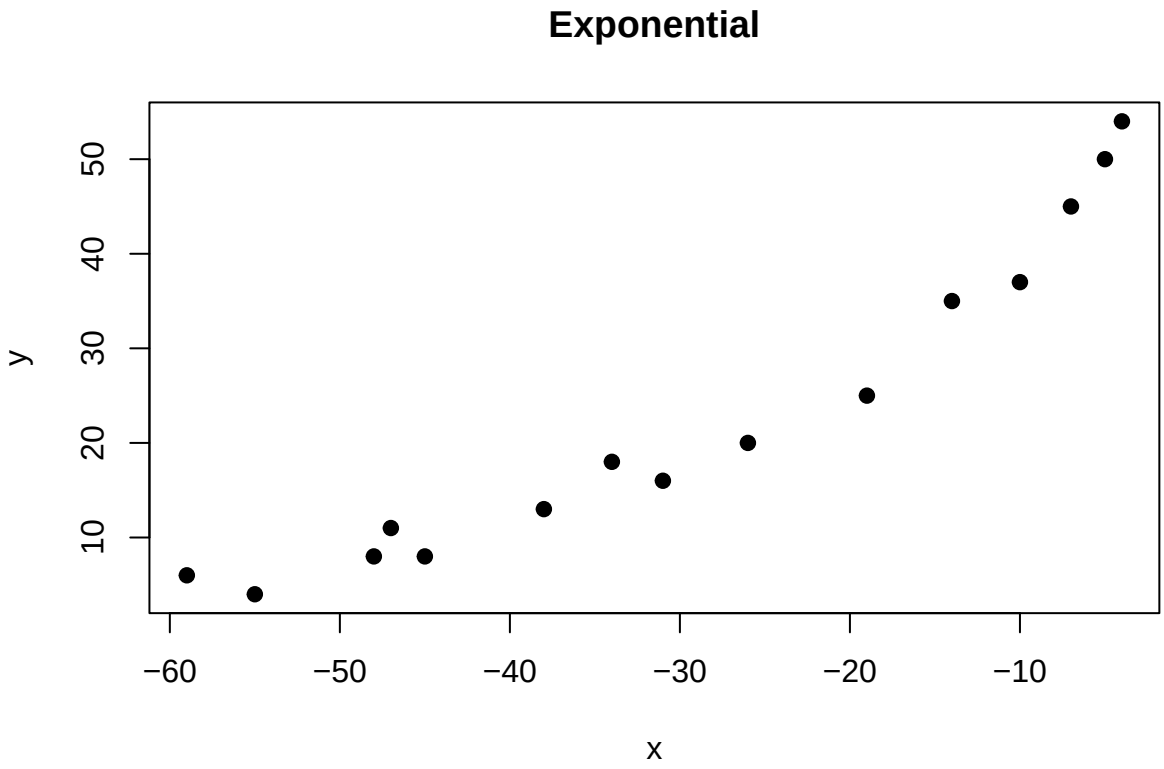
Example:



4.7.2 Exponential Relationships

Exponential relationships are those in which one variable is related to another variable raised to a power. This means that as the dependent variable increases, the dependent variable increases exponentially, presenting a very rapid curved increase.

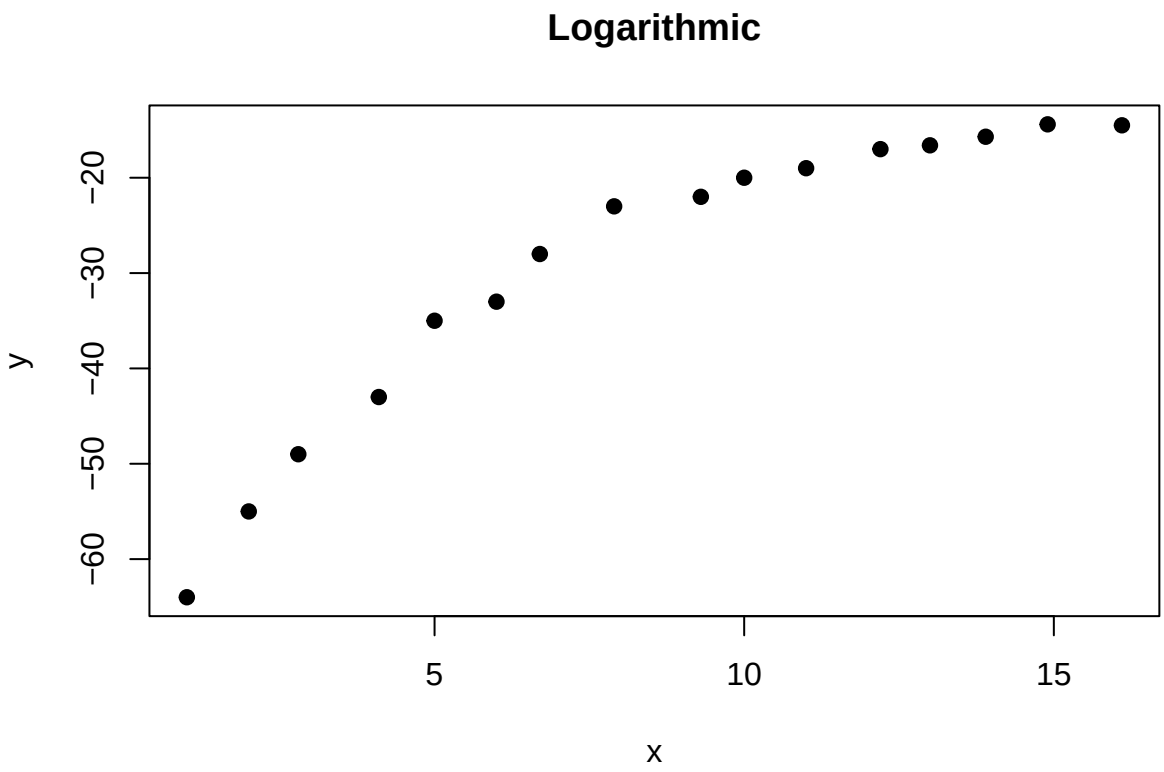
Example:



4.7.3 Logarithmic Relationships

Logarithmic relationships are those in which one variable is related to the logarithm of another variable. The general shape is a curve that grows more slowly as x increases and eventually approaches a horizontal asymptote.

Example:


Logarithmic
y
−20
−30
−40
−50
−60
5
10
15
x

Exercises

For each of the following questions, consider the following data on distances between two places (in miles) and travel time (in minutes) by car.

Distance(in Miles)	**Travel Time(in Minutes)**
3.2	10.2
12.4	18.0
3.6	12.3
7.2	16.1
13.0	19.4
9.4	17.0
5.1	12.0

1. Identify the explanatory and the response variable.
2. Create a scatter plot of the variables.
3. Find the correlation coefficient.
4. Describe the correlation (positive/negative?,strong/moderate/weak?)
5. Find $\bar{x}$, $\bar{y}$, s_x, and s_y.
6. Find the least squares regression equation.
7. Interpret the value of the slope.
8. Predict the travel time for two places that are 5.1 miles apart. What is the residual value?
9. Would it make sense to predict the travel time for two places that are 160 miles apart? Why?
10. Can you think of any lurking variable that may be present in this study?

5 Probability

5.1 Terminology

Probability: The likelihood of a random event occurring in a long series of trials or experiments.

Event: Getting a specific outcome. A "success" is obtained if we end up getting the outcome that we wanted. Events within a sample space are typically denoted as **A,B,C,**....

Sample Space (S): A set of all possible outcomes (events) that can occur within a specific trial or experiment. Sample spaces can be discrete or continuous.

- **Discrete Sample Space:** A sample space comprised of a finite number of outcomes. For example, the sample space for rolling a six-sided die: $S = \{1, 2, 3, 4, 5, 6\}$. If we have a discrete sample space and every outcome within the sample space is equally likely, then we can find the probability of a particular event using the following formula.

$$P(A) = \frac{\#\ possible\ success\ of\ A}{overall\ total\ possibilites}.$$

- **Continuous Sample Space:** A sample space comprised of outcomes that take values within intervals or ranges. For example, the heights of all young adults, the pour volume of a certain machine, the weights of rocks at a quarry.

Probability Model: A mathematical representations or frameworks used to describe uncertain events and their associated probabilities.

Discrete Probability Model: A mathematical representation used to describe and analyze situations where the sample space is discrete.

Continuous Probability Model: A mathematical representation used to describe and analyze situations where the sample space is continuous.

5.2 Notation

We refer to the probability of event A as $P(A)$.

For example, is the event was tossing heads when tossing a coin, a correct notation could be $P(heads)$.

5.3 Probability Rules

5.3.1 Basic Rules

Some basic axioms (accepted truths) of probability are:

- The assignment of a probability to an event A is always between 0 and 1 ($0 \leq P(A) \leq 1$)
- The probability of the entire sample space must be 1 ($P(S) = 1$)

Example:

Consider a fair four-sided die. So we have the discrete probability model:

Roll	1	2	3	4
$P(x)$	0.25	0.25	0.25	0.25

The sample space S is $S = \{1, 2, 3, 4\}$, because those are all the possible outcomes.

According to this model the probabilities of each event in the sample space are:

- $P(1) = 0.25$
- $P(2) = 0.25$.
- $P(3) = 0.25$.
- $P(4) = 0.25$.

Additionally, the probability of the entire sample space can be represented as:

$$P(S) = P(1) + P(2) + P(3) + P(4) = 0.25 + 0.25 + 0.25 + 0.25 = 1$$

5.3.2 Probability Rules for Multiple Events

5.3.2.1 Types of Events

Disjoint events: Two events that can never occur at the same time. For example, tossing one coin once and getting heads and tails at the same time, or rolling a die and getting a two and a six at the same time.

Complementary events: The complement of an event is the set of all other outcomes in the sample space that are not the event. For example, the complement of getting heads when tossing a coin is getting tails, or the complement of getting a one when rolling a die is getting either a two, a three, a four, a five or a six.

Independent events: Two events that do not affect the occurrence of the other. "If event A occurs, it does not affect the occurrence of event B." For example, tossing a coin and getting heads and rolling a die and getting a one. If we toss coin and get heads, it does not affect our chances of rolling a die and getting a one. Another example is tossing a coin twice and getting heads first and tails after. If we toss our coin and we get heads, this does not affect the chances that we toss it again and we get tails.

Dependent events: The occurrence of one event affects the probability of the occurrence of another. For example, suppose a bag has 3 red balls and 6 green balls. Two balls are drawn from the bag one after the other. Suppose we want to get two red balls in a row, given that once a ball is drawn from the bag, it is not reinserted. If we draw the first ball and we get a red one, this affect our chances of getting red again because the number of red balls in the bag has changed (now there is only two red balls left instead of 3).

5.3.2.2 Situations and Rules

To calculate the probability of multiple events occurring we follow certain probability rules. The different situations and the rules applied in each of them are summarized in the table below.

Situation	Rule
Union of Events	Addition Rule
Complement of an event	Complement Rule
Intersection of events	Multiplication Rule
Conditional probabilities	Conditional rule

5.3.2.3 Probability Rules for Multiple Events

For the purpose of this explanation we will consider our four-sided fair die and suppose we roll the die twice and compute the sum of the outcomes.

We would have a total of 16 outcomes for the two individual rolls, which visually are:

First Roll	1	1	1	1	2	2	2	2	3	3	3	3	4	4	4	4
Second Roll	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
Sum	2	3	4	5	3	4	5	6	4	5	6	7	5	6	7	8

Now, consider the probability experiment of interest to be the computation of the sum of two individual rolls of this fair four-sided die.

Then, the sample space for these sums is: $S = \{2, 3, 4, 5, 6, 7, 8\}$ and it follows the following discrete probability model:

Sum	2	3	4	5	6	7	8
P(x)	1/16	2/16	3/16	4/16	3/16	2/16	1/16

Note: The following rules can be applied to any two events **A** and **B**, and some can also be extrapolated to more than just two events.

5.3.2.3.1 Complement Rule

The complement rule states that the probability of the complement B of event A is 1 minus the probability of A. We use it to calculate the complement of an event. This is set off by the language "not" or "but".

Complement rule formula:

$$\text{P (not A)} = 1 - \text{P(A)}$$

Example:

What is the probability of getting anything **but** a sum of 5 for the above example?

Solution:

We use the complement rule. To apply it, we subtract the probability of getting a sum of 5 from 1.

$P(\text{not } 5) = 1 - \frac{4}{16} = \frac{12}{16} = \frac{3}{4}$

5.3.2.3.2 Multiplication Rule for Independent Events

We use it to calculate the probability of the intersection of two independent events, meaning that one outcome has no influence on the next one.

Multiplication rule for independent events formula:

$$\text{P(A and B)} = \text{P(A)} \times \text{P(B)}$$

This represents what is referred to as an **intersection**, and the intersection of events is set off by the language of "and".

Example:

What is the probability of rolling a 1 on the first roll *and* rolling a 1 on the second roll in the considered example?

Solution:

We use the multiplication rule for independent events. We multiply the probability of the first event times the probability of the second one.

$\text{P}(1^{1st} \text{ and } 1^{2nd}) = \text{P}(1^{1st}) \times \text{P}(1^{2nd})$

$\text{P}(1^{1st} \text{ and } 1^{2nd}) = \frac{1}{4} \times \frac{1}{4} = \frac{1}{16}$

This holds true for each of our 16 possible pairs. Each pair is equiprobable thanks to independence.

5.3.2.3.3 Addition Rule

This rule represents what is referred to as a **union**, and the union of events is set off by the language of "or". This thought of as either event **A or B**, or **both A and B.**

The addition rule for probability can be divided into

- **Disjoint Addition Rule**
- **General Addition Rule**

Addition rule for disjoint events: In the case that the events are disjoint of one another (have no outcomes in common), then the P(A and B) = 0. That is, the probability that both events occur at the same time cannot occur.

Addition rule for disjoint events formula:

$$P(A \text{ or } B) = P(A) + P(B)$$

Example:

What is the probability of obtaining a sum of 3.

Solution:

There are two possible ways in which this can happen: a 1 in the first roll and a 2 in the second roll or a 2 in the first roll and a 1 in the second roll. Then we will define event A to be rolling a 1 and then a 2 and event B to be rolling a 2 and then a 1.

P(A or B)=P(A)+P(B)=$\frac{1}{16} + \frac{1}{16} = \frac{1}{8}$

Addition rule for non-disjoint events: In the case that the events are not disjoint of one another, there is an intersection.

Addition rule for non-disjoint events formula:

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

We subtract the probability of the intersection of two events to avoid the double count of their outcomes in common.

Example:

Define event A as obtaining a sum of 2 or 3 and event B as obtaining a sum of 3 or 4. What is the probability of obtaining a event A or B?

Solution:

A={2,3}

B={3,4}

Because 3 is an outcome in common, the intersection of A and B is A and B={3}

To avoid double counting the probability of getting a sum of 3, we add the probabilities of both events and subtract the probability of the event in common.

P(A or B)= P(A)+P(B)-P(A and B)= $\frac{3}{16} + \frac{5}{16} - \frac{2}{16} = \frac{6}{16}$

5.3.2.3.4 Multiplication Rule for Dependent Events

We use this rule to calculate the probability of the intersection of two dependent events, meaning that the occurrence of the first event affects the probability of the occurrence of the second event.

Multiplication rule for dependent events formula:

$$P(\text{A and B}) = P(\text{A})P(\text{B}|\text{A})$$

Note: The last term in the formula above, P(B|A), is read as, the probability of the event B given that the event A has occurred. This will set up what is known as a **conditional probability** and will be discussed next.

Example:

Suppose we need to select two students from a class of 10 students with 4 girls and 6 boys. What is the probability of randomly selecting two boys?

Solution:

We use the multiplication rule for dependent events. We multiply the probability of selecting a boy (6/10) by the probability of selecting another boy, given that one was already selected (5/9). Because we already selected a boy we only have 5 boys left for our second selection out of 9 remaining students.

P(boy) P(boy | 1 boy was already selected) $= \frac{6}{10} \times \frac{5}{9}$

Note: In these cases, we need to be careful on the probability of the second event because the first event is affecting its probability of occurrence.

5.3.2.3.5 Conditional Probability

Here we will calculate the probability of an event B, given that another event A has previously occurred. That is, how does knowing that one event has occurred influence the probability of another.

Conditional probability formula:

$$P(\text{B}|\text{A}) = \frac{P(A \text{ and } B)}{P(A)} \text{ with } P(A) > 0$$

We will learn more about conditional probabilities in the next section, **Probabilities with Two-Way Tables**.

5.3.3 Probability With Two-Way Tables

Recall the two-way table discussion from the categorical data chapter.

To learn probabilities with two-way tables we will consider the following situation.

Suppose we recorded information about fast food orders of food and drink for the last 25 customers that came to a restaurant. The available food options were a chicken sandwich, a double burger or chicken nuggets. The available drinks were coke, juice or water. The recorded information is expressed below.

	Chicken Sandwich	**Double Burger**	**Nuggets**	**Total**
Coke	4	9	5	18
Juice	0	2	1	3
Water	1	3	0	4
Total	5	14	6	25

Two-way tables can be used to calculate marginal, intersections, and conditional probabilities.

5.3.3.1 Marginal Probabilities

Marginal probabilities are the chances of getting an event over the overall total.

Examples:

1. What is the probability that a customer ordered a chicken sandwich?To solve this problem, we look at the total number of customers who ordered a chicken sandwich and we divide it by the overall total.

   Solution:

   $\frac{5}{25} = \frac{1}{5} = 0.20$

2. What is the probability that a customer ordered coke?

   Solution:

   $\frac{18}{25} = 0.72$

5.3.3.2 Probabilities of Intersections

To find the probability of a intersection, we can examine the cell of the two-way table and divide it by the overall total.

Examples:

1. What is the probability that a customer ordered a double burger with water?

   Solution:

   $\frac{3}{25} = 0.12$

2. What is the probability that a customer ordered nuggets with coke?

   Solution:

   $\frac{5}{25} = 0.20$

5.3.3.3 Conditional Probabilities

Now we will deal with the chance of an event A, given that we know B, so instead of looking at the overall total we will be considering the total of observations that belong to the given condition.

Examples:

1. What is the probability that a customer ordered a chicken sandwich given that they ordered coke?To solve this problem, we look at the number of customers who ordered a chicken sandwich with a coke and we divide it by the number of people who ordered coke.

   Solution:

   $\frac{4}{18} \approx 0.22$

2. What is the probability that a customer ordered coke given that they ordered chicken nuggets?

   Solution:

   $\frac{5}{6} \approx 0.83$

5.4 Further Exploration

This section is optional, and it is a deeper investigation of probability theory.

5.4.1 Probability as a Long Run Procedure

Probability: The proportion of a random outcome in the long run.

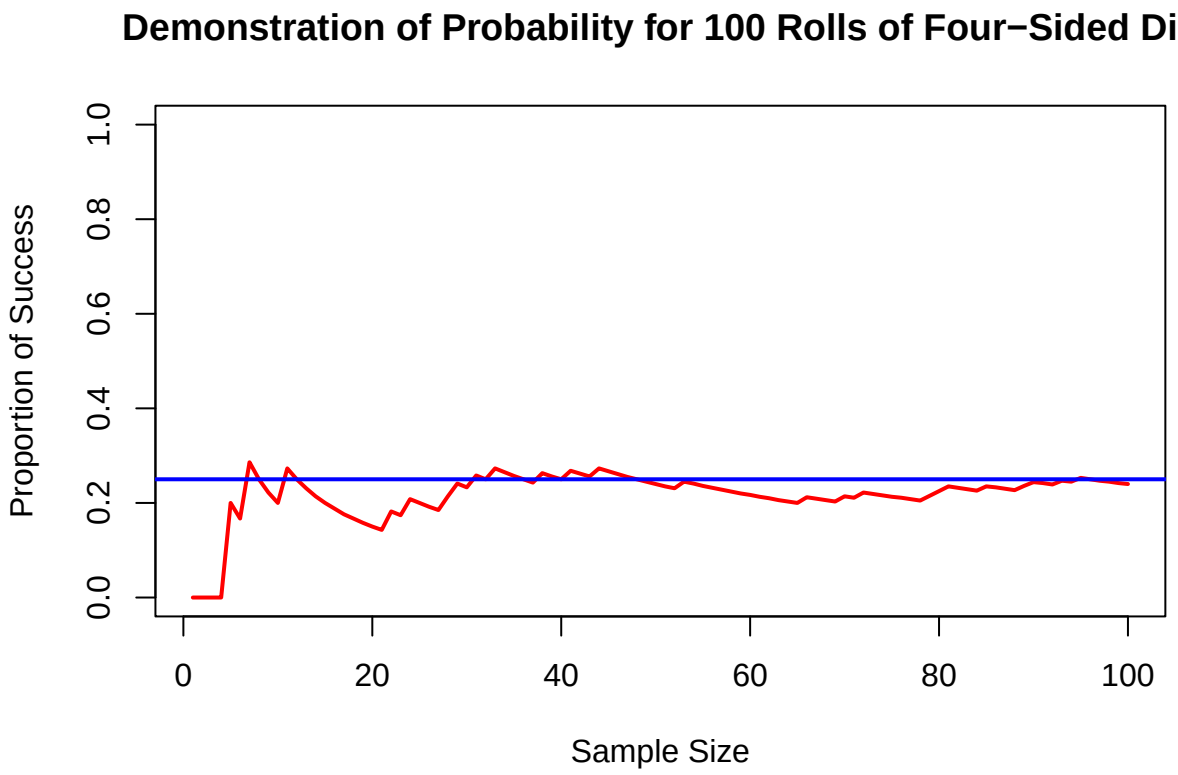
Random Outcome an outcome is uncertain, but none-the-less follows a regular pattern in the long run.

Illustration:

Consider a fair four-sided die. Imagine that we roll this die many times and keep track of successes and failures. Here is a sample of our rolls, where n is the roll number and X is tracking our success and failures, and prop is our proportion of successes:

```
      n    Outcome x         prop
 [1,] "1"  "1"     "Failure" "0"
 [2,] "2"  "3"     "Failure" "0"
 [3,] "3"  "4"     "Failure" "0"
 [4,] "4"  "2"     "Success" "0.25"
 [5,] "5"  "1"     "Failure" "0.2"
 [6,] "6"  "4"     "Failure" "0.167"
 [7,] "7"  "1"     "Failure" "0.143"
 [8,] "8"  "1"     "Failure" "0.125"
 [9,] "9"  "1"     "Failure" "0.111"
[10,] "10" "2"     "Success" "0.2"
```

Now imagine we decide to roll the die again and again, keeping track of the proportion of successes out of our trial number. If we truly have a fair die, then the proportion of successes should approach *0.25* and more importantly stay close. Visually,



In this plot, the x-axis represents the number of rolls and the y-axis represents the proportion of successes.

Exercises

1. Identify the event and the sample space in the following situations. Then, solve the probabilities.
   (a) What is the probability of rolling a three or a five when rolling a 6-sided fair dice?
   (b) What is the probability of getting an even number when drawing a card from a standard deck of 52 cards?

2. Suppose we have a bag with 6 red balls, 4 green balls and 2 white balls. Classify the following events.
   (a) Selecting two red balls in a row without replacement.
   (b) Selecting anything other than a white ball.
   (c) Selecting one red ball and one green ball with replacement.
   (d) Selecting one red ball and the tossing heads when flipping a coin.

3. Calculate the probabilities from the events in the previous exercise.

4. Suppose we collected data about all employees from a company and expressed them as probabilities. Use the information below to calculate probabilities.
   P(owns a car)=0.89
   P(owns a car and had children)=0.66
   P(does not own a car and has a children)=0.02
   (a) What is the probability that a randomly selected employee does not own a car?
   (b) What is the probability that a randomly selected employee does not own a car and has children?
   (c) What is the probability that a randomly selected employee has children given that they own a car?
   (d) What is the probability that a randomly selected employee has children?
   (e) What is the probability of randomly selecting two employees who own a car?

5. In a writing event, we gathered information about writer's nationalities and genre, and we expressed it in the two way table below.

	Romance	**Fiction**	**Thriller**	**Total**
British	10	4	18	32
American	20	21	11	52
Canadian	5	7	4	16
Total	35	32	33	100

   Calculate the following probabilities.
   (a) What is the probability of selecting an American writer?
   (b) What is the probability of selecting a writer that writes romance?
   (c) What is the probability of selecting an American writer or that they write romance?
   (d) What is the probability of selecting a British writer that writes fiction?
   (e) What is the probability of selecting a British writer given that they write fiction?
   (f) What is the probability of selecting a writer that is not Canadian?
   (g) What is the probability of selecting a Canadian writer given that they write thriller?
   (h) What is the probability of selecting a writer who writes romance or fiction or thriller?

6 Probability Distributions

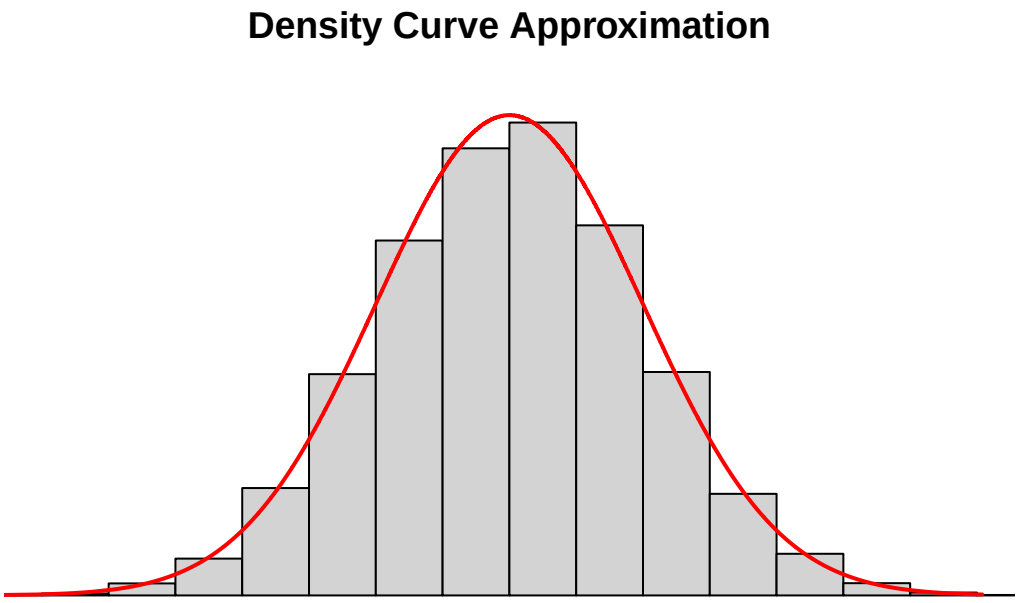
In the chapter, we will discuss some common probability distributions for both discrete and continuous distributions, by assigning a random variable X that represents a mapping from the events in a sample space to a numerical outcome.

6.1 Common Proability Distributions

6.1.1 The Normal Distribution

Density Curve: A mathematical model used to approximate real world data. A density curve must have an entire area under the curve to be one, and must always be on or above the axis.

There are many density curves of interest but for now, we will concentrate on the Normal Distribution.



The Normal Distribution: A bell shaped and symmetric distribution that is specified by its mean μ and standard deviation σ. The mean μ will specify the center of the distribution while the standard deviation σ will determine the spread.

Note: All Normal Distributions are proportional to their standard deviations.

We will use the notation $N(\mu, \sigma)$ to represent the shape, center, and spread of Normal Distributions.

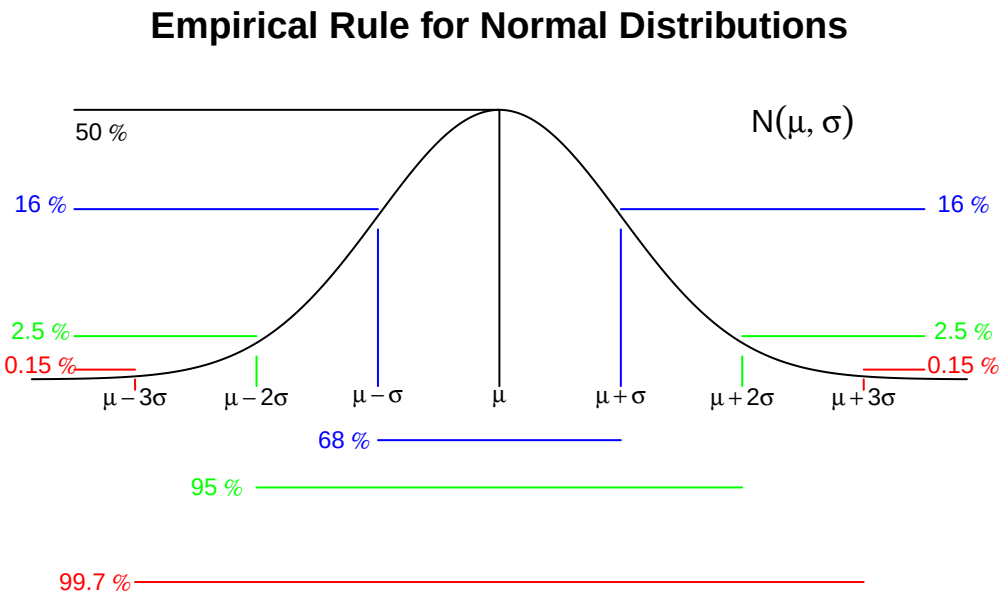
Properties of the Normal Distribution:

- It is a bell shaped
- It is symmetric

- It is unimodal
- The curve does never touch or cross the x-axis
- It has mean μ and standard deviation σ (the mean μ will specify the center of the distribution while the standard deviation σ will determine the spread)
- The mean, the median and the mode are approximately equal
- The notation N(μ, σ) represents the shape, center, and spread of Normal Distributions
- All Normal Distributions are proportional to their standard deviations
- The area under the curve is a representation of probabilities
- The area under the entire curve is equal to 1 or 100%
- All Normal Distribution follow the 68%-95%-99.7% rule

6.1.1.1 The 68%-95%-99.7% Rule

The 68%-95%-99.7% Rule, also known as the **Empirical Rule**, is used as an approximation as can be illustrated with the following image.



This rule states the following:

- Roughly 68% of observations fall within one standard deviation from the mean, $\mu \pm 1\sigma$
- Roughly 95% of observations fall within two standard deviations from the mean, $\mu \pm 2\sigma$
- Roughly 99.7% of observations fall within three standard deviations from the mean, $\mu \pm 3\sigma$

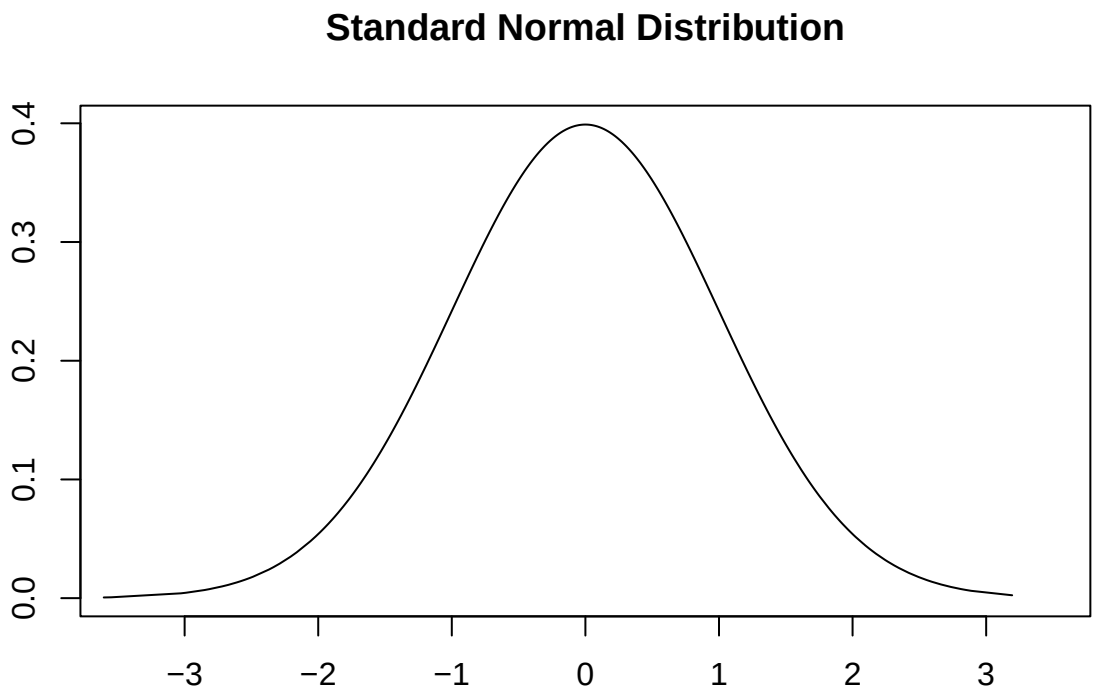
Note: These percentages represent the probabilities of finding observations below, above or between one, two or three standard deviations from the mean.

6.1.1.2 Standardization and The Standard Normal Distribution

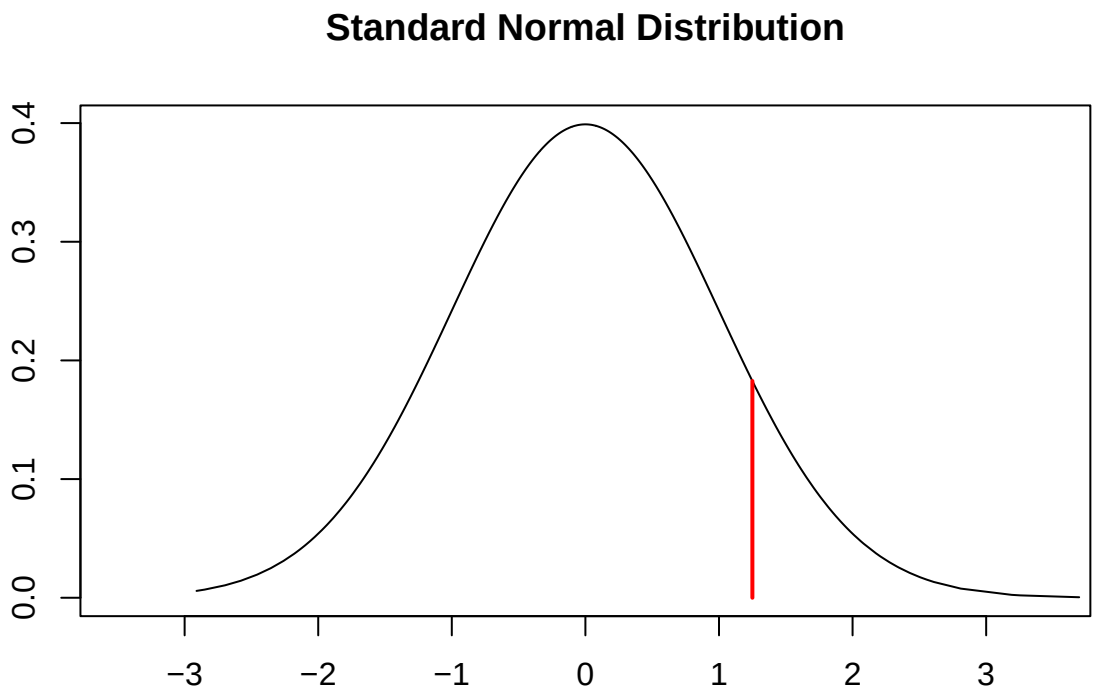
The Standard Normal Distribution: A Normal Distribution with a mean of zero and a standard deviation of one. That is $N(\mu = 0, \sigma = 1)$. Cumulative Proportions, which represent areas under the density cure to the left of the **z-score**, are found in **The Standard Normal Distribution Table** or through the use of R.

Z-score: A statistical measure that quantifies the distance of a particular data point from the mean of a distribution in terms of standard deviations. It indicates how many standard deviations an observation is above or below the mean. It is also known as the standard score.

The Standard Normal Distribution is demonstrated below.



A z-score of 1.25 represents a data point that is 1.25 standard deviations above the mean. This is visualized below.



Standardization

When working with real data, finding distributions that are naturally described by a Standard Normal Curve (distribution with a mean of 0 and standard deviation of 1) are very hard to find. However, because all Normal Curves are proportional to their standard deviations, we can obtain standard z-scores for any Normal Distribution with mean other than 0 and standard deviation other than 1 using a technique called **standardization**.

Standardization: A re-scaling of any Normal Distribution to the Standard Normal Distribution. To standardize a value is to measure the number of standard deviations away the value is from the mean. For Normal Distributions, we will refer to the standardized value as a **z-score**.

If we allow x to be the value of an observation, then for Normal Distribution we will find the z-score using the standardization formula.

Standardization formula:

$$z = \frac{x - \mu}{\sigma}. \tag{1}$$

where - z = the z-score
- x = our data point
- μ = the mean of the distribution
- σ = the standard deviation of the distribution

Note: This formula represents the deviation away from the mean compared to the average deviation away from the mean. Standardizing values allows us to put distances into perspective so we can gauge what is close to the mean and what is far away.

For Normal Distributions, standardizing values, also known as finding z-scores, will take any Normal Distribution with its own mean and standard deviation and shift and re-scale it to the **Standard Normal Distribution**.

Example:

When finding probabilities using the Standard Normal Distribution, R will automatically scale the Normal Distribution for us when we use the corresponding code. The following will only give us the probability of being less than or equal to the observed value **x**. That is, if **X** follows a Normal Distribution with mean μ and standard deviation σ, then the following will find $P(X \leq x)$

To calculate probabilities under the Normal Distribution we use the following R code.

$$\text{pnorm(q= , mean = , sd =)}$$

where: - q = our data point.
- mean = the mean of the distribution.
- sd = the standard deviation of the distribution.

Next, we will examine different kinds of scenarios using the Normal Distribution and finding the correct probability.

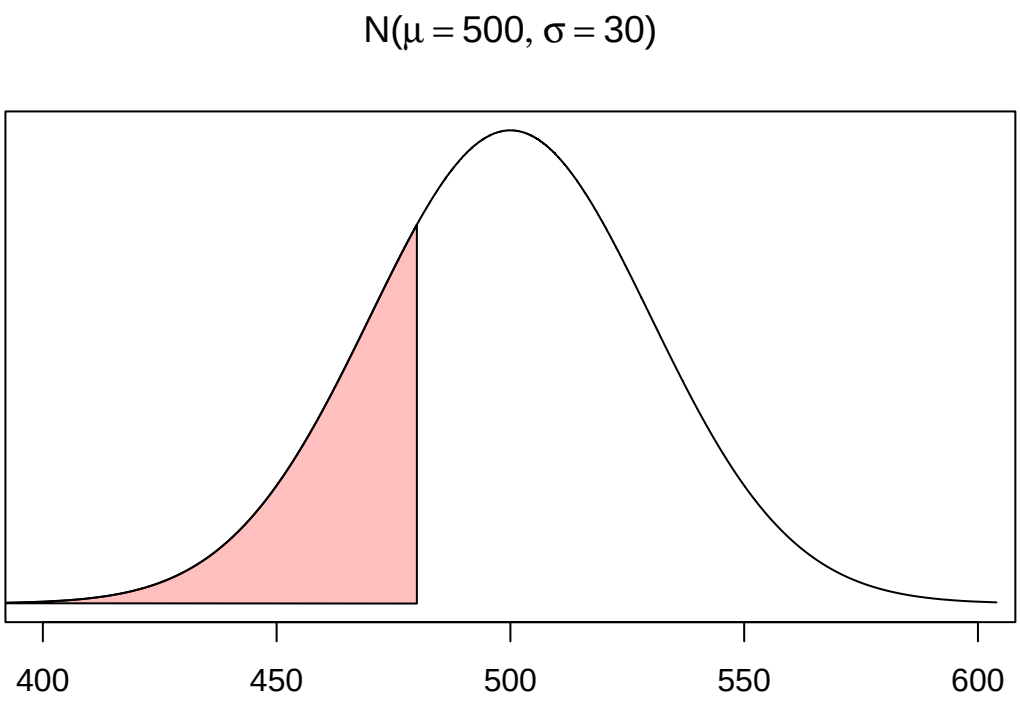
Cumulative Probabilities $P(X \leq x)$

Example:

Suppose that the weight of a certain distribution of fish is Normally distributed with a mean of 500 grams and a standard deviation of 30 grams. What is the probability that a randomly selected fish will weigh less than 480 grams?

Solution:

We wish to find this probability, which is visualized with the shaded region below:



In R, rounding to 4 decimal places,

```
round(pnorm(q=480,mean=500,sd=30),4)
```

```
[1] 0.2525
```

Alternatively, using the standardization technique, we would find the standardized value by:

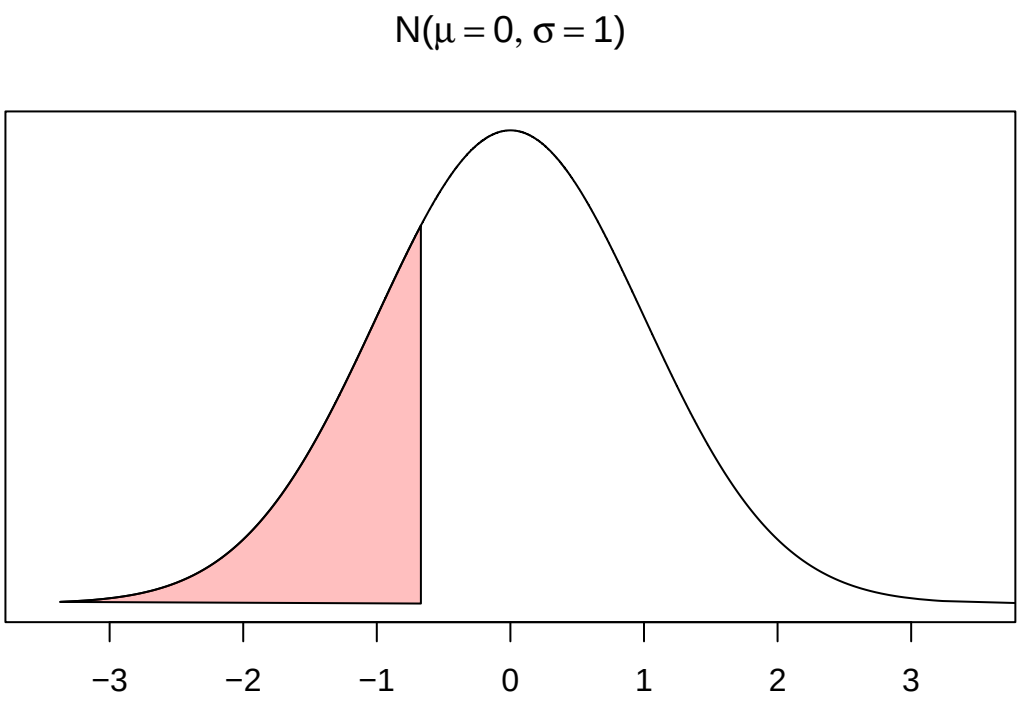
$$z = \frac{x-\mu}{\sigma} = \frac{480-500}{30} = \frac{-2}{3}.$$

That is, we are $\frac{2}{3}$ of a standard deviation below the the mean $\mu = 500$. So using the Standard Normal Distribution, we can see that $Z \leq -\frac{2}{3}$ would have a probability of 0.2525, obtained with the following code.

```
round(pnorm(-2/3),4)
```

```
[1] 0.2525
```

And can be visually represented in the Standard Normal Curve as:



Greater Than Probabilities $P(X \geq x)$

Example:

Consider the above example concerning the weight of fish in grams. What is the probability that we randomly select a fish that weighs more than 450 grams?

Solution:

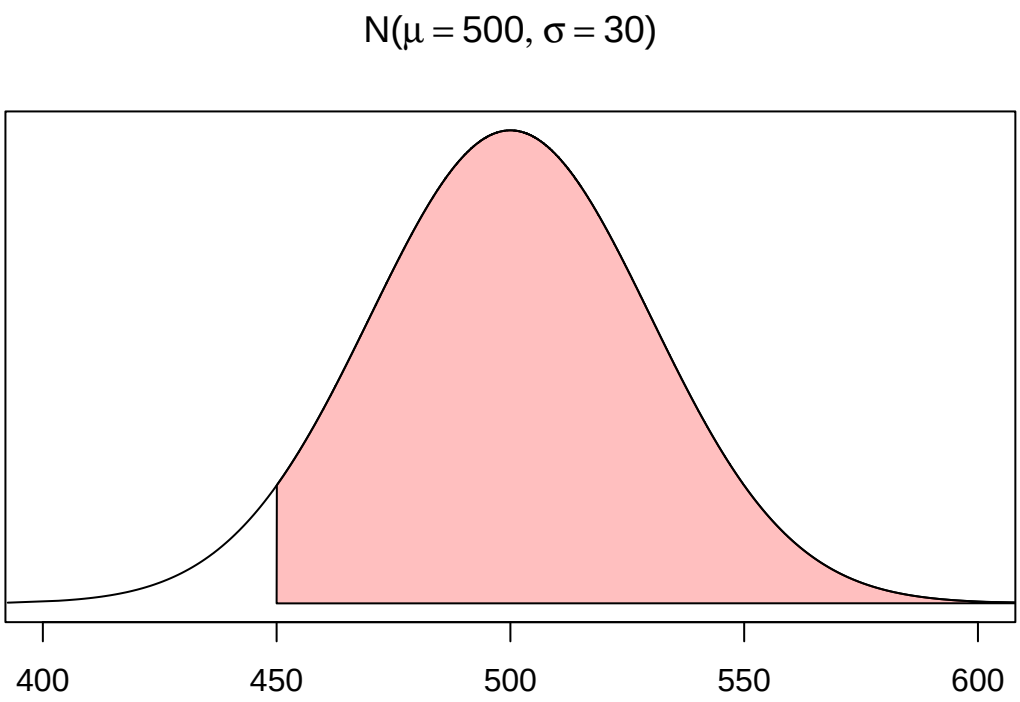
Here, we will need to use the complement probability law to our less than 450 gram event.

In R, rounding to 4 decimal places,

```
1-round(pnorm(q=450,mean=500,sd=30),4)
```

```
[1] 0.9522
```

Visualized by:



Note: The Standardized value for 450 grams on the distribution above is $z = -1.67$.

Between Two Values $P(x_1 \leq X \leq x_2)$

For our "between" problems we will use two cumulative probabilities and subtract these probabilities. Take care to ensure you subtract the smaller probability from the larger to ensure you answer is positive.

Example:

What is the probability that we randomly select a fish that weighs between 450 and 520 grams?

Solution:

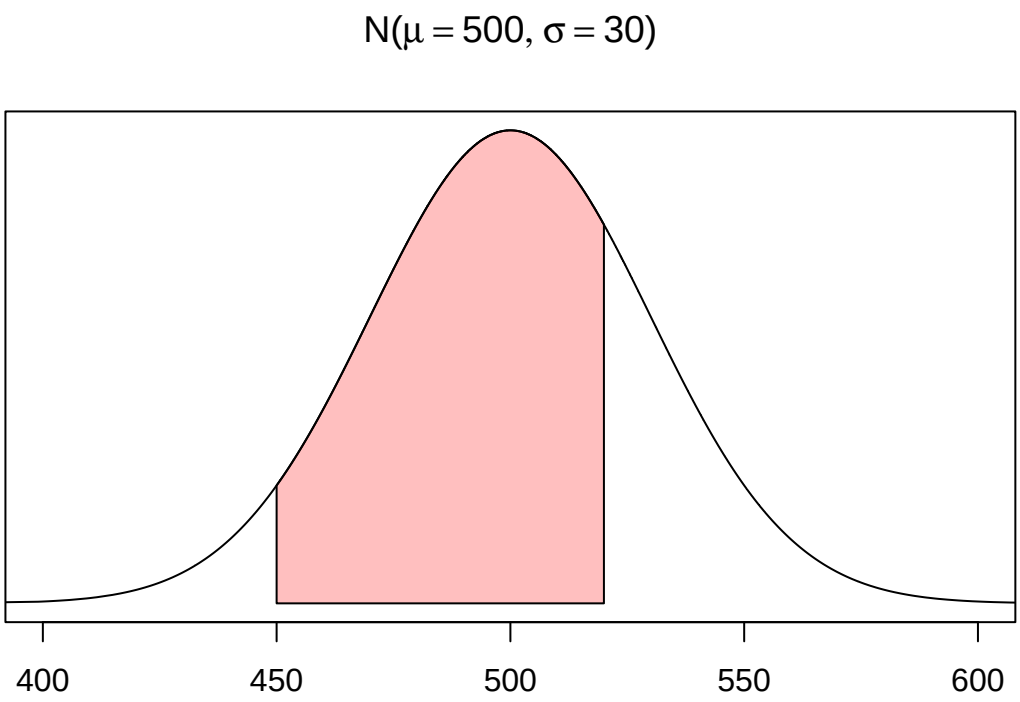
In this case, we calculate the probabilities of obtaining a weight that is less than each of these values and then subtract the smaller one from the greater one.

In R, rounding to 4 decimal places,

```
x1<-round(pnorm(q=450,mean=500,sd=30),4)
x2<-round(pnorm(q=520,mean=500,sd=30),4)
x2-x1
```

```
[1] 0.6997
```

This can be visually represented as:



Backwards Problems

In terms of our "backwards" problems, we will start with a proportion and work backwards to find us the value that will generate the appropriate proportion. This will entail using z-scores and Standard Normal Distribution again in conjunction. These are also referred to as finding a value given a proportion.

Example:

Consider the distribution of distribution of the weights of fish as discussed above. What is the top 10% of fish weights?

Solution:

In this case we want the top 10% which will correspond to the cumulative proportion of 0.9000 or 90%. Since we have a Normal Distribution, we find a z-score that is the closest we can get to 0.9000 using the cumulative proportions from the Standard Normal Distribution.

In R, qnorm is used to represent a quantile,

```
round(qnorm(.9000),2)
```

```
[1] 1.28
```

We know the mean of the distribution as well as the standard deviation, so we are merely looking for the value of x that will give us the standardized value of 1.28.

To do this we will use the standardization formula: $z = \frac{x-\mu}{\sigma}$

Then, we rearrange it to solve for x:

$$x = z \times \sigma + \mu$$

Continuing in R,

```
x = 1.28*30+500
x
```

```
[1] 538.4
```

Alternatively, in R,

```
round(qnorm(.9000,mean=500,sd=30),1)
```

```
[1] 538.4
```

Answer: Top 10% of fish weights is 538.4.

6.1.2 Binomial Distribution

Binomial Distribution: A discrete probability model that will assign the probabilities to the counts of success within a fixed number of trials n. It responds to questions such as, "what is the probability of getting a 3 four times if we roll a die ten times?"

Binomial Distribution Conditions:

- A fixed number of trials (n)
- Dichotomy, that is, only two possible outcomes {Success,Failure}
- There is independence within each trial
- The probability of success, $P(Success) = p$, stays the same within each of the trials

The sample space for a Binomial Distribution will always be $X = 0, 1, \ldots, n$.

The Binomial Distribution has the probability model:

$$\binom{n}{x} p^x (1-p)^{n-x} \text{ for } x = 0, 1, \ldots, n$$

where: - n is the number of trials
- x is the number of successes
- p is the probability of success
- $\binom{n}{x} = \mathrm{C}_x^n = \frac{n!}{x!(n-x)!}$ is the combination formula

Combination: A counting technique used to find all the possible ways of obtaining a specific outcome.

Note: The $n!$ refers to a factorial and represents multiply the all whole numbers from n down to 1.

Example:

$5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$

In R,

```
factorial(5)
```

```
[1] 120
```

Note: $0! = 1$

The combination can be found in R as well by using the choose function. Consider $\binom{n=10}{x=4}$,

```
choose(10,4)
```

```
[1] 210
```

However, we will use R to calculate binomial probabilities automatically.

Calculating Binomial Probabilities

We will explore two cases of binomial probabilities: **exact** and **cumulative**. A description of each case and their respective R codes are provided below.

Case 1: Assigning Exact Probabilities to an Event

Example:

What is the probability of rolling a die and getting a 3 exactly four times if we roll a die ten times?

Solution:

$n = 10$

$x = 4$

$p = \frac{1}{6}$

$$P(x = 4) = \binom{10}{4} \frac{1}{6}^4 \frac{5}{6}^6 \approx 0.0543$$

In R, the code for these cases is given by

$$\text{dbinom(x=, size=, prob=)}$$

where: - x = the number of successes that we want
- size = the number of trials
- prob = the probability of success

In this case, rounding to four decimal places,

```
round(dbinom(x=4,size=10,prob=1/6),4)
```

```
[1] 0.0543
```

Then, the probability of rolling a 3 exactly four times if we roll a die ten times is 0.0543.

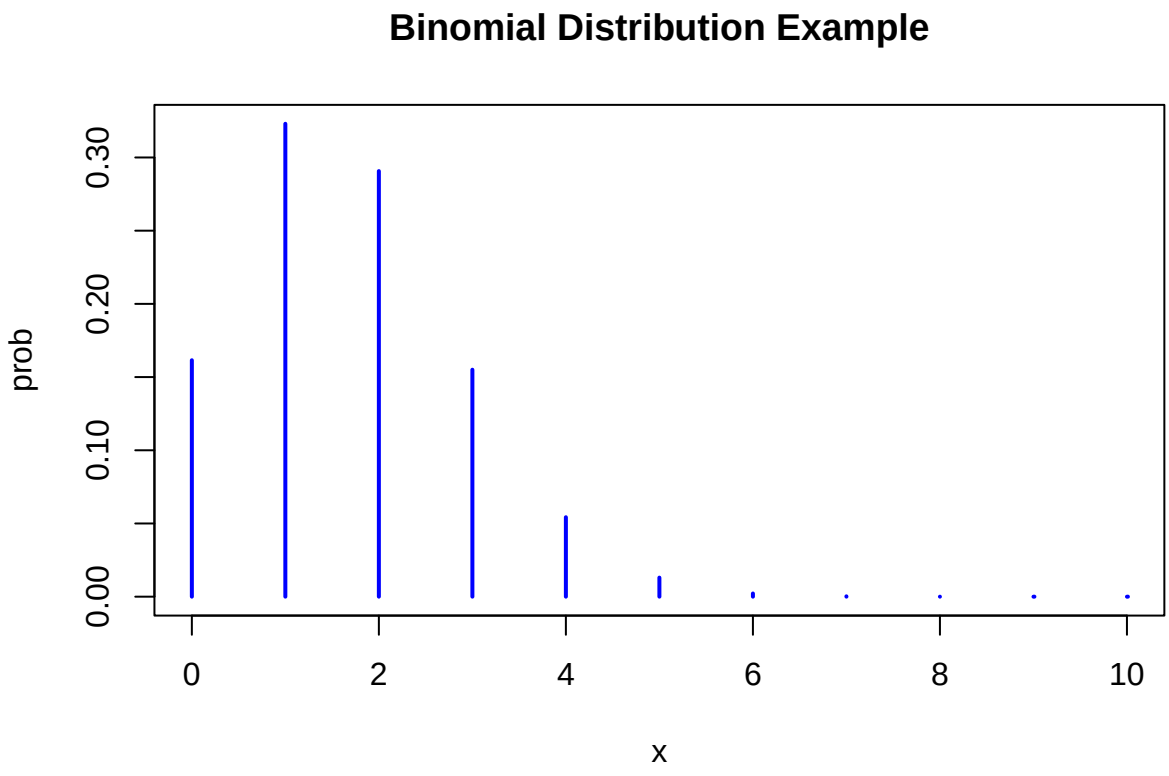
Further, the following code will illustrate the probabilities of getting a 3 exactly X number of times.

```
# Define n
n=10

# Define the Probability of Success
p=1/6

# Create the Sample Space
x<-seq(0,n,1)
```

```r
# Assign Exact Probabilities to Each Event
prob<-dbinom(x,n,p)

# Display the Probabilities
cbind(x,round(prob,4))
```

```
      x
 [1,]  0 0.1615
 [2,]  1 0.3230
 [3,]  2 0.2907
 [4,]  3 0.1550
 [5,]  4 0.0543
 [6,]  5 0.0130
 [7,]  6 0.0022
 [8,]  7 0.0002
 [9,]  8 0.0000
[10,]  9 0.0000
[11,] 10 0.0000
```

```r
# Plot the Probability Mass Function
plot(x,prob,typ="h", main="Binomial Distribution Example", col="blue", lwd=2)
```



Case 2. Finding Cumulative Probabilities for a Binomial Distribution

The R codes for these cases are given below.

- For "less than or equal to" x successes, $P(X \leq x)$:

pbinom(q=, size=, prob=)

where: - q = the number of successes that we want
- size = the number of trials
- prob = the probability of success

- For "more than" x successes, $P(X > x)$:

1-pbinom(q=, size=, prob=)

Note: We are using the complement rule here.

Examples:

1. What is the probability of rolling a die and getting a 3 less than two times if we roll a die ten times? That is, $P(X < 3)$

Note: This is strictly less than 3. Then, the argument for q is 2. Notice that $P(X < 3) = P(X \leq 2)$.

Solution:

```
round(pbinom(q=2,size=10,prob=1/6),4)
```

```
[1] 0.7752
```

2. What is the probability of rolling a die and getting a 3 at least twice if we roll a die ten times?

Note: $P(X \geq 2) = 1 - P(X = 1)$, where here we will apply the complement rule.

Solution:

```
1-round(pbinom(q=1,size=10,prob=1/6),4)
```

```
[1] 0.5155
```

6.2 Other Common Probability Distributions

In the following, we will give some other common probability distributions, but will refrain from going to much detail.

6.2.1 The Uniform Distribution

The Uniform Distribution is a distribution in which the probability of success is equally likely for all elements in the sample.

Example:

A sports team organized a giveaway for their fans in which the prize was a jersey. A total of 160 fans participated in the giveaway, each of them purchasing one giveaway ticket. Ann was one of the participants. What is the probability that Ann wins the giveaway?\

Solution:

P (win) = $\frac{1}{160}$

Note: A Uniform Distribution can either be discrete or continuous. An example of a Discrete Uniform Distribution would be a fair 4-sided die, while an example of a Continuous Uniform Distribution may be to select a number within the interval [0,1].

6.2.2 The Bernoulli Distribution

A Bernoulli Distribution is a discrete distribution for either success or failure for one trial or experiment. The $P(Success) = p$ and, thanks to the complement law, the $P(Failure) = 1 - p$.

Example:

What is the probability of rolling a 3 on a fair 6-sided die and the probability of not rolling a 3?

Solution:

$\frac{1}{6}$ and $\frac{5}{6}$ respectively.

6.2.3 The Geometric Distribution

The Geometric Distribution is a discrete distribution that is counting the number of trials up to and including the first success. The trials are independent of one another and the probability of success p stays the same within each trial.

6.2.4 The Negative Binomial Distribution

The Negative Binomial Distribution is a discrete distribution that concerns itself with the number of trials up to and including the r^{th} success. The trials are independent of one another and the probability of success p stays the same within each trial.

Note that if $r = 1$, then the Negative Binomial Distribution is the Geometric Distribution.

6.2.5 The Poisson Distribution

The Poisson Distribution is a discrete distribution that looks at approximating the Binomial Distribution when n is large and p is small. It can be thought of as modeling the number of successes within a fixed interval of time or space.

Exercises

1. The monthly salaries of employees in a company are Normally distributed with a mean of $5,000 and a standard deviation of $600.

   (a) What is the probability of randomly selecting an employee with a salary of less than $4000?

   (b) What is the probability of randomly selecting an employee with a salary higher than $5500?

   (c) What is the probability of randomly selecting an employee with a salary between $5000 and $6000?

2. A restaurant owner knows that on average, 25% of customers who enter the restaurant will order dessert. If 20 customers enter the restaurant today, what is the probability that exactly 5 of them will order dessert?

3. For quality control, a company samples 15 products at each production run. If the probability of an item being defective is 0.03, what is the probability that at most one item is defective?

4. A solitaire player has won 196 out of the 230 games that he has played in his career. What is the probability that he wins at least 4 of the next 5 solitaire games?

7 Sampling Distributions and The Central Limit Theorem

The information inside of this chapter is the foundation for remaining topics in this booklet. The topics of this chapter will be:

1. Parameter vs Statistic
2. The Law of Large Numbers
3. Sampling Distributions

7.1 Parameter vs Statistic

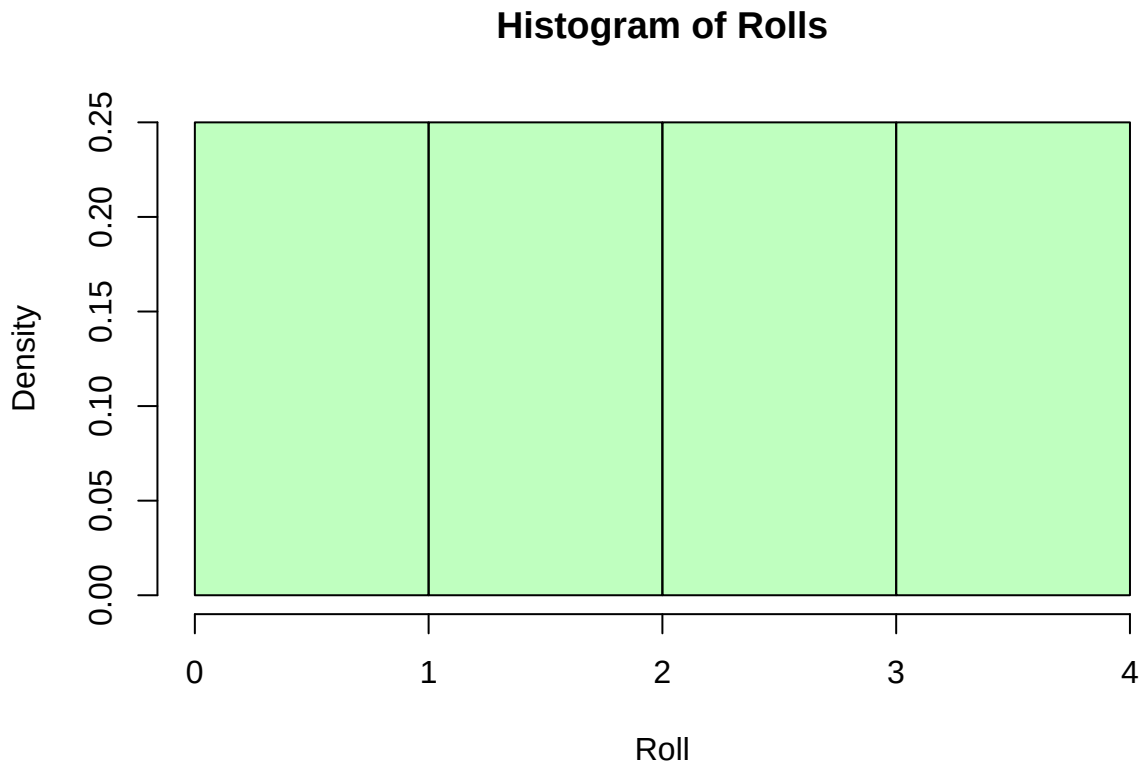
A **population distribution** is the distribution of a variable taken by all the individuals. The values that describe our population are referred to **parameters**. The specific parameter depends on our research question, but typically we represent these values with either μ, σ, or p, for the mean, standard deviation, and proportion respectively. The values of the parameters are typically unknown and we will use our sample statistics to estimate these unknown values.

What follows is an example that illustrates the population distribution in its entirety.

Consider the individual rolls of a fair four-sided die as a population distribution, with probability model:

x	1	2	3	4
$\mathbf{P}(x)$	0.25	0.25	0.25	0.25

and distribution plot given by the following histogram:



The mean $\mu = 2.5$ is our population mean, and represents the expected roll on a four-sided die.

The standard deviation of the population distribution is $\sigma = 1.118$.

However, in most real world cases, the population parameters are unknown, and thus, we attempt to estimate them utilizing sample statistics.

Illustration:

Suppose we do not know the real values of the population parameters so we roll the die 5 times to collect a simple random sample of size $n = 5$, to try to estimate them. Suppose we got the following rolls: 4, 3, 1, 4, and 2.

Calculating the mean ($\bar{x}$) and standard deviation (s) using R we get the following results.

```
# Creating the Variable
rolls<-c(4,3,1,4,2)
rolls
```

```
[1] 4 3 1 4 2
```

```
# Calculating the Sample Mean
xbar<-mean(rolls)
xbar
```

```
[1] 2.8
```

```
# Calculating the Sample Standard Deviation
s=sd(rolls)
s
```

```
[1] 1.30384
```

These, would act as estimates for our unknown population parameters.

Note: These sample estimates ($\bar{x} = 2.8$ and $s = 1.30384$) are not exactly equal to the true value of the parameters μ and σ, but just representative of a guess.

Next, we will explore what will happen if we increase our sample size.

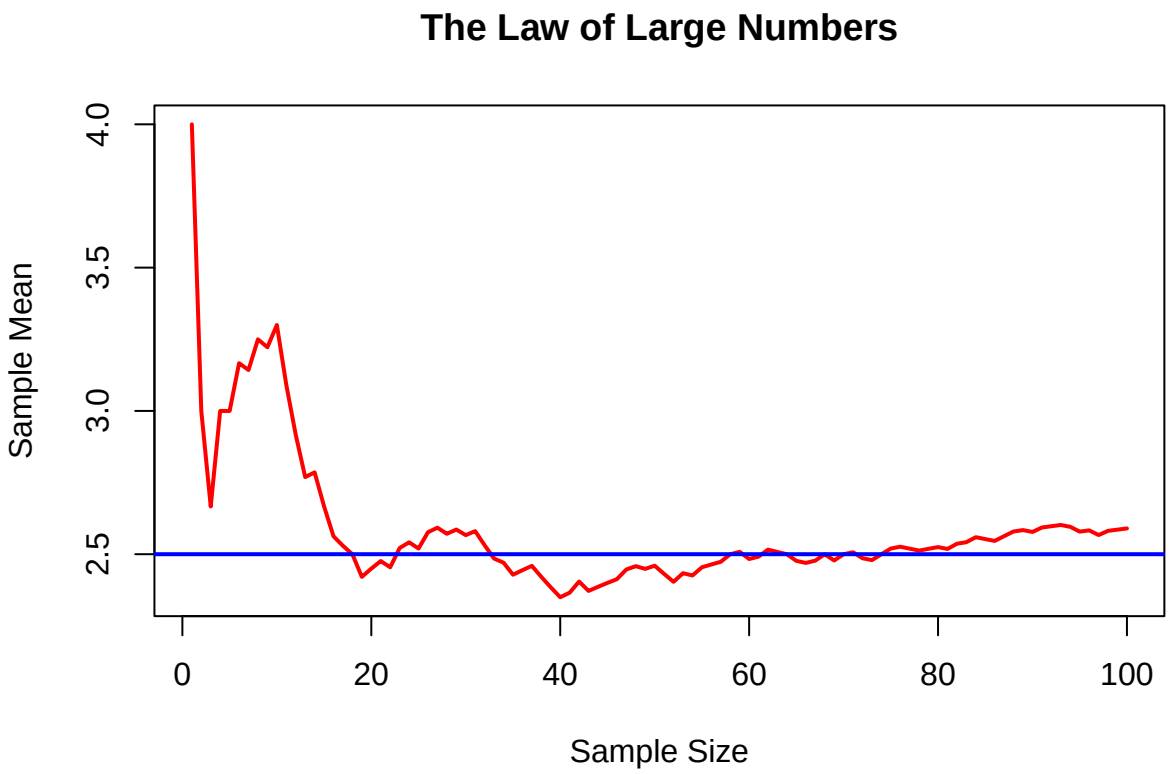
7.2 The Law of Large Numbers

Considering a simple random sample from any population with finite mean μ, the **Law of Large Numbers** states that as the number of observations drawn increases, our sample mean $\bar{x}$ tends to get closer and closer to the true value of the population parameter μ.

Illustration:

Considering the example of our four-sided die, suppose we roll it many, many, times and each time, we record the outcome of each roll constructing many simple random samples (by adding each new roll to the previous sample. We will then keep track of the sample mean $\bar{x}$ through each of our samples.

The following plot is a simulation of the situation described above. The population parameter μ is shown by the blue line, and the sample means are shown by the red line. The x-axis represents the sample size and the y-axis represents the sample mean.



Note: This is essentially a limiting process. As our sample size increases, our statistic approaches the value of the parameter.

7.3 Sampling Distributions and The Central Limit Theorem

A **Sampling Distribution** is the distribution of a statistic that is taken for all possible samples of the size n from the population of interest.

In general, the procedure to build up a Sampling Distribution is:

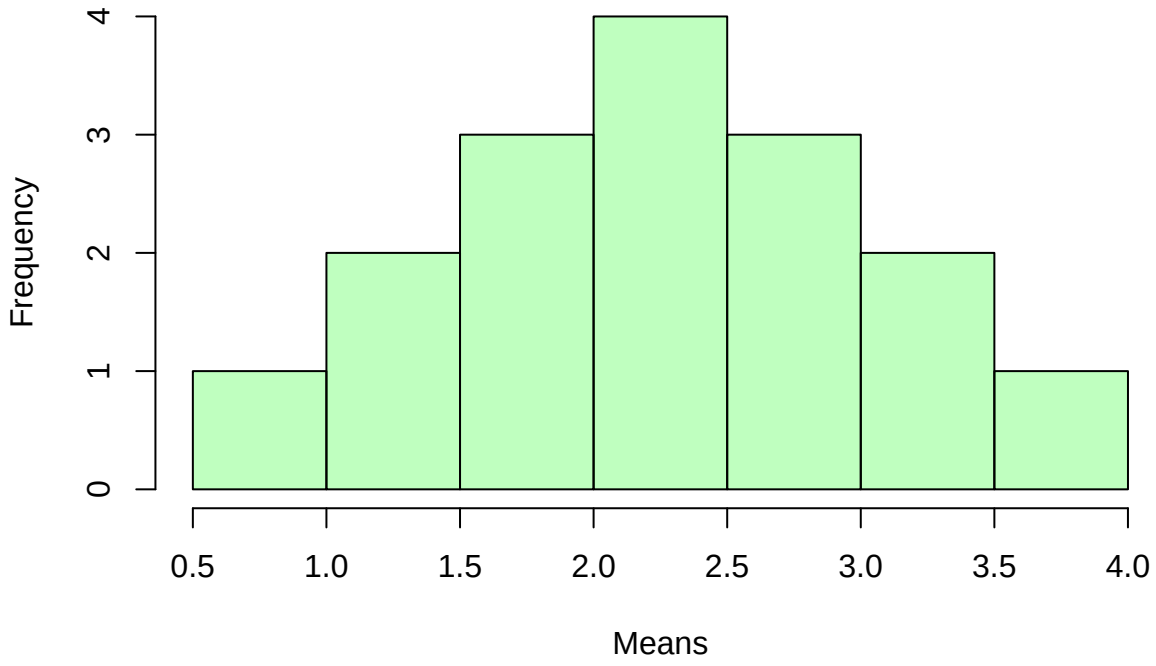
1. Take many SRS's of size n from the population of interest
2. Within each selected sample calculate your statistic of interest
3. Create a histogram of all the statistics gathered from your samples

Illustration:

Consider the example dealing with the four-sided die. If we roll the four-sided die twice we recognize that there are 16 possible samples of size 2, holding the following sample means:

	Samples															
First Roll	1	1	1	1	2	2	2	2	3	3	3	3	4	4	4	4
Second Roll	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
Sample Mean	1	1.5	2	2.5	1.5	2	2.5	3	2	2.5	3	3.5	2.5	3	3.5	4

The sample means of each of these samples of size two set up the Sampling Distribution of the Sample Mean of Samples of Size $n = 2$. Since this is all possible samples of size two, this represents the full Sampling Distribution of Samples of Size 2, and it is plotted below:



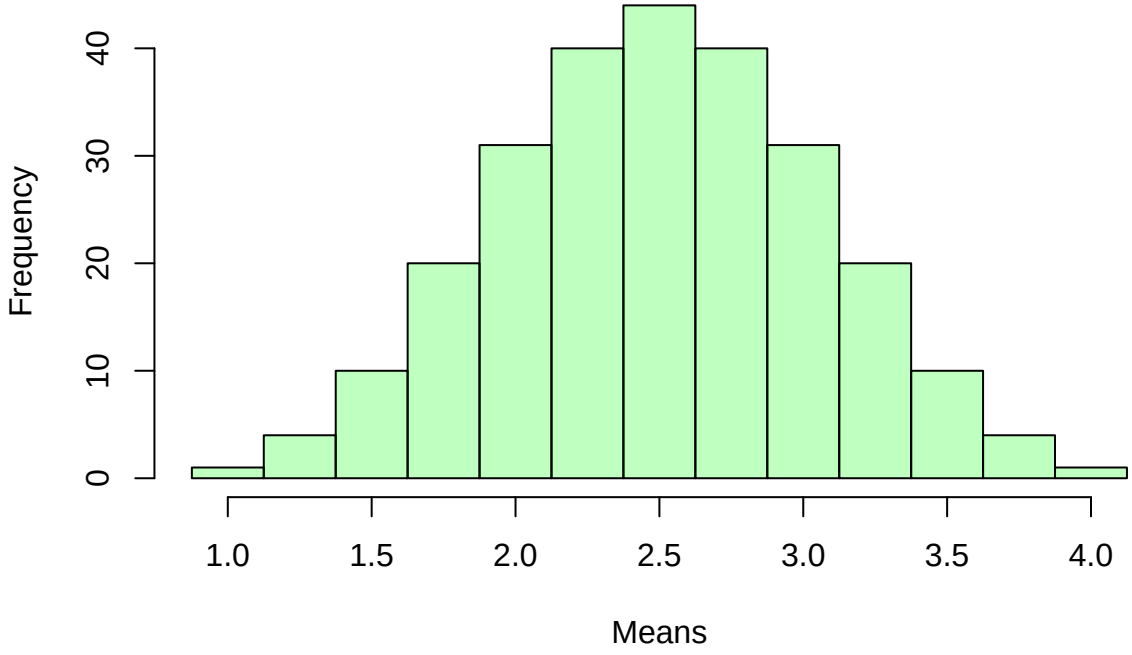
To describe distributions we use:

1. Shape
2. Center
3. Spread

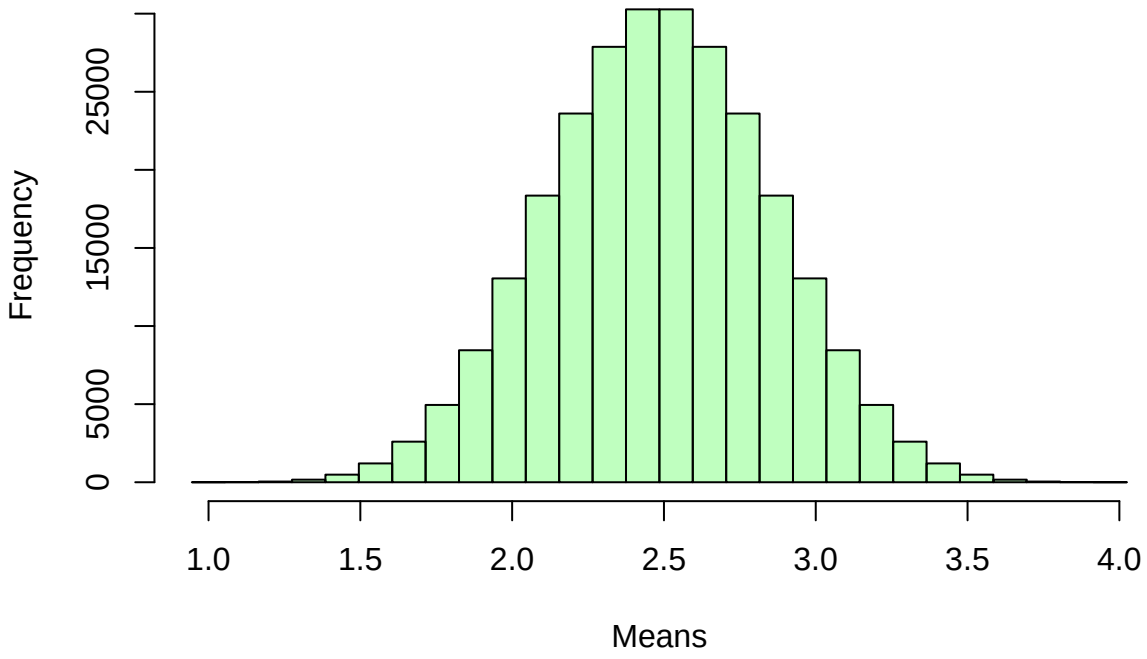
In this case, we have:

1. Shape: A symmetric distribution starting to look Normal.
2. Center: A sample mean of 2.5, which is equal to the population parameter. This happens because our sample mean $\bar{x}$ is known as an unbiased estimator.
3. Spread: A standard deviation of 0.9393, which is smaller than the population distribution, in part due to our law of large numbers.

To introduce and begin to illustrate the **Central Limit Theorem** we will increase our sample size to n=4, and repeat the process. The following is the Sampling Distribution of the Sample Mean of Samples of Size 4:



Increasing the sample size to $n = 9$ and repeating the process we get the following plot:



Conclusion:

As we increase the sample size, the histograms start to look more bell-shaped, they are centered at the population mean μ of 2.5.

Note: We get more clustered about the mean and therefore have a smaller standard deviation than samples of individual rolls. Averages are less variable than individuals.

This was an illustration of **The Central Limit Theorem**. Stated formally:

Drawing a simple random sample from any population with finite mean μ and standard deviation σ, when the sample size n is sufficiently large, then the Sampling Distribution of the Sample Mean is approximately Normal with mean μ and standard deviation σ. That is,

$$\bar{X} \sim N(\mu, \tfrac{\sigma}{\sqrt{n}})$$

The Central Limit Theorem gives us the shape, center, and spread of the Sampling Distribution of the sample mean of samples of size n. The symbol $\sim$ is read as "distributed as".

Exercises

1. What is a Sampling Distribution?
2. State The Central Limit Theorem.
3. Suppose a large population of American adults have an average income of \$50,000 and a standard deviation of \$10,000. A polling organization wished to take an SRS of size n=100 from this population and calculate the sample mean $\bar{x}$.
   (a) What is the shape of the Sampling Distribution of the Sample Mean?
   (b) What is the mean of the Sampling Distribution of the Sample Mean?
   (c) What is the standard deviation of the Sampling Distribution of the Sample Mean?
   (d) What is probability that the average of a sample of size 100 from this population will yield an average that is less than \$45,000?

8 Inference Methods

Utilizing the Central Limit Theorem as the foundation for this chapter, we will introduce the concepts and applications of Statistical Inference.

Statistical Inference: The process of drawing conclusions about a population parameter based on a sample statistic.

Inference Methods: Techniques used to draw these conclusions from Statistical Inference.

8.1 Overview of Inference Methods

Inference methods will rely heavily on a Sampling Distribution. There are two common methods used in Statistical Inference:

1. Confidence Intervals
2. Tests of Hypothesis

Statistical Inference can be done for different types of parameters, and each parameter has their own different formulas. The two parameters that we will explore in this booklet are proportions and means. Also inference processes can be classified into one or two sample processes. We will also explore one type of Chi-Squared tests.

8.1.1 Confidence Intervals

Confidence intervals are intended to construct an interval or range of values to which the true population parameter may belong through the use of a sample statistic.

Confidence Intervals are comprised of two things:

- **A Confidence Level.** Confidence levels represent the success rate of the method. That is, in the construction of every confidence interval of this type, we will have constructed an interval that contains the true population parameter. The typical levels of confidence are 90%,95%, and 99%.
- **An Interval of estimation.** Their general form is *point estimate* $\pm$ *margin of error*, in which the point estimates serves as the initial guess and the lower and upper bounds of the interval are given by that estimate plus and minus a margin of error, which depends strictly in the confidence level and the parameter we are trying to estimate.

Interpretations of a Confidence Interval

The general interpretation of a confidence interval is:

"We are ____% confident, the true population parameter lies within the calculated interval."

Behavior of Confidence Intervals

- The greater the sample size, the smaller the margin of error.
- The greater the level of confidence, the larger the margin of error.
- The smaller the level of confidence, the smaller the margin of error.

8.1.2 Tests of Hypothesis

Tests of Hypothesis are intended to assess the validity of a claim about a population parameter. They are not intended to prove a claim, but to find evidence to support one of two opposing ideas. They are also often defined as **Tests of Statistical Significance**.

Statistical Significance: A statistic that is such a large number of standard deviations away from an assumed value that it has a small likelihood of occurring.

Steps to perform tests of significance:

1. Determine the type of test
   - Left tailed test: if we believe the parameter is smaller than suggested.
   - Right tailed test: if we believe the parameter is higher than suggested.
   - Two-tailed test: if we believe it is different than states but we do not specify if lower or higher.
2. State the claims (the representation of two conflicting ideas, an initial claim and an alternative claim)
   - The **Null Hypothesis (H_0)** represents a statement of equality, is always assumed to be true and is written in terms of the population parameter.
   - The **Alternative Hypothesis (H_a)** represents what the researcher believes to be true, it is anything other than an equality ($<, >, or\ \neq$), and it is used to test the Null Hypothesis. The Alternative will be determined by the type of test: left tailed, right tailed or two tailed, for each symbol respectively.
3. Calculate a Test Statistic

   Test Statistic: A standardized value. We will measure the distance our observed statistic is from the hypothesized value in terms of the number of standard deviations.
4. Calculate a P-Value

   P-value: The probability of observing the statistic obtained or that of one more extreme under the assumption that the null hypothesis were true.
5. Make a Decision

   Decisions are based on comparing the p-value to a predetermined level of significance α. The value of α represents a gauge on the p-value that will help us determine what is considered a large or small probability. The value of α is calculated as 1-(Level of Confidence).
   - If P-value $< \alpha$ - Reject the H_0.
   - If P-Value $> \alpha$ - Fail to Reject the H_0.
6. Interpret the decision to state a conclusion
   - If H_0 was rejected: "There is enough statistical evidence to reject the null and conclude that (alternative hypothesis)".
   - If H_0 was not rejected: "There is not enough statistical evidence to suggest that (alternative hypothesis)".

Note on chapter format: Each subsection dealing with hypothesis tests includes the code for calculating the corresponding p-value. These codes are repetitive, since they are always the same depending on the Sampling Distribution of each case. This was thought in case readers wished to alter the order of learning these Inference Methods.

8.2 One-Sample Statistical Inference

Here we will explore Inference Methods for single samples.

8.2.1 One-Sample Statistical Inference for Proportions

Parameter of interest: p, the true population proportion.

Statistic to be used: is $\hat{p}$, the sample proportion.

Conditions for Inference:

- We work with a simple random sample from the population of interest.
- The sample size is sufficiently large (at least 10 successes and 10 failures).
- The Sampling Distribution is approximately Normal with mean p and standard deviation $\sqrt{\frac{p(1-p)}{n}}$.

When the sample size is sufficiently large, then the **Sampling Distribution** of the Sample Proportion is approximately described by

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

8.2.1.1 Confidence Intervals for Proportions

Idea: Use the sample statistic $\hat{p}$ to find a range of values to which the true population proportion may belong to.

Formula:

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

where: - $\hat{p}$ is the sample proportion
- α is the significance level. (100-confidence level) / 100 = α
- $z_{\alpha/2}$ is the critical z value
- n is the sample size

Note: $SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$ is an approximation of the population standard deviation, and it is called standard error of the sample proportion.

The table below provides the most typical $z_{\alpha/2}$ values.

Confidence Level	**Significance Level**	$z_{\alpha/2}$
90%	0.10	1.645
95%	0.05	1.96
99%	0.01	2.576

Any $z_{\alpha/2}$ value can be calculated with R using

qnorm(1-alpha/2)

Example:

Suppose for an election year we surveyed 400 people and found out that 37% stated they would vote for candidate A. Construct a 95% confidence interval for the true population proportion that is expected to vote for candidate A in the elections.

Solution:

$\hat{p} = 0.37$

$\alpha = (100 - 95)/100 = 0.05$

$z_{\alpha/2} = 1.96$

$n = 400$

Confidence Interval:

$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.37 \pm 1.96\sqrt{\frac{0.37(1-0.37)}{400}}$

In R,

```
# Lower Bound:
0.37-1.96*sqrt(0.37*(1-0.37)/400)
```

```
[1] 0.3226852
```

```
# Upper Bound:
0.37+1.96*sqrt(0.37*(1-0.37)/400)
```

```
[1] 0.4173148
```

```
# or
prop.test(148, 400, conf.level=0.95, correct=F)$conf.int
```

```
[1] 0.3241320 0.4183412
attr(,"conf.level")
[1] 0.95
```

Answer: The 95% confidence interval is $(0.3227, 0.4173)$

Interpretation: We are 95% confident that the true proportion of people who will vote for candidate A is between 0.32 and 0.42.

8.2.1.2 Hypothesis Tests for Proportions

Idea: Use the sample statistic $\hat{p}$ to evaluate the null hypothesis and reach a conclusion.

In these cases,

1. Type of test

   It can be:

   - Left tailed test
   - Right tailed test
   - Two-tailed test

2. Claims

- Null Hypothesis: $H_0 : p = p_0$, where p_0 is the hypothesized proportion.
- Alternative Hypothesis: It must be one of the following:
  - $H_a : p < p_0$
  - $H_a : p > p_0$
  - $H_a : p \neq p_0$

3. The Test Statistic

   Test Statistic: A value from the Normal Distribution that depends on the sample proportion, the claimed proportion, and the sample size. The test statistic formula is:

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

4. The P-Value

   The p-value can be calculated with R, and it depends on the alternative hypothesis and the test statistic. The codes to calculate p-values depending on the alternative are given by the following table.

$H_a : p < p_0$	$H_a : p > p_0$	$H_a : p \neq p_0$
pnorm(*test statistic*)	1-pnorm(*test statistic*)	2×(1-pnorm(*test statistic*))

5. Decision

6. Conclusion

Example:

It has been claimed that the proportion of incoming freshmen who start college as an undecided major is 0.3. An education researcher believes this number is erroneous and decides to collect data on the matter. He took a random sample of 186 incoming freshmen of different institutions and found out that 62 were undeclared majors. Conduct a hypothesis test at the 0.1 significance level to test the claim of the researcher.

Solution:

Type of test: two-tailed (indicated by the word "erroneous")

$H_0 : p = 0.30$

$H_a : p \neq 0.30$

$p_0 = 0.30$

$\hat{p} = 62/186 = 0.33$

$n = 186$

Test statistic:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.33 - 0.30}{\sqrt{\frac{0.30(1-0.30)}{186}}}$$

In R,

```
#Test Statistic:
tstat = (0.33-0.3)/(sqrt((0.3*(1-0.3))/186))
tstat
```

```
[1] 0.8928286
```

```r
#P-value:
2*(1-pnorm(tstat))
```

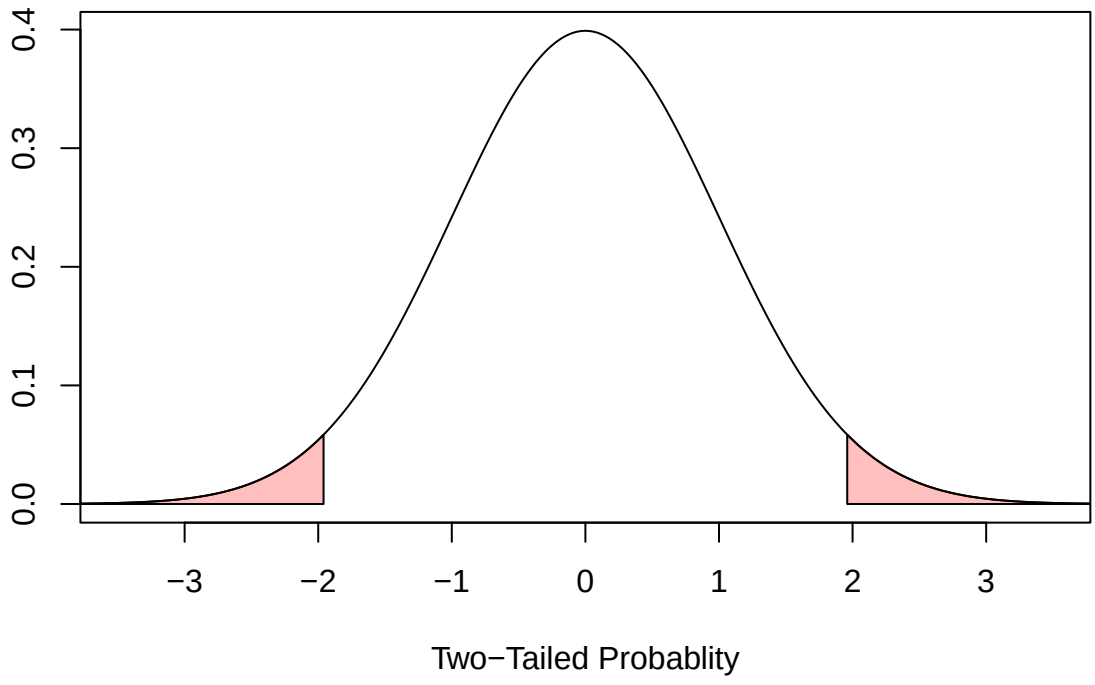
```
[1] 0.371949
```

Decision: Do not reject H_0.

Conclusion: Because the p-value is greater than the significance level, we do not reject the null and conclude that there is not enough evidence to suggest that true proportion of incoming freshmen that start college as undeclared major is different than 0.30.

Visualization:

For a two tailed test, we can plot the situation in the following manner, where the white region under the curve is called the **non-rejection region** and the cyan region under the curve is called the **rejection region**. These regions are determined by the significance level and represent probabilities. When the p-value is less than the significance level, the probability of observing the statistic obtained or one more extreme under the assumption of the null were true, will fall in the rejection region, and thus we will conclude to reject the null. When the p-value is greater than the significance level, the probability of observing the statistic obtained or one more extreme under the assumption of the null were true, will fall in the non-rejection region, and thus we will conclude to not reject the null, which is the case here.



8.2.2 One-Sample Statistical Inference for Means

Parameter of interest: μ

Statistic to be used: $\bar{x}$

Conditions for Inference:

- We work with a simple random sample from the population of interest.
- The population of interest is Normally distributed with mean μ and standard deviation σ.
- The procedure for inference for means depends on whether the population standard deviation is known or not. There are two cases:
    1. The population standard deviation is known. In these cases we use Z-Procedures.
    2. The population standard deviation is unknown. In these cases we use T-Procedures.

8.2.2.1 Case 1: Large Sample Z-Procedures for Inference for Means

- Confidence Intervals
- Hypothesis Tests

In these cases, the Sampling Distribution of the Sample Means is approximately described by

$$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

Thus, the Normal Distribution will be used to construct confidence intervals and perform tests of hypothesis.

8.2.2.1.1 Confidence Intervals for Means - Z-Procedures

Idea: Use the sample statistic $\bar{x}$ to find a range of values to which the true population mean may belong to.

Formula:

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

where: - $\bar{x}$ is the sample mean
- σ is the population standard deviation
- $z_{\alpha/2}$ is the critical z value
- α is the significance level
- n is the sample size

Note: $SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ is an approximation of the population standard deviation, and it is called standard error of the sample mean.

Example:

Suppose 10 children with the flu entered a hospital and their body temperature (in degrees Fahrenheit) was recorded at the check in. Their temperatures were recorded in the following table.

100.2	98.6	99.4	102.0	99.3	97.1	104.0	98.9	96.7	100.3

From previous records, the hospital knows that the population standard deviation of children body temperature at check in is 4.1 degrees. Estimate the mean body temperature at check in for children with the flu with a 99% confidence interval.

Solution:

Case: Population standard deviation is known $\rightarrow$ Use a Z-Procedure

In R,

```r
flu <- c(100.2,98.6,99.4,102.0,99.3,97.1,   104.0,98.9, 96.7,   100.3)
flu
```

```
 [1] 100.2  98.6  99.4 102.0  99.3  97.1 104.0  98.9  96.7 100.3
```

```r
xbar = mean(flu)
xbar
```

```
[1] 99.65
```

```r
sd = 4.1
z = qnorm(1-0.01/2)
n = 10

xbar-z*sd/sqrt(n)
```

```
[1] 96.31035
```

```r
xbar+z*sd/sqrt(n)
```

```
[1] 102.9896
```

Answer: The 99% confidence interval is $(96.31, 102.99)$.

Interpretation: We are 99% confident that the mean body temperature for children with the flu at check in is between 96.31 and 102.99 degrees.

8.2.2.1.2 Hypothesis Tests for Means - Z-Procedures

Idea: Use the sample statistic $\bar{x}$ to evaluate the null hypothesis and reach a conclusion.

In these cases,

1. Type of test

   It can be:

   - Left tailed test
   - Right tailed test
   - Two-tailed test

2. Claims

   - Null Hypothesis: $H_0 : \mu = \mu_0$, Where μ_0 is the hypothesized mean.
   - Alternative Hypothesis: It must be one of the following:
     - $H_a : \mu < \mu_0$
     - $H_a : \mu > \mu_0$
     - $H_a : \mu \neq \mu_0$

3. The Test Statistic

   Test Statistic: A value from the Normal Distribution that depends on the sample mean, the population standard deviation, and the sample size. The test statistic formula is:

$$z = \frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}}$$

4. The P-Value

   The p-value can be calculated with R, and it depends on the alternative hypothesis and the test statistic. The codes to calculate p-values depending on the alternative are given by the following table.

$H_a : \mu < \mu_0$	$H_a : \mu > \mu_0$	$H_a : \mu \neq \mu_0$
pnorm(*test statistic*)	1-pnorm(*test statistic*)	2×(1-pnorm(*test statistic*))

5. Decision

6. Conclusion

Example:

A certain institution records from previous years suggest that the average completion time of their Spanish exam is 87.4 minutes with a standard deviation of 17.2 minutes. A researcher believes the average completion time is greater than that so he took a random sample of 16 students' exams time from this month and found a mean value of 96.9 minutes. Assume that the time to complete the Spanish exam follows a Normal Distribution, and conduct a hypothesis test to test whether the researcher's claim at the 90% confidence level.

Solution:

Type of test: right-tailed test(indicated by the word "greater")

$H_0 : \mu = 87.4$

$H_a : \mu > 87.4$ $\bar{x} = 96.9$

$\mu_0 = 87.4$

$\sigma = 17.2$

$n = 16$

Test statistic:

$z = \frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{96.9-87.4}{\frac{17.2}{\sqrt{16}}}$

In R,

```
#Test statistic:
tstat = (96.9-87.4)/(17.2/sqrt(16))
tstat
```

```
[1] 2.209302
```

```
#P-value
(1-pnorm(tstat))
```

```
[1] 0.01357681
```

Decision: Reject H_0.

Conclusion: Because the p-value is less than the significance level (0.10), we decide to reject the null and conclude that there is enough evidence to suggest that the true mean solving time for the Spanish exam of this institution is greater than 87.4 minutes at the 0.1 significance level.

8.2.2.2 Case 2: T-Procedures for Inference for Means

- Confidence Intervals
- Hypothesis Tests

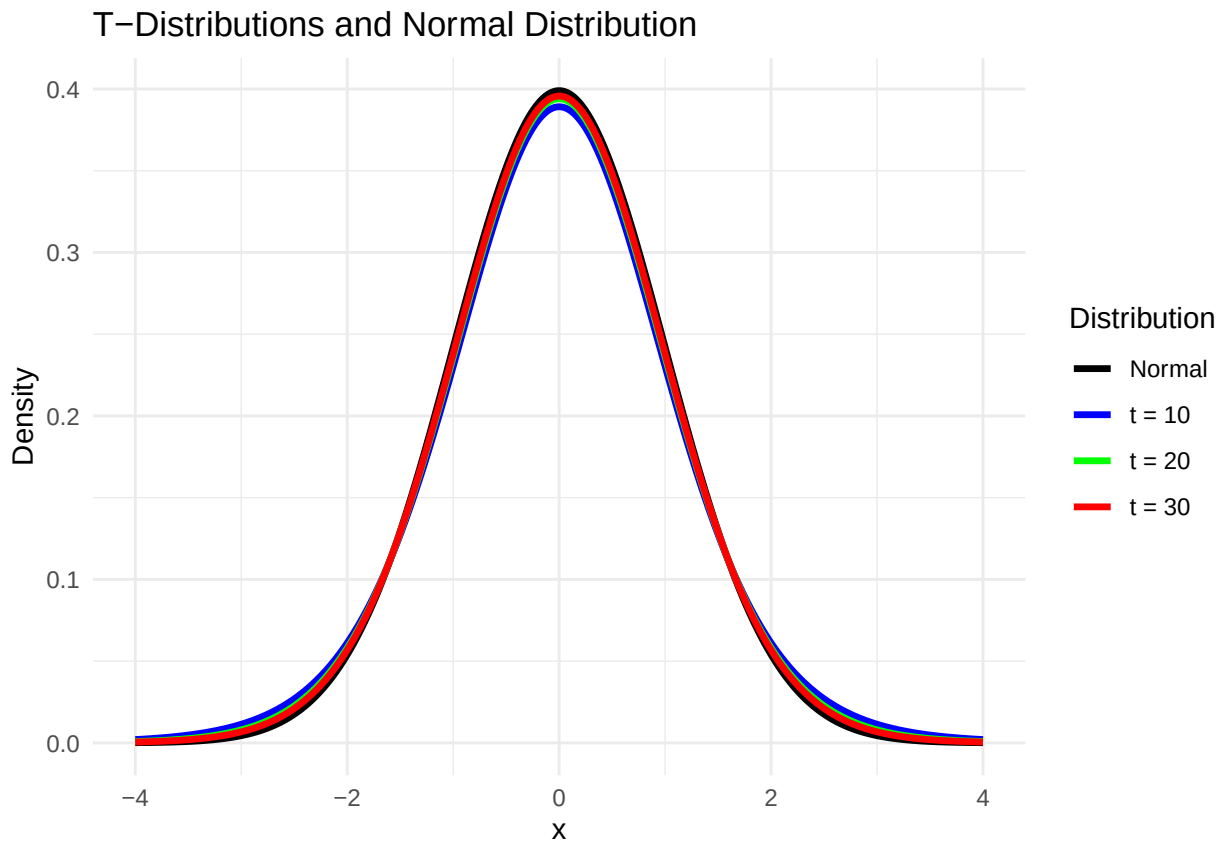
The Student T-Distribution: The Student T-Distribution is an approximation of the Normal Distribution that is used when the population standard deviation is unknown, which is the case in most real-world situations. It has heavier tails than the Normal Distribution, so it is a little bit more conservative than it. As the sample size becomes larger, the T-Distribution approaches the Normal Distribution.

The T-Distribution is determined by its **degrees of freedom (df)**. For the one sample procedures the degrees of freedom are determined by the sample size minus one: $n - 1$.

Degrees of freedom: The number or logically independent values from the sample.

Visualization: The following graph is a representation of T-Distributions with 10, 20 and 30 degrees of freedom.

```
Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use ‘linewidth‘ instead.
This warning is displayed once every 8 hours.
Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
generated.
```



We trust T-Procedures when:

- n<40 and the Sampling Distribution is somewhat symmetric with no outliers.

- n>40.
- σ is unknown and it is estimated by the sample standard deviation s.

8.2.2.2.1 Confidence Intervals for Means - T-Procedures

Idea: Use the sample statistic $\bar{x}$ to find a range of values to which the true population mean may belong to.

Formula:

$$\bar{x} \pm t_{\alpha/2,df} \frac{s}{\sqrt{n}}$$

where: - $\bar{x}$ is the sample mean
- s is the sample standard deviation
- $t_{\alpha/2,df}$ is the critical t value
- α is the significance level
- n is the sample size

Note: $SE_{\bar{x}} = \frac{s}{\sqrt{n}}$ is an approximation of the population standard deviation, and it is called standard error of the sample mean.

Example:

A special education center is interested in the mean solving time of 9-piece puzzles for children with down syndrome. 8 children with down syndrome were asked to solve a 9-piece puzzle and the mean solving time was found to be 208 seconds with a standard deviation of 72 seconds. Estimate the true mean solving time of 9-piece puzzles for children with down syndrome with a 95% confidence interval.

Solution:

Case: Population standard deviation is unknown → Use a T-Procedure

In R,

```
xbar = 208
s = 72
alpha=0.05
n = 8
t = qt(1-alpha/2,n-1)

# Lower Bound
xbar-t*s/sqrt(n)
```

```
[1] 147.8065
```

```
# Upper Bound
xbar+t*s/sqrt(n)
```

```
[1] 268.1935
```

Answer: The 95% confidence interval is $(147.80, 268.20)$.

Interpretation: We are 95% confident that the true mean solving time for 9-piece puzzles for children with down syndrome is between 147.80 and 268.20 seconds.

8.2.2.2.2 Hypothesis Tests for Means - T-Procedures

Idea: Use the sample statistic $\bar{x}$ to evaluate the null hypothesis and reach a conclusion.

In these cases,

1. Type of test

   It can be:

   - Left tailed test
   - Right tailed test
   - Two-tailed test

2. Claims

   - Null Hypothesis: $H_0 : \mu = \mu_0$, where μ_0 is the hypothesized mean.
   - Alternative Hypothesis: It must be one of the following:
     - $H_a : \mu < \mu_0$
     - $H_a : \mu > \mu_0$
     - $H_a : \mu \neq \mu_0$

3. The Test Statistic

   Test Statistic: A value from the Student T-Distribution that depends on the sample mean, the population standard deviation, and the sample size. The test statistic formula is:

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

4. The P-Value

   The p-value can be calculated with R, and it depends on the alternative hypothesis, the test statistic and the degrees of freedom. The codes to calculate p-values depending on the alternative are given by the following table.

$H_a : \mu < \mu_0$	$H_a : \mu > \mu_0$	$H_a : \mu \neq \mu_0$
pt(*test statistic, df*)	1-pt(*test statistic, df*)	2×(1-pt(*test statistic, df*))

5. Decision

6. Conclusion

Example:

A study suggests that the mean weight of adult hummingbirds that inhabit in a particular south American jungle is 4.6 grams. A researcher believes their mean weight is higher, so to test this claim, he recorded the weights of 12 adult hummingbirds from this South American jungle. His data is summarized in the table below.

4.6	3.8	4.7	4.5	5.0	3.9	3.9	4.2	4.4	4.8	5.0	3.7

Conduct a hypothesis test at the 0.01 significance level to test this claim.

Solution:

Type of test: right-tailed (indicated by "lower")

$H_0 : \mu = 4.6$

$H_a : \mu < 4.6$

Continuing in R,

```r
grams<-c(4.6,3.8,4.7,4.5,5.0,3.9,3.9,4.2,4.4,4.8,5.0,3.7)
xbar = mean(grams)
xbar
```

```
[1] 4.375
```

```r
s = sd(grams)
mu = 4.6
n = length(grams)
alpha=0.01
qt(1-alpha/2, n-1)
```

```
[1] 3.105807
```

```r
# Test Statistic:
tstat = (xbar-mu)/(s/sqrt(n))
tstat
```

```
[1] -1.667776
```

```r
# P-value:
p_value=pt(tstat, n-1)
p_value
```

```
[1] 0.06177253
```

```r
# or

t.test(grams,mu=4.6,conf.level = .90, alt="less")
```

```
    One Sample t-test

data:  grams
t = -1.6678, df = 11, p-value = 0.06177
alternative hypothesis: true mean is less than 4.6
90 percent confidence interval:
     -Inf 4.558941
sample estimates:
mean of x
    4.375
```

Decision: Fail to Reject H_0.

Conclusion: Because the p-value is greater than the level of significance, there is not enough evidence to suggest that the true mean weight of adult hummingbirds is lower than 4.6 grams, at the 0.01 significance level.

8.3 Two-Sample Statistical Inference

Here we will explore Inference Methods for the comparisons of two samples.

8.3.1 Difference of Proportions

Idea: Compare the proportions of two different samples (often distinguished by a categorical variable) by taking the difference between their sample proportions. The responses of the variable will be classified in the dichotomous situation of {Success,Failure}.

Parameter of interest: $p_1 - p_2$.

Sample statistic to be used: $\hat{p}_1 - \hat{p}_2$

Conditions for Inference:

- We work with two simple random samples from the two independent population of interest (one sample of each population).
- The sample size is sufficiently large (at least 10 successes and 10 failures).
- The Sampling Distribution is approximately Normal with mean $p_1 - p_2$ and standard deviation $\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

The differences in proportions will fall under the heading of a Large-Sample Z-Procedure. That is, we will work with the Standard Normal Distribution.

8.3.1.1 Confidence Intervals for Difference in Proportions

Idea: Use the sample statistic $\hat{p}_1 - \hat{p}_2$ to find a range of values to which the true difference of the population proportions may belong to.

Sampling Distribution:

When the sample size is sufficiently large, then the Sampling Distribution of the Difference of Sample Proportions is approximately described by

$$\hat{p}_1 - \hat{p}_2 \sim N\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$$

Formula:

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

where: - $\hat{p}_1$ is the sample proportion for sample 1
- $\hat{p}_2$ is the sample proportion for sample 2
- n_1 is the sample size for sample 1
- n_2 is the sample size for sample 2
- α is the significance level. $(100 - confidence)/100 = \alpha$
- $z_{\alpha/2}$ is the critical z value

Note: $SE_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$ is an approximation of the population standard deviation, and it is called standard error of the sample proportion.

Example:

A researcher is comparing the effect of a new acne medication on teenagers. A random sample of 1000 teenagers that were searching for medication for acne relief were given the new medication, while a random sample of 500 teenagers that were searching for acne relief were given a placebo. Of the teenagers given the new medication, 768 saw relief from acne in one month. Of the teenagers that were given the placebo, 210

saw relief from acne after a month. The researcher is interested in obtaining a 95% Large Sample Confidence Interval on the differences in the proportions of between the new medication and the placebo.

Solution:

$\hat{p}_1 = \frac{768}{1000} = 0.768$

$\hat{p}_2 = \frac{210}{500} = 0.420$

$n_1 = 1000$

$n_2 = 500$

$\alpha = 0.05$

$z_{\alpha/2} = 1.96$

Confidence Interval:

$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} = (0.768 - .420) \pm (1.96)\sqrt{\frac{0.768(1-0.768)}{1000} + \frac{0.420(1-0.420)}{500}}\backslash$

In R,

```
# Lower Bound
(0.768-0.420)-(1.96)*sqrt(((0.768*(1-0.768))/1000)+(0.420*(1-0.420))/500)
```

```
[1] 0.297442
```

```
# Upper Bound
(0.768-0.420)+(1.96)*sqrt(((0.768*(1-0.768))/1000)+(0.420*(1-0.420))/500)
```

```
[1] 0.398558
```

Answer: The 95% confidence interval is $(0.30, 0.40)$.

Interpretation: We are 95% confident that true difference in the proportion of the new medication and the placebo in the relief given to teenagers is between 0.30 and 0.40.

8.3.1.2 Test of Hypothesis for Difference in Proportions

Idea: Use the sample statistic $\hat{p}_1 - \hat{p}_2$ to evaluate the null hypothesis and reach a conclusion.

When the sample size is sufficiently large, then the Sampling Distribution of the Difference of Sample Proportions is approximately described by

$$\hat{p}_1 - \hat{p}_2 \sim N\left((p_1 - p_2)_0, \sqrt{\bar{p}(1-\bar{p})(\frac{1}{n_1} + \frac{1}{n_2})}\right)$$

In these cases,

1. Type of test

   It can be:

   - Left tailed test
   - Right tailed test
   - Two-tailed test

2. Claims

- Null Hypothesis: $H_0 = p_1 - p_2 = (p_1 - p_2)_0$, where $(p_1 - p_2)_0$ is the hypothesized difference in proportions. It is typically 0, unless stated otherwise.
- Alternative Hypothesis: It must be one of the following:
  - $H_a = p_1 - p_2 < (p_1 - p_2)_0$
  - $H_a = p_1 - p_2 > (p_1 - p_2)_0$
  - $H_a = p_1 - p_2 \neq (p_1 - p_2)_0$

3. The Test Statistic

   Under the assumption that the null is true, we create a pooled proportion to be used in the calculation of the test statistic.

   $$\bar{p} = \frac{\#successes_1 + \#successes_2}{n_1 + n_2}$$

   This will yield a value from the Normal Distribution that depends on the sample difference in proportions, the claimed difference, and the sample sizes. The test statistic formula is:

   $$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)_0}{\sqrt{\bar{p}(1-\bar{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$$

4. The P-Value

   The p-value can be calculated with R, and it depends on the alternative hypothesis and the test statistic. The codes to calculate p-values depending on the alternative are given by the following table.

$H_a = p_1 - p_2 < (p_1 - p_2)_0$	$H_a = p_1 - p_2 > (p_1 - p_2)_0$	$H_a = p_1 - p_2 \neq (p_1 - p_2)_0$
pnorm(*test statistic*)	1-pnorm(*test statistic*)	2×(1-pnorm(*test statistic*))

5. Decision

6. Conclusion

Example:

Consider the example above concerning the comparison of the new medication and the placebo for treating acne. Do we have evidence that the new medication will increase the relief from acne over the placebo? Use a level of significance $\alpha = 0.05$.

Solution:

Type of test: right-tailed (indicated by the word "increase")

$H_0 = p_1 - p_2 = 0$ (there is no difference)

$H_a = p_1 - p_2 > 0$

$\hat{p}_1 = \frac{768}{1000} = 0.768$

$\hat{p}_2 = \frac{210}{500} = 0.420$

$n_1 = 1000$

$n_2 = 500$

$\alpha = 0.05$

$(p_1 - p_2)_0 = 0$

Continuing in R,

```r
# Pooled Proportion
pooledp<-(710+210)/(1000+500)
pooledp
```

```
[1] 0.6133333
```

```r
# Test Statistic
tstat<-((0.768-0.420)-0)/sqrt((pooledp*(1-pooledp)*(1/1000+1/500)))
tstat
```

```
[1] 13.04674
```

```r
# P-value
pvalue<-1-pnorm(tstat)
pvalue
```

```
[1] 0
```

```r
# or

phat1<-768/1000
phat2<-210/500
n1<-1000
n2<-500
alpha<-0.05
p_0<-0

# Pooled Proportion
pooledp<-(710+210)/(n1+n2)
pooledp
```

```
[1] 0.6133333
```

```r
# Test Statistic
tstat<-((phat1-phat2)-p_0)/sqrt((pooledp*(1-pooledp)*(1/n1+1/n2)))
tstat
```

```
[1] 13.04674
```

```r
# P-value
pvalue<-1-pnorm(tstat)
pvalue
```

```
[1] 0
```

Decision: Reject H_0.

Conclusion: Because the p-value is less than the significance level, there is enough evidence to suggest that the new medication will increase the relief from acne over the placebo.

8.3.2 Difference of Means

Differences of means is used when we want to compare the same attribute (quantitative variable) of two independent populations.

Parameter of interest: $\mu_1 - \mu_2$. That is, the differences between the two means of the population

Statistic to be used: $\bar{x}_1 - \bar{x}_2$. That is, the difference of the sample means.

Conditions for Inference:

- We work with two simple random samples, one from each population.
- The distributions of the samples have similar shapes.

When the sample size is sufficiently large, then the Sampling Distribution of the Difference of Two Sample Means is approximated by

$$\bar{x}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

The degrees of freedom for these T-Procedures will be:

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

Note: These procedures are software intensive and hand calculation should not be performed.

8.3.2.1 Confidence Intervals for Differences of Means

Idea: Use the sample statistic $\bar{x}_1 - \bar{x}_2$ to find a range of values to which the true mean difference between the populations may belong to.

Formula:

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

where: - $\bar{x}_1$ is the sample mean for sample 1
- $\bar{x}_2$ is the sample mean for sample 2
- s_1^2 is the sample variance for sample 1
- s_2^2 is the sample variance for sample 2
- n_1 is the sample size of sample 1
- n_2 is the sample size of sample 2
- α is the significance level
- $t_{\alpha/2}$ is the critical t value

Note: $SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ is the standard error, used as an estimate of the standard deviation of the Sampling Distribution of the Differences of Two-Sample Means.

Example:

Researchers are interested in a new method for teaching a History class to students. The new method consists of virtual reality instruction, where each student is involved with a virtual tour of an historic battle site. The researches took 12 students and randomly divided them into two groups of six. The first group of students works with the virtual reality instruction while the second group reads about the historic battle site with the conventional textbook approach. A common test was giving to both groups and the results, out of 100 points, are given below. Construct a 90% Confidence Interval for the true difference of the means between the two groups.

Virtual Reality Test Score	95	87	90	91	85	85
Textbook Test Score	90	76	98	60	80	84

Solution:

In R,

```
Virtual<-c(95,87, 90, 91, 85, 85)
Textbook<-c(90,76,98,60,80,84)
diff<-mean(Virtual)-mean(Textbook)

# Calculation of the Degrees of Freedom
df=(var(Virtual)/length(Virtual)+var(Textbook)/length(Textbook))^2/
  ((1/(length(Virtual)-1))*(var(Virtual)/length(Virtual))^2+
     (1/(length(Textbook)-1))*(var(Textbook)/length(Textbook))^2)
df
```

```
[1] 5.901464
```

```
# Standard Error of the Difference of the Two Sample Means
SE<-sqrt(var(Virtual)/length(Virtual)+var(Textbook)/length(Textbook))
SE
```

```
[1] 5.544266
```

```
# Critical T Value

alpha=0.1
t<-qt(1-alpha/2,5.9015)
t
```

```
[1] 1.949004
```

```
# Confidence Interval

diff-t*SE
```

```
[1] -3.305798
```

```
diff+t*SE
```

```
[1] 18.3058
```

```
# or

test<-t.test(Virtual,Textbook,conf.level=0.90)
test$conf.int
```

```
[1] -3.30581 18.30581
attr(,"conf.level")
[1] 0.9
```

Answer: The 90% confidence interval is $(-3.306, 18.306)$

Interpretation: We are 90% Confident, that the true difference in the means of the test scores for the Virtual Reality Instruction and the Classic Textbook Instruction is within the calculated interval. In another words, it is estimated that students instructed with virtual reality will score between 18.31 units more and 3.31 units less than students instructed with a classical textbook.

8.3.2.2 Hypothesis Test For Difference of Means

Idea: Use the sample statistic $\bar{x}_1 - \bar{x}_2$ to evaluate the null hypothesis and reach a conclusion.

In this case,

1. Type of test

   It can be:

   - Left tailed test
   - Right tailed test
   - Two-tailed test

2. Claims

   - Null Hypothesis: $H_0 = \mu_1 - \mu_2 = (\mu_1 - \mu_2)_0$, where $(\mu_1 - \mu_2)_0$ is the hypothesized difference of the means and is typically zero unless stated otherwise.
   - Alternative Hypothesis: It must be one of the following:
     - $H_a : \mu_1 - \mu_2 < (\mu_1 - \mu_2)_0$
     - $H_a : \mu_1 - \mu_2 > (\mu_1 - \mu_2)_0$
     - $H_a : \mu_1 - \mu_2 \neq (\mu_1 - \mu_2)_0$

3. The Test Statistic

   Test Statistic: A value from the T-Distribution that depends on the hypothesized mean, the sample statistic, the sample variances and the sample sizes. The test statistic formula is:

   $$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

4. The P-Value

   The p-value can be calculated with R, and it depends on the alternative hypothesis, the test statistic and the degrees of freedom. The codes to calculate p-values depending on the alternative are given by the following table.

$H_a : \mu_1 - \mu_2 < (\mu_1 - \mu_2)_0$	$H_a : \mu_1 - \mu_2 > (\mu_1 - \mu_2)_0$	$H_a : \mu_1 - \mu_2 \neq (\mu_1 - \mu_2)_0$
pt(*test statistic, df*)	1-pt(*test statistic, df*)	2×(1-pt(*test statistic, df*))

5. Decision

6. Conclusion

Example:

Consider the above example regarding virtual reality instruction vs classical textbook instruction. Do you have significant evidence to say that the methods differ?

Solution:

Type of test: two-tailed (indicated by "lower")

$H_0 = \mu_{Virtual} - \mu_{Textbook} = 0$

$H_a = \mu_{Virtual} - \mu_{Textbook} \neq 0$

Continuing in R,

```
Virtual<-c(95,87, 90, 91, 85, 85)
Textbook<-c(90,76,98,60,80,84)
diff<-mean(Virtual)-mean(Textbook)

n1 <- length(Virtual)
n2 <- length(Textbook)
var1 <- var(Virtual)
var2 <- var(Textbook)

# Degrees of Freedom
df <- (var1/n1+var2/n2)^2/((1/(n1-1))*(var1/n1)^2+(1/(n2-1))*(var2/n2)^2)
df
```

```
[1] 5.901464
```

```
# Test Statistic
tstat <- (diff-0)/sqrt((var1/n1)+(var2/n2))
tstat
```

```
[1] 1.352749
```

```
# P-Value
pvalue <- 2*(1-pt(tstat,df))
pvalue
```

```
[1] 0.2256616
```

```
# or
t.test(Virtual,Textbook,mu=0,conf.level = 0.90, alternative = "two.sided")
```

```
    Welch Two Sample t-test

data:  Virtual and Textbook
t = 1.3527, df = 5.9015, p-value = 0.2257
alternative hypothesis: true difference in means is not equal to 0
90 percent confidence interval:
 -3.30581 18.30581
sample estimates:
mean of x mean of y
 88.83333  81.33333
```

Decision: Fail to Reject H_0..

Conclusion: We do not have sufficient evidence to say that the difference in the test scores for the virtual reality instruction is different from that of the classical instruction with the textbook.

8.3.3 Matched Pairs

Matched pairs procedures consist of only two treatments being applied to either the same individual or sets of individuals as similar as possible. A one sample T-Procedure will be conducted on the differences of the pairs of the treatments conducted on the subjects.

Parameter of interest:μ_{diff}

Statistic to be used: $\bar{x}_{diff}$.

Conditions for Inference:

- We work with a simple random sample from the population of interest.
- The population of interest is Normally distributed with mean μ and standard deviation σ.
- We assume the population standard deviation is unknown and we use a T-Procedure.

8.3.3.1 Confidence Intervals for Matched Pairs

Idea: Use the sample statistic $\bar{x}_{diff}$ to find a range of values to which the true difference of the population proportions may belong to.

Formula:

$$\bar{x}_{diff} \pm t_{\alpha/2}\frac{s_{diff}}{\sqrt{n}}$$

where

$\bar{x}_{diff}$ is the mean of the differences between the treatments for the samples

s_{diff} is the standard deviation of the differences

n is the sample size

α is the level of significance

$t_{\alpha/2}$ is the critical t value

Note: $\frac{s_{diff}}{\sqrt{n}} = SE_{\bar{x}_{diff}}$ is the standard error of the mean of the differences and it is used as an estimate of the standard deviation on the Sampling Distribution.

Example:

A weight loss clinic is touting the results of it treatment as saying that "Individuals experience a loss in weight after a month of being on the new treatment." They would be interested in calculating a 95% Confidence Interval for the true mean difference in weight loss after the month on the new treatment. A group of 6 subjects were used in the study and their weights (in lbs.) prior to the treatment and after the month on the treatment are given below.

	Subject 1	**Subject 2**	**Subject 3**	**Subject 4**	**Subject 5**	**Subject 6**
Before Treatment	243	279	210	195	213	254
After Treatment	237	275	208	199	214	250
Difference(Before-After)	6	4	2	-4	-1	4

Solution:

In R,

```r
Before<-c(243, 279,210,195,213,254)
After<-c( 237,275,208, 199,214,250)
Differences<-Before-After

xbar<-mean(Differences)
s<-sd(Differences)
n<-length(Differences)
alpha<-0.05
t<- qt(0.975,5)

# Lower Bound
xbar-t*s/sqrt(n)
```

```
[1] -2.060436
```

```r
# Upper Bound
xbar+t*s/sqrt(n)
```

```
[1] 5.727103
```

```r
# or

test<-t.test(Differences,conf.level=0.95)
test$conf.int
```

```
[1] -2.060436  5.727103
attr(,"conf.level")
[1] 0.95
```

```r
# or
testpaired<-t.test(Before,After,mu=0,conf.level=0.95,paired=T)
testpaired$conf.int
```

```
[1] -2.060436  5.727103
attr(,"conf.level")
[1] 0.95
```

Answer: The 95% Confidence Interval is $(-2.06, 5.73)$

Interpretation: We are 95% confident, that the true mean difference in weight after a month on the new treatment is 5.73 lbs less and 2.06 lbs more than the initial weight.

8.3.3.2 Test of Hypothesis for Matched Pairs

Idea: Use the sample statistic $\bar{x}_{diff}$ to evaluate the null hypothesis and reach a conclusion.

In this case,

1. Type of test

   It can be:

   - Left tailed test

- Right tailed test
- Two-tailed test

2. Claims

   - Null Hypothesis: $H_0 = \mu_{diff} = \mu_{(diff)0}$, where $\mu_{(diff)0}$ is the hypothesized difference. It is typically 0, unless stated otherwise.
   - Alternative Hypothesis: It must be one of the following:
     - $H_0 = \mu_{diff} < \mu_{(diff)0}$
     - $H_0 = \mu_{diff} > \mu_{(diff)0}$
     - $H_0 = \mu_{diff} \neq \mu_{(diff)0}$

3. The Test Statistic

   Test Statistic: A value from the T-Distribution that depends on the hypothesized mean of the differences, the mean and standard deviation of the sample differences and the sample size. The test statistic formula is:

$$t = \frac{\bar{x}_{diff} - \mu_{(diff)0}}{\frac{s_{diff}}{\sqrt{n}}}$$

4. The P-Value

   The p-value can be calculated with R, and it depends on the alternative hypothesis, the test statistic and the degrees of freedom. The codes to calculate p-values depending on the alternative are given by the following table.

$H_0 = \mu_{diff} < \mu_{(diff)0}$	$H_0 = \mu_{diff} > \mu_{(diff)0}$	$H_0 = \mu_{diff} \neq \mu_{(diff)0}$
pt(*test statistic, df*)	1-pt(*test statistic, df*)	2×(1-pt(*test statistic, df*))

5. Decision

6. Conclusion

Example:

Consider the example above concerning the the treatment for weight loss. Does the clinic have evidence to say, that on average, people will lose weight after a month on the new treatment at the $\alpha = 0.05$ significance level?

Solution:

Type of test: right-tailed (indicated by the word "lose", this indicates the difference will be greater than 0).

$H_0 = \mu_{diff} = 0$

$H_a = \mu_{diff} > 0$

Continuing in R,

```
Before<-c(243, 279,210,195,213,254)
After<-c( 237,275,208, 199,214,250)
Differences<-Before-After

xbar<-mean(Differences)
s<-sd(Differences)
n<-length(Differences)
alpha<-0.05
```

```r
t<- qt(0.975,5)

# Test Statistic
tstat = (1.83-0)/(s/sqrt(n))

#P-Value
pvalue = 1-pt(tstat, n-1)
pvalue
```

```
[1] 0.1405084
```

```r
# or
t.test(Differences,mu=0,alternative = "greater",conf.level = 0.95)
```

```
    One Sample t-test

data:  Differences
t = 1.2103, df = 5, p-value = 0.1401
alternative hypothesis: true mean is greater than 0
95 percent confidence interval:
 -1.218946       Inf
sample estimates:
mean of x
 1.833333
```

```r
# or
t.test(Before,After,mu=0,alternative = "greater",conf.level = 0.95,paired = T)
```

```
    Paired t-test

data:  Before and After
t = 1.2103, df = 5, p-value = 0.1401
alternative hypothesis: true mean difference is greater than 0
95 percent confidence interval:
 -1.218946       Inf
sample estimates:
mean difference
       1.833333
```

Decision: Fail to reject the H_0.

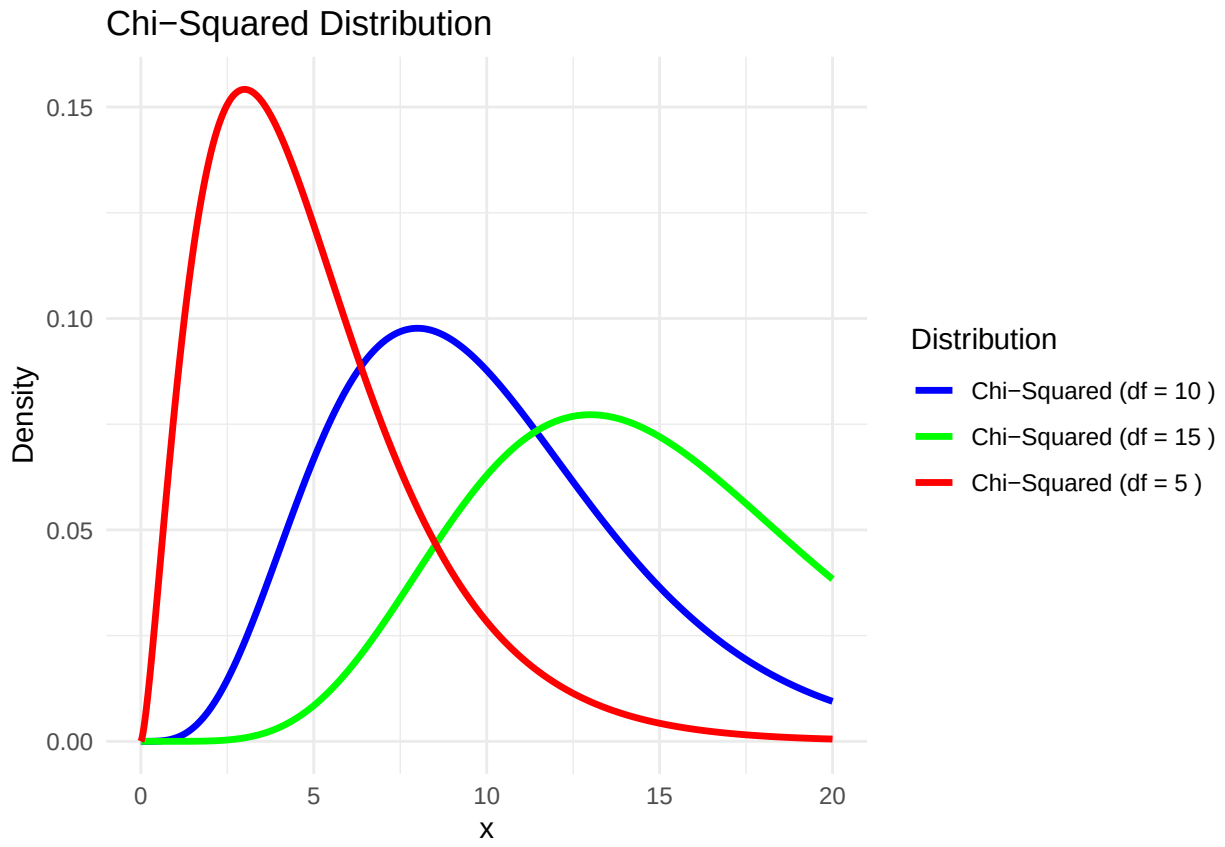
Conclusion: At the 0.05 significance level, there is insufficient evidence to support that, on average, individuals lose weight after a month on the new treatment.

8.4 Chi-Squared Tests

8.4.1 Chi-Squared Tests of Hypothesis

The Chi-Squared distribution (χ^2): It is a right skewed distribution dependent on its degrees of freedom. Its mean is equal to its degrees of freedom.

Visualization: The following graph is a representation of Chi-Squared Distributions with 5, 10 and 15 degrees of freedom.



Chi-Squared Tests: These are tests for multiple comparisons. There are three common types of hypothesis using the Chi-Squared Distribution, and they respond to questions such as the following:

- Independence Test – is there any relationship between two variables? (One SRS considered).
- Goodness of Fit Test – is a particular probability model good? (One variable considered).
- Test of Homogeneity – Is there a similar behavior with respect to a categorical variable? (Multiple SRS considered).

General idea for all Chi-squared Tests: Compare the observed to the expected counts under that assumption that a null hypothesis is true.

Expected counts formula:

$$n \times p$$

where: - n is the sample size.
- p is the probability under the null hypothesis.

Conditions:

- Each cell has an expected count of at least 5.

In this booklet, we will only learn about the Chi-Squared test for independence.

8.4.1.1 Independence Test

Idea: To evaluate whether the observed frequencies deviate significantly from the expected frequencies, suggesting evidence for a relationship between the variables.

We will work with two way tables for the Independence test. We will start with an observed counts table, and as part of the process of calculating the test statistic we will construct the expected table.

1. Type of test
   - It is a two-tailed test.
2. Claims
   - Null Hypothesis: There is no relationship between the two variables.
   - Alternative Hypothesis: There is a relationship between the two variables.
3. The Test Statistic

   To obtain the test statistic we need to get the expected table.

   To get expectations we use:

   $$expected = \frac{(row\ total)(column\ total)}{table\ total}$$

   Each intersection has a contribution to the test statistic that is obtained working with both observed and expected counts. The test statistic will be the sum of all contributions.

   To get individual contributions we use:

   $$contribution = \frac{((Observed\ count)_{cell} - (Expected\ count)_{cell})^2}{(Expected\ count)_{cell}}$$

   The full Test Statistic is:

   $$X^2 = \Sigma_{over\ all\ cells} \left(\frac{((Observed\ count)_{cell} - (Expected\ count)_{cell})^2}{(Expected\ count)_{cell}} \right)$$

4. The P-Value

   The p-value can be calculated with R using:

   1-pchisq({textittest statistic, degrees of freedom)

   The degrees of freedom are $df = (\#rows - 1)(\#columns - 1)$.

5. Decision
6. Conclusion

Example:

A counselor at a local high school, took a random sample of 500 high school students and administered a survey. Two of the questions on the survey were “How do you feel about the school’s cafeteria food?” (Like, Dislike, No opinion) and “Do you ride the bus home from school?”. The counselor is interested in conducting a test of independence on the two variables (cafeteria food and ride home}. The table below represents the summarized counts.

	Bus	Not Bus	Total
Like	170	30	200
Dislike	130	70	200
No opinion	50	50	100
Total	350	150	500

Solution:

Type of test: Independence test.

H_0 =There is no relationship between the two variables (independence).

Ha =There is a relationship between the two variables.

The degrees of freedom is $df = 2 \times 1 = 2$.

A table of the expected counts is given by:

	Bus	Not Bus	Total
Like	$\frac{350 \times 200}{500} = 140$	60	200
Dislike	140	60	200
No opinion	70	30	100
Total	350	150	500

A table of each cell contribution to the χ^2-distribution's test statistic X^2 is given by:

	Bus	Not Bus
Like	$\frac{170-140}{140} = 6.43$	15
Dislike	0.714	1.667
No opinion	5.714	13.333

The value of the test statistic is the sum of each cells contribution, thus

$X^2 = \Sigma_{over\ all\ cells} contributions = 6.43 + 15.00 + 0.71 + 1.67 + 5.71 + 13.33 = 42.86$

Continuing in R,

```
# P-Value
1-pchisq(42.858,2)
```

```
[1] 4.937459e-10
```

Alternatively, in R, we can also perform

```
Ride<-c(rep("Bus",170),rep("Bus",130),rep("Bus",50),
                rep("Not Bus",30),rep("Not Bus",70)
        ,rep("Not Bus",50))
Cafeteria<-c(rep("Like",170),rep("Dislike",130),
rep("No Opinion",50),rep("Like",30),rep("Dislike",70),
rep("No Opinion",50))
tab<-table(Cafeteria,Ride)
chisq.test(tab)
```

```
    Pearson's Chi-squared test

data:  tab
X-squared = 42.857, df = 2, p-value = 4.94e-10
```

Decision: Reject H_0.

Conclusion: Because the p-value is smaller than the significance level, there is enough evidence to support the idea that there is a relationship between the two variables.

Exercises

1. The final round of a reality show is taking place tomorrow and the vote of the audience will define the winner. A random sample of 50 people was taken, and 54% of them stated they will vote for participant A. Construct a 95% confidence interval for the proportion of people who will vote for participant A in the final tomorrow.

2. An online clothing store conducted a satisfaction survey aiming to find out if over 80% of their customers are satisfied with their online shopping experience. They took a sample of 100 customers and found that 85% of the surveyed customers stated satisfaction. Conduct a hypothesis test to see if there exists significant evidence that the proportion of satisfied customers is higher than 0.80 at the 0.05 significance level.

3. The owner of a bakery is interested in the average check out time per client. To investigate this, he took a random sample of 36 clients and found a mean check out time of 1.5 minutes. It is known that the standard deviation is 0.4 minutes. Construct a 95% confidence interval for the average check out time at this bakery.

4. A computer systems engineer has created a new algorithm and he is interested in the mean running time. He ran the algorithm 81 times and found a mean of 14.2 seconds with a standard deviation of 1 second. Construct a 90% confidence interval for the running time of this algorithm.

5. Using the same information as in question 4, is there enough statistical evidence to suggest that the mean running time of this algorithm is less than 14.5 seconds at the 0.10 significance level?

6. A wine manufacturer has raised concerns about the volume of wine in their 1000 ml wine bottles. To investigate, they randomly sample 25 wine bottles and find that the mean volume is 999.6 ml. The standard deviation is known to be 0.6 ml. Test at the 0.01 significance level if the mean volume of wine per bottle is different than 1000 ml.

7. A researcher is interested in comparing the proportion of people that commute to work by driving in Atlanta, GA to the proportion of people that commute to work by driving in Lexington, Ky. A random sample of 542 adults were selected in Atlanta with 312 claiming to commute to work by driving. A separate random sample of 630 adults in Lexington is selected, of which 415 claim to commute to work by driving. The researcher would like to perform a test of hypothesis using a level of significance of $\alpha = 0.05$, whether the proportion of adults commuting to work in Atlanta is less than the proportion that commute to work by driving for the adults of Lexington.

8. Researchers are interested in the thought that cardio exercise will more beneficial to weight loss than that of strength training. To this end a Researcher selected 10 participants at their local gym, that classified themselves as dieting to loss weight. Five of the participants had cardio as their primary exercise regimen, while the other five individuals declared strength training as their primary source of exercise. The weights (in lbs) of all 10 participants was measured and then remeasured after 4 weeks. The differences in the weights for each of the 10 participants is listed below. Do we have evidence that cardio is more beneficial to weight loss than that of strength training?

Cardio	5	7	-4	3	5	5
Strength	7	6	8	-1	8	4

9. It is claimed that students that take a particular standardized exam (out of 200 points) a second time generally score better than on their first attempt. Researches are wanting to construct a 95% Confidence Interval for the true mean of the difference in the first and second attempt. To this end, a random sample of 6 individuals that took the exam a second time are selected and the scores from the first attempt and the second attempt are given by the following table.

	Student 1	Student 2	Student 3	Student 4	Student 5	Student 6
First Attempt	129	140	109	157	128	138
Second Attempt	145	156	119	165	145	167‘

10. A survey was given to a large class of college students. The class is large and can be thought of as representing a random sample of all students at the university. Among the questions asked on the survey were "What is your school classification?" (Freshman, Sophomore, Junior, Senior), and "Do you have a positive or negative opinion on the university?". The table below summarizes the counts.

	Freshman	Sophomore	Junior	Senior	Total
Positive Opinion	400	110	125	94	729
Negative Opinion	95	102	225	99	521
Total	495	212	350	193	1250

Conduct a test of independence to see if there is evidence of a relationship between the variables *Classification* and *Opinion*.

Appendix - Basics of R

At its core, R can be thought of as calculator. This section will demonstrate some of the basic mathematical functions as well as some details concerning variable types and their creation in R.

Basic Arithmetic in R

The basic arithmetic functions in R are treated similar to those found in other calculators or software packages. Below are some examples.

```
# Addition

2+3
```

```
[1] 5
```

```
# Subtraction

5-2
```

```
[1] 3
```

```
# Multiplication

3*2
```

```
[1] 6
```

```
# Division

9/3
```

```
[1] 3
```

```
# Exponents

3**2
```

```
[1] 9
```

```
#or

3^2
```

```
[1] 9
```

```
# Square Roots

sqrt(4);
```

```
[1] 2
```

The exponential and logarithmic functions are demonstrated below.

```
# Exponential

exp(1)
```

```
[1] 2.718282
```

```
#Logarithmic

log(exp(1))
```

```
[1] 1
```

We are also allotted trigonometric functions in R that are evaluated in radians.

```
# Pi

pi
```

```
[1] 3.141593
```

```
# Sine

sin(pi/2)
```

```
[1] 1
```

```
# Cosine

cos(pi)
```

```
[1] -1
```

```
# Tangent

tan(pi/4)
```

```
[1] 1
```

```
# Inverse Sine

asin(1)
```

```
[1] 1.570796
```

```r
# Inverse Cosine

acos(-1)
```

```
[1] 3.141593
```

```r
# Inverse Tangent

atan(1)
```

```
[1] 0.7853982
```

Variable Creation

To assign names to variables we are limited to upper and lower case letters, dots, underscores and numbers. Variables names must start with a letter.

We can create a vector called x inside of R with the code x<-c($\cdots$), by separating numerical values within the vector with commas.

```r
# Creating a Numeric Variable

x<-c(1,2,3,4,5)
x
```

```
[1] 1 2 3 4 5
```

To create variables that are not numerical the process is similar but we write the observations between quotations marks. For example, suppose we wanted to create a list of five colors.

```r
# Creating a Categorical Variable

colors<-c("red","orange","yellow","green","blue")
colors
```

```
[1] "red"    "orange" "yellow" "green"  "blue"
```

We can also create sequences of numerical values with an arbitrary starting point, end point, and a step size.

```r
# Creating Sequences of Values

seq(1,5,1)
```

```
[1] 1 2 3 4 5
```

```r
seq(1,5,0.5)
```

```
[1] 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0
```

We can also repeat values by indicating the value that we want to repeat and the number of times we want it to be repeated.

```r
# Repeating Values

rep(0,5)
```

```
[1] 0 0 0 0 0
```

```r
rep(10,3)
```

```
[1] 10 10 10
```

```r
rep("pink",4)
```

```
[1] "pink" "pink" "pink" "pink"
```

Vector Arithmetic

In R, we can manipulate vectors to perform quick calculations or to do component wise arithmetic.

```r
# Creating the Vector

x<-c(2,3,4)
x
```

```
[1] 2 3 4
```

```r
# Performing Calculations

sum(x)
```

```
[1] 9
```

```r
prod(x)
```

```
[1] 24
```

```r
y<-2*x
y
```

```
[1] 4 6 8
```

```r
x+y
```

```
[1]  6  9 12
```

```r
x-y
```

```
[1] -2 -3 -4
```

```r
# Slicing The Array or Vector

x[1];x[2:3];x[c(1,3)];x[-3]
```

```
[1] 2
```

```
[1] 3 4
```

```
[1] 2 4
```

```
[1] 2 3
```

```r
y[y<8 & y>4]
```

```
[1] 6
```

Matrices and Data Frames

In R, we can create matrices and data frames. To create a matrix, we proceed similarly to that of vector creation, with the only difference being that we can generate rows and columns using the same data type such as character strings, numerical values, or logical statements. Data frames in R is merely the data displayed in a table format that do not require every entry to be of the same type.

```r
# Creating a Matrix

nrows=3
ncols=2
X<-matrix(seq(1,6,1),nrows,ncols)
X
```

```
     [,1] [,2]
[1,]    1    4
[2,]    2    5
[3,]    3    6
```

```r
# Checking to See if it is a Matrix

is.matrix(X)
```

```
[1] TRUE
```

With R, we can also slice or extract certain elements within our vectors, matrices, and data frames.

```r
# Selecting the First Column of X

X[,1]
```

```
[1] 1 2 3
```

```
# Selecting the First Row of X

X[1,]
```

```
[1] 1 4
```

```
# Selecting the 2 Row, 1 Column Element

X[2,1]
```

```
[1] 2
```

We can also transpose and calculate row and column sums along with with row and column means.

```
# Performing Matrix Calculations

t(X)
```

```
     [,1] [,2] [,3]
[1,]    1    2    3
[2,]    4    5    6
```

```
rowSums(X)
```

```
[1] 5 7 9
```

```
colSums(X)
```

```
[1]  6 15
```

```
rowMeans(X)
```

```
[1] 2.5 3.5 4.5
```

```
colMeans(X)
```

```
[1] 2 5
```

To create a data frame, we will use the example of a vector string from above and introduce some ages.

```
# Creating a Data Frame

Y<-data.frame(Classification,Age=c(rep(15,2),rep(16,3),rep(17,4),rep(18,2)))
Y
```

```
   Classification Age
1         Freshman  15
2         Freshman  15
3        Sophomore  16
4        Sophomore  16
5        Sophomore  16
6           Junior  17
7           Junior  17
8           Junior  17
9           Junior  17
10          Senior  18
11          Senior  18
```

```
# Check to see if it is Data Frame

is.data.frame(Y)
```

```
[1] TRUE
```

```
# Locating Values within the Data Frame

Y[,1];Y[3,1]
```

```
 [1] "Freshman"  "Freshman"  "Sophomore" "Sophomore" "Sophomore" "Junior"
 [7] "Junior"    "Junior"    "Junior"    "Senior"    "Senior"

[1] "Sophomore"
```

We are able to select specific subsets of the data frame using the code below.

```
# Locating all the Positions of the Juniors in the Data Frame
Y[Classification=="Junior",]
```

```
  Classification Age
6         Junior  17
7         Junior  17
8         Junior  17
9         Junior  17
```

Reading in Data

CSV Files

To read in data from a .csv file (comma separated values), we will use the following to create a variable called "data" to store the information found in the file. The extra parameter of *header*=TRUE is used to signify variable names within the file. If the csv file does not contain a header, then the FALSE may be used.

1. If we know the extension of the file we can use

data<-read.csv('file extension',header=TRUE)

Note: The file extension can be found using the *file.choose()* command, selecting the desired file and the copying and pasting the path that appears as the output.

2. If we do not know the extension of the file we can use

data<-read.csv(file.choose(),header=TRUE)

This option will allow us to browse our files to choose the appropriate one.

Text Files

To read in data from a .txt file we will use a similar approach with a slightly different code that will adapt to these file types.

1. If we know the extension of the file we can use

data<-read.table(file.choose(),header=TRUE)

2. If we do not know the extension of the file we can use

data<-read.table(file.choose(),header=TRUE)

Additional Notes

By using the *attach(data)* command, we are able to use the variable names in our read in file.

If the data has not been attached, we use *data$"variable name"* to work with the desired variable.

Coding

In this section, we will briefly discuss **function declaration**, **for loops** and **if-else statements**.

Function Declaration

The declaration of a function in R will follow the following syntax.

name<-function(parameters){

body of function

return(output)

}

Here is an example of calculating the mean of a set of data.

```
# Creating a Function

x<-c(10,20,30)
average<-function(vec){
  a<-sum(vec)/length(vec)
  return(a)
}

average(x)
```

```
[1] 20
```

```
# Corroborating the Function Works
mean(x)
```

```
[1] 20
```

For Loops

The syntax of a **for loop** is as follows:

for(index in sequence){

body of loop

}

Here is an example of a for loop that adds 2 to each natural number between 1 and 5, included, and prints the output.

```
# Creating a For Loop
for (i in 1:5) {
  result <- i + 2
  print(paste(result))
}
```

```
[1] "3"
[1] "4"
[1] "5"
[1] "6"
[1] "7"
```

If-Else Statements

The syntax of an **if-else statement** is as follows:

if (Boolean statement){

body

}

else

{

alternate output

}

Here is an example of an if-else statement within a for loop that adds 2 to each natural number between 1 and 5 that is greater than 3, and leaves numbers that are lower than or equal to 3 unchanged.

```
# Creating an If-Else Statement
for (i in 1:5) {
  if (i > 3) {
    result <- i + 2
    print(paste(result))
  } else {
    print(paste(i))
  }
}
```

```
[1] "1"
[1] "2"
[1] "3"
[1] "6"
[1] "7"
```