MBA 8350: Analyzing and Leveraging Data The Course Companion

Scott Dressler

2021-06-02

Contents

VERITAS
UNITAS
CARITAS
1842
VILLANO
UNIVERSI
Villanova School of

Preface

> "In ancient times they had no statistics so they had to fall back on lies."
>
> — Stephen Leacock

About this book...

This course companion is a collection of lecture notes I have compiled over my years of teaching *Analyzing and Leveraging Data* (MBA 8350). I have custom made this material to ensure that all of the relevant topics of the course are included within these pages so we **DO NOT** need an additional textbook.

The purpose of this course companion is to accomplish three goals.

1. Introduce you to the foundational theory and application of statistical methods
2. Introduce you to a cutting-edge software language for statistical computing and data visualization (R)
3. Do away with the need for a formal textbook

While there are certain chapters devoted solely to the second goal, new tricks on how to use R will be scattered throughout the chapters whenever we apply a new statistical method. By the end of the course, we will be not only have a clear understanding of statistical methods but have the working knowledge of an extremely powerful software language with which to apply it!

Acknowledgements

This course companion would not be possible without the many students I have had the pleasure of teaching statistics to at VSB. Their questions, comments, and corrections to previous editions of these notes have made a significant contribution to what this product is. I am not going to call this a *final* product,

because I hope my current and future students will continue to help me improve this.

This course companion was also made possible thanks to a Villanova School of Business 2021 Teaching Innovation Grant.

Chapter 1

Introduction

This chapter is designed to motivate the use of statistics, because we are constantly bombarded with statistics on a daily basis (whether we like it or not). We will start with some motivating examples and an introduction to some important terminology and equations used to calculate descriptive measures (e.g., mean, standard deviation, etc.). We will start throwing around some data and R code in the applications, but we formally get into these details later. You should skim over those details at first and focus more on the results and conclusions provided. You can always come back to this chapter and focus on replicating the code once we cover R in later chapters.

1.1 The "Big Picture" of Statistics

Question 1:

A wholesaler has an inventory of 100,000 light bulbs and wants to market them. What can be said about the lifespan of the light bulbs in this inventory?

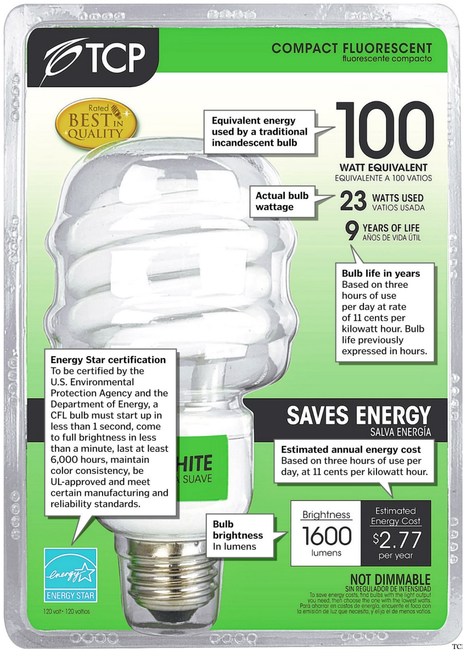
Question 2:

An economist wants to forecast the future state of output in the economy. What variables (i.e., leading indicators) should be included in her model? How much confidence should we place in her predictions?

Question 3:

A meteorologist wants to predict the path of a hurricane. How confident can we be in her predicted path of the storm?

\begin{figure}



Equivalent energy used by a traditional incandescent bulb

Actual bulb wattage

Bulb life in years Based on three hours of use per day at rate of 11 cents per kilowatt hour. Bulb life previously expressed in hours.

Energy Star certification To be certified by the U.S. Environmental Protection Agency and the Department of Energy, a CFL bulb must start up in less than 1 second, come to full brightness in less than a minute, last at least 6,000 hours, maintain color consistency, be UL-approved and meet certain manufacturing and reliability standards.

Estimated annual energy cost Based on three hours of use per day, at 11 cents per kilowatt hour.

Bulb brightness In lumens

Figure 1.1: How do we know so much about this unused lightbulb?

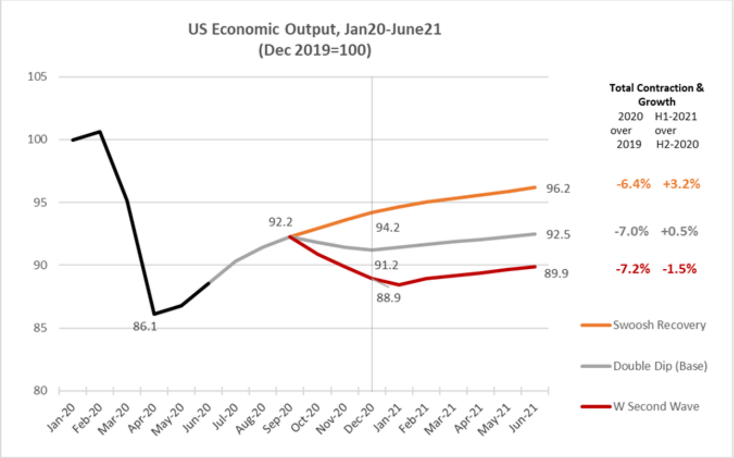
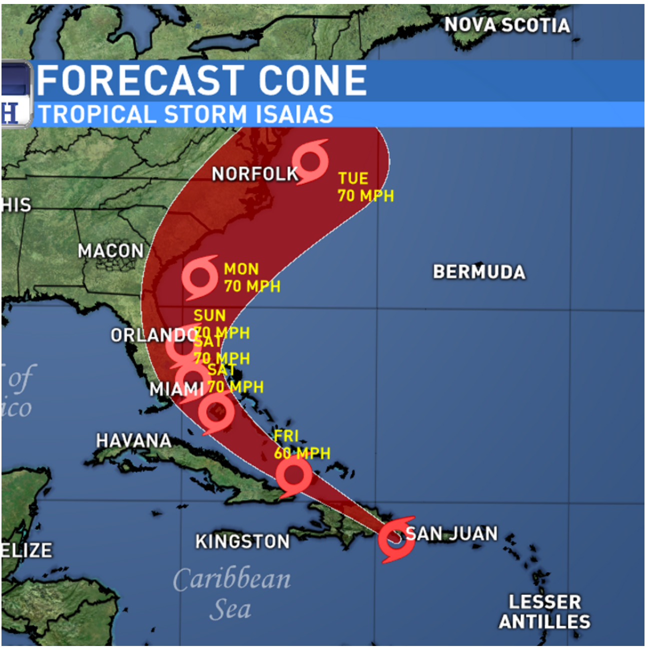


Figure 1.2: The Dismal Science

{



}

\caption{The *Cone of Uncertainty*} \end{figure}

These and many more relevant questions can be answered (as best as possible) using **statistics**.

Statistics is the branch of mathematics that transforms data into useful information for decision makers. Statistics are used to summarize data, draw conclusions, make forecasts, and make better sense of the world and the decisions we make within it.

Statistics is broken into two related branches.

Descriptive statistics are used to summarize, analyze, and present data. Since the values of a dataset are in hand (i.e., observable), the descriptive statistics of a dataset are *facts.*

Inferential statistics use the descriptive statistics from the data to draw conclusions about a larger group of observations that at the moment are *impossible* (or too expensive) to observe. Since this larger group of data is not in hand, the inferential statistics that stem from an analysis are *predictions.*

1.2 The Vocabulary of Statistics

- A **variable** is a characteristic of an item or group that one wishes to analyze.
- **Data** are the different values of the variable observed and recorded.
- An **operational definition** establishes a meaningful use of the variable. Simply put, we need to ensure the data sufficiently captures what you want to analyze.
- A **population** consists of all items you want to draw a conclusion from. The issue with a population is that it is the entire universe of observations that you are interested in, but they can never be fully observed.
  - Sometimes a population is too costly to collect and analyze. For example, you won't call up every single voter for an election poll.
  - Sometimes a population is impossible to collect because some observations have yet to be determined. For example, the population of end-of-day indices for the S&P 500 includes every observation that has ever existed as well as every observation **that has yet to exist**.
- A **sample** is the portion (i.e., subset) of a population selected for analysis. These are our observations in hand.
- A **statistic** is a characteristic of a sample. Since we can observe the sample (i.e., our data), these are our descriptive statistics.
- A **Parameter** is a characteristic of a population. Since we cannot observe the population, the best we can do is draw inferential statistics (or predictions) about them. While the value of a parameter exists, we would have to be omniscient in order to know it. The best we can do is use our sample statistics to construct an *educated guess* of what this value might be.

Recall the problem of the wholesaler who has a supply of light bulbs. It would be great if we could state what the *average lifespan* of the light bulbs are, but that would require timing every light bulb until they burn out. This isn't very useful.

The seven terms stated above translate to our light bulb example as follows:

Term	Our light bulb problem
Variable	The lifespan of a light bulb
Data	The light bulbs that you actually plugged in and recorded the time it takes until burnt out
Operational Definition	The lifespan *in minutes*

Term	Our light bulb problem
Population	The entire group of light bulbs (all 100,000 of them)
Sample	The subset of the population selected for analysis. Sometimes referred to as the *data sample*.
Statistic	The average lifespan of every light bulb **in the sample**
Parameter	The average lifespan of every light bulb **in the population**

Inferential statistics allow us to describe the parameter of a population by using the corresponding statistic of a sample. We will **never** be able to truly know the population parameter, because the information available in the sample is all we got.

How do we know if the sample statistic is a GOOD predictor of the population parameter? The kicker is that since we cannot observe the population, the only thing we can do is try our best to ensure that the characteristics of the sample are the same as the population. This has to do with sample selection - a very important topic that will be addressed soon. First, we start by discussing the descriptive measures of data.

1.3 Descriptive Measures

This section summarizes the measures we use to describe data samples.

- **Central Tendency**: the central value of a data set
- **Variation**: the dispersion (scattering) around a central value
- **Shape**: the distribution pattern of the data values

These measures will be used repeatedly in our analyses, and will affect how confident we are in our conclusions.

To introduce you to some numerical results in R, we will continue with our light bulb scenario and add some actual data. Suppose we sampled 60 light bulbs from our population, turned them on, and timed each one until it burned out. If we recorded the lifetime of each light bulb, then we have a dataset (or data sample) of 60 observations on the lifetimes of light bulbs. This is what we will be using below.

1.3.1 Central Tendency

The **Arithmetic** or **sample mean** is the average value of a variable within the sample.

$$\bar{X} = \frac{\text{Sum of values}}{\text{Number of observations}} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

```
# the mean of our light bulb sample:
# First we load the data set and this will give us 60 observations of the lifespan of

load("data/Lightbulb.Rdata")
list(Lifetime)
```

```
## [[1]]
##  [1]  858.9164  797.2652 1013.5366 1064.8195  874.2275  825.1137  897.0879  924.0998
## [11]  955.1281  977.2073  888.1690  826.6483  776.7479  877.5691  998.7101  892.8178
## [21] 1082.9650 1034.9549  784.5026  919.2082 1049.1824  923.5767  907.7295  890.3758
## [31] 1009.7146  890.3709  930.9597  809.9274  919.9381  793.7455  919.9824  948.8593
## [41]  955.3873  833.2762  892.4969  973.1861  913.7650  928.6057  940.7637  964.434
## [51]  831.5395  967.2442 1030.7598  857.5421  889.3689 1094.1440  927.7684  730.9976
```

```
(mean(Lifetime))
```

```
## [1] 907.5552
```

The average lifetime of our 60 ($n = 60$) observed light bulbs is 908 hours.

The **median** is the middle value of an ordered data set.

- If there is an odd number of values in a data set, the median is the middle value
- If there an even number, median is the average of the two middle values

```
# the median of our light bulb sample:
(median(Lifetime))
```

```
## [1] 902.4087
```

The median lifetime of our 60 observed light bulbs is 902 hours.

The **mode** is the most frequently appearing value in a set of observations. The mode might not exist if there is a set of unique, non-repeating observations.

```
# There isn't a built in function for the mode in R... but we can create one.
```

```
Mode <- function(x) {
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]
}

(Mode(Lifetime))
```

```
## [1] 858.9164
```

The most commonly observed lifetime of our 60 observed light bulbs is 859 hours. Note that this doesn't say anything really about how many times this value has been observed, just that it has been observed more than any other value. We will not be using this measure as much as the previous two.

Percentiles break the ordered values of a sample into proportions.

- Quartiles split the data into 4 equal parts
- Deciles split the data into 10 equal parts
- In general, the pth percentile is given by: $(p * 100)^{th} = p(n+1)$

A percentile delivers an observed value such that a determined proportion of observations are less than or equal to that value. You can choose any percentile value you wish. For example, the code below calculates the 4th, 40th, and 80th percentiles of our light bulb sample.

```
# You can generate any percentile (e.g. the 4th, 40th, and 80th) using the quantile function:
(quantile(Lifetime,c(0.04, 0.40, 0.80)))
```

```
##       4%      40%      80% 
## 787.8301 888.8889 966.4164
```

This result states that 4% of our observations are less that 788 hours, 40% of our observations are less than 889 hours, and 80% of our observations are less than 966 hours. Note that the median (being the middle-ranked observation) is by default the 50th percentile.

```
(quantile(Lifetime,0.50))
```

```
##      50% 
## 902.4087
```

The main items of central tendency can be laid out in a Five-Number Summary:

- Minimum
- First Quartile (25th percentile)

- Second Quartile (median)
- Third Quartile (75th percentile)
- Maximum

```
summary(Lifetime)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   731.0   857.3   902.4   907.6   955.2  1094.1
```

1.3.2 Variation

The **sample variance** measures the average (squared) amount of dispersion each individual observation has around the sample mean. This is a very important measure in statistics, so take some time to understand exactly what this equation is calculating. In particular, X is a variable and X_i is an arbitrary single observation from that group of data. Once the mean ($\bar{X}$) is calculated, $X_i - \bar{X}$ is the difference between a single observation of X and the overall mean of X. Sometimes this difference is negative ($X_i < \bar{X}$) and sometimes this difference is positive $X_i > \bar{X}$ - which is why we need to square these differences before adding them all up. Nonetheless, once we obtain the average value of these differences, we get a sense of the amount of dispersion these individual observations are scattered around the sample average. If this value was zero, then *every* observation of X is equal to $\bar{X}$. The greater the value is from zero, the greater the average dispersion of individual values around the mean.

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

The **sample standard deviation** is the square-root of the sample variance and measures the average amount of dispersion in the same units as the mean. This essentially is done to back-out the fact that we had to square the differences of $X_i - \bar{X}$, so the variance is technically denoted in *squared units.*

$$S = \sqrt{S^2}$$

```
(var(Lifetime))
```

```
## [1] 6235.852
```

```
(sd(Lifetime))
```

```
## [1] 78.96741
```

The variance of our sample of light bulb lifetimes is 6236 squared-hours. After taking the square root of this number, we can conclude that the standard deviation of our sample is 79 hours. Is this standard deviation big or small? The answer to this comes when we get to statistical inference.

Discussion

- The term $(X_i - \bar{X})$ is squared because individual observations are either above or below the mean by design. If you don't square the terms (making the negative numbers positive) then they will sum to zero by design.
- The term $(n-1)$ appears in the denominator because this is a *sample* variance and not a *population* variance. In a population variance equation, $(n-1)$ gets replaced with n because we know the population mean. Since we had to estimate the population mean (i.e., used the sample mean), we had to deduct one **degree of freedom**. We will talk more about degrees of freedom later, but the rule of thumb is that we deduct a degree of freedom every time we build a sample statistic (like sample variance) using another sample statistic (like sample mean).

The **coefficient of variation** is a relative measure which denotes the amount of scatter in the data relative to the mean.

The coefficient of variation is useful when comparing data on variables measured in different units or scales (because the CV reduces everything to percentages).

$$CV = \frac{S}{\bar{X}} * 100\%$$

Take for example the Gross Domestic Product (i.e., output) for the states of California and Delaware.

```r
library(readxl)
CARGSP <- read_excel("data/CARGSP.xls")
DENGSP <- read_excel("data/DENGSP.xls")

CGDP <- CARGSP$CGDP
DGDP <- DENGSP$DGDP

(mean(CGDP))
```

```
## [1] 2094764
```

```r
(sd(CGDP))
```

```
## [1] 397103.9
```

```
(mean(DGDP))
```

```
## [1] 56996.58
```

```
(sd(DGDP))
```

```
## [1] 12395.78
```

A quick analysis of the real annual output observations from these two states between the years 1997 and 2020 suggest that the average annual output of California is 2,094,764 million dollars (with a standard deviation of 397,104 million) and that of Delaware is 56,997 million dollars (with a standard deviation of 12,396 million). These two states have lots of differences between them, and it is difficult to tell which state has more volatility in their output.

If we construct coefficients of variation:

```
(sd(CGDP)/mean(CGDP))*100
```

```
## [1] 18.95698
```

```
(sd(DGDP)/mean(DGDP))*100
```

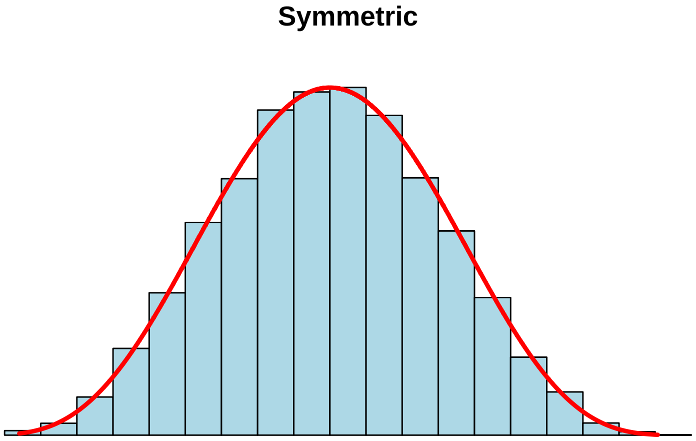
```
## [1] 21.74828
```

We can now conclude that Delaware's standard deviation of output is almost 22% that of its' average output, while California's standard deviation is 19%. This would suggest that Delaware has the more volatile output, relatively speaking.

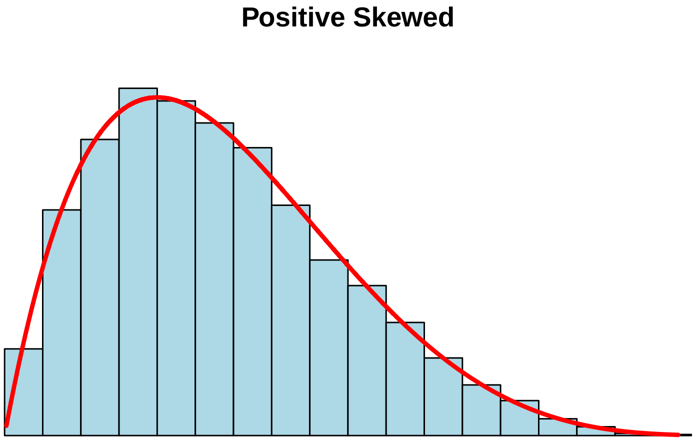
1.3.3 Measures of shape

Comparing the mean and median of a sample will inform us of the skewness of the distribution.

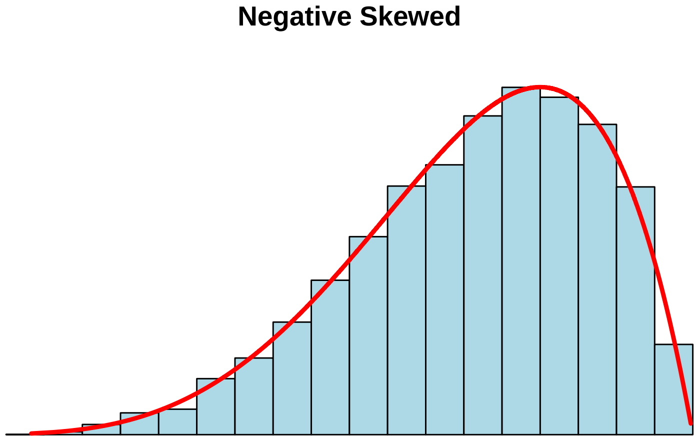
- mean = median: a *symmetric* or *zero-skewed* distribution.



- mean > median: a *positive-skewed* or *right-skewed* distribution
  - the right-tail is pulled in the positive direction



- mean < median: a *negative-skewed* or a *left-skewed* distribution
  - the left-tail is pulled in the negative direction



The degree of skewness is indicative of outliers (extreme high or low values) which change the shape of a distribution.

1.3.4 Covariance and Correlation

While we won't be examining relationships between different variables until later on in the course, we can easily calculate and visualize these relationships.

The **covariance** measures the strength of the relationship between two variables. This measure is similar to a variance, but it can be either positive or negative depending on how the two variables move in relation to each other.

$$cov(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

The **coefficient of correlation** transforms the covariance into a relative measure

$$corr(X,Y) = \frac{cov(X,Y)}{S_X S_Y}$$

The correlation transformed the covariance relationship into a measure between -1 and 1

- $corr(X,Y) = 0$: There is no relationship between X and Y
- $corr(X,Y) > 0$: There is a positive relationship between X and Y - meaning that the two variables tend to move in the same direction
- $corr(X,Y) < 0$: There is a negative relationship between X and Y - meaning that the two variables tend to move in the opposite direction

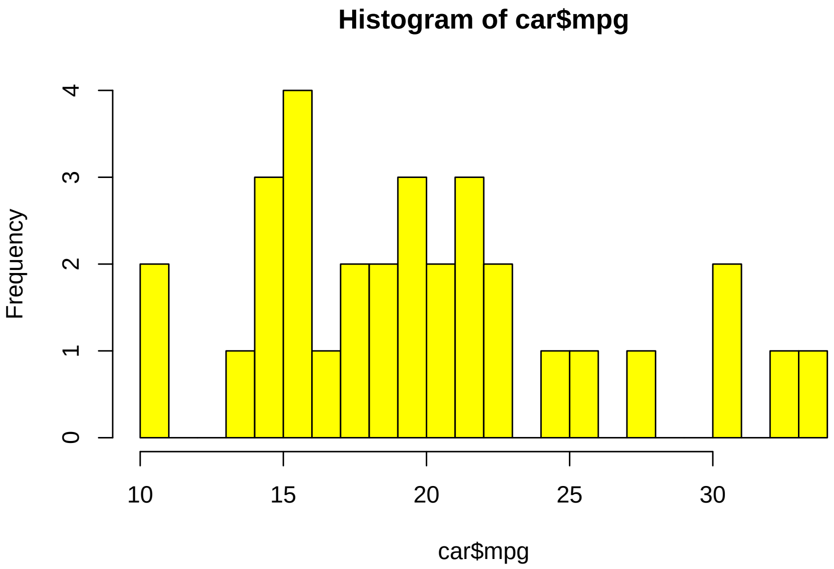
Extended Example:

This chapter concludes with a summary of all of the descriptive measures we discussed. Consider a dataset that is internal to R (called mtcars) that contains characteristics of 32 different automobiles. We will focus on two variables: the average miles per gallon (mpg) and the weight of the car (in thousands of pounds).

```
car <- mtcars # This loads the dataset and calls it car

# Lets examine mpg first:
summary(car$mpg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.40   15.43   19.20   20.09   22.80   33.90
```

```
hist(car$mpg,20,col = "yellow")
```



```
# Variance:
(var(car$mpg))
```

```
## [1] 36.3241
```

```
# Standard deviation:
(sd(car$mpg))
```

```
## [1] 6.026948
```

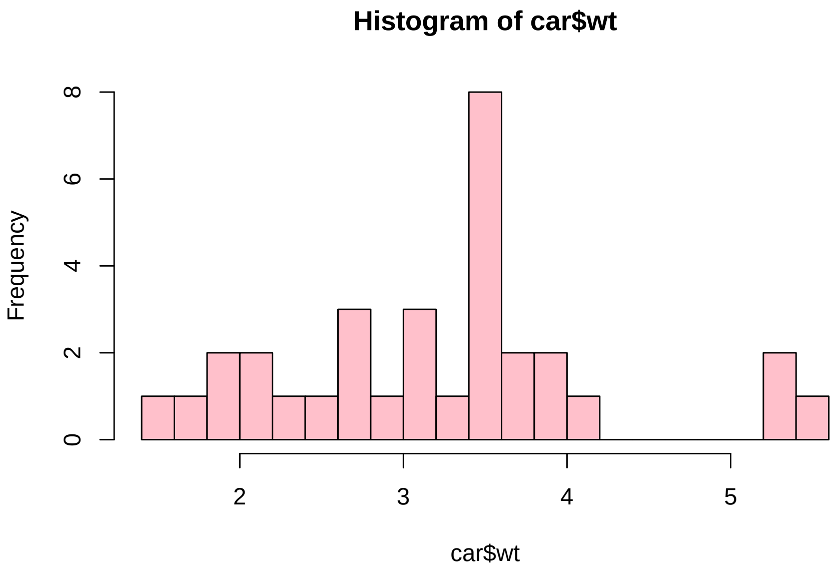
The above analysis indicates the following:

- The sample average MPG in the sample is 20.09, while the median in 19.20. This indicates that there is a slight positive skew to the distribution of observations.
- The lowest MPG is 10.4 while the highest is 33.90.
- The first quartile is 15.43 while the third is 22.80. This delivers the *inter-quartile range* (the middle 50% of the distribution)
- The standard deviation is 6.03 which delivers a 30 percent coefficient of correlation.

```
## Lets now examine weight:
summary(car$wt)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.513   2.581   3.325   3.217   3.610   5.424
```

```
hist(car$wt,20,col = "pink")
```



```
# Variance:
(var(car$wt))
```

```
## [1] 0.957379
```

```
# Standard deviation:
(sd(car$wt))
```

```
## [1] 0.9784574
```

The above analysis indicates the following:

- The sample average weight in the sample is 3.22 thousand pounds, while the median in 3.33. This indicates that there is a slight negative skew to the distribution of observations.

- The lowest weight is 1.51 while the highest is 5.42.

- The first quartile is 2.58 while the third is 3.61.

- The standard deviation is 0.99 which also delivers a 30 percent coefficient of correlation.

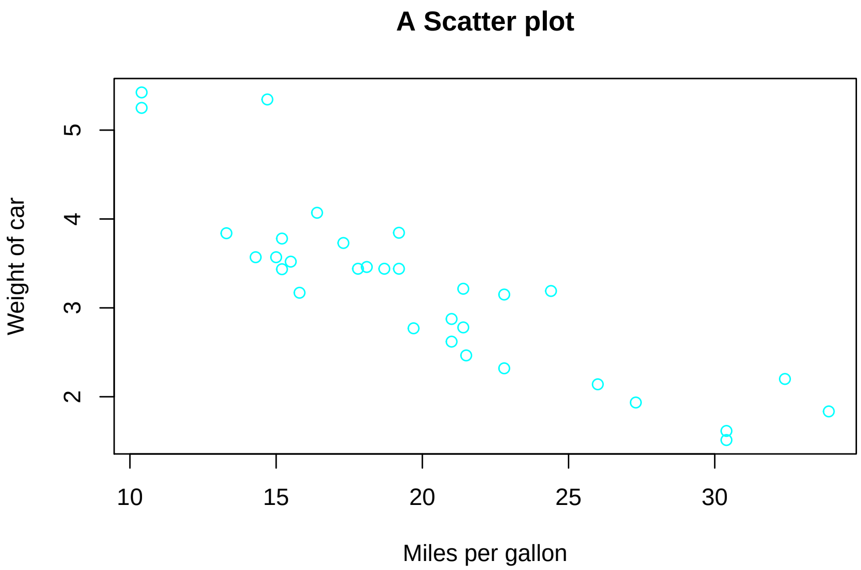
```
# We can now look at the relation between mpg and weight

(cov(car$mpg,car$wt))
```

```
## [1] -5.116685
```

```
(cor(car$mpg,car$wt))
```

```
## [1] -0.8676594
```

```
plot(car$mpg,car$wt,
     xlab = "Miles per gallon",
     ylab = "Weight of car",
     main = "A Scatter plot",
     col = "cyan")
```



The negative correlation as well as the obviously negative relationship in the scatter-plot between the weight of a car and its miles per gallon should make intuitive sense - heavy cars are less efficient.

The Punchline

Suppose we want to learn about a relationship between a car's weight and its fuel efficiency. Our sample is 32 automobiles, but our population is EVERY automobile (EVER). We would like to say something about the population mean Weight and MPG.

How does the sample variance(s) give us confidence on making statements about the population mean when we're only given the sample? That's where inferential statistics comes in. Before we get into that, we will dig into elements of collecting data (upon which our descriptive statistics are based on) and using R (with which we will use to calculate our descriptive statistics using our collected data).

Chapter 2

Data Collection and Sampling

Always remember the ultimate goal of inferential statistics: We want to say something important about the characteristics (parameters) of a population without ever observing the entire population. Therefore, the best thing we can do is to draw a sample (i.e., subset) from the population and use it to calculate characteristics (statistics) of a sample and draw inference on the population parameters.

The reason why we can say something about a population parameter of interest solely by looking at the statistics from a sample is because we are under the assumption that *the sample has the same characteristics of the population*. In other words, we say that the sample average is a good guess for the population average, the sample standard deviation is a good guess for the population standard deviation, etc. This is not an assumption that is simply made by wishful thinking. In fact, there is an entire field of statistics devoted to proper *sample selection*. We won't spend a lot of time on this very important matter, but we will discuss a few sampling methods so you can rest assured that our crucial assumption of similar sample and population characteristics has a reasonable chance of holding.

2.1 Sampling Distributions

Recall that a **sample** is the subset of a **population** selected for analysis.

We are forced to analyze a sample rather than a population because:

1. selecting a sample is less time-consuming than selecting the population

2. selecting a sample is less costly
3. the resulting analysis is less cumbersome and more practical
4. sometimes obtaining the sample is *impossible*! So the sample is the best we can do.

When making statements on the population parameters using the sample statistics, we are drawing **statistical inference**. In order for this inference to be reasonable, we must assume that the characteristics of the sample (i.e., the sample statistics) are reasonably close to the characteristics of the population (i.e., the population parameters). The problem with this assumption is that since we will never see the population, we will never be able to verify if the statistics are reasonably close to the parameters. This chapter discusses several different methods of drawing a sample from a population, as well as their pros and cons. The bottom line is that all of these methods attempt to get a sample to be the best possible subset of the population.

Failing to obtain a sample with the same characteristics as the population can fatally flaw a statistical analysis. If the sample statistics are not close to the population parameters, you are potentially over-representing and/or under-representing important aspects of the population. When the sample statistics do not coincide with the population parameters, then the statistics are said to be *biased.* When this bias stems from a faulty sample, then this is called **sampling bias**.

2.2 Sampling Bias - two examples

There are quite a few glaring examples of sampling bias in history. One of them has to do with a rather famous photo:

2.2.1 Dewey Defeats Truman?

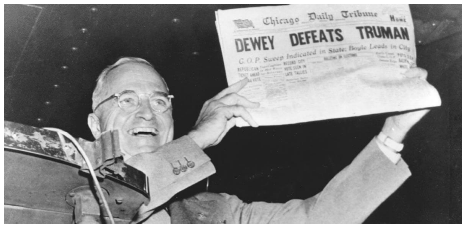


Figure 2.1: Dewey Defeats Truman? (1948)

After defeating Thomas Dewey with a comfortable margin of 303 electoral college votes to Dewey's 189, President Harry Truman holds up a Chicago Daily Tribune stating the exact opposite. While the Truman Library would like to think that this iconic photo is an example of tenacity, perseverance, self-confidence, and success - it's actually a result of *sampling bias*.

The reporting error stems from the fact that the newspaper conducted a poll using phone numbers obtained from a list of vehicle registrations. Most people didn't have phones in 1948, and the people that were being polled had both phones and automobiles. This skewed the sample distribution to wealthy, white, males - which was obviously not sharing similar views with the overall voting population.

2.2.2 98.6?

Everybody knows the following *fact* about their body...

> The average human body temperature is 98.6 degrees F (37 degrees C).

Is it really? To put this issue in the context of our terminology, the *average human body temperature* is a population parameter. The population here is every human that has ever lived and ever will live (i.e., an unobtainable sample). This average is actually a sample average obtained by a German physician in 1851 - a time believed by many current physicians to be one where many suffered from chronic infectious diseases resulting in inflammation and mild fevers. Current studies are suggesting the average human body temperature is more like one degree lower than previously thought.

Now to be clear, there is a bit of a semantic argument about this last example. Some empiricists do not call this necessarily a sampling bias issue in 1851, because if a large portion of the population did regularly suffer from mild fevers then the sample was an accurate subset of the population at the time. Of course, if one is saying that the 1851 estimate of 98.6 degrees F is a representation of the *current* population - then that can be regarded as sampling bias.

2.3 Sampling Methods

The sampling process begins by defining the **frame** - a listing of items that make up the population. A frame could be population lists, maps, directories, etc. For our Truman example above, the frame was incorrectly chosen to be a list of registered vehicle owners (so the poll was doomed from the start).

A sample is drawn from a frame. The sample could be a **nonprobability sample** or a **probability sample**. The items in a nonprobability sample are selected without knowing their probabilities of selection. This is usually done out of convenience (e.g., all voluntary responses to a survey or selecting the top or bottom of a frame). While these samples are quick, convenient, & inexpensive, they most likely suffer from selection bias. We can perform (albeit, incorrectly) statistical analyses on these samples, but we are going to restrict attention to probability samples which are selected based on known probabilities.

2.3.1 Simple random sampling

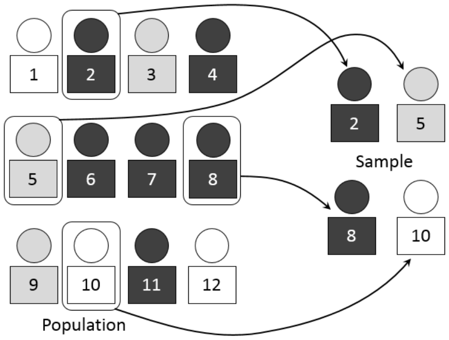


Figure 2.2: Simple Random Sampling

In a **simple random sample**, every item in a frame has an equal chance of being selected.

The chance (or probability) of being selected depends on if you're selecting…

- With replacement (1/N chance for all)
- Without replacement (1/N, 1/(N-1), 1/(N-2), …)

Examples of simple random sampling methods:

- Fishbowl methods
- random number indexing

Advantages:

- Simple random sampling is associated with the minimum amount of sampling bias compared to other sampling methods.
- If the sample frame is available, selecting a random sample is very easy.

Disadvantages:

- Simple random sampling requires a list of all potential respondents (sampling frame) to be available beforehand - which can be costly.
- The necessity to have a large sample size (i.e., lots of observations) can be a major disadvantage in practical levels

2.3.2 Systematic Sampling

A systematic sample begins with partitioning the N items in a frame into n groups of k items

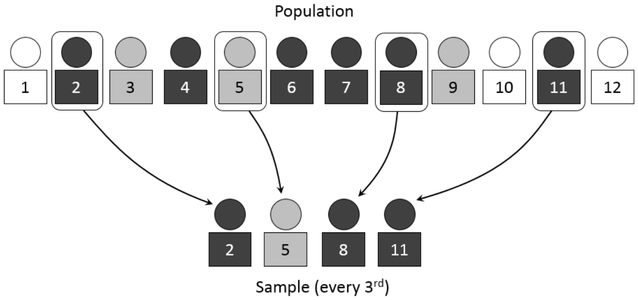


Figure 2.3: Systematic Sampling

$$k = \frac{N}{n}$$

- randomly select a number from 1 through k
- select the kth member from each of the n groups

For example: Suppose you want a sample n=40 out of N=800.

- Divide the population into k=20 groups.
- Select a number from 1-20 (e.g. 8)
- Sample becomes items 8,28,48,68,88,...

Advantages

- it will approximate the results of simple random sampling
- it is cost and time efficient

Disadvantages

- it can be applied only if the complete list of a population is available
- the sample will be biased if there are periodic patterns in the frame

2.3.3 Stratified Sampling

A stratified sample divides the N items in the frame into important sub-populations (strata)

- Each strata groups items according to some shared characteristic (gender, education, etc.)

Once these strata are constructed. A researcher selects a simple random sample from each strata and combines.

Advantages

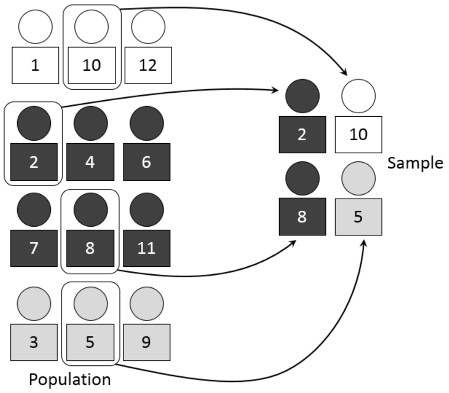


Figure 2.4: Stratified Sampling

- it is superior to simple random sampling because it reduces sampling error and ensures a greater level of representation
- ensures adequate representation of all subgroups
- when there is homogeneity within strata and heterogeneity between strata, the estimates can be as precise (or even more precise) as with the use of simple random sampling

Disadvantages

- requires the knowledge of strata membership
- process may take longer and prove to be more expensive due to the extra stage in the sampling procedure

2.3.4 Cluster Sampling

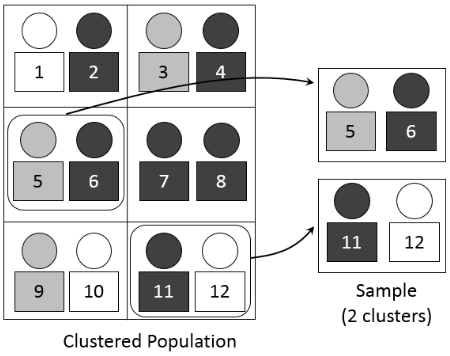


Figure 2.5: Cluster Sampling

Cluster Sampling occurs when you break the sample frame into specific groups (i.e., clusters) and then randomly select several clusters as your sample. An example of this method is the consumer price index (CPI) which is a measure of inflation calculated by the Bureau of Labor Statistics in the U.S. (US BLS). When trying to estimate the overall change in a *basket* of consumption goods across the US, the BLS breaks the US into metropolitan statistical areas (MSAs) and treats each one as a cluster. The BLS then goes and prices the various goods in the clusters selected for analysis.

Advantages:

- the most time-efficient and cost-efficient probability design for large geographical areas
- easy to use
- larger sample size can be used due to increased level of accessibility of perspective sample group members

Disadvantages

- requires group-level information to be known
- commonly has higher sampling error than other sampling techniques
- may fail to reflect the diversity in the sampling frame

2.4 Sampling in Practice

The sampling methods above deal with situations in which the population is hypothetically obtainable, but it is not feasible due to time or resource constraints. For example, a company could run a concrete election poll by calling up *every single registered voter* (the population), but that would cost too much and take too long. What happens in the situation where the population is unobtainable, meaning that at any point in time there will be some obtainable portion of the population because it hasn't occurred yet. For example, if I wanted to analyze US unemployment rates, I couldn't possibly consider *future* rates that haven't been observed yet. In situation like these, one must take time to consider exactly what population you want to draw inferences from and draw their sample accordingly.

A quick example of data sampling in my own research is as follows. Some of my research deals with how bank lending responds to changes in the stance of monetary policy.[1] Since bank lending data is coming in daily, it is clear that the entire population is unobtainable. However, selecting a sample is not simply *collect as many observations as possible* because we must be clear

[1] Dave, Chetan, Scott J. Dressler, and Lei Zhang, (2013). The bank lending channel: a FAVAR analysis. *Journal of Money, Credit, and Banking* 45(8). 1705-1720.

about what population we want to actually talk about. In my example, I want to talk about how bank lending responds to monetary policy shocks *in normal times*. This means that observations in the sample cannot be impacted by episodes where monetary policy differed from what is currently considered normal. This restricts my sample to be after World War 2 and before episodes of unconventional monetary policy (i.e., anything post-2007).

What happens if characteristics of the population potentially changes? That's easy - you repeat the analysis with an updated sample and acknowledge that you are drawing inferences on a potentially different population. That is what I am currently researching. In particular, I am determining how bank lending responds to monetary policy under unconventional monetary policy practices of paying interest on excess reserves. This requires a data sample of observations appearing after 2007.

2.5 Sampling and Sampling Distributions

This chapter concludes with the hypothetical concept of a sampling distribution. Understanding this concept is crucial to understanding the entire point of inferential statistics.

Recall that we want to make statistical inferences that use statistics calculated from samples to estimate parameters of the population.

Plain statistics draws conclusions about the sample (those are facts) while statistical inference draws conclusions about the population.

In practice, a single sample is all that is selected. In other words, once you construct your sample it is all of the observations you have to use.

Since the actual observations inside your sample were selected at random, then the sample you constructed is in fact a *random* sample.

If the random observations were drawn from a sample frame, what was the resulting random sample drawn from? The answer is a **sampling distribution**.

2.5.1 An Application

Consider the scenario discussed above where we want to determine the population average human body temperature. At a particular point in time, the population is every human. As the particular points in time change, new births implies that the population is changing as well! Clearly the overall population is unobtainable - so we need to draw a sample.

Suppose we decide on a sample size of 10,000 adults. Regardless of the sampling method chosen from the list above, we arrive at a data sample of 10,000 observations of human body temperatures. Since these individuals were selected *randomly*, then the sample mean calculated from the *random sample* is itself *random*. If we randomly draw another sample of 10,000 observations, we can get another sample average. We can do this repeatedly, getting a different sample average for every sample randomly drawn.

Note that this is purely hypothetical because we would never draw numerous samples… but bear with me.

We have established that our sample was a random draw from our population. Therefore, the sample mean calculated from our random sample is itself a random draw from a sampling distribution.

Think of a sampling distribution as a histogram showing you the outcomes of all possible sample means and their frequency of appearing. This distribution will have characteristics of its own. The mean of this distribution would be the mean value of all possible sample means. The standard deviation would be the amount of average dispersion all individual sample means around the overall mean.

What we will soon see is that this sampling distribution will be the foundation to inferential statistics. To see this, we will combine this concept of a sampling distribution with something called the Central Limit Theorem (CLT). The CLT is so important, it deserves its own chapter. However, before we get to that conceptual stuff, we will first get into the practical stuff. Namely, an introduction to the R project for Statistical Computing.

Chapter 3

Getting Started with R

This chapter is designed to get R on your machine (Section 1), introduce you to some basic commands for data and variable manipulation (Section 2), and introduce you to some introductory data visualization (Section 3). We will also be using a companion software called Rstudio which will make our interaction with R much more pleasant.

After the basics are covered in this chapter, you should be able to go back to Chapter 2 and have a better understanding of some of the code there. We will be learning additional R commands as they become needed in the subsequent chapters. By the end of the course, you should have a pretty solid understanding of working your way around R.

Let's get started!

3.1 The R Project for Statistical Computing

R is an open source (i.e. *free*) programming language and software environment for statistical computing. The R language is widely used among statisticians and data analysis. As of July 2020, R ranks 8th in the TIOBE index, a measure of popularity of programming languages.

In addition to using R, we will also be using RStudio. RStudio is an integrated development environment (IDE) for R. The desktop version of R studio is also free, and comes with many useful features. In fact, the document you are reading was formatted in R studio.

This section will walk you through downloading, installing, and preparing R and RStudio for our purposes.

Note: this section is intended for installation on either a PC or a Mac. For those of you using Chromebooks (which uses a Linux operating system), you will not be able to install these programs unless you partition your hard drive to run purely in Linux. If you don't know what that means, then I don't suggest you go down that route. However, you are still in luck because Rstudio also offers a free cloud-based version of R that will serve the purpose of this course. You can sign up here: https://www.rstudio.com/products/cloud/.

Another Note: For those of you that have PCs or Macs and simply *do not* wish to download and install R and Rstudio, you are welcome to sign up to the Rstudio cloud as well. This will also be handy if you have a pesky work computer with a rigid fire wall. However, you should note that the cloud version of R and Rstudio can only be accessed if you have an internet connection.

A Final Note: The cloud-based version of Rstudio is relatively new and going through changes. As of the writing of this chapter, they are discussing discontinuing the *free* version and switching over to a fee-based scenario. They are also limiting the computation time for the free version, and this may be a constraint for us. I therefore ask that you contact me to discuss things if you are planning on taking the cloud route.

3.2 Downloading and installing R

The first step to get R onto your machine is to go to the website, download the correct version of R, and install.

The website is https://www.r-project.org/.

3.2.1 Choosing a *Mirror*

Since R is open source, there are many different servers around the world where you can download it. You are welcome to choose any mirror you wish, but you may want to be sure that you know the national language of whichever country you select. I was boring and simply chose a mirror in Pittsburgh because it was closest to my location.

3.2.2 Download and install the correct version

R is available for PCs, Macs, and Linux systems. You will most likely want one of the first two options. Be sure to choose the option in the top box that offers *Precompiled binary distributions*.

For Macs:

- Click on the "Download R for (Mac) OS X" link at the top of the page.
- Click on the file containing the latest version of R under "Files."
- Save the .pkg file, double-click it to open, and follow the installation instructions.

For PCs:

- Click on the "Download R for Windows" link at the top of the page.
- Click on the "install R for the first time" link at the top of the page.
- Click "Download R for Windows" and save the executable file somewhere on your computer. Run the .exe file and follow the installation instructions.

Once this is complete, you will never need to actually open R. We will using RStudio to communicate with R.

3.2.3 Downloading and installing RStudio

While R is open source, RStudio is a company that sells versions of its IDE. In short, it is an easy-to-use interface that makes working with R easier. We will be using the free version of RStudio Desktop which is available here:

https://rstudio.com/products/rstudio/download/

For Macs:

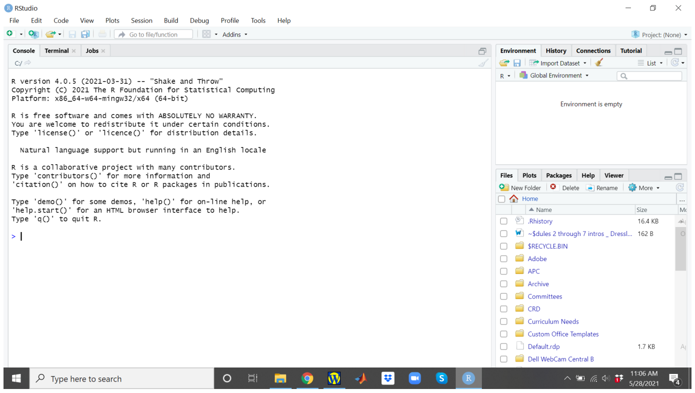
- Click on "Download RStudio Desktop."
- Click on the version recommended for your system, or the latest Mac version, save the .dmg file on your computer
- Double-click it to open, and then drag and drop it to your applications folder.

For PCs:

- Click on "Download RStudio Desktop."
- Click on the version recommended for your system, or the latest Windows version, and save the executable file.
- Run the .exe file and follow the installation instructions.

3.2.4 Taking Stock

If done correctly, you should now be able to open Rstudio and see a screen that looks like this:



```
R version 4.0.5 (2021-03-31) -- "Shake and Throw"
Copyright (C) 2021 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> |
```

Environment is empty

Name	Size
.Rhistory	16.4 KB
~$dules 2 through 7 intros _ Dressl...	162 B
$RECYCLE.BIN	
Adobe	
APC	
Archive	
Committees	
CRD	
Curriculum Needs	
Custom Office Templates	
Default.rdp	1.7 KB
Dell WebCam Central B	

Figure 3.1: Welcome to R!

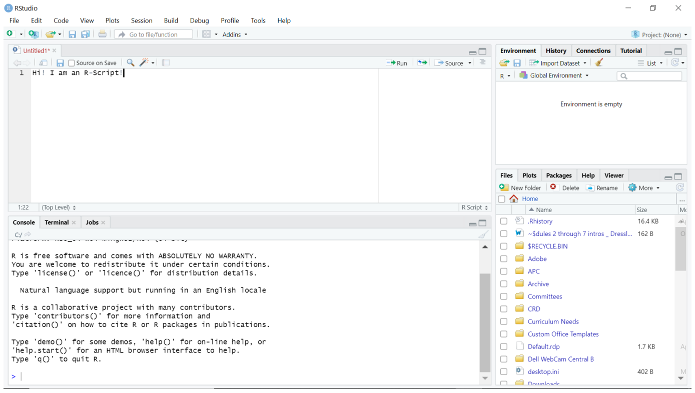
The window on the left is your *Console* which is exactly what you would see if you opened up R instead of Rstudio.[1] The window on the upper-right is your *Global Environment*. It will show you all of the data sets, variables, and result objects that are currently in R and available to you. Note that it is currently empty because we haven't done anything yet. The window on your bottom-right has several useful tabs that let us look at our folder directory (as shown) as well as any figures we generate and R packages at our disposal.

This is the default mode of Rstudio. You can input commands into the console right at the ">" and R will execute them line by line. This is fine if you wish to execute one single command, but it becomes tedious if we have a series of commands we need to execute before we arrive at our desired result. We can therefore alter this default mode by adding R-scripts.

Clicking on the green plus in the upper left of the screen will give you the option of opening an R-script. An R-script window will now appear and take up half of what was once our console space. An R-script is really nothing more than a text file. We can type several commands in sequence without running them line by line (which we would need to do if we typed them into the console). Once the commands are typed out, we can highlight them all and hit that run button on top. The commands get sent to the console and you're off...

The picture above is just a quick example of what an R-script can do. Line 3 tells R to plot all of the variables in a dataset called *mtcars*. Highlighting that line and hitting the run button sends the command to the console below, and

[1] Note that we will never open R by itself, because it is easier to communicate with R through Rstudio.



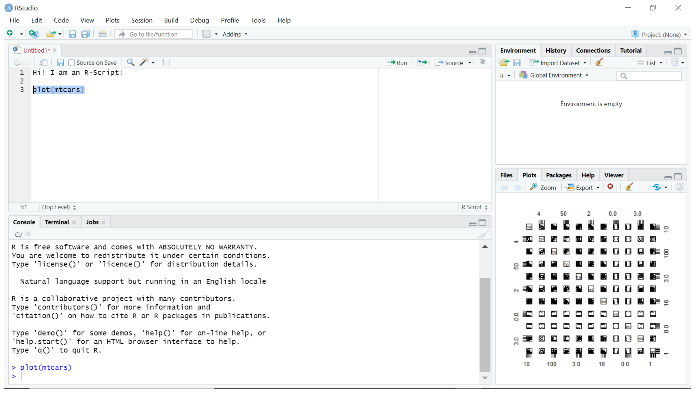
```
Hi! I am an R-Script!
```

Environment is empty

Name	Size
.Rhistory	16.4 KB
~$dules 2 through 7 intros _ Dressl...	162 B
$RECYCLE.BIN	
Adobe	
APC	
Archive	
Committees	
CRD	
Curriculum Needs	
Custom Office Templates	
Default.rdp	1.7 KB
Dell WebCam Central B	
desktop.ini	402 B

```
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

>
```

Figure 3.2: An R-script!



```
Hi! I am an R-Script!

plot(mtcars)
```

Environment is empty

```
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> plot(mtcars)
>
```

Figure 3.3: Running Commands from R-scripts!

the plot figure shows up in the Plots window. That's that!

3.2.5 Installing *Packages*

R is essentially a super-powered calculator. In order for it to be able to do some of the sophisticated things we will be doing in the course, we need to install source code called *packages*.

Whenever you need a package, all you need to do is type:

```
install.packages("name of package")
```

Once this is done, the package is installed in your version of R and you will never need to install it again. You will need to *unpack* the packages each time you want to use them by calling a *library* command, but we will get to that later.

The first R-script you will run as part of your first assignment is called *Install_Packages.R*. The executable portion of the code looks like this:

```
install.packages( c("AER", "car", "dplyr",
"fastDummies", "readxl", "xtable", "vars",
"WDI", "xts", "zoo", "wooldridge") )
```

This is a simple command that asks R to download 11 packages and install them. Download and import the R-script, highlight the portion above, and hit the *Run* tab at the top of the upper-left window of RStudio. A bunch of notifications will appear on the R console (lower-left window) while the list of packages will be downloaded from the mirror site you selected earlier and installed. This can take some time (about 30 mins) depending on your internet connection, so it is advised to do this when you can leave your computer alone for awhile.

3.3 Coding Basics

Now that your software is ready to go, this section introduces you to how R likes to be talked to. Note that the subsequent chapters are full of commands that you will need to learn when the time comes. In the meantime, here are just a few general pointers.

R is what is known as a line command computing language - meaning that it doesn't need to compile code prior to execution. That being said, try the following command at the prompt in your console (>):

```
12 + 4
```

```
## [1] 16
```

See? Just a big calculator.

3.3.1 Assigning Objects

We declare variable names and other data objects by *assigning* things names. For example, we can repeat the calculation above by first assigning some variables the same numbers:

```
BIG <- 12
SMALL <- 4
(TOTAL <- BIG + SMALL)
```

```
## [1] 16
```

Notice that all of these variable names should now be in your global environment (upper-right window). The reason why 16 was returned on the console is because we put the last command in parentheses. That is the *print to screen* command.

You might be asking why R simply doesn't use an equal sign in stead of the assign sign. The answer is that we will be assigning names to output objects that contain much more than a single number. Things like regression output is technically *assigned* a name, so we are simple being consistent.

3.3.2 Listing, Adding, and Removing

We can list all objects in our global environment using the list command: ls()

```
ls()
```

```
##   [1] "AL"          "alpha"       "AUTO"        "B1"          "Bhat0"       "Bhat1"       "BIG
##   [8] "car"         "CARDATA"     "CARDATA2"    "CARGSP"      "CDdata"      "CGDP"        "CM'
##  [15] "CREG"        "D"           "DENGSP"      "df"          "DGDP"        "DS"          "DTF
##  [22] "e"           "eps"         "EPS"         "Fcrit"       "fit"         "fitpoints"   "Fst
##  [29] "grid.lines"  "h"           "hprice1"     "i"           "i1"          "i2"          "j"
##  [36] "k"           "left"        "LEFT"        "LFT"         "Lifetime"    "Lifetime1"   "Lif
##  [43] "m"           "M"           "MDAT"        "Mode"        "mtcars"      "mu"          "MUL
##  [50] "n"           "N"           "P"           "PARK"        "probability" "Pval"        "R"
##  [57] "R2r"         "R2u"         "Rate"        "REG"         "REG1"        "REG2"        "REG
##  [64] "REG4"        "RES"         "Revenue"     "RHT"         "right"       "RIGHT"       "RRE
##  [71] "S"           "SBhat1"      "Sig"         "sigma"       "SMALL"       "t"           "t_v
##  [78] "tcrit"       "TOTAL"       "tstat"       "UREG"        "wage1"       "x"           "X"
##  [85] "x.pred"      "X1"          "X2"          "X3"          "Xbar"        "Xcrap"       "xfi
##  [92] "xtick"       "xy"          "Y"           "y.pred"      "Y1"          "Y2"          "Y3'
##  [99] "yfit"        "Yhat"        "Yz"          "Z"           "z.pred"      "Zcrit"       "Zst
```

As we already showed, we can add new variables by simply assigning names to our calculations.

```
TOTAL.SQUARED <- TOTAL^2
```

If you ever wanted to remove some variables from your global environment, you can use the remove command: rm(*name of variable*)

```
rm(TOTAL.SQUARED)
```

3.3.3 Loading Data

R can handle data in almost any format imaginable. The main data format we will consider in this class is a trusty old MS Excel file. There are two ways to load data...

1. The Direct Way

Once you locate a data file on your computer, you can direct R to import the file and assign it any name you want. The example below imports a dataset of automobile sales called AUTO_SA.xlsx and names it *CARDATA*.

```
library(readxl)
CARDATA <- read_excel("data/AUTO_SA.xlsx")
```

The term *"data/AUTO_SA.xlsx"* is the exact location on my computer for this data file. It is recommended that you put all of you data files somewhere easy to access. Like a single folder directly on your C drive.

2. The Indirect (but easy) Way

You can also import data directly into R through Rstudio.

1. Use the files tab (bottom-right window) and locate the data file you want to import.
2. Left-click on file and select *Import Dataset...*
3. The import window opens and previews your data.
4. If everything looks good, hit *Import* and your done.

Note that the import window in step 3 has a *code preview* section which is actually writing the code needed to import the dataset. It will look exactly like what your code would need to look like in order to import data the direct way. You can refer to that for future reference.

3.3.4 Manipulating Data

You should now have a dataset named *CARDATA* imported into your global environment. You can examine the names of the variables inside the dataset using the list command - only this time we reference the name of the dataset.

```
ls(CARDATA)
```

```
## [1] "AUTOSALE" "CPI"      "DATE"     "INDEX"    "MONTH"    "YEAR"
```

When referencing a variable within a dataset, you must reference both the names of the dataset and variable so R knows where to get it. The syntax is:

Dataset$Variable

For example, if we reference the variable *AUTOSALE* by stating that it is in the CARDATA dataset.

CARDATA$AUTOSALE

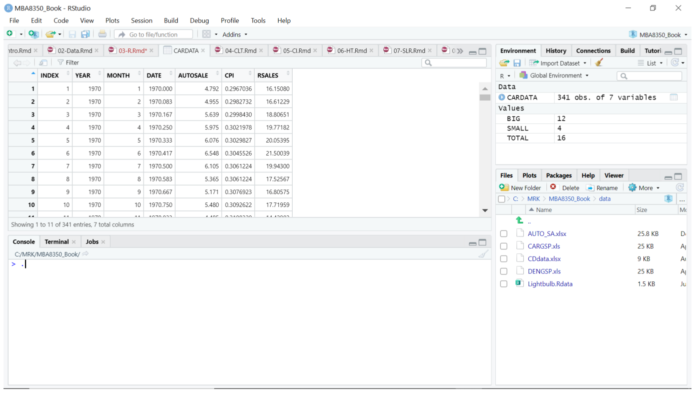
We can now manipulate and store variables within the dataset by creating variables for what ever we need. For example, we can create a variable for real autosales by dividing autosales by the consumer price index (CPI).

```
CARDATA$RSALES <- CARDATA$AUTOSALE / CARDATA$CPI
ls(CARDATA)
```

```
## [1] "AUTOSALE" "CPI"      "DATE"     "INDEX"    "MONTH"    "RSALES"   "YEAR"
```

3.3.5 Subsetting Data

Sometimes our dataset will contain more information than we need. Let us narrow down our dataset to see how we can get rid of unwanted data. You should see a little Excel looking icon to the left of the name CARDATA up in the global environment window. If you click on it, you should see the following:



	INDEX	YEAR	MONTH	DATE	AUTOSALE	CPI	RSALES
1	1	1970	1	1970.000	4.792	0.2967036	16.15080
2	2	1970	2	1970.083	4.955	0.2982732	16.61229
3	3	1970	3	1970.167	5.639	0.2998430	18.80651
4	4	1970	4	1970.250	5.975	0.3021978	19.77182
5	5	1970	5	1970.333	6.076	0.3029827	20.05395
6	6	1970	6	1970.417	6.548	0.3045526	21.50039
7	7	1970	7	1970.500	6.105	0.3061224	19.94300
8	8	1970	8	1970.583	5.365	0.3061224	17.52567
9	9	1970	9	1970.667	5.171	0.3076923	16.80575
10	10	1970	10	1970.750	5.480	0.3092622	17.71959

Showing 1 to 11 of 341 entries, 7 total columns

```
C:/MRK/MBA8350_Book/
> .
```

```
Data
CARDATA   341 obs. of 7 variables
Values
  BIG     12
  SMALL   4
  TOTAL   16
```

C: > MRK > MBA8350_Book > data

Name	Size
..	
AUTO_SA.xlsx	25.8 KB
CARGSP.xls	25 KB
CDdata.xlsx	9 KB
DENGSP.xls	25 KB
Lightbulb.Rdata	1.5 KB

Figure 3.4: A Dataset in R

Thinking of the data set as a matrix with 341 rows and 7 columns will help us understand the code needed to select specific portions of this data.

Note that the variable MONTH cycles from 1 to 12 indicating the months of the year. Suppose we only want to analyze the 12th month of each year (i.e., December). We can do this by creating a new dataset that keeps only the rows associated with the 12 month.

```
CARDATA2 <- CARDATA[CARDATA$MONTH==12,]
```

What the above code does is treat the dataset CARDATA as a matrix and lists it as [rows,columns]. The rows instruction is to only keep rows where the month is 12. The columns instruction is left blank, because we want to keep all columns.

3.4 Data Visualization

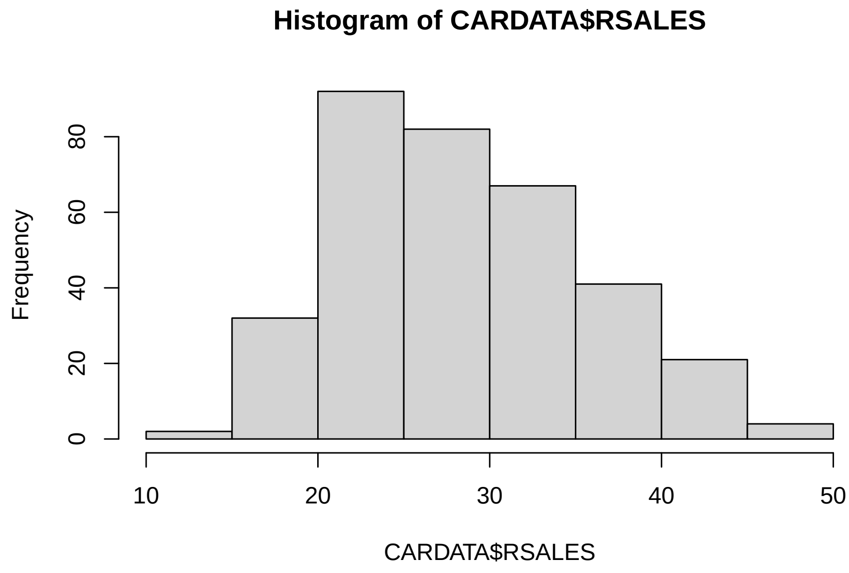
R is absolutely brilliant when it comes to data visualization, and this section will only scratch the surface. We will go over some basic data visualizations using the built-in features of R. There are a lot of resources out there that covers a separate R package called ggplot. It's a handy package, but knowing the features discussed here will be sufficient for our course as well as give you some background that will help you push ggplot farther (if need be).

3.4.1 Histograms

A histogram breaks data observations into bins (or breaks) and shows the frequency distribution of these bins. We will use this to consider probability distributions, but it also helps us get an idea of the distributional properties of any data sample.

Let us continue to analyze the car dataset we created above:

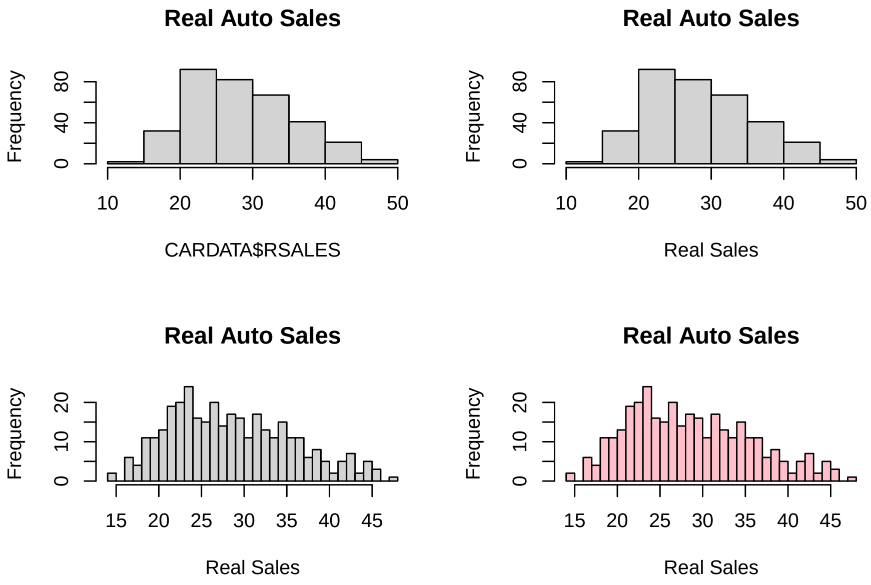
```
hist(CARDATA$RSALES)
```



We can fancy this up by changing the title (main), labels (xlab), number of bins (breaks), and color (col). We will do this one at a time by creating a 2 by

2 set of figures using the par(mfrow=c(2,2)) command. This command *partitions* the plot window into a 2x2 series of subplots.

```
par(mfrow=c(2,2))
hist(CARDATA$RSALES,main = "Real Auto Sales")
hist(CARDATA$RSALES,main = "Real Auto Sales",
     xlab = "Real Sales")
hist(CARDATA$RSALES,main = "Real Auto Sales",
     xlab = "Real Sales",
     breaks = 40)
hist(CARDATA$RSALES,main = "Real Auto Sales",
     xlab = "Real Sales",
     breaks = 40,
     col = "pink")
```

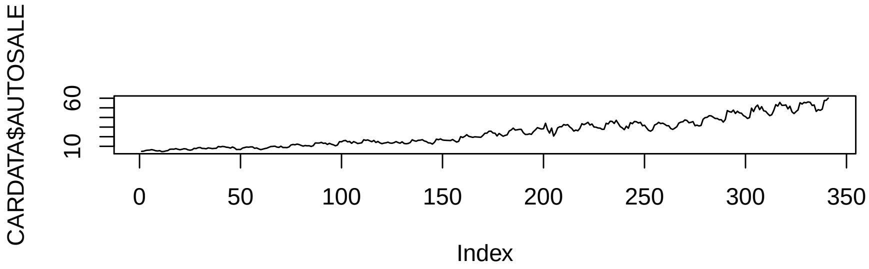


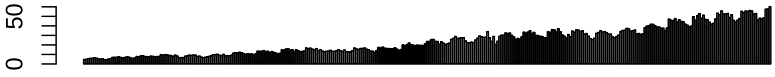
3.4.2 Line, bar, and Scatter Plots

The plot command can visualize the relationship between two variables or just one variable in order. The barplot command is similar to a single-variable plot.

We can look at the nominal sales data in a line plot by specifying the type of plot as “l”. A barplot delivers the same information, but just looks different.

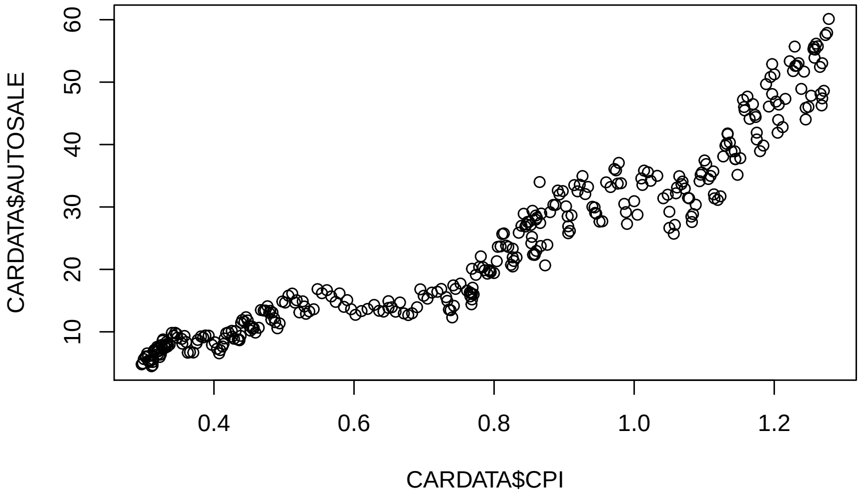
```
par(mfrow=c(2,1))
plot(CARDATA$AUTOSALE, type = "l")
barplot(CARDATA$AUTOSALE)
```





We can look at relationships using the default values of the plot command.

```
plot(CARDATA$CPI,CARDATA$AUTOSALE)
```

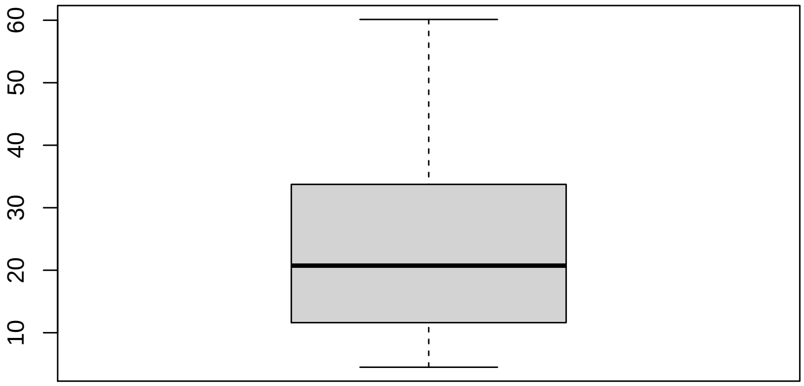


You will see plenty of these plots throughout these notes, and they will get increasingly more sophisticated with titles, colors, etc.

3.4.3 Boxplots

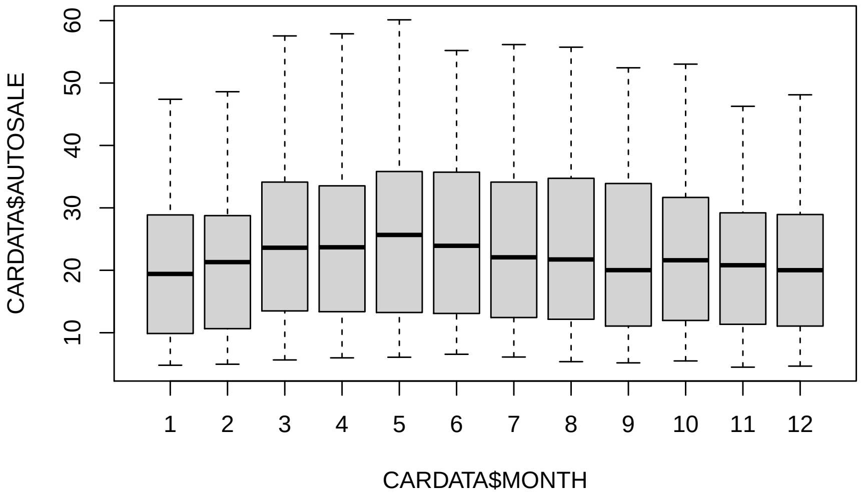
Box Plot illustrate the minimum, the 25th, 50th (median), 75th percentiles and the maximum. It is useful for visualizing the spread of the data.

```
boxplot(CARDATA$AUTOSALE)
```



We can also examine these five numbers within groups according to some other variable. Lets look at this breakdown of auto sales per month of the year.

```
boxplot(CARDATA$AUTOSALE~CARDATA$MONTH)
```



3.4.4 Much more out there

While this basically covers most of the plots we will need for the course, there is a ton more out there. The interested reader can consult a *free* book on the matter.

https://rkabacoff.github.io/datavis/

However deep you want to go, I hope you have seen that data visualization in R is a heck of a lot easier than in Excel.[2]

[2] For example, a histogram in MS Excel takes about 20 minutes for me to create for each one!

Chapter 4

The Central Limit Theorem

The **Central Limit Theorem** (henceforth, CLT) is one of the most important conceptual parts of inferential statistics. It is the essential reason why we can make educated guesses regarding the parameters of a population using information on the statistics of a sample. The CLT will going on in the background of every subsequent chapter of this course companion. Doing statistics without fully understanding the CLT is simply going through the motions. You will not be able to fully appreciate inferential statistics without knowing what is really going on beneath the hood.

4.1 The CLT (Formally)

Recall the concept of sampling distribution from chapter 2. For every randomly selected sample (i.e., a subset of the population), you can calculate a sample mean. If you were to repeatedly collect random samples and record their sample means, then you would be able to construct a *sampling distribution* of the sample mean values. Looking at the frequency of values (i.e., a frequency distribution) would give you an idea of where you think the mean value from the next sample you would randomly draw will be. The statistical properties of this sampling distribution is where the educated guessing is coming from.

So here is the CLT formally…

> The central limit theorem states that if you have a population with mean μ and standard deviation σ and take sufficiently large random samples of size n from the population with replacement, then the distribution of the sample means will be approximately normally distributed.

There are some finer details to note.

- Given the population parameters μ and σ, the resulting sampling distribution will be a normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$.

- This will hold true regardless of whether or not the source population is normal, provided the sample size is sufficiently large (usually $n > 30$).

- If the population distribution is normal, then the theorem holds true even for samples smaller than 30.

- This means that we can use the normal probability distribution to quantify uncertainty when making inferences about a population mean based on the sample mean.

Now, the CLT can be proven - but I think it's better to illustrate the CLT with a couple of examples.

4.2 Application 1: A Sampling Distribution with a Known Population

The first application presents sampling distributions for a random process where we know the underlying process of the population: The rolling of two die.

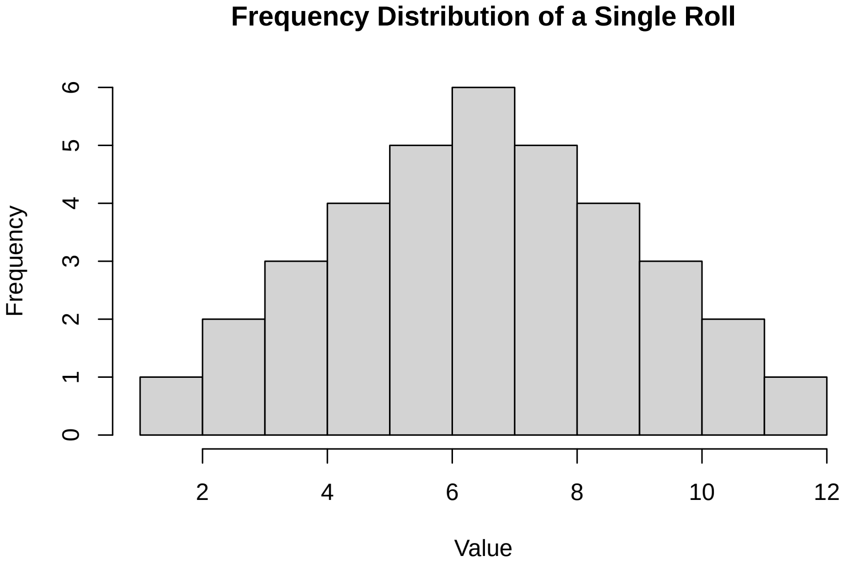
Suppose you worked for a gaming commission and placed in charge of making sure the dice at a casino were fair. **We know** that the (population) average roll of 2 fair die is 7 while the standard deviation is 2.45.[1]

It wouldn't be fair for you to test a set of dice by rolling them once because there is a large probability of rolling a number other than 7. In particular, there are 36 possible outcomes of rolling two die and only 6 of those outcomes equal 7. This means that although 7 is the highest probability single outcome, there is a much higher probability of rolling a number other than 7 (*ever play craps?*).

[1] The mean of a single dice throw is 3.5,

$$3.5 = (1 + 2 + 3 + 4 + 5 + 6)/6$$

and the expected value of two independent dice is the sum of expected values of each die. Standard deviation can be calculated using this mean value and the formula presented earlier.



The figure above is the population distribution of rolling two die. The average (mean) value is 7, the range of possible outcomes are between 2 and 12, and the standard deviation is a number that represents the dispersion of individual values around the mean. If you were to roll two die, then the outcome of that roll is conceptually a draw from this distribution.

Since we don't want to wrongfully accuse the casino of cheating, we need to roll the dice a few times to get an idea of what the average roll value is. If it is fair dice, then we know they will roll a 7 on average - but that means we would need to roll the dice an *infinite* amount of times to achieve this. To be realistic, lets settle on a number of rolls to be generally given by n. If we choose $n = 5$, then that means we roll the dice 5 times, record the roll each time, and then record the average. This is a sample average of a sample of size 5. We could do this for $n = 10$, $n = 30$, $n = 300$, etc.

The figure is illustrating four potential *sampling distributions*. For example, if you were to collect a sample of 5 rolls, then you would technically be drawing a sample average from the distribution in the upper left. On the other hand, if you decide to roll the dice 300 times, then you are technically drawing a sample average from the distribution in the lower-right.

There are two main takeaways from the above illustration.

1. Sampling distributions *appear* to approximate normal distributions. The normal distribution is the classic *bell-curve* distribution that tends to mysteriously show up in empirical analyses. The CLT is the reason why. Note that even though the original distribution didn't look like a normal distri-

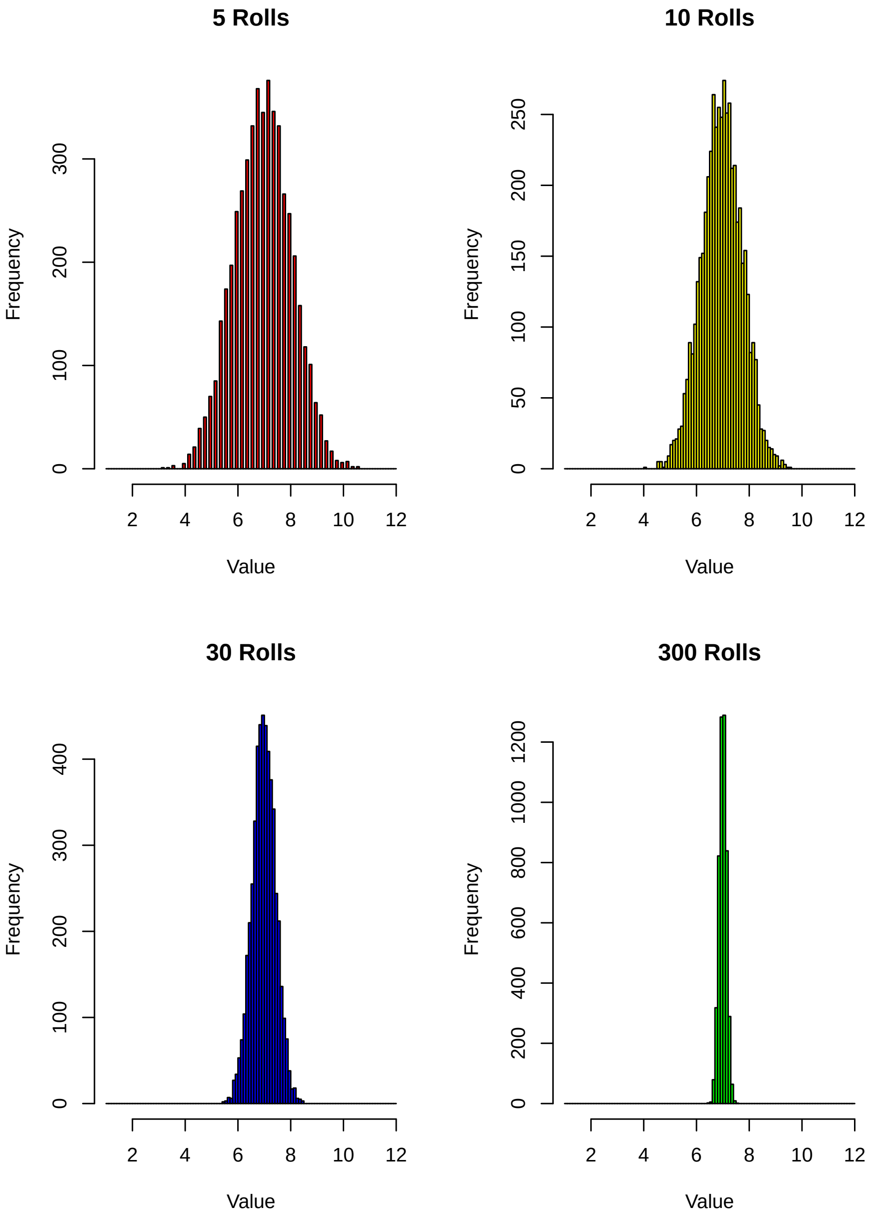


Figure 4.1: Sampling Distributions

bution at all, you still can construct sampling distributions that appear normal. This holds regardless of the initial population distribution (check out the video about rabbits and dragons on the course website if you don't believe me).

2. Sampling distributions become *more* normal and have a lower standard deviation when the sample size gets bigger. Notice that as the sample size goes up, the distributions become narrower. This means that when there is a big sample size there is a very low probability that your going to see sample averages near 2 or 12. This should make sense: If you roll two dice 300 times and take the average, there is no way you are going to record a sample average of 2 unless you roll *snake eyes* 300 times in a row. As the sample size increases, the *extreme* events start getting diluted. This reduces the standard deviation of the sampling distribution.
3. The sampling distributions (for $n \geq 30$) are distributed normal with mean μ and standard deviation $\sigma/\sqrt{n}$. Technically this means that your random sample will produce a random outcome (a sample mean) which we denote $\bar{X}$.

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

You can see these two properties in the four sampling distributions illustrated above. All four sampling distributions are centered around 7, which is the population mean. As sample size gets larger, the sampling distributions get *narrower* around the population mean. This illustrates why a larger sample has a better shot at becoming a better representation of the population.

4.3 Application 2: A Sampling Distribution with an Unknown Population

In most applications, we will not be as lucky as in the first application and we will know nothing about the underlying population. We won't know the distributional properties of the population, we won't know any of the population parameters… nothing. The beauty of the CLT is that this doesn't matter. We can still apply the CLT to set the stage for statistical inference.

In light of school closings back in 2020, the city of Philadelphia considered sending out $100 EBT cards to every student registered in public school.

A key question at the beginning of deliberation is how much would this policy cost?

- There are 352,272 families in Philadelphia, and the city has records on how many students are registered in public schools.

  - Suppose it is too costly (at the initial stage) to determine the total number of children.

- If we knew the average number of children registered per family, we can get an estimate of the cost of the policy.

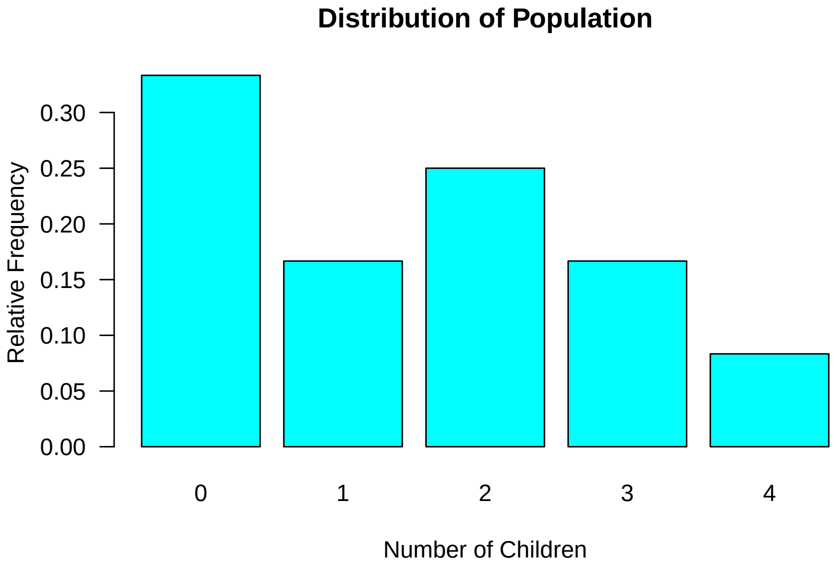
Suppose we are *omniscient*...

- The **POPULATION** average number of children per family is...

$$\mu = 1.5$$

$$\sigma = 1.38$$

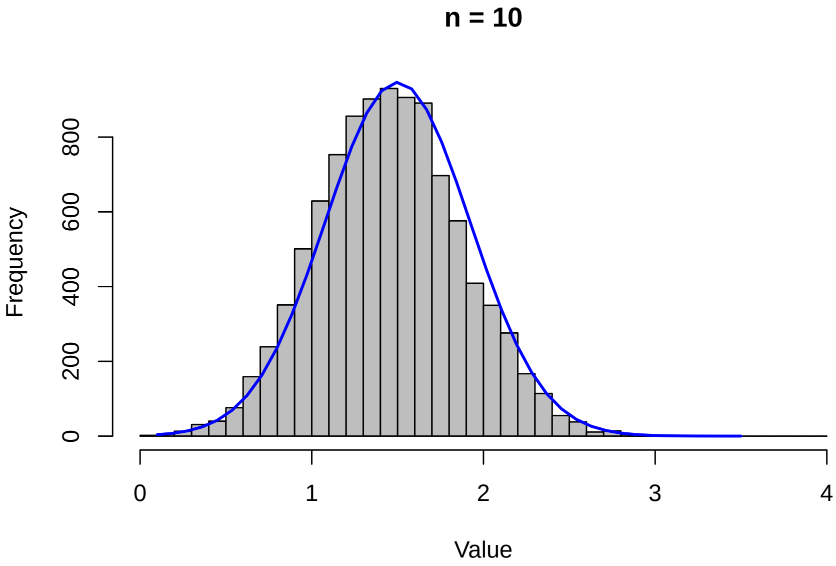
NOTE: We do not know these population parameters. I am simply stating them here so we can refer to them later for verification. In reality, we will **never** know these population parameters. That's why we need inferential statistics.

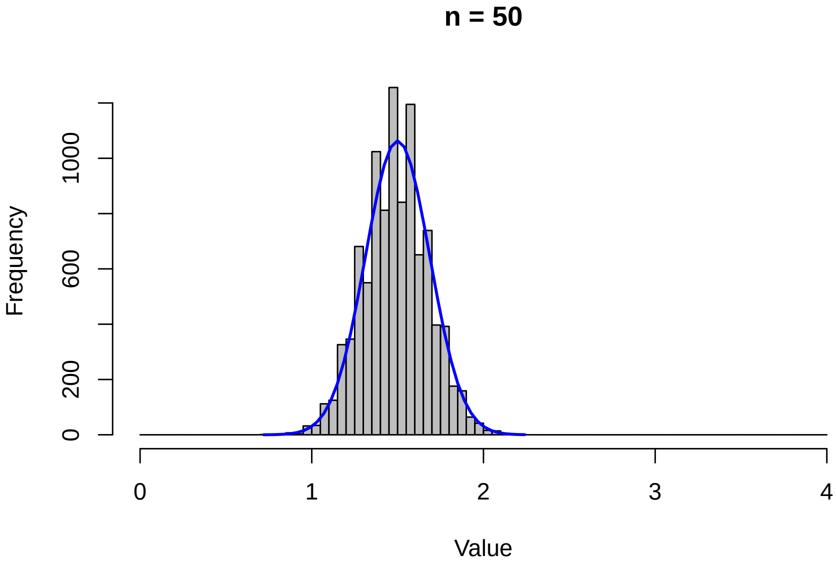


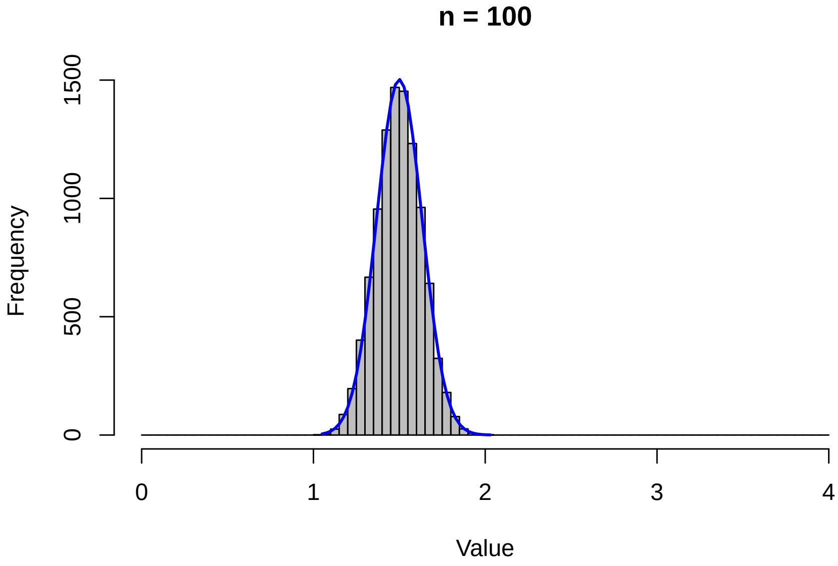
4.3.1 The Sample

- Since it is too costly to examine the entire population (at the initial stage), we draw a single sample.
- We use the sample to calculate sample statistics
- Since the sample is randomly drawn from the population, the sample statistics are randomly drawn from a sampling distribution.

The characteristics of the sampling distribution depends on the sample size n.







The figures above show sampling distributions of various sample sizes. Note that all of these distributions are centered around the same number (of 1.5), and the dispersion around the mean is getting smaller as n is getting larger. In other words, the standard deviation $\sigma/\sqrt{n}$ is getting smaller because n is getting larger (while σ remains unchanged).

4.4 The Punchline

Once you determine a sample size (n), you get **one random draw** from the appropriate sampling distribution.

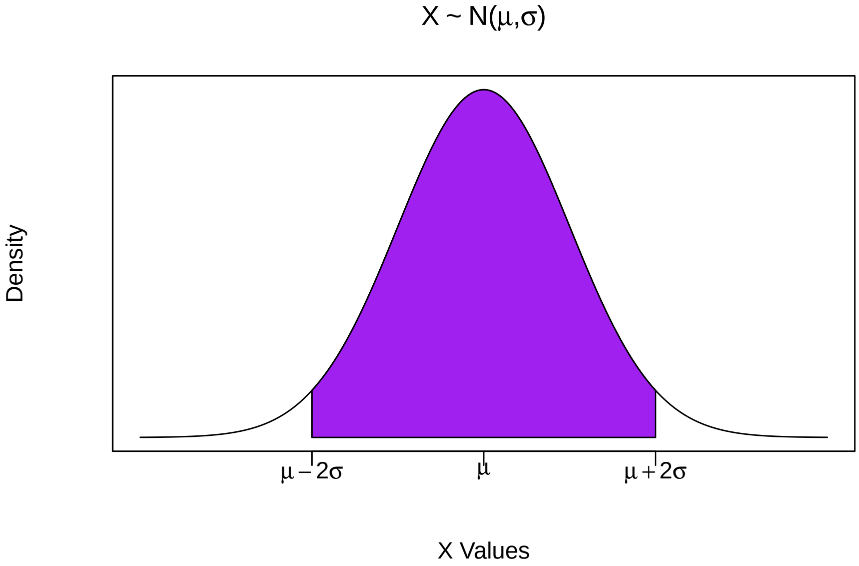
- The distribution is approximately *normal*
- The mean is μ
- The standard deviation $\sigma/\sqrt{n}$

What does this buy us? The answer is *everything* if we want to apply any form of *confidence* (i.e., stating a probability of occurring).

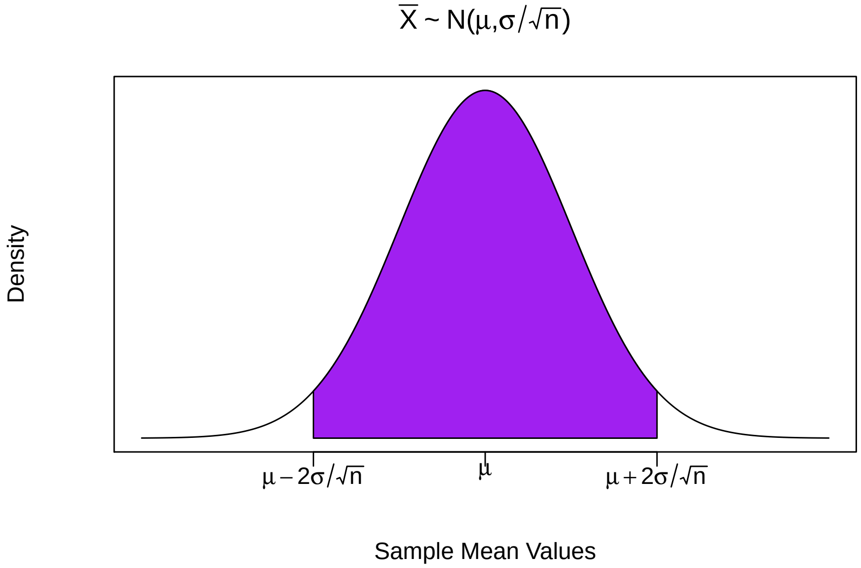
The reason is that the normal distribution has a lot of useful properties.

1. The distribution is symmetric. The shape of the distribution to the right of the mean is identical to the shape of the distribution to the left of the mean.
2. Approximately 95% of all possible outcomes are within 2 standard deviations of the mean.

To illustrate these two properties, consider the generic normal distribution illustrated below. You can easily see the symmetry of the distribution, while the shaded area represents 95% of the distribution. In probability terms, 95% of the area of the probability distribution means that there is a 95% chance of drawing a value within this range.



So what does this really buy us? Consider the application above about the Philadelphia policy where we would have in reality have no idea what the population parameters (μ, σ) are, or what the population distribution even looks like. However, the CLT says that if we decide on a sample size n, then we will draw from a sampling distribution that is a normal distribution with mean mu and standard deviation $\sigma/\sqrt{n}$.



So what we know is that once we draw a random sample and construct a sample mean, we can say with 95% confidence that that sample mean was drawn from the shaded region of the above distribution. We know what the sample mean value is because we just calculated it. What we don't know is what $/mu$ is. However, we can construct a probabilistic range (a *confidence interval*) around where we think this population parameter lies. This is where we are going next.

Chapter 5

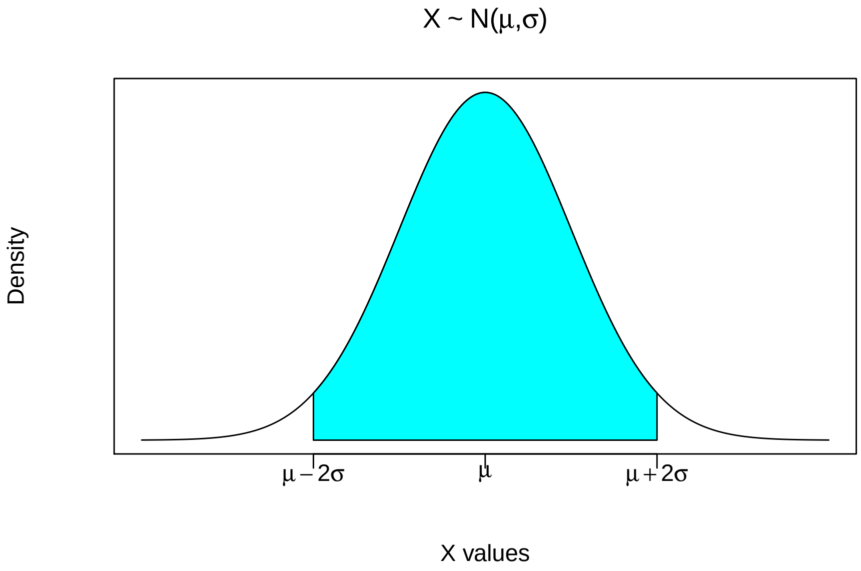
Confidence Intervals

With the concept of the Central Limit Theorem (CLT) under our belts, we can discuss our first application of statistical inference. The main concept is that when a statistician discusses *confidence*, they are actually saying how likely something is going to happen. In other words, 95% confidence in a statement means that something is going to happen 95 out of 100 replications. We only get one shot (not 100 replications), so it is the likelihood that it's going to happen in this one shot.

In order for us to get into statistical inference, we need to first have a refresher on probability. Thanks to the CLT, our probability calculations are going to come from the normal probability distribution.

5.1 A Refresher on Probability

Suppose I give you a random variable (X) and tell you that this random variable comprises a normal distribution with an *arbitrary* mean equal to μ and an *arbitrary* standard deviation equal to σ. We can denote this generally as $X \sim N(\mu, \sigma)$ and we can draw this generally as



This picture is the normal probability density of this random variable. It is very much like a histogram, only we can consider a continuum of possible numbers (i.e., unlimited number of histogram bins). A normal probability density has several useful properties.

1. It is centered at the mean. It is a **symmetric** distribution with 50% of the probability being on either side of the mean.

2. Like all probability distributions, it must add up to 1 (or 100%). This is like saying that the probability of reaching into a hat full of numbers and pulling out a number between positive and negative infinity is equal to 100%.

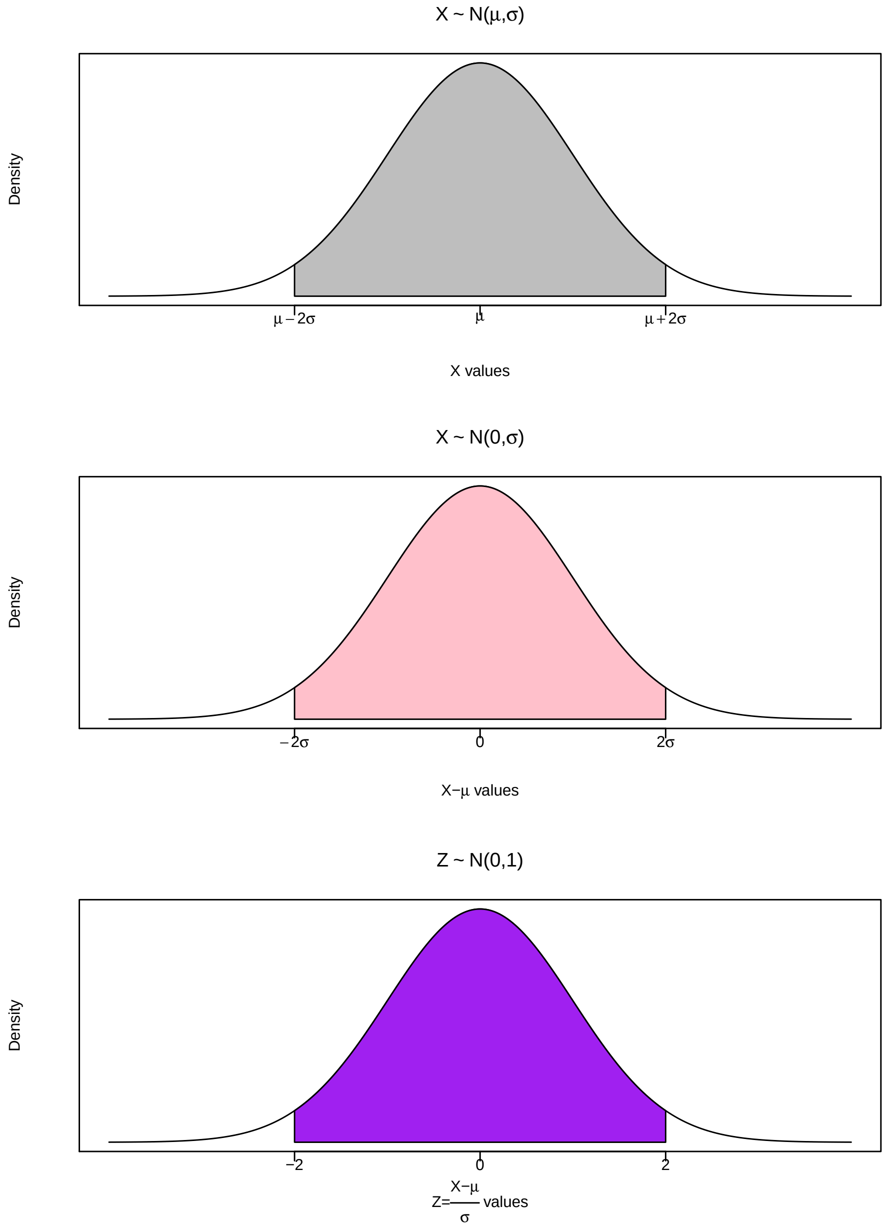
3. A normal distribution has the nice property that approximately 95% of the density area is between two standard deviation above and below the mean. This is the shaded area in the above figure. It roughly states that if you reached into a bowl full of numbers that comprised this distribution, then you have a 95% chance of pulling out a number between $\mu - 2\sigma$ and $\mu + 2\sigma$.

This is very useful, but for our purposes we need to take this arbitrary normal distribution and transform it into a **standard normal distribution**. We do this by applying what is known as a Z-transformation:

$$Z = \frac{X - \mu}{\sigma}$$

The figure below illustrates how this transformation changes an otherwise arbitrary normal distribution. The top figure is the arbitrary random variable with a mean of μ and a standard deviation of σ ($X \sim N(\mu, \sigma)$). The second figure shows what happens to the distribution when we subtract the mean from every number in the distribution. This effectively shifts the distribution such that it is now centered around zero, so we now have a normally distributed random variable with a mean of zero and a standard deviation of σ ($X - \mu \sim N(0, \sigma)$). The third figure shows what happens when we divide every number in the distribution by σ. Recall that σ is a positive number that can be greater or less than one. If you divide a number by a number less than one then the number gets bigger. If you divide a number by a number greater than one then the number gets smaller. This means that dividing every number by σ will either increase or decrease the dispersion of values such that the standard deviation is equal to one. A normally distributed random variable with a mean of zero and a standard deviation is said to be a standard normal random variable ($Z \sim N(0, 1)$).

Note that this transformation shifts the distribution, but **does not** change its properties. This was done on purpose to get you to see that a standard normal transformation shifts the mean and alters the dispersion, but does not change the facts that the distribution is still symmetric, still adds to one, and still has the property that 95% of the probability area is between 2 standard deviations to the right and left of the mean.



What does this transformation do? It takes a normally distributed random variable with arbitrary μ and σ, and transforms the distribution into one with mean 0 and standard deviation 1.

Why is this useful? It can easily be used for numerical probability calculations - but this isn't as useful nowadays since we have computers. However, this transformation will be essential to put the normal distribution on the same level as other distributions we will soon encounter.

5.1.1 Application 1

Suppose there exists a bowl of numbered cards. The numbers on these cards comprises a normal distribution where the mean value is 5, and the standard deviation is 3.

$$X \sim N(5,3)$$

We now have everything we need to calculate the probability of any outcome from this data-generating process. For example, suppose we wanted to determine the probability of reaching into this bowl and picking a number between 2 and 3. In probability terms:

$$Pr(2 \leq x \leq 3)$$

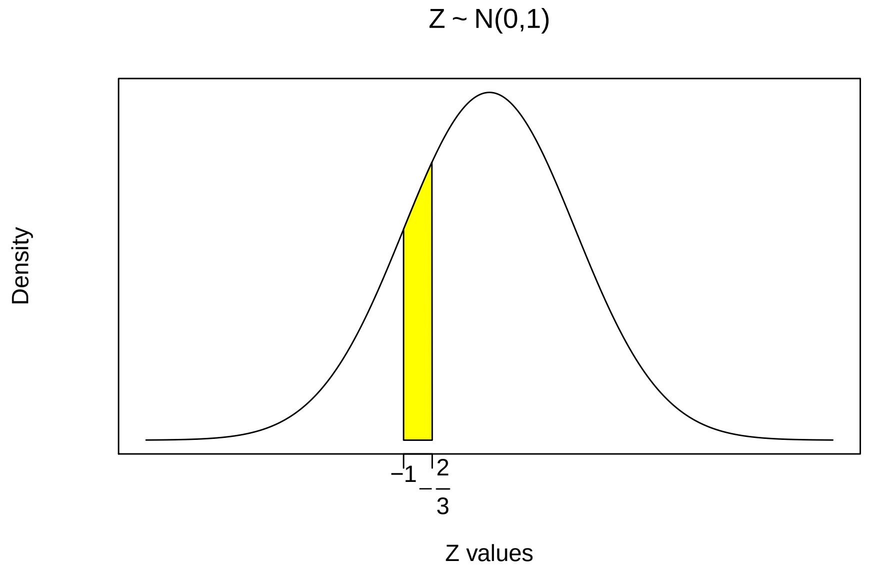
1. First we perform a standard normal transformation $Z = \frac{X-\mu}{\sigma} = \frac{X-5}{3}$, so our probability question gets transformed as well:

$$Pr(2 \leq x \leq 3) = Pr\left(\frac{2-5}{3} \leq \frac{x-5}{3} \leq \frac{3-5}{3}\right)$$

This delivers the same probability question, only in standard normal terms:

$$Pr(2 \leq x \leq 3) = Pr\left(-1 \leq z \leq -\frac{2}{3}\right)$$

2. Next we illustrate exactly what this probability question looks like in our distribution. In other words, indicate what *slice* of the distribution answers the probability question. This slice is illustrated in the figure below by shading in the probability area of the distribution between -1 and $-\frac{2}{3}$.



3. Finally, we calculate the probability in R. Now this is where the illustration above will help get us organized, because we can exploit the distributional properties of symmetry and the distribution summing to one. This is important because we can use R to calculate the same number in several different ways. All of these routes to the answer are acceptable, so we will go through them all here.

First thing to do is introduce you to the R command "pnorm"

```
pnorm(q, mean = 0, sd = 1, lower.tail = TRUE)
```

The command requires a number (*quantity*) for the variable q. It will then use a normal distribution with a mean of 0 and a standard error of 1 (*by default*) and calculate the area to the **left of the number q**. Note that this is the default action which is given by "lower.tail = TRUE". If you want to turn off this default action, then you need to set "lower.tail = FALSE" and the command will calculate the area to the **right of the number q**. For example, we can calculate $Pr(z \leq -1)$ or the area to the left of -1.

```
pnorm(-1)
```

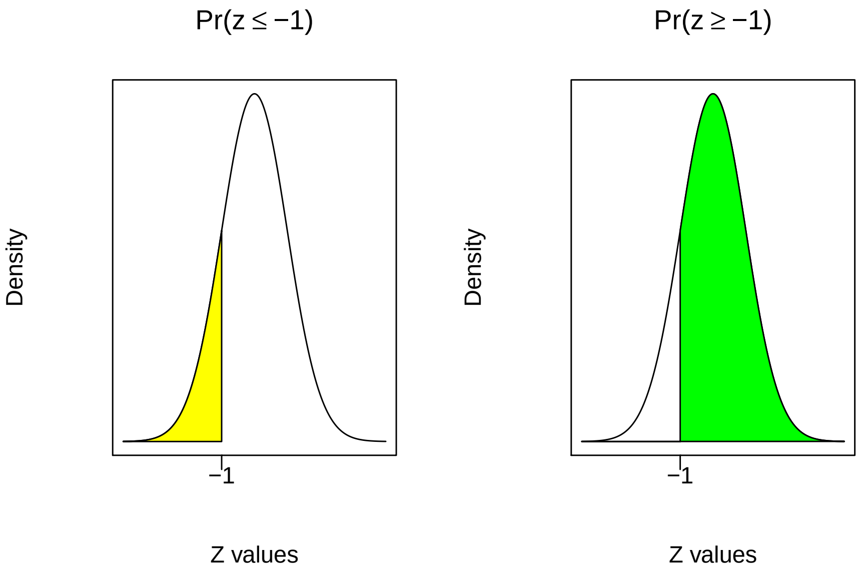
```
## [1] 0.1586553
```

We could also calculate $Pr(z \geq -1)$ or the area to the right of -1.

```
pnorm(-1,lower.tail = FALSE)
```

```
## [1] 0.8413447
```

These two probability areas sum to 1 (as they should), and are illustrated below. The left figure illustrates that 15.9% of the area under the distribution is to the left of -1, so you have a 15.9% chance of picking a number less than or equal to -1. Conversely, the right figure illustrates that 84.1% of the area under the distribution is to the right of -1, so you have a 84.1% chance of picking a number greater than or equal to -1.



Now that we know how R likes to calculate probabilities, we can use it to determine $Pr(-1 \leq z \leq -\frac{2}{3})$ which is the shaded slice of the distribution in the previous figure.

1. Using the *default* setting: suppose you want to calculate all of the probabilities using the default setting of calculating areas to the left. The shaded slice of the distribution is then the difference between the area to the left of $-\frac{2}{3}$ and the area to the left of -1.

$$Pr\left(-1 \leq z \leq -\frac{2}{3}\right) = Pr\left(z \leq -\frac{2}{3}\right) - Pr\left(z \leq -1\right)$$

```
pnorm(-2/3)-pnorm(-1)
```

```
## [1] 0.09383728
```

2. Removing the *default* setting. If you want to calculate probabilities from the right (which might come in handy), then the same slice of the distri-

bution is the difference between the area to the right of -1 and the area to the right of $-\frac{2}{3}$.

$$Pr\left(-1 \leq z \leq -\frac{2}{3}\right) = Pr(z \geq -1) - Pr\left(z \geq -\frac{2}{3}\right)$$

```
pnorm(-1,lower.tail = FALSE)-pnorm(-2/3,lower.tail = FALSE)
```

```
## [1] 0.09383728
```

3. Exploiting that the area sums to 1. Yet another way to arrive at the same answer is to calculate the area to the left of -1, the area to the right of $-\frac{2}{3}$, and arrive at the slice by subtracting these areas from 1.

$$Pr\left(-1 \leq z \leq -\frac{2}{3}\right) = 1 - Pr(z \leq -1) - Pr\left(z \geq -\frac{2}{3}\right)$$

```
1-pnorm(-1)-pnorm(-2/3,lower.tail = FALSE)
```

```
## [1] 0.09383728
```

As you can see, each procedure delivers the same answer - you have a 9.4% chance of picking a number from a standard normal distribution between -1 and $-\frac{2}{3}$.

Note that this is the same answer to the original question (before we transformed the distribution). The take away from this exercise is that there are plenty of straightforward ways of calculating probabilities in R, and we will be making a fair amount of use of them.

$$Pr\left(-1 \leq z \leq -\frac{2}{3}\right) = 0.094$$

5.1.2 Application 2

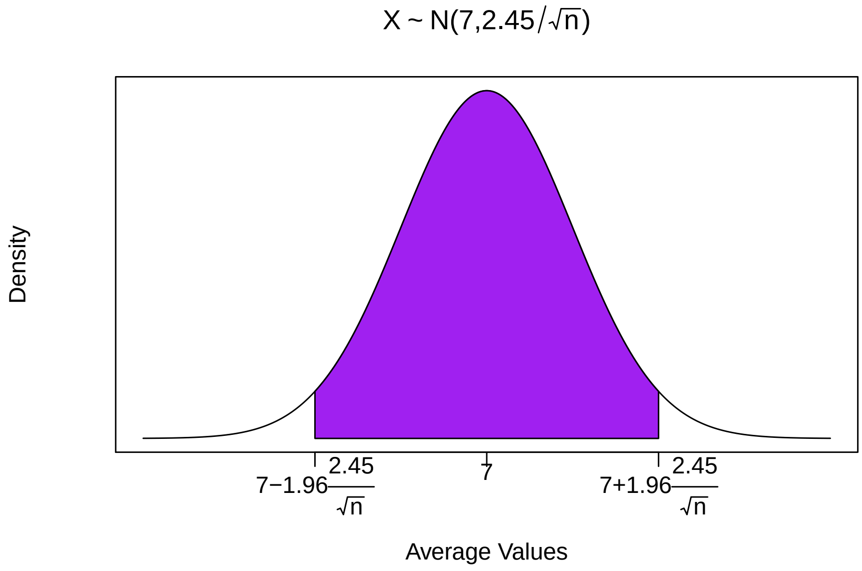
Let us look deeper into our dice example. In particular, if I were to roll two fair dice a n number of times and calculated the average, what range of values should I expect to see?

Recall that the distribution of the population has a mean of 7 and a standard deviation of 2.45. This means that for $n \geq 30$, the *sampling distribution* is normal and given by

$$\bar{X} \sim N\left(7, \; \frac{2.45}{\sqrt{n}}\right)$$

Recall that in this (rare) example, we *know* the population parameters. Therefore, we can build a range where we expect sample averages to reside.

```r
X <- seq(-4,4,0.01)
Y <- dnorm(X)

plot(X,Y,type="n",xlab="Average Values",ylab = "Density",
     yaxt = "n", xaxt = "n", main =  TeX('$X \\sim N(7,2.45 / \\sqrt{n})'))
xtick<-seq(-2, 2, by=2)
axis(side=1, at=xtick, labels = FALSE)
text(x=xtick, par("usr")[3],
     labels = c(TeX('$7-1.96\\frac{2.45}{\\sqrt{n}}$'),7,
                TeX('$7+1.96\\frac{2.45}{\\sqrt{n}}$')),
     pos = 1, xpd = TRUE)
i <- X >= -2 & X <= 2
lines(X, Y)
polygon(c(-2,X[i],2), c(0,Y[i],0), col="purple")
```



So if we collected a sample on $n = 100$, meaning we rolled two dice 100 times, recorded the total each time, and calculated the mean value, then...

$$Pr\left(7 - 1.96\frac{2.45}{\sqrt{100}} \leq \bar{X} \leq 7 - 1.96\frac{2.45}{\sqrt{100}}\right) = 0.95$$

```
Z = qnorm(0.975,lower.tail = FALSE)
n = 100
mu = 7
sigma = 2.45

(LFT = mu - Z * sigma / sqrt(n))
```

```
## [1] 7.480191
```

```
(RHT = mu + Z * sigma / sqrt(n))
```

```
## [1] 6.519809
```

$$Pr(6.52 \leq \bar{X} \leq 7.48) = 0.95$$

This means that with 95% confidence, the single outcome from your *experiment* will be within 6.52 and 7.48 if the die you are rolling are in fact fair.[1]

As stated earlier, this example is rare because we know the population parameters. When we don't, we reverse engineer the probability statement so we can take what we know (the sample statistics) and use them to say something about what we don't. This is known as a confidence interval.

5.2 Deriving a Confidence Interval

Recall when we randomly draw a sample from a sampling distribution and use it to calculate a sample mean ($\bar{X}$), we essentially state that a sample mean is a random variable. Since the CLT states that the sampling distribution is a normal distribution, then this further states that the sample mean is a normally distributed random variable with a mean of μ and a standard deviation of $\frac{\sigma}{\sqrt{n}}$.

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

We can apply our standardization trick so we have

$$Z = \frac{\bar{X} - \mu}{(\sigma/\sqrt{n})}$$

[1] Note that the code used a command called *qnorm*. This gets described further below.

Suppose we have a sample of size n, calculated a sample mean $\bar{X}$, and that we *know* the population standard deviation σ (more on this later). If we know these values, then we can use the standard normal distribution as well as the normalization above to draw statistical inference on the population parameter μ.

What can we say about μ*?*

- Since our sample has the same characteristics as the population, we would like to say $\mu = \bar{X}$ (i.e., $Z = 0$), but this is not likely. Recall the dice example discussed earlier, while 7 (the population mean) is the most likely average of a sample, there is a larger likelihood of a sample average *close* to, but not exactly equal to the population mean.
- Since $\bar{X}$ is a single draw from a normal distribution, we can construct a *probabilistic* range around μ. This range requires an arbitrary level of confidence $(1 - \alpha)$ - which provides bounds for the Z distribution (i.e., it gives us the area under the curve).

We therefore start with a probabilistic statement using a standard normal distribution:

$$Pr\left(-Z \leq \frac{\bar{X} - \mu}{(\sigma/\sqrt{n})} \leq Z\right) = 1 - \alpha$$

This states (in general terms) that the probability of realizing a value of $\frac{\bar{X}-\mu}{(\sigma/\sqrt{n})}$ drawn from a standard normal distributed random variable to be between the values -Z and Z is equal to $1 - \alpha$. To put this into context, suppose I set $\alpha = 0.05$ so $1 - \alpha = 0.95$ implies that I am looking for something that will occur with 95% probability. Recall the normal distribution has the nice property that 95% of the probability space is approximately between two standard deviations above and below the mean. By approximately, I mean it is actually 1.96 and not 2.[2]

Finding these numbers requires another R command: qnorm.

```
qnorm(p, mean = 0, sd = 1, lower.tail = TRUE)
```

Just like how pnorm takes a quantity and returns a probability, qnorm takes a probability and returns a quantity.

```
qnorm(0.025)
```

```
## [1] -1.959964
```

```
qnorm(0.975,lower.tail=FALSE)
```

[2] This explains the 1.96 used in the dice application above.

```
## [1] -1.959964
```

$$Pr\left(-1.96 \leq \frac{\bar{X}-\mu}{(\sigma/\sqrt{n})} \leq 1.96\right) = 0.95$$

Now back to our statement for a general α and Z:

$$Pr\left(-Z \leq \frac{\bar{X}-\mu}{(\sigma/\sqrt{n})} \leq Z\right) = 1-\alpha$$

Given that the only thing we **do not** know is the population parameter μ, we can rearrange the inequalities inside the probability statement to deliver a probabilistic range where we think this parameter will reside.

$$Pr\left(-Z \leq \frac{\bar{X}-\mu}{(\sigma/\sqrt{n})} \leq Z\right) = 1-\alpha$$

$$Pr\left(-Z\frac{\sigma}{\sqrt{n}} \leq \bar{X}-\mu \leq Z\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

$$Pr\left(-\bar{X}-Z\frac{\sigma}{\sqrt{n}} \leq -\mu \leq -\bar{X}+Z\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

$$Pr\left(\bar{X}-Z\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}+Z\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

This statement is a **confidence interval**, which can be written concisely as

$$\bar{X}-Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}+Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$$

or

$$\bar{X} \pm Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$$

It explicitly states that given the characteristics of the sample $(\bar{X}, n, \sigma)$ and an arbitrary level of confidence that gives us the probability limits from the standard normal distribution $(Z_{\frac{\sigma}{2}})$, then we can build a range of values where we can state with $(1 - \alpha) * 100$ confidence that the population parameter resides within.

Welcome to statistical inference!

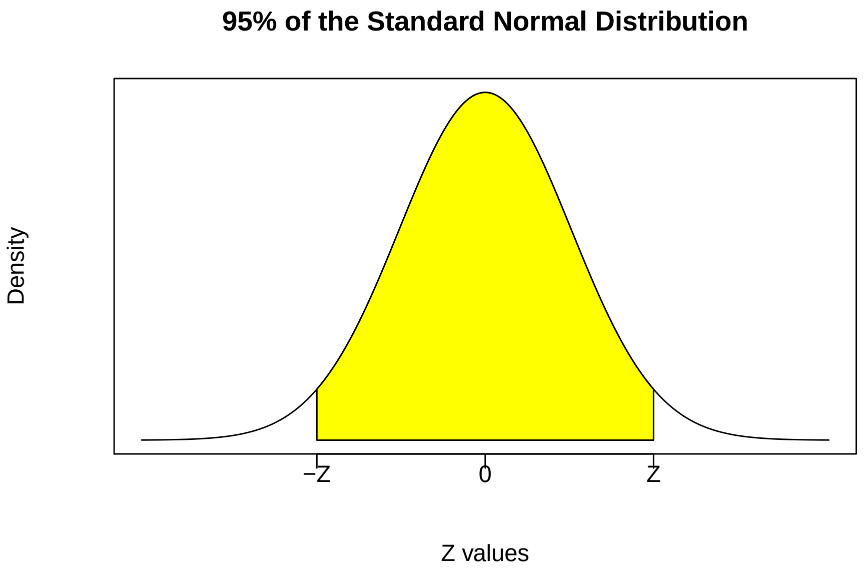
5.2.1 Application 3

A paper manufacturer produces paper expected to have a mean length of 11 inches, and a known standard deviation of 0.02 inch. A sample of 100 sheets is selected to determine if the production process is still adhering to this length. If it isn't, then the machine needs to go through the costs of being taken off line and recalibrated. The sample was calculated to have a average value of 10.998 inches.

$$\bar{X} = 10.998, \quad n = 100, \quad \sigma = 0.02$$

Construct a 95% confidence interval around the average length of a sheet of paper in the population.

1. Since we want 95% confidence, we know that $\alpha = 0.05$ and we need the critical values from a standard normal distribution such that 95% of the probability distribution is between them. These critical values were calculated previously to -1.959964 and 1.959964 and are illustrated below. Note that since the shaded region is 95% of the *central* area of the distribution, we are chopping of 5% of the *total* area from both tails combined. That means 2.5% is chopped off of each tail.



2. Now using the *positive* critical Z value in our confidence interval equation, we have:

$$\bar{X} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

$$10.998 \pm 1.96 \frac{0.02}{\sqrt{100}}$$

Using R for the calculations:

```
Xbar = 10.998
n = 100
Sig = 0.02
alpha = 0.05
Z = qnorm(alpha/2,lower.tail = FALSE)
(left = Xbar - Z * Sig / sqrt(n))
```

```
## [1] 10.99408
```

```
(right = Xbar + Z * Sig / sqrt(n))
```

```
## [1] 11.00192
```

$$10.99408 \leq \mu \leq 11.00192$$

Conclusion: I am 95% confident that the mean paper length in the population is somewhere between 10.99408 and 11.00192 inches.

Note that any value within this range is equally likely!

5.2.2 What if we want to change confidence?

If we want to increase the confidence of our statement to 99% or lower it 90%, then we change α and calculate a new critical Z value. Everything else stays the same.

```
alpha = 0.01 # increase confidence to 99%
Z = qnorm(alpha/2,lower.tail = FALSE)
(left = Xbar - Z * Sig / sqrt(n))
```

```
## [1] 10.99285
```

```
(right = Xbar + Z * Sig / sqrt(n))
```

```
## [1] 11.00315
```

```
alpha = 0.10 # decrease confidence to 90%
Z = qnorm(alpha/2,lower.tail = FALSE)
(left = Xbar - Z * Sig / sqrt(n))
```

```
## [1] 10.99471
```

```
(right = Xbar + Z * Sig / sqrt(n))
```

```
## [1] 11.00129
```

What happens to the size of the confidence interval when we increase our *confidence*?

5.3 What to do when we do not know σ

In most instances, if we know nothing about the population parameter μ then we know nothing about any of the other parameters (like σ). In this case, we are forced to use our best guess of σ. Since we are assuming that our sample has the same characteristics of the population, then our best guess for σ is the sample standard deviation S.

Put plainly, we substitute the statistic (S) for the population parameter (σ). Because S is an estimate of σ, this will slightly change our probability distribution. In particular, If $\bar{X}$ is normally distributed as per the CLT, then a standardization using S instead of σ is said to have a t distribution with $n-1$ degrees of freedom

$$t = \frac{\bar{X} - \mu}{(S/\sqrt{n})}$$

Note that this looks almost exactly like our Z transformation, only with S replaced for σ. However, this statistic is said to be drawn from a distribution with $n-1$ degrees of freedom. We mentioned degrees of freedom before, and we stated that we lose a degree of freedom when we build statistics on top of each other. In other words, we lose a degree of freedom for every statistic we use to calculate another statistic. Consider the standard deviation equation needed to calculate S.

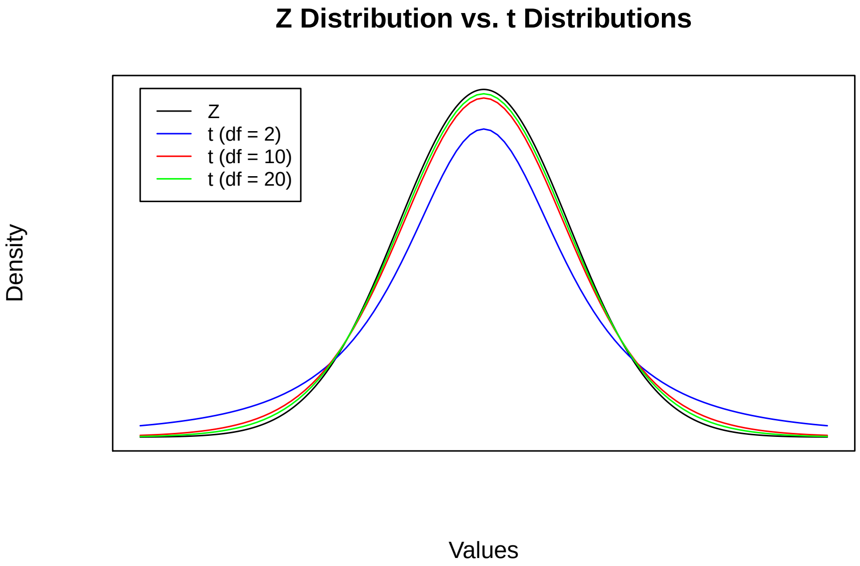
$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

The equation states that the sample mean ($\bar{X}$) is used to calculate the sample standard deviation. This means one statistic is used to calculate a subsequent statistic… and that is why we lose one degree of freedom.

5.3.1 t distribution versus Z distribution…

A t distribution and Z distribution have very much in common: they are both symmetric, both centered at a mean of 0, and both sum to one (because they are both probability distributions). The main difference is that a t-distribution has *fatter tails* than a Z distribution, and the fatness of the tails depends on the degrees of freedom (which in turn depends on the sample size).

The figure below compares the standard normal (Z) distribution with several t distributions that differ in degrees of freedom. Notice that tail thickness of the t distributions are inversely related to sample size. As the the degrees of freedom get larger (i.e., the larger the sample size), the closer the t distribution gets to the Z distribution. This is because as n gets larger, S becomes a better estimate of σ.



$$\bar{X} - Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

$$\bar{X} - t_{(\frac{\alpha}{2}, df=n-1)} \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{(\frac{\alpha}{2}, df=n-1)} \frac{S}{\sqrt{n}}$$

In a nutshell, the only difference encountered when not knowing σ is that we have a slightly different probability distribution (which requires knowing the degrees of freedom and uses a different R command). The new R commands are `qt` and `pt` which requires degrees of freedom but otherwise has all of the same properties of `qnorm` and `pnorm` discussed above.

```
pt(q, df, lower.tail = TRUE)
qt(p, df, lower.tail = TRUE)
```

5.3.2 Application 4

Suppose you manage a call center and just received a call from Quality Control asking for the *average call length* at your facility. They are asking for the average call length in the population, so the best you can do is provide a confidence interval around this population parameter. You select a random sample of 50 calls from your facility and calculate a sample average of 5.8 minutes and a **sample standard deviation** of 2.815 minutes.

$$\bar{X} = 5.8, \quad n = 50, \quad S = 2.815$$

Calculate a 95% confidence interval around the population average call length.

```
Xbar = 5.8
n = 50
df = n-1
S = 2.815
alpha = 0.05
t = qt(alpha/2,df,lower.tail = FALSE)
(left = Xbar - t * S / sqrt(n))
```

```
## [1] 4.999986
```

```
(right = Xbar + t * S / sqrt(n))
```

```
## [1] 6.600014
```

With 95% confidence, the population average call length is between 5 minutes and 6.6 minutes.

As before, if we want to change our level of confidence then we change α and recalculate the t statistic. Notice that the relationship remains that a lower confidence level delivers a narrower confidence interval.

```
alpha = 0.01 # increase confidence to 99%
t = qt(alpha/2,df,lower.tail = FALSE)
(left = Xbar - t * S / sqrt(n))
```

```
## [1] 4.733108
```

```
(right = Xbar + t * S / sqrt(n))
```

```
## [1] 6.866892
```

```
alpha = 0.10 # decrease confidence to 90%
t = qt(alpha/2,df,lower.tail = FALSE)
(left = Xbar - t * S / sqrt(n))
```

```
## [1] 5.132563
```

```
(right = Xbar + t * S / sqrt(n))
```

```
## [1] 6.467437
```

5.4 Determining Sample Size

It was previously stated that the sample size should always be as big as possible in order to deliver the most *precise* conclusions. This isn't always a satisfactory answer, because collecting observations might be possible (but costly).

How big should n be?

Selecting an appropriate sample size could be determined by many constraints

- budget, time, … (things that cannot really be dealt with statistically)
- *acceptable* sampling error (we can deal with this)

Recall our confidence interval equation:

$$\bar{X} - Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

or

$$\bar{X} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

The term $Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ is one-half the width of the confidence interval. This is called the *sampling error* (or *margin of error*).

$$e = \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

In our previous exercises, we were given a sample size (n) and used our calculations to determine the width of the confidence interval ($2e$). If we instead wanted to fix the margin of error, then we can let the above identify determine how big our sample size needs to be.

$$n = \left(\frac{Z_{\frac{\alpha}{2}} \sigma}{e} \right)^2$$

Going back to our call center example, suppose that quality control demanded a 95% confidence interval with a 15 second (0.25 minute) margin of error. This means that the 95% confidence interval can only be 0.5 minutes wide. How many calls need to be in the sample?

```
alpha = 0.05
Z = qnorm(alpha/2,lower.tail = FALSE)
Xbar = 5.8
Sig = 2.815
```

```
e = 0.25

(n = (Z*Sig/e)^2)
```

```
## [1] 487.0493
```

```
# Round up since you can't have a fraction of an observation
ceiling(n)
```

```
## [1] 488
```

Our analysis indicates that if you want this particular a margin of error, then you will need to collect a sample of 488 calls.

You might have noticed that we did something a bit incorrect in the last exercise. We specified a Z distribution and called the sample standard deviation σ. Note that only in these sort of applications that determine a sample size is this permissible. The reason is because a sample standard deviation obviously depends on the sample in question. We therefore need to assume that the standard deviation is fixed when calculating the sample size (even though this isn't the case). Once you determine a sample size, then you collect a sample, calculate the sample standard deviation, and calculate the appropriate confidence interval. The margin of error should be reasonably close to what was required.

5.5 Concluding Applications

5.5.1 Light Bulbs (Last Time)

Let's go back one last time to our light bulb example

```
load("C:/Data/MBA8350/Lightbulb.Rdata")
(n = length(Lifetime))
```

```
## [1] 60
```

```
(Xbar = mean(Lifetime))
```

```
## [1] 907.5552
```

```
(S = sd(Lifetime))
```

```
## [1] 78.96741
```

We have the following information from our sample:

$$\bar{X} = 907.6, \quad n = 60, \quad S = 78.967$$

Use the above information to calculate a 95% confidence interval around the population average lifespan of the light bulbs you have left to sell. You can put this on information on the box!

```
alpha = 0.05
df = n-1
t = -qt(alpha/2,df)

(left = Xbar - t * S / sqrt(n))
```

```
## [1] 887.1557
```

```
(right = Xbar + t * S / sqrt(n))
```

```
## [1] 927.9546
```

5.5.2 Returning to the Philadelphia School Policy Application

Let us return to the Philadelphia school policy example to provide one final discussion of a confidence interval. This application may appear redundant, but it is intended to provide an alternative approach to the confidence interval concept. It has helped students in the past, so it might do some good.

In early February 2020, the city of Philadelphia considered sending out $100 EBT cards to every student registered in public school due to the school closings brought on by the pandemic.

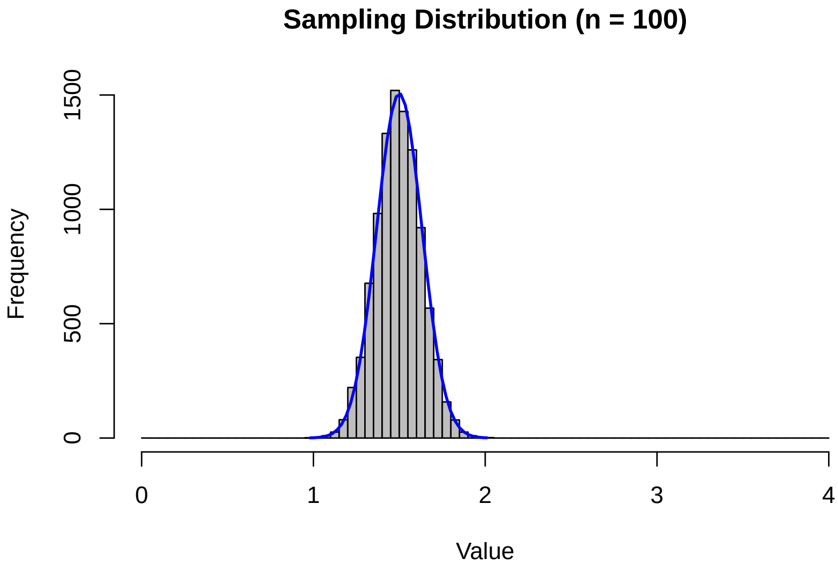
How much would this policy cost?

The Frame

There are 352,272 families in Philadelphia, and the city has records on how many students are registered in public schools.

However, suppose it is too costly (at this stage) to determine the total number of children registered in public schools. If we knew the average number of children registered per family, we can get an estimate of the cost of the policy.

Since it is too costly to examine the entire population (at the moment), we draw a single sample and use the sample to calculate sample statistics. Since the sample is randomly drawn from the population, the sample statistics are randomly drawn from a sampling distribution.



Your Sample

Once you determine a sample size (n), you get **one random draw** from the appropriate sampling distribution.

- The distribution is approximately *normal*
- The mean is μ
- The standard deviation $\sigma/\sqrt{n}$

We use this information and our sample characteristics to say something about the population parameters…

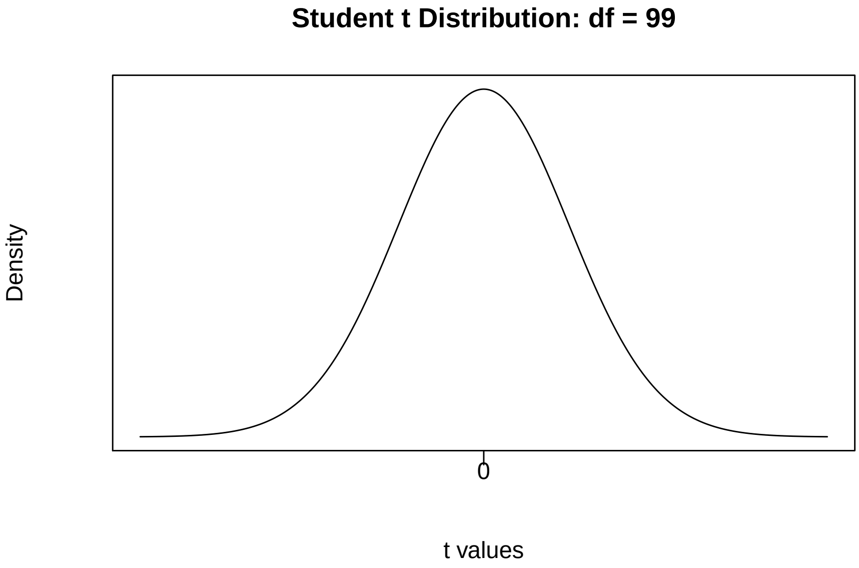
Suppose you select a sample of $n = 100$ families and calculate

$$Xbar = 1.7$$

$$S = 1.5$$

Since we have an *estimate* of the population standard deviation from our sample, our sampling distribution is now a t distribution with $n - 1 = 99$ degrees of freedom.

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{1.7 - \mu}{1.5/\sqrt{100}}$$



What we know...

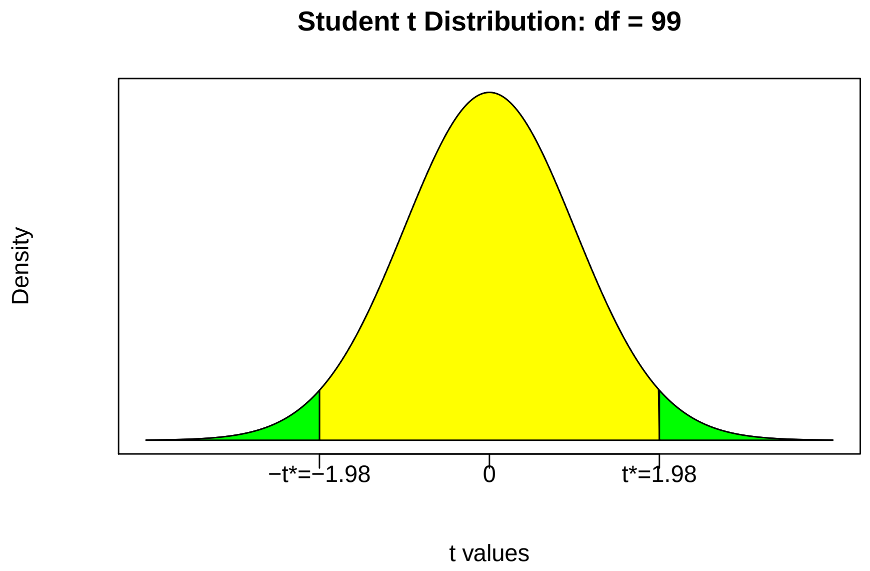
- CLT: the true population average is the central point of our sampling distribution
- We can choose an *arbitrary* level of confidence $(1 - \alpha)$ to limit where we think our statistic from a single draw will fall.

$$Pr(-t^* \leq \frac{1.7 - \mu}{1.5/\sqrt{100}} \leq t^*) = 1 - \alpha$$

Suppose we want 95% confidence $(\alpha = 0.05)$

```
(tcrit <- qt(0.05/2,99))
```

```
## [1] -1.984217
```



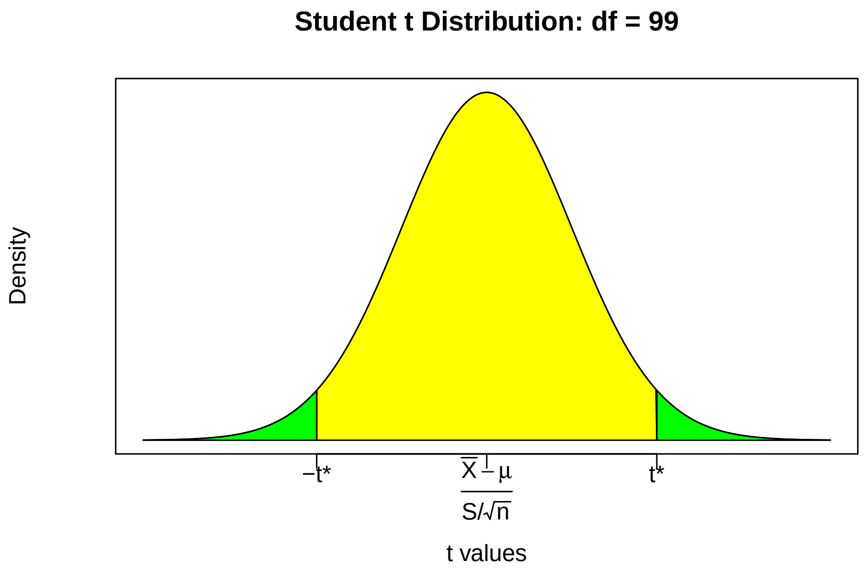
What we DON'T know...

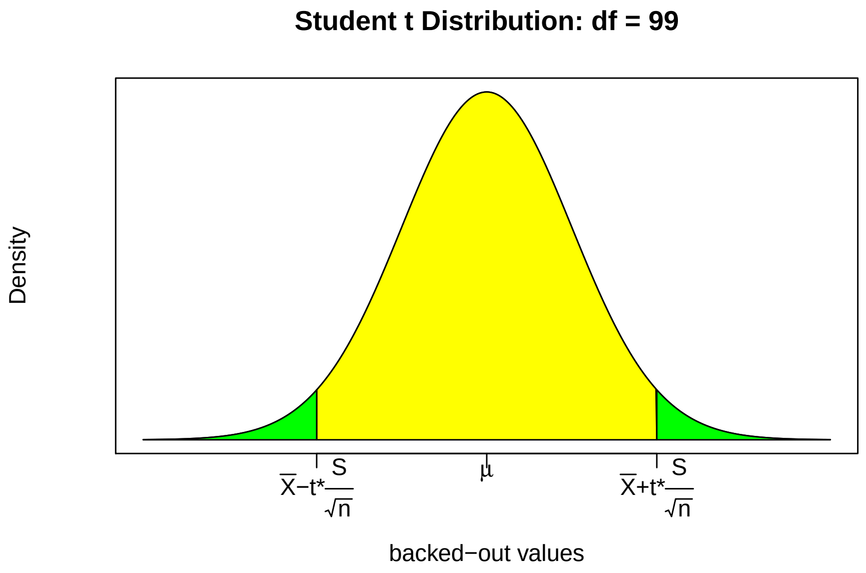
- We don't know the numerical value of μ...
- We don't know where our value of $\bar{X}$ falls in relation to μ
  - $\bar{X} = \mu$?
  - $\bar{X} > \mu$?
  - $\bar{X} < \mu$?

The fact that we don't know where $\bar{X}$ is in relation to μ is why we end up with an *interval* around where we think the population parameter resides.

$$Pr(\bar{X} - t^* \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t^* \frac{S}{\sqrt{n}}) = 1 - \alpha$$

$$Pr(1.7 - 1.98 \frac{1.5}{\sqrt{100}} \leq \mu \leq 1.7 + 1.98 \frac{1.5}{\sqrt{100}}) = 0.95$$





```
Xbar = 1.7; S = 1.5; n = 100; AL = 0.05
tcrit <- -qt(AL/2,n-1)

(LFT <- Xbar - tcrit * S / sqrt(n))
```

```
## [1] 1.402367
```

```
(RHT <- Xbar + tcrit * S / sqrt(n))
```

```
## [1] 1.997633
```

$$Pr(1.40 \leq \mu \leq 2.00) = 0.95$$

With 95% confidence, the average number of children per family is between 1.4 and 2.

Total cost is between…

$$1.4 * \$100 * 352,272 = \$49,318,080$$

and

$$2 * \$100 * 352,272 = \$70,454,400$$

Chapter 6

Hypothesis Tests

Confidence intervals determine a range where our population mean resides given the characteristics of a sample and a desired level of confidence. Recall that the population parameter can be *anywhere* within the range dictated by a confidence interval.

Hypothesis testing is a similar inferential method, but it approaches the problem from the opposite direction.

1. You start with an *unambiguous claim* on the value of the population parameter. This claim is nonarbitrary, and dictated from either theory or a past observation.

2. You test to see if the sample statistics are consistent with the claim (or refute it)

The general idea is that you begin with some *nonarbitrary* statement on what value you believe (or do not believe) the population parameter to be. You then test if the characteristics of your sample suggest that it is likely or not that a population with your proposed parameter values generated a sample similar to the one you currently have.

If this seems a bit vague at the moment, it will hopefully be more concrete soon. The main thing to keep in mind is that hypothesis tests are quite simple and structured. Once you learn how to perform one hypothesis test - you can essentially perform them all. This chapter guides you through some basic steps that once mastered - you'll have a powerful tool of statistical inference under your belt.

6.1 Anatomy of a Hypothesis Test

A hypothesis test begins with a claim about the value of particular a population parameter.

This statement takes the form of a **null hypothesis**

$$H_0 : \mu = x$$

This statement gets contrasted against an **alternative hypothesis**

$$H_1 : \mu \neq x$$

The null hypothesis (H_0) represents a belief of a population parameter that you would like to *disprove*, while the alternative hypothesis (H_1) is the opposite of the null and represents a claim you would like to show.

A hypothesis test uses the characteristics of the sample to determine if the statement about the population parameter in the null appears consistent (or inconsistent) with the characteristics of the sample. Recall that we are still under the assumption that the characteristics of the sample are similar to the **true** characteristics of the population. Therefore, if the sample characteristics are inconsistent with the statement in the null hypothesis, then you are likely to **reject the null hypothesis**. This means that the null hypothesis does not capture the **true** characteristics of the population (because the sample does). If the sample characteristics are *similar* to those stated in the null hypothesis, then you do not have evidence to reject the null and you conclude to **not reject the null hypothesis**.

In other words, if you *reject the null*, you have statistical evidence that H_1 is correct (and the null hypothesis cannot be correct). If you *do not reject the null*, you have failed to prove the alternative hypothesis. Note that failure to prove the alternative does NOT mean that you have proven the null. In other words, there IS a difference between *do not reject* and *accept*!

This distinction between *do not reject* and *accept* cannot be emphasized enough. First, if anyone concludes that they accept the null in this class - you will get marked incorrect. If you conclude to accept the null outside of this class - then people will suspect that you don't fully understand what you are talking about. Second, we can never say accept the null because it is simply too strong of a statement to make regarding a population parameter.

Suppose we believe that the population mean life span of our light bulbs is x hours. A hypothesis test will give us a specific way of testing this belief, and allows us to conclude whether or not this statement is consistent with our sample.

$$H_0 : \mu = x \quad versus \quad H_1 : \mu \neq x$$

We will discuss how to formally conduct a hypothesis tests in a bit. For now, lets compare these hypotheses with the confidence interval we calculated in the previous section.

$$887 \leq \mu \leq 928$$

Recall that our confidence interval states that with 95% confidence, the population average life span of the light bulbs is *somewhere* between 887 hours and 928 hours - meaning that any value within this range is equally likely. It also states that there is only a 5% chance that the population parameter lies outside of this range.

If we were to test that the population average lifespan was 1000 hours,

$$H_0 : \mu = 1000 \quad versus \quad H_1 : \mu \neq 1000$$

then our confidence interval would give us evidence to **reject** the null because there would be less than a 5% chance for the null to be true. The sample characteristics and the statement in the null hypothesis are therefore inconsistent.

If we were to test that the population average lifespan was 900 hours, then we will reach a different conclusion.

$$H_0 : \mu = 900 \quad versus \quad H_1 : \mu \neq 900$$

Since 900 is a value *inside* our confidence interval, then we would not have evidence to reject the null and we therefore conclude **do not reject** the null. The reason why we never say accept is that while 900 is within the confidence interval, there are also a *continuum* of other values in there. The true population mean might be 901, 900.0001, 910, etc. If you were to *accept* the null, then you are explicitly stating that the population parameter is exactly 900 - we do not have enough evidence for this.

Note that while we are seeing a clear connection between hypothesis tests and confidence intervals, hypothesis tests can get more sophisticated than this. It is therefore worthwhile to consider a formal solution methodology.

6.2 Two methods for conducting a hypothesis test (when σ is known)

We will consider two equivalent methods for conducting a hypothesis test. The first is called the *rejection region* method and is very similar to confidence intervals. The second is the *p-value* method and delivers some very useful results that we will be using for the rest of the course. We will consider these methods in turn.

6.2.1 Rejection Region Method

All hypothesis tests start with a statement of the null and alternative hypotheses. The null makes an explicit statement regarding the value of a population parameter (μ). Once this value of μ is established under the null, all hypothesis tests construct a *test statistic under the null.* In particular, assuming we know the population standard deviation σ, we can construct a *Z statistic under the null* using our familiar z-transformation:

$$Z = \frac{\bar{X} - \mu}{(\sigma/\sqrt{n})}$$

Note that we already know values from the sample ($\bar{X}$, n, σ) and we have a *hypothesized* value of μ from the null hypothesis. We can therefore directly calculate this Z-value *under the null.* This would be the value of a Z-statistic given our sample characteristics under the *assumption* that our null hypothesis is correct.

The rejection region method takes this Z-statistic under the null and sees where it falls in a standard normal sampling distribution. The sampling distribution of a test statistic is first divided into two regions…

1. A region of rejection (a.k.a., critical region) - values of the test statistic that are unlikely to occur **if the null hypothesis is true**.
2. A region of nonrejection - values of the test statistic that are likely to occur if the null hypothesis is true, so they are *consistent with the null hypothesis.*

The regions of rejection and nonrejection are identified by determining *critical values* of the test statistic. These are particular values of the sampling distribution that divides the entire distribution into rejection and nonrejection regions. This is why some people refer to the *rejection region* approach as the *critical value* approach.

Once the different regions are established, then you simply see where the calculated test statistic under the null falls. If it falls inside the rejection region, then you **reject the null** because the characteristics of the sample are *too inconsistent* with the population parameter stated inside the null hypothesis. If it falls inside the nonrejection region, then you **do not reject the null** because the characteristics of the sample are such that the population parameter stated inside the null is *possible* (but we can't say if it's necessarily true).

The Steps of a Hypothesis Test (Rejection Region Method)

1. State the null and alternative hypotheses
2. Calculate a test statistic under the null
3. Determine the rejection and nonrejection regions of a standardized sampling distribution
4. Conclude (reject or do not reject)

Application 1

Suppose a fast-food manager wants to determine whether the waiting time to place an order has changed from the previous mean of 4.5 minutes. We want to see if our sample characteristics are consistent with an average of 4.5 minutes or not. We start with a statement of the two hypotheses.

$$H_0 : \mu = 4.5 \quad versus \quad H_1 : \mu \neq 4.5$$

The null explicitly states that $\mu = 4.5$, so we can use this value to construct a test statistic using the information from our sample. Suppose that a sample of $n = 25$ observations delivered a sample mean of $\bar{X} = 5.1$ minutes. Suppose further that we know the population standard deviation of the wait time process to be $\sigma = 1.2$. We can calculate a test statistic under the null.

```
mu = 4.5
Xbar = 5.1
Sig = 1.2
n = 25

(Zstat = (Xbar - mu)/(Sig/sqrt(n)))
```

```
## [1] 2.5
```

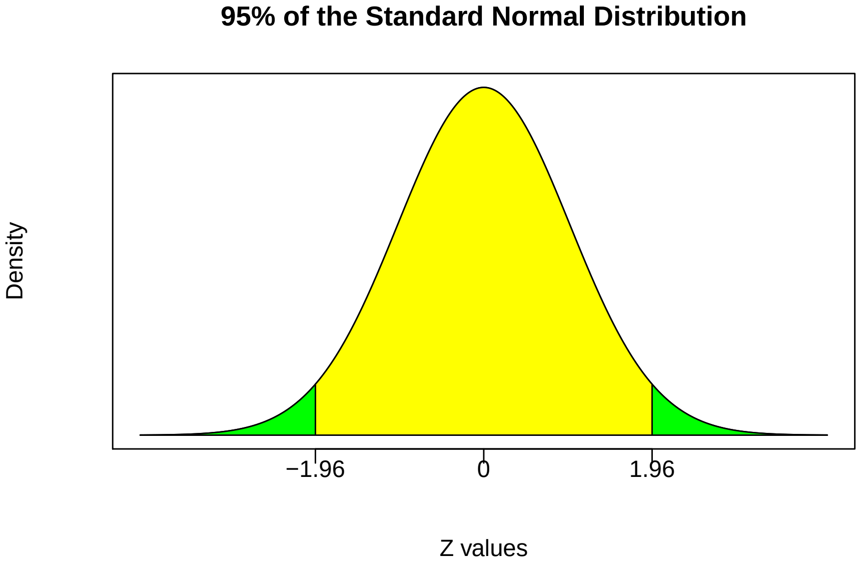
$$Z = \frac{\bar{X} - \mu}{(\sigma/\sqrt{n})} = \frac{5.1 - 4.5}{(1.2/\sqrt{25})} = 2.5$$

The next step involves taking a standard normal sampling distribution and breaking it up into regions of rejection and nonrejection. Before we do this, lets take a step back and think about what it means for a sample to be consistent or inconsistent with the null hypothesis. If you had a sample average ($\bar{X}$) that was the same value as the parameter value stated in the null hypothesis (μ) then you could state that the value of μ in the null hypothesis is *very consistent* with the characteristics of the sample. In fact, you can't be more consistent than having $\bar{X}$ and μ being the exact same number. Furthermore, a test statistic under the null would be equal to zero because ($\bar{X} - \mu$) is in the numerator. This implies that a test statistic under the null equal to zero is very consistent with the null hypothesis, and you will therefore *not reject* the null hypothesis. However, the farther away the test statistic under the null gets from zero, the farther away it gets from the center of the *do not reject region* and the closer it gets to one of the *rejection regions*.

Lets illustrate this supposing that we want to test the hypothesis at the 95% confidence level ($\alpha = 0.05$). The central 95% of a standard normal distribution is given by

$$Pr(-1.96 \leq Z \leq 1.96) = 0.95$$

This means that if you reached into a bowl of numbers comprising a standard normal distribution, then 95% of the time you will draw a number between -1.96 and 1.96. The remaining numbers outside of this range will show up only 5% of the time. These regions are the bases for confidence intervals and also the bases for the nonrejection and rejection regions.



The yellow-shaded region is centered on zero and represents the *nonrejection region.* It tells you that if you calculate a test statistic under the null to be between -1.96 and 1.96, then you do not have enough evidence to reject the null. However, the green-shaded regions are the *rejection regions.* It tells you that if you calculate a test statistic under the null that is greater than 1.96 or less than -1.96, then you have enough evidence to reject the null (with 95% confidence). In other words, it is so unlikely to have the null be correct while simultaneously randomly selecting a sample with the observed sample characteristics, so we conclude that the statement in the null cannot be true. In the fast food example above, the test statistic under the null of 2.5 falls in the rejection region. This means that we can *reject* the null hypothesis of $\mu = 4.5$ minutes with 95% confidence. In other words, we are 95% confident that the population mean is some number other than 4.5 minutes.

Changing the level of confidence (α)

The hypothesis test above was concluded under a specified 95% confidence level ($\alpha = 0.05$). This level of confidence effectively delivered our rejection and nonrejection regions. So… what happens when we change α?

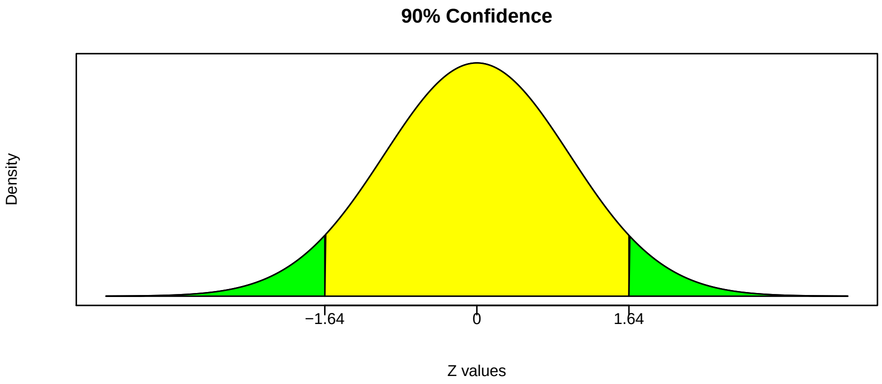
The first thing to understand is that the level of confidence does not impact the hypotheses or the test statistic under the null. The **only** thing the level of confidence impacts is the shaded regions in the sampling distribution. The figure below illustrates rejection and nonrejection regions for α values of 0.10, 0.05, and 0.01. Note that as α gets smaller, the size of the nonrejection region gets larger. This means that *do not reject* is becoming a more likely conclusion.

This should make sense because *do not reject* is a wishy-washy conclusion, while *reject* is definitive. Do not reject states that the null may or may not be true - it's a punt! Therefore, if you are placing more confidence on your conclusion, the more likely you are to make the wishy-washy conclusion.

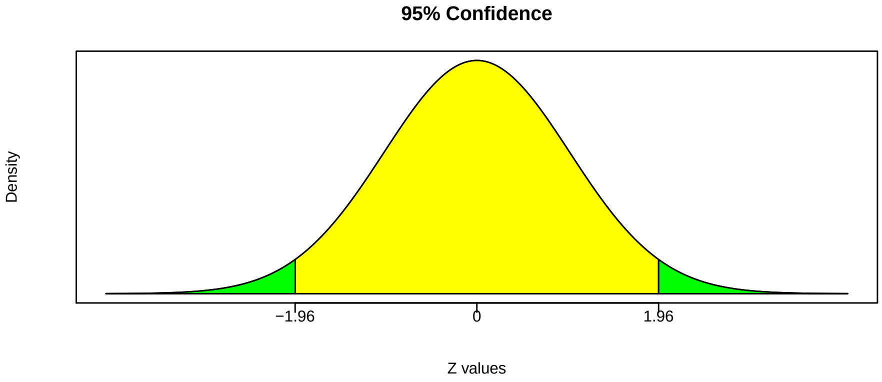
The fast food example had a test statistic under the null of 2.5. This test statistic falls in the rejection region for both 90% and 95% levels of confidence. This suggests that if you can reject a null hypothesis with a certain level of confidence, then you can automatically reject at all lower levels of confidence. However, the do not reject region under 99% confidence is given by

$$Pr(-2.58 \leq Z \leq 2.58) = 0.99$$

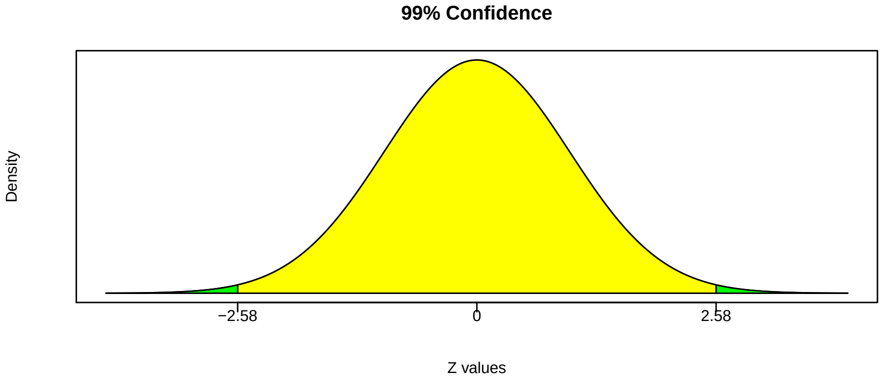
and our test statistic of 2.5 falls inside it. We therefore conclude that we *do NOT reject* the null with 99% confidence. In other words, we do not have enough evidence to say that the null hypothesis is incorrect at 99% confidence, so we conclude that it *may or may not* be true (i.e., we punt).



90% Confidence



95% Confidence



99% Confidence

Using the rejection region method, we were able to reject the null with 95% confidence ($\alpha = 0.05$) but unable to reject with 99% confidence ($\alpha = 0.01$). This begs the question as to the highest confidence level at which we can reject the null. We know it is some confidence level between 95% and 99%, but what is it exactly? We can use the rejection region approach multiple times by choosing various values of α and narrow things down, or we can conduct the hypothesis test using the *p-value* approach.

6.2.2 P-value Approach

The P-value is an extremely useful and often misunderstood number. I therefore have THREE equivalent ways of explaining it. Each one works - so just stick with the one that works for you. Before we get to those, lets talk explicitly about what we mean when we make statements based on confidence.

When using a sample statistic to draw conclusions about a population parameter, there is always the risk of reaching an incorrect conclusion. In other words, you can make an **error**.

When making a conclusion about a hypothesis test, one can either reject a null hypothesis or not. Therefore, there are two possible types of errors to be made.

1. A **Type I error** occurs when a researcher incorrectly rejects a true hypothesis. (*You rejected something that shouldn't have been rejected.*)
2. A **Type II error** occurs when a researcher incorrectly fails to reject a false hypothesis. (*You did not reject something that should have been rejected.*)

The **acceptable** probability of committing either one of these errors depends upon an arbitrary confidence level α. To be precise, when you reject a hypothesis with 95% confidence, then you are implicitly stating that you are accepting a 5% chance of being wrong. That is where $\alpha = 0.05$ (or 5% comes from). If you decrease α to 0.01 (or 1%), then you can reject a hypothesis with 99% confidence and implicitly accept a 1% chance of being wrong. The kicker is that the more you decrease the probability of committing a type I error, the more you increase the chance of not rejecting a hypothesis that you should be rejecting (a type II error). For example, if you want a conclusion with 100% confidence, then you will *never* reject a hypothesis no matter how wrong it actually is.[1]

The main take away from the previous statement is that α states the *acceptable* probability of committing a type one error. Recall in our fast food example that we rejected the null hypothesis with 95% confidence (i.e., a 5% acceptable probability of being wrong), but we did not reject the null hypothesis with 99% confidence (i.e., a 1% acceptable probability of being wrong). This means that the *actual* probability of committing a type one error is somewhere in between 0.05 and 0.01 (i.e., 5% and 1%). This actual probability of committing a type I error is called the **p-value**.

```
# Fast Food Example Revisited
mu = 4.5
```

[1] This point touches on the idea of confidence in statistics. If you want me to make a statement with 100% confidence, then I'll simply say *anything can happen* because it is a statement that has zero chance of being wrong.

```
Xbar = 5.1
Sig = 1.2
n = 25

(Zstat = (Xbar - mu)/(Sig/sqrt(n)))
```

```
## [1] 2.5
```

```
# 95% confidence:
alpha = 0.05
(Zcrit = qnorm(alpha/2,lower.tail = FALSE))
```

```
## [1] 1.959964
```

```
# 99% confidence:
alpha = 0.01
(Zcrit = qnorm(alpha/2,lower.tail = FALSE))
```

```
## [1] 2.575829
```

```
# p-value:
(Pval = pnorm(Zstat,lower.tail = FALSE)*2)
```

```
## [1] 0.01241933
```

```
# Actual confidence level:
((1-Pval)*100)
```

```
## [1] 98.75807
```

The calculations regarding the fast food example were repeated and continued to include a p-value. Recall that the null hypothesis stated that the population mean was equal to 4.5, and the test statistic under the null is equal to 2.5. The critical values marking the boundaries between the do not reject region and the reject regions was ± 1.96 for $\alpha = 0.05$ and ± 2.58 for $\alpha = 0.01$. Our test statistic falls inside the rejection region for $\alpha = 0.05$ and inside the nonrejection region for $\alpha = 0.01$. Our test statistic falls *right on the boundary* of a rejection and nonrejection region when $p = 0.0124$. This is the p-value of the problem. It states that you can reject the null hypothesis with *at most* 98.76% confidence and you will incur a 1.24% chance of being wrong. As expected, it is between 5% and 1% and gives you a tailor-made confidence interval for the hypothesis test at hand.

The definitions of a P-value

> The p-value is the probability of getting a test statistic equal to or more extreme than the sample result, given that the null hypothesis (H_0) is true.

While this is the technical definition of a p-value, it is a bit vague. There are some roughly equivalent definitions that might be easier to digest.

1. The p-value is the probability of committing a Type I error. If the p-value is greater than some arbitrarily given α, then you cannot reject the null.
2. The p-value is the probability that your null hypothesis is *correct*. The HIGHEST level of confidence at which you can reject the null is therefore $1 - p$.

6.3 Two-sided vs One-sided Test

$$H_0 : \mu = 4.5 \quad versus \quad H_1 : \mu \neq 4.5$$

The hypothesis test considered above is known as a **two-sided** test because the null gets rejected if the mean of the sample is either significantly greater than or less than the value stated in the null hypothesis. If you notice from the illustrations above, a two-sided test has **TWO** rejection regions - one in each tail (hence the name). Note that this is also why we calculated critical values using half of the value of α and doubled the calculated probability value in order to arrive at a p-value.

In the fast food example above, suppose we want to show that the service time *increased*. In other words, we want statistical evidence that the wait time actually increased from a previous time of 4.5 minutes. We can provide statistical evidence by rejecting a null hypothesis that the new population average wait time is 4.5 minutes *or less*. This scenario delivers us a **one-sided hypothesis test**.

$$H_0 : \mu \leq 4.5 \quad versus \quad H_1 : \mu > 4.5$$

As the name implies, a one-sided hypothesis test only has one rejection region. This means that the entire value of α is grouped into either the right or left tail. The tail containing the rejection region depends upon the exact specification of the hypothesis test.

The hypothesis test above is called a *right-tailed* test because the rejection region is in the right tail. To see this, consider several hypothetical sample averages and see if they are consistent with the null $\mu \leq 4.5$.

- Suppose you observe $\bar{X} = 4$. Is this sample average consistent with $\mu \leq 4.5$?
- What about $\bar{X} = 2$?
- What about $\bar{X} = 1$?

Your answer should be yes to all of these sample averages. In fact, *any* sample average less than or equal to 4.5 is consistent with $\mu \leq 4.5$.

Next, recall the test statistic under the null:

$$Z = \frac{\bar{X} - 4.5}{(\sigma/\sqrt{n})}$$

For any of the hypothetical values considered above (4, 2, or 1), the test statistic would be a negative number. Since we said that all of these sample averages are consistent with the null being true, then we would never reject the null in any of these instances. Therefore, the rejection region cannot be in the left tail because that is where the negative values of the distribution reside. The rejection region must therefore be in the right tail. Only when a sample average is sufficiently greater than 4.5 is when we can consider rejecting the null.

Now that we already have the null and alternative hypotheses down as well as the test statistic under the null, the next step is to determine the critical value that divides the distribution into rejection and nonrejection regions.

```
# Fast Food Example Revisited
mu = 4.5
Xbar = 5.1
Sig = 1.2
n = 25

(Zstat = (Xbar - mu)/(Sig/sqrt(n)))
```

```
## [1] 2.5
```

```
# 95% confidence:
alpha = 0.05
(Zcrit = qnorm(alpha,lower.tail=FALSE))
```

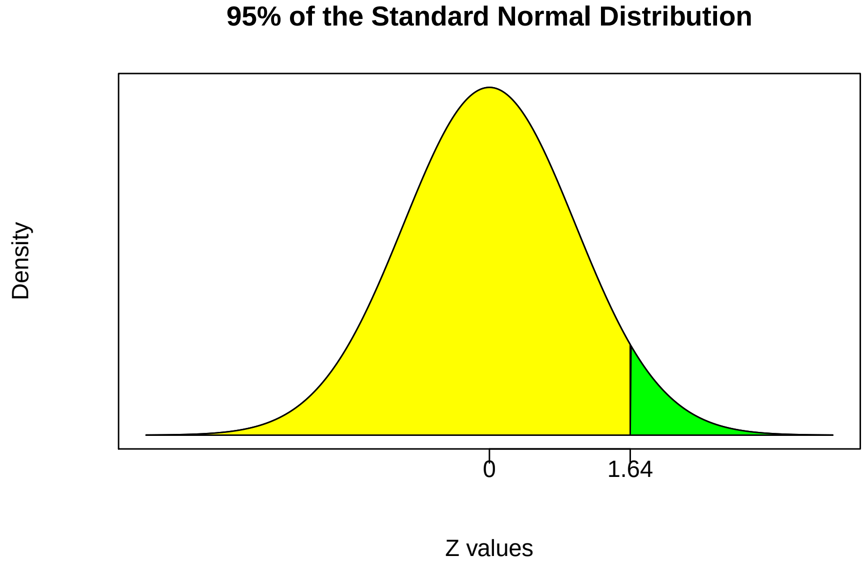
```
## [1] 1.644854
```

```
# P-value:
(Pval = pnorm(Zstat,lower.tail=FALSE))
```

```
## [1] 0.006209665
```

```
# highest Confidence Level for rejection:
((1-Pval)*100)
```

```
## [1] 99.37903
```



If we conducted this hypothesis test at the 95% confidence level ($ =0.05), you will see that the rejection region is the 5% of the curve in the right tail. That means you reject all test statistics greater than or equal to 1.64. Since our test statistic is 2.5, we can reject with 95% confidence. We can also conduct this hypothesis test using the p-value approach which delivers a p-value of 0.0062. This means that if we reject the null, we only incur a 0.62% chance of being wrong. This equivalently means that we can reject the null with up to 99.38% confidence.

6.4 Conducting a hypothesis test (when σ is unknown)

When the population standard deviation (σ) is unknown, it must be estimated. Just like with confidence intervals, When you replace σ with it's estimate S, you change the distribution from Z to t (and need to mind the degrees of freedom).

That's the only difference

Let's go through some applications when σ is unknown. You will see that the only difference is that we use a t distribution with $n-1$ degrees of freedom to calculate rejection / nonrejection regions and p-values.

Application 2

The Saxon Home Improvement Co. has had a mean per sales invoice of $120 over the last 5 years and would like to know if the mean amount per sales invoice has significantly changed. This is enough information to state our hypotheses for a two-sided test.[2]

$$H_0 : \mu = 120 \quad versus \quad H_0 : \mu \neq 120$$

You collected a sample of 12 observations, and concluded that the sample mean was $112.85 and the sample standard deviation was $20.80.

$$\bar{X} = 112.85, \quad n = 12, \quad S = 20.80$$

This information allows us to calculate a t-test statistic under the null. The only difference is that we now have a sample standard deviation (S) were we once had a population standard deviation (σ).

```r
Xbar = 112.85
n = 12
S = 20.80
mu = 120

(t = (Xbar - mu) / (S/sqrt(n)))
```

```
## [1] -1.190785
```

$$t = \frac{\bar{X} - \mu}{(S/\sqrt{n})} = \frac{112.85 - 120}{\left(20.80/\sqrt{12}\right)} = -1.19$$

Now that we have our test statistic, we need to determine if it falls into our nonrejection or rejection regions. The important thing to realize is that these regions are now part of a t distribution with 11 $(n - 1)$ degrees of freedom. If we consider 95% confidence…

```r
alpha = 0.05
(tcrit = qt(alpha/2,n-1,lower.tail=FALSE))
```

```
## [1] 2.200985
```

[2] Note the language - *significantly changed* means that the value could have either gone up or down. This is why it is a two-sided test.

The calculations suggest that the nonrejection region is between ±2.2. Since our test statistic falls within this region, we do not reject the null. This implies that we do not have evidence that the population average sales invoice has significantly changed from $120 with 95% confidence. The conclusion is therefore *do not reject*.

We could also calculate a p-value for the test:

```
(Pval = pt(t,n-1)*2)
```

```
## [1] 0.2588003
```

```
# Highest confidence interval for rejection:
((1-Pval)*100)
```

```
## [1] 74.11997
```

Notice here that the p-value states that if we were to reject the null, then we would incur a 25.88% chance of being wrong. This means that we could only reject the null with 74.12% confidence.

Note that the calculations uses a new R command: `pt(q,df)`. This command calculates the probability under a t distribution the same way the `pnorm(q)` command calculates the probability under a standard normal distribution. In addition, I again encourage you to always visualize the distribution and explicitly draw the rejection and nonrejection regions. This is extremely helpful when first getting started. Below you will also see a note I wrote for a previous class reinforcing how R likes to calculate probabilities. It is for reference if needed.

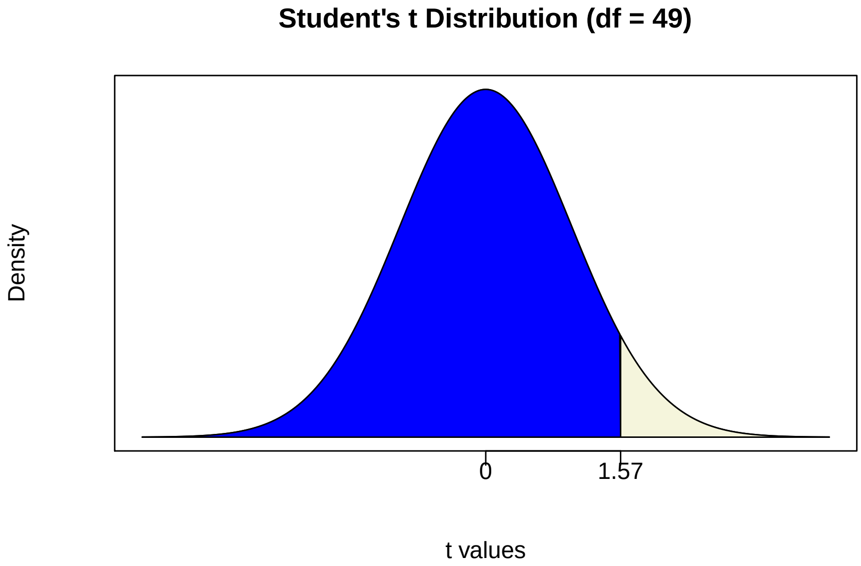
6.5 Appendix: A note on calculating P-values

6.5.1 The Problem

Suppose you were performing a right-tailed hypothesis test (using a t distribution) and you arrived at a test statistic under the null of 1.57. This means that the rejection region is in the right tail, and if you wished to calculate the p-value, then it would be the area of the curve to the right of 1.57.

If you have a sample of $n = 50$, then you would use a t-distribution with 49 or $(n - 1)$ degrees of freedom.

An illustration is below:



6.5.2 How to calculate p-values

In case you haven't noticed by now, R has a default way of calculating probability areas...

IT ALWAYS CALCULATES AREAS FROM THE LEFT!

In other words, the default is to give you the area to the left of a number...

```
(Pval = pt(1.57,49))
```

```
## [1] 0.9385746
```

Don't be annoyed about this, because all software does this (including Excel).

We can use this default to calculate the p-value (i.e. the area to the right of 1.57) in THREE different ways by relying on two properties of our probability distributions.

Property 1: The distribution is centered at zero and symmetric.

This means that the area to the right of 1.57 is the same as the area to the left of -1.57. So we can use the pt function with the default setting to this effect:

```
(Pval = pt(-1.57,49))
```

```
## [1] 0.06142544
```

Property 2: The distribution always adds up to one.

This means that you have a 100% chance of pulling a number between negative and positive infinity. So if you use 1.57 and the default setting which gives you the area to the left, then subtract that number from 1 to get the area to the right:

```
(Pval = 1-pt(1.57,49))
```

```
## [1] 0.06142544
```

Final Option: Undo the default setting...

The full command for calculating a p-value from a t-distribution (for our purposes) is as follows:

pt(q, df, lower.tail = TRUE)

Note that q is the *quantity*, and df is the *degrees of freedom*. All other entries (if not specified) go to their default values. This is where *lower.tail* comes in. It is set to *TRUE* by default, meaning that whatever number you input for q, you will get the area to the left. If you change this entry to *FALSE*, then the default is switched off and you will calculate the area to the right.

```
(Pval = pt(1.57,49,lower.tail = FALSE))
```

```
## [1] 0.06142544
```

Notice that all three ways of calculating a p-value give you the exact same result. Therefore, you do not need to master all three - just pick whichever method works best for you.

Chapter 7

Simple Linear Regression

Suppose you have two homes that are the same in *every way* except for size. Our intuition would suggest that bigger homes cost more (*all else equal*) so we would expect that there is a positive relationship between house size and house price.

Saying *bigger homes cost more* is a **qualitative** statement because all we are saying is that the relationship between house size and house price is positive. What if we want to make a **quantitative** statement? In other words, while we are fairly confident that the actual house price (say, in dollars) will increase for every unit increase in house size (say, an additional square foot) - we want to know exactly what this *average-price-per-square-foot* is.

A **Regression** can measure the relationship between the mean value of one variable and corresponding values of other variables. In other words, it is a statistical technique used to explain average movements of one (dependent) variable, as a function of movements in a set of other (independent) variables.

This chapter will discuss the estimation, interpretation, and statistical inference of a *simple* linear regression model, which means that we will attempt to explain the movements in a dependent variable by considering **one** independent variable. This is the simplest regression model we can consider in order to understand what is going on under the hood of a regression. The next chapter will extend this analysis to *multiple* regression models where the only real difference is that the number of independent variables are greater than one.

7.1 A Simple Linear Regression Model

A Linear Regression model is a line equation.

The simplest example of a line equation is:

$$Y_i = \beta_0 + \beta_1 X_i$$

The *betas*, β_0 and β_1 are called line coefficients.

- β_0 is the *constant* or *intercept term*
- β_1 is the *slope* term - it determines the change in Y given a change in X

$$\beta_1 = \frac{Rise}{Run} = \frac{\Delta Y_i}{\Delta X_i}$$

7.1.1 What does a regression model imply?

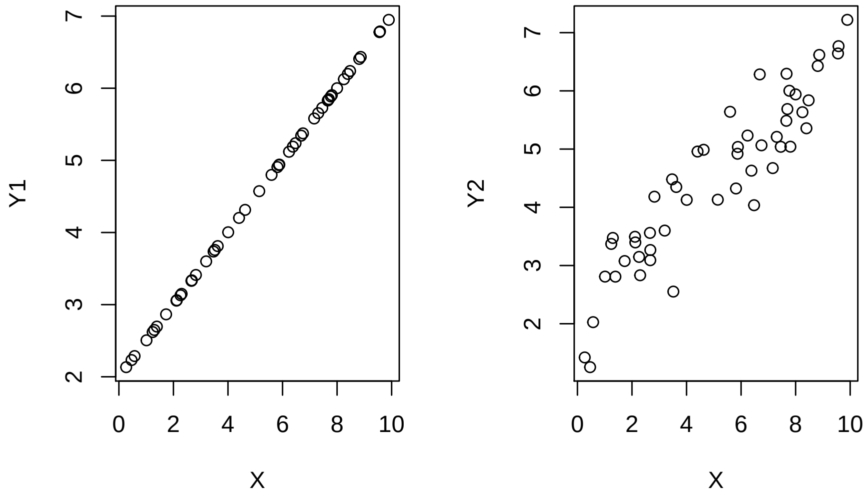
$$Y_i = \beta_0 + \beta_1 X_i$$

When we write down a model like this, we are imposing a huge amount of assumptions on how we believe the world works.

First, there is the **Direction of causality**. A regression implicitly assumes that changes in the independent variable (X) *causes* changes in the dependent variable (Y). This is the ONLY direction of causality we can handle, otherwise our analysis would be confounded (**what causes what**) and not useful.

Second, The equation assumes that information on the independent variable (X) is all the information you need to explain the dependent variable (Y). In other words, if we were to look at pairs of observations of X and Y on a plot, then the above equation assumes that all observations (data points) line up exactly on the regression line.

This would be great if the observations look like the figure on the left, but not if they look like the figure on the right.



It would be extremely rare for the linear model (as detailed above) to account for all there is to know about the dependent variable Y...

1. There might be other independent variables that explain different parts of the dependent variable (i.e., multiple dimensions). *(more on this next chapter)*

2. There might be measurement error in the recording of the variables.

3. There might be an incorrect functional form - meaning that the relationship between the dependent and independent variable might be more sophisticated than a straight line. *(more on this next chapter)*

4. There might be purely random and therefore totally unpredictable variation in the dependent variable.

This last item can be easily dealt with!

Adding a stochastic error term (ε_i) to the model will effectively take care of all sources of variation in the dependent variable (Y) that is not explicitly captured by information contained in the independent variable (X).

7.1.2 The ***REAL*** Simple Linear Regression Model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

The Linear Regression Model now explicitly states that the explanation of the dependent variable (Y_i) can be broken down into two components:

- A **Deterministic** Component: $\beta_0 + \beta_1 X_i$
- A **Random / Stochastic / Explainable** Component: ε_i

Lets address these two components in turn.

The Deterministic Component

$$\hat{Y}_i = \beta_0 + \beta_1 X_i$$

The deterministic component delivers the expected (or *average*) value of the dependent variable (Y) given a values for the coefficients (β_0 and β_1) and a value of the dependent variable (X).

Since X is given, it is considered *deterministic* (i.e., non-stochastic)

In other words, the deterministic component determines the mean value of Y associated with a particular value of X. This should make sense, because the average value of Y is the best guess.

Technically speaking, the deterministic component delivers the *expected value of Y conditional on a value of X* (i.e., a conditional expectation).

$$\hat{Y}_i = \beta_0 + \beta_1 X_i = E[Y_i|X_i]$$

The Unexpected (*Garbage Can*) Component

$$\varepsilon_i = Y_i - \hat{Y}_i$$

Once we obtain the coefficients, we can compare the observed values of Y_i with the expected value of Y_i conditional on the values of X_i.

The difference between the true value (Y_i) and the expected value ($\hat{Y}_i$) is by definition... *unexpected!*

This unexpected discrepancy is your **prediction error** - and everything your deterministic component cannot explain is deemed *random* and *unexplainable.*

If a portion of the dependent variable is considered random and unexplainable - then it gets thrown away into the *garbage can* (ε_i).

This is a subtle but crucial part of regression modeling...

- Your choice of the independent variable(s) dictate what you believe to be important in explaining the dependent variable.
- The unimportant (or random) changes in the dependent variable that your independent variables cannot explain end up in the garbage can *by design.*
- Therefore, **the researcher** essentially chooses what is important, and what gets thrown away into the garbage can.

YOU are the researcher, so **YOU** determine what goes into the garbage can!

7.2 Application: Predicting House Price Based on House Size

Let's consider an application where we attempt to explain the price of a house (in thousand US$) by the size of a house (in square feet). We start by establishing the theory and relating it to our statistical terminology.

The Population Regression Function:

$$Price_i = \beta_0 + \beta_i Size_i + \varepsilon_i$$

- $Price_i$ is the dependent variable (Y_i)
- $Size_i$ is the independent variable (X_i)
- The equation above is the true (but unknown), population regression function.
- The coefficients (β_0 and β_1) are the population regression coefficients!
  - They are the coefficients you would obtain if you had *every* possible observation (i.e., the population)
  - This ain't gonna happen…
- We need to obtain the estimated, sample regression coefficients. To do this, we need to collect a sample of observations.

The Sample

In order to obtain *sample estimates* of our regression model above, we must obtain a sample of observations. We collect a (random) sample of size n. This is where the subscript i comes in - indicating that in general, each individual observation can be identified as $i = 1, ..., n$. The sample estimates are based on the sample.

Hypothetically, we can obtain different estimated coefficients for every different sample… but we will address that later.

To facilitate this application, we will use a data set internal to R, called *hprice1*.

```r
data(hprice1,package='wooldridge')
ls(hprice1)
```

```
##  [1] "assess"   "bdrms"    "colonial" "lassess"  "llotsize" "lotsize"  "lprice"   "lsqrft"   '
## [10] "sqrft"
```

This data set contains 88 observations of homes where each home has 10 pieces of information called *variables*. We are only concerned with two variables at the moment - the house price (*price*) and the house size (*sqrft*).

```
summary(hprice1$price)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   111.0   230.0   265.5   293.5   326.2   725.0
```

```
summary(hprice1$sqrft)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    1171    1660    1845    2014    2227    3880
```

We know that:

1. the average house price is $293,500
2. 50% of the observations are between the 1st and 3rd quartiles of $230,000 and $326,200
3. the minimum house price in the sample is $111,000
4. the maximum house price in the sample is $725,000. You can look at the summary output for the size variable and make similar statements.

The Sample Regression Function

We combine our Population Regression Function (PRF) and our data sample to estimate a *Sample Regression Function* (SRF).

$$Price_i = \hat{\beta}_0 + \hat{\beta}_1 Size_i + e_i$$

The difference between the SRF and the PRF are very important.

1. The PRF coefficients are *population parameters* while the SRF coefficients are *sample statistics*. In other words, the SRF coefficients are actual numbers that correspond to our sample, and we use them to draw inference on the things we really want to talk about - the PRF coefficients.
2. The difference between the SRF residual (e_i) and the PRF residual (ε_i) is along the same lines as the difference between the SRF and PRF coefficients. The SRF residual contains the unexplained variability of the dependent variable in the sample while the PRF residual theoretically contains the unexplained variability in the population.

We will get into the details about how these regression estimates can be obtained later. Right now, lets just arrive at our estimates and shed light on the big picture.

```
Y <- hprice1$price
X <- hprice1$sqrft

REG <- lm(Y~X)
coef(REG)
```

```
## (Intercept)           X
##   11.204145    0.140211
```

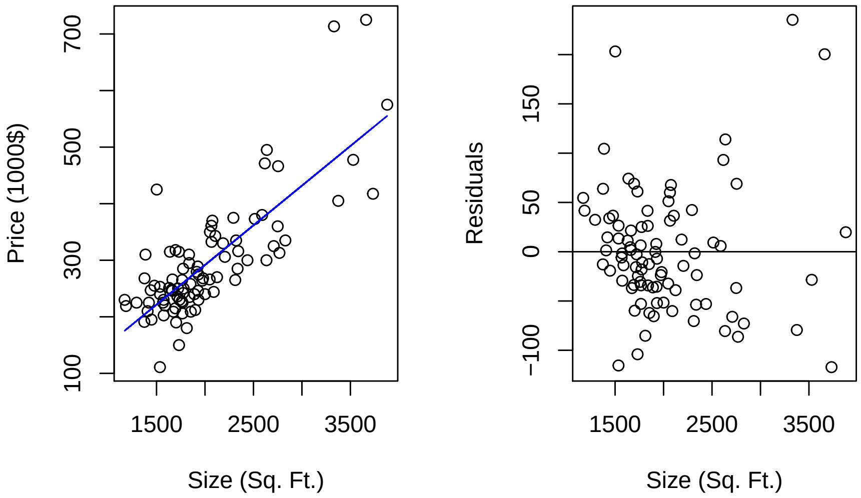
Our regression estimates are $\hat{\beta}_0 = 11.2$ and $\hat{\beta}_1 = 0.14$. This delivers a prediction equation from our SRF as:

$$Pr\hat{i}ce_i = 11.2 + 0.14 Size_i$$

Where $\hat{Y}_i = Pr\hat{i}ce_i$ is the expected house price conditional on a particular size.

We can illustrate the results of the regression as follows:

```
par(mfrow = c(1,2))
plot(X, Y,
     xlab = "Size (Sq. Ft.)",
     ylab = "Price (1000$)")
lines(X,fitted(REG),col = 'blue')
plot(X,residuals(REG),
     xlab = "Size (Sq. Ft.)",
     ylab = "Residuals")
abline(h = 0)
```



In the left figure, the *dots* are a scatter-plot of the actual observations of house price (Y) and house size (X) while the blue line is our estimated regression which delivers the expected house price for each observation of house size. Note that every time an actual house price is different than the expected value from the regression - then that difference is considered *unexpected* and ends up in the *garbage can* (residual). The residual values are in the right figure. Note that the residual values are centered around the zero line - this means that the unexpected component of house price is equal to zero *on average*.

Analysis of the SRF

We can get plenty of mileage out of our estimated SRF.

1. We can interpret the estimated coefficients (one at a time) to get a sense of how house size influences house price.

- $\hat{\beta}_0 = 11.2$ is the estimated intercept term. Mathematically, it is the expected value of the dependent variable conditional on the independent variable being 0 ($E[Y_i|X_i = 0] = 11.2$). In the context of this problem, we are saying that the *expected price of a house that has 0 square feet in size is 11.2 thousand dollars.* If that sounds funny to you... it should. The take away is that an intercept term always has a mathematical interpretation, but it might not always make sense. The key is if an independent value of zero (i.e., $X = 0$) makes sense.

- $\hat{\beta}_1 = 0.14$ is the estimated slope term. Mathematically, it is the expected change in value of the dependent variable given a unit-increase in the independent variable ($\Delta Y_i / \Delta X_i = 0.14$). In the context of this problem, we are saying that the *expected price of a house will increase by 0.14 thousand dollars ($140) for every (square-foot) increase in house size.* If you were a realtor, you can now state that somebody looking for a home would be paying $140 per square foot of house size *on average*.

2. We can use the model for forecasting purposes.

To illustrate a forecast, suppose you came across a 1,800 square-foot house with a selling price of $250,000. Does this seem like a fair price? In order to answer this question with our estimated results, we simply plug 1800 square-feet as a value for our independent variable and arrive at an expected price conditional on this house size.

$$\widehat{Price}_i = 11.2 + 0.14(1800) = 263.6$$

```r
Bhat0 <- summary(REG)$coef[1,1]
Bhat1 <- summary(REG)$coef[2,1]

(Yhat = Bhat0 + Bhat1 * 1800)
```

```
## [1] 263.5839
```

Our regression forecast states that an 1,800 square-foot house should have an *average* price of $263,000. Since this is more than the $250,000 of the house in question, then the regression model suggests that this is a fair price.

Discussion

While our model appears pretty useful, we must always be mindful of the limitations of our model. Namely, our regression **assumes** that house size is the **only** thing that matters when predicting house price. Our candidate house is more than $10,000 below the average 1,800 square-foot house price in the sample, but this might be due to very relevant things that our model considers *unpredictable*.

- Is the house located next to the town dump?
- Is the house built on top of an ancient burial ground?
- Does it have a really ugly kitchen?
- Does the roof leak?

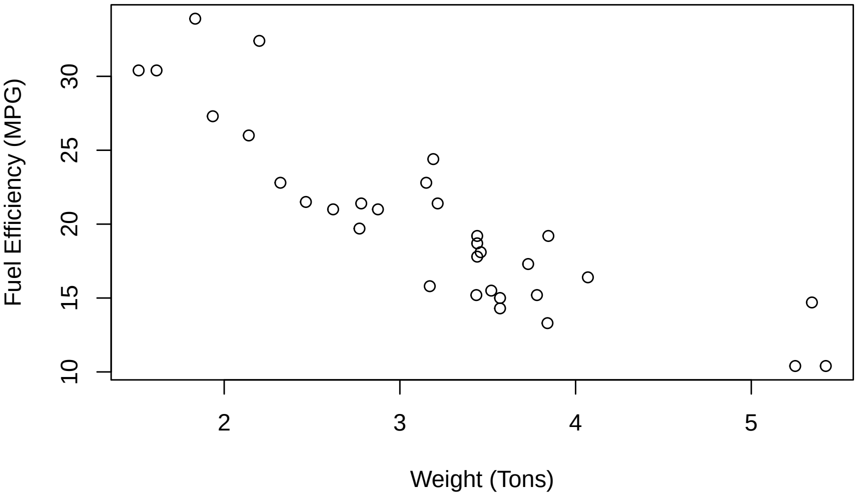
The bottom line is that one should always view our regression estimates within the lens of its limitations. This isn't to say that the estimates are *incorrect* or *wrong*, because they are actually quite useful. However, understanding how far one can take regression results is important.

7.3 Ordinary Least Squares (OLS)

Ordinary Least Squares (OLS, for short) is a popular method for estimating the unknown parameters in a linear regression model. OLS chooses the parameters of a linear function by minimizing the sum of the squared differences between the observed dependent variable (values of the variable being observed) in the given data set and those predicted by the linear function.

To illustrate this, consider a sample of observations and a PRF:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$



Look at the figure and try to imagine the *best fitting* straight line that goes through all observations in the scatter plot. This line has two features: and intercept term ($\hat{\beta}_0$) and a slope term ($\hat{\beta}_1$). Which values would you assign?

We can go about this a little more formally. First, if we had values for $\hat{\beta}_0$ and $\hat{\beta}_1$, then we can determine the residual (error) for each pair of Y_i and X_i.

$$e_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)$$

We can sum across all observations to get the *total error*

$$\sum_i e_i = \sum_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)$$

The problem we face now is that error terms can be both positive and negative. That means they will start to wash each other out when we sum them up and we therefore get an incomplete measure of the total error. To prevent the positive and negative error terms from washing each other out, we square each of the terms. This makes the negative errors positive, while the positive errors stay positive.[1]

$$\sum_i e_i^2 = \sum_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

Notice that this function now states that we can calculate the sum of squared errors for any given values of $\hat{\beta}_0$ and $\hat{\beta}_1$. We can therefore find the *best* values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that deliver the *lowest* sum of squared errors. The line that delivers the lowest squared errors is what we mean by the best line.

$$min \sum_i e_i^2 = min_{(\hat{\beta}_0, \hat{\beta}_1)} \sum_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

This function is called an *objective function*, and we can minimize the sum of squared errors by taking first-order conditions (i.e., the derivative of the objective function with respect to $\hat{\beta}_0$ and $\hat{\beta}_0$).

$$\hat{\beta}_1 = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i (X_i - \bar{X})^2} = \frac{cov(X,Y)}{var(X)}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Where a ‘bar’ term over a variable represents the mean of that variable (i.e., $\bar{X} = \frac{1}{n}\sum_i X_i$)

These two equations are important. The first equation states that the slope of the line equation ($\hat{\beta}_1$) is equal to the ratio between the covariance of Y and X and the variance of X. Remember that a covariance measures how two variables systematically move together. If they tend to go up at the same time,

[1] Note: this is where the *sum of squared errors* comes in.

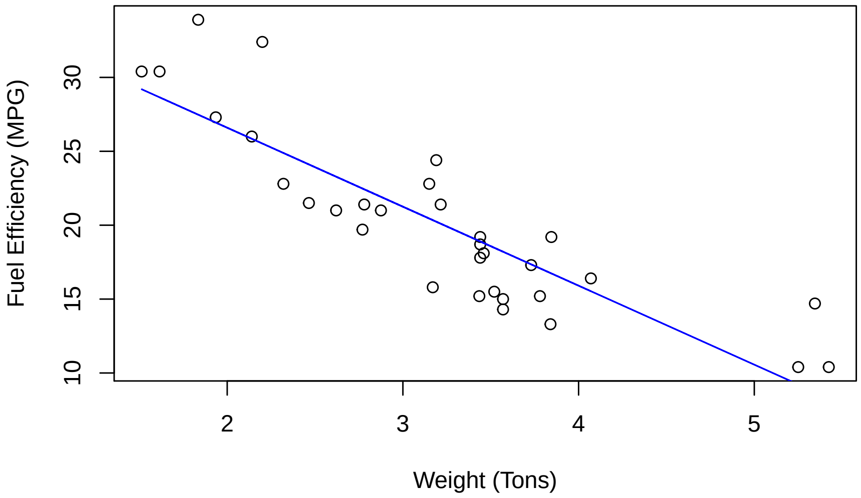
then they have a positive covariance. If they tend to go down - a negative covariance. If they do not tend to move together in any systematic way, then they have zero covariance. This systematic movement is precisely what helps determine the slope term. The second equation states that with $\widehat{\beta}_1$ determined, we can determine $\widehat{\beta}_0$ such that the regression line goes through the means of the dependent and independent variables.

Lets see what these estimates and the resulting regression line look like.

```
REG <- lm(Y~X)
coef(REG)
```

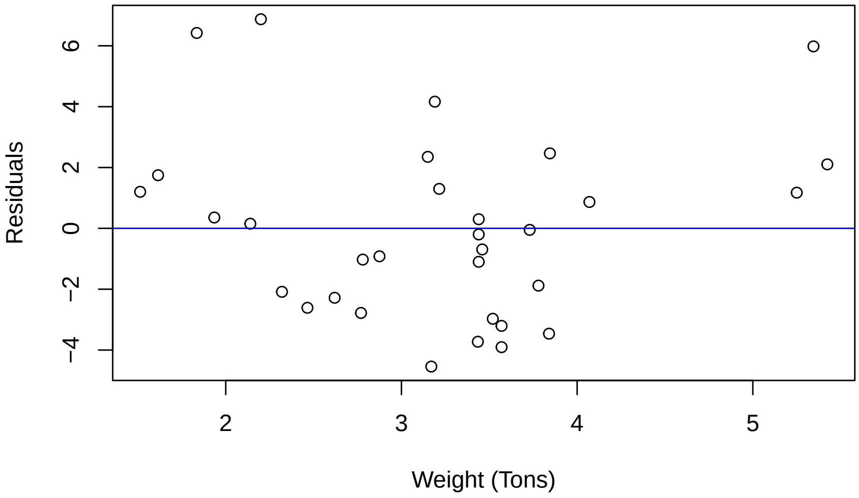
```
## (Intercept)           X
##   37.285126   -5.344472
```

```
plot(X,Y,
     xlab = "Weight (Tons)",
     ylab = "Fuel Efficiency (MPG)")
lines(X,fitted(REG),col='blue')
```



Now you probably imagined a line that looked kinda like this, but we know that this line (with these coefficients) is the absolute best line that minimizes the total difference between the observations (the *dots*) and the predictions (the *line*). Any other line we could draw would have a larger sum of squared errors. We can see what this difference looks like by looking at the residuals.

```
plot(X,residuals(REG),
     xlab = "Weight (Tons)",
     ylab = "Residuals")
abline(h = 0,col="blue")
```



Notice that these residual values are distributed both above and below the zero line. If you were to sum them all up - then you get zero ALWAYS. It is what the mathematical problem is designed to do!

$$\sum_i e_i = 0$$

This mathematical outcome is actually important. First, if the residuals or *forecast errors* sum up to zero, then that means that they have a *mean* that is also 0 ($\bar{e} = 0$). This means that they are zero *on average*, so the *expected value* is zero!

$$E[e_i] = 0$$

If the expected value of the forecast error is zero, then this means that our regression line is correct *on average*. If we think about it, this is the best we can ask for out of a regression function.

7.3.1 B.L.U.E.

OLS is a powerful estimation method that delivers estimates with the following properties.

1. They are the **BEST** in a minimized mean-squared error sense. We just showed this.
2. They are **LINEAR** insofar as the OLS method can be quickly used when the regression model is a linear equation.
3. They are **UNBIASED** meaning that the sample estimates are true estimates of the population parameters.

Therefore, **BEST**, **LINEAR**, **UNBIASED**, **ESTIMATES** is why the output of an OLS method is said to be **B.L.U.E.**

7.4 Decomposition of Variance

Using our regression estimates and sample information, we can construct one of the most popular (and most abused) measures of *goodness of fit* for a regression. We will construct this measure in pieces.

First, the **total sum of squares** (or TSS) can be calculated to measure the total variation in the dependent variable:

$$TSS = \sum_{i=1}^{N} (Y_i - \bar{Y})^2$$

This expression is similar to a variance equation (without averaging), and since the movements in the dependent variable are ultimately what we are after, this measure delivers *the total variation in the dependent variable that we would like our model to explain.*

Next, we can use our regression estimates to calculate an **estimated sum or squares** (or ESS) which measures the total variation in the dependent variable that our model *actually* explained:

$$ESS = \sum_{i=1}^{N} (\hat{Y}_i - \bar{Y})^2$$

Note that this measure uses our conditional forecasts from our regression model in place of the actual observations of the dependent variable.

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

Finally, we can use our regression estimates to also calculate a **residual sum or squares** (or RSS) which measures the total variation in the dependent variable that our model *cannot* explain:

$$RSS = \sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{N} e_i^2$$

Note that this is a measure of the variation in the garbage can, and the garage can is where all of the variation in the dependent variable that your model cannot explain ends up.

7.4.1 The R^2

Our regression breaks the variation in Y_i (the TSS) into what can be explained (the ESS) and what cannot be explained (the RSS). This essentially means $TSS = ESS + RSS$. Furthermore, our OLS estimates attempt to maximize the ESS and minimize the RSS. This delivers our first measure of how well our model explains the movements in the dependent variable or *goodness of fit*

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

This **coefficient of determination** or R^2 should be an intuitive measure. First, it is bounded between 0 and 1. If the measure is 0 then the model explains **NOTHING** and all variation is in the garbage can. If the measure is 1 then the model explains **EVERYTHING** and the garbage can is empty. Any number in between is simply the proportion of the variation in the dependent variable explained by the model.

```
REG3 <- lm(price ~ sqrft, data = hprice1)
summary(REG3)$r.squared
```

```
## [1] 0.6207967
```

```
pander(summary(REG3))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	11.2	24.74	0.45	0.65
sqrft	0.14	0.01	11.87	0

Table 7.2: Fitting linear model: price ~ sqrft

Observations	Residual Std. Error	R^2	Adjusted R^2
88	63.62	0.62	0.62

Returning to our house price application above, you can see that our coefficient of determination (R^2) is 0.62.[2] This states that approximately 62 percent of the variation in the prices of homes in our sample is explained by the size of the house (in square feet), while the remaining 38 percent is *unexplained* by our model and shoved into the garbage can. That is all it says… no more and no less.

7.4.2 What is a *good* R^2?

Is explaining 62 percent of the variation in house prices *good*? The answer depends on what you want the model to explain. We know that the house size explains a majority of the variation in house prices while *all other* potential independent variables will explain at most the remaining 38 percent. If you want to explain everything there is to know about house prices, then an R^2 of 0.62 leaves something to be desired. If you only care to understand the impact of size, then the R^2 tells you how much of the variation in house prices it explains. There really isn't much more to it than that.

7.4.3 Standard Error of the Estimate

$$S_{YX} = \sqrt{\frac{RSS}{n-2}} = \sqrt{\frac{\sum_{i=1}^{N} e_i^2}{n-2}}$$

The standard error of the estimate is much like a standard deviation equation. However, while the standard deviation measures the variability around a mean, the standard error of the estimate measures the variability around the prediction line.

Note that the denominator of this measure is $n-2$. The reason that we are *averaging* the sum of squared errors by $n-2$ is because we lost **two degrees of freedom**. Recall that we lose a degree of freedom whenever we need to estimate something based on other estimates. When we consider how we calculated the residuals in the first place,

$$e_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 \, X_i$$

[2] Note that this number is sometimes called the *multiple* R^2

you will see that we had to estimate **two** line coefficients before we can determine what the prediction error is. That is why we deduct two degrees of freedom.[3]

7.5 Assumptions of the Linear Regression Model

An empirical regression analysis always begins with a statement of the population regression function (PRF). The PRF explicitly states exactly how you (the researcher) believes the independent variable is related to the dependent variable. One thing to be clear about when stating a PRF is that you are imposing a great deal of assumptions on how the world works. If your assumptions are correct, then the PRF is a reasonable depiction of reality and OLS will uncover accurate estimates of the PRF parameters. If your assumptions are incorrect, then the estimates are highly unreliable and might actually be misleading.

Verifying the assumptions of a linear regression model is a majority of the work involved with an empirical analysis, and we will be doing this for the rest of the course. Before getting into the details of *how* to verify the assumptions, we first need to know what they are.

One should note that these are not assumption of our model, because these assumption are actually imposed on our model. These assumptions are made on reality - at least on the relationship between the dependent and independent variables that actually occurs in the world.

The main assumptions of a linear regression model that we will focus on are as follows.

1. Linearity: the true relationship (in the world) is in fact linear. This assumption must hold because you are estimating a linear model (hence the assumption is imposed).
2. Independence of Errors: the forecast errors (e_i) are not correlated with each other
3. Equal Variance (*homoskedasticity*): the variance of the error term is constant
4. Normality of Errors: the forecast errors comprise a normal distribution

7.5.1 Linearity

If we write down the following PRF:

[3] NOTE: this line of reasoning implies that we will lose more degrees of freedom when we estimate models with more independent variables... later.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$$

we are explicitly assuming that this accurately captures the real world. In particular,

- The relationship between Y_i and $X_1 i$ is in fact linear. This means that the a straight-line (i.e., a constant slope) fits the relationships between the dependent variable and independent variables better than a nonlinear relationship.

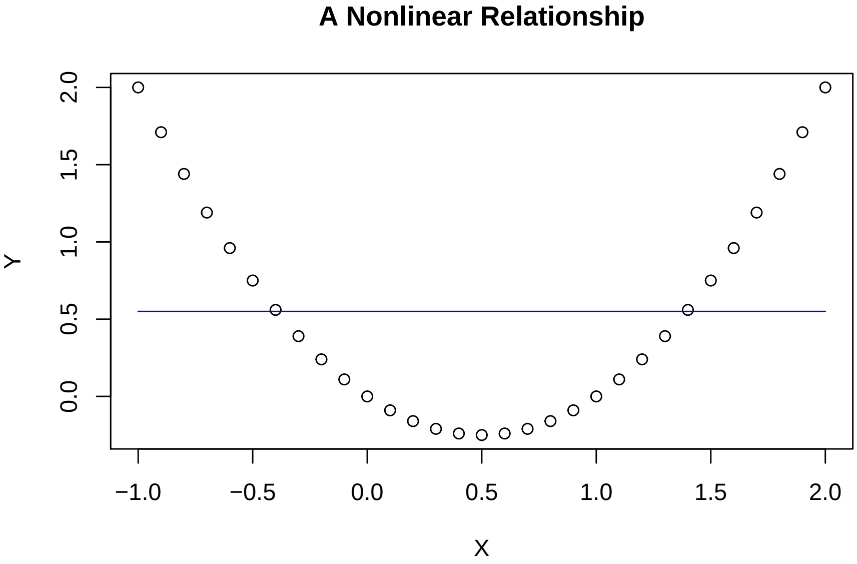
- The error term (i.e., the garbage can) is additive, meaning that the forecast errors are separable from the forecasts.

If these assumptions differ from the relationship that is going on in reality, then our model will suffer from *bias.* The SRF estimates will not be good representations of the PRF parameters, and they should not be interpreted as such.

There is an entire section devoted to relaxing the linearity assumption later on, but consider a stark example to illustrate a violation of this assumption. In particular, suppose you consider a simple linear regression model to uncover the relationship between a dependent variable and an independent variable.

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

However, suppose the true relationship (shown in the figure) is clearly nonlinear, and the blue line in the figure is the estimated (linear) SRF. As the line suggests, it is horizontal suggesting that the linear relationship between Y and X is *zero.* This doesn't mean that there is no relationship - because there clearly is. However, our assumption of this relationship being linear is incorrect because the results tell us that there is no *linear* relationship.



7.5.2 Independence of Errors

Serial correlation exists when an observation of the error term is correlated with past values of itself. This means that the errors are not independent of each other.

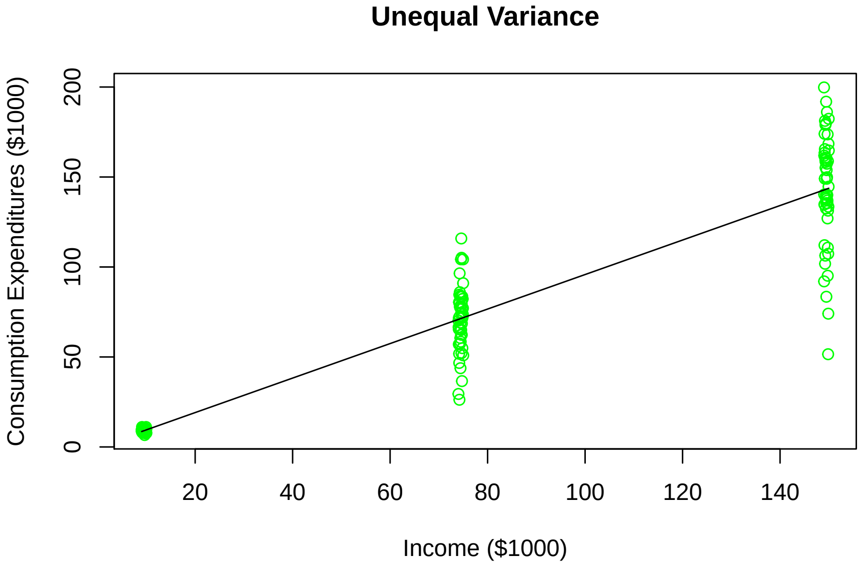
$$\varepsilon_t = \rho\varepsilon_{t-1} + \nu_t$$

If this is the case, the model violates the idea that the errors are completely unpredictable. If we would be able to view our past mistakes and improve upon our predictions - why wouldn't we?[4]

7.5.3 Equal Variance

The error term must have a constant variance throughout the range of each independent variable because we will soon see that the confidence we place in our estimates are partially determined by this variance. We are unable to change our confidence in the estimates throughout the observed range of independent values - it is one size fits all. Therefore, the size of the errors (i.e., the dispersion in the garbage can) must be constant throughout.

[4] Note that serial correlation is potentially a problem in time-series data (i.e., data that must be in chronological order).



Suppose you wanted to estimate how much of an additional dollar of income the population would spend on consumption.[5] Your data set has 50 household observations from each of three annual income levels: $10,000, $75,000, and $150,000 as well as their annual consumption expenditures. As the figure illustrates, households earning around $10,000 a year all have roughly the same consumption level (because they all save very little). As income levels increase, you see more *dispersion* in consumption expenditures because more income is paired with more options. Households earning $150,000 annually could choose to save a majority of it or even go into debt (i.e., spend more than $150,000). This data could be used to estimate a regression line (illustrated in black), but you can see that the model looks like it does a poorer and poorer job of predicting consumption expenditures as the income levels increase. This means that the forecast errors are increasing as income levels increase, and this is *heteroskedasticity.* We will briefly come back to potential solutions to this later in the advanced topics section.

7.5.4 Normality of Errors

We know that OLS will produce forecast errors that have a mean of zero as well as a variance that is as low as possible by finding the *best fitting* straight line. The assumption that these are now the two moments that can be used to describe a normal distribution comes directly from the Central Limit Theorem and the concept of a sampling distribution. Recall that the *population* error

[5] In economics this is called the *Marginal Propensity to Consume* and is an important measure for considering who should and should not get hit with a tax.

term is zero on average and has some nonzero variance. A random sample of these error terms should have similar characteristics, as well as comprising a normal distribution.

7.6 Statistical Inference

Once the assumptions of the regression model have been verified, we are able to perform statistical inference. Since we are now dealing with a regression model, not only are we able to calculate confidence intervals and conduct hypothesis tests on the population coefficients, but we are able to perform statistical inference on the forecasts of the model as well.

7.6.1 Confidence Intervals (around population parameters)

Recall our earlier formula for calculating a confidence interval in a single-variable context:

$$Pr\left(\bar{X}-t_{(\frac{\alpha}{2},df=n-1)}\frac{S}{\sqrt{n}}\leq\mu\leq\bar{X}+t_{(\frac{\alpha}{2},df=n-1)}\frac{S}{\sqrt{n}}\right)=1-\alpha$$

We used the CLT to ultimately state that $\bar{X}$ was drawn from a normal distribution with a mean of μ and standard deviation $\sigma/\sqrt{n}$ (but we only have S which makes this a t distribution). This line of reasoning is *very* similar to what we have with regression analyses.

First, $\hat{\beta}$ is an estimate of β just like $\bar{X}$ is an estimate of μ. However, the standard error of the sampling distribution of $\hat{\beta}$ is derived from the standard deviation of the residuals.

$$S_{\beta}=\frac{S_{YX}}{\sum(X_i-\bar{X})^2}$$

This means that we construct a *standardized* random variable from a t distribution with $n-2$ degrees of freedom.

$$t=\frac{\hat{\beta}-\beta}{S_{\beta}}$$

We have already derived a confidence interval before, so we can skip to the punchline.

$$Pr\left(\hat{\beta} - t_{(\frac{\alpha}{2},df=n-2)}S_{\beta} \leq \beta \leq \hat{\beta} + t_{(\frac{\alpha}{2},df=n-2)}S_{\beta}\right) = 1 - \alpha$$

This is the formula for a confidence interval around the *population* slope coefficient β given the estimate $\widehat{beta}$ and the regression characteristics. It can also be written compactly as before.

$$\hat{\beta} \pm t_{(\frac{\alpha}{2},df=n-2)}S_b$$

Recall our regression explaining differences in house prices given information on house sizes.

```
pander(summary(REG3))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	11.2	24.74	0.45	0.65
sqrft	0.14	0.01	11.87	0

Table 7.4: Fitting linear model: price ~ sqrft

Observations	Residual Std. Error	R^2	Adjusted R^2
88	63.62	0.62	0.62

The information included in the regression summary is all that is needed for us to construct a 95 percent ($\alpha = 0.05$) confidence interval around the *population* slope coefficient β_1.

```
# Back out all of the needed information:

Bhat1 <- summary(REG3)$coef[2,1]
SBhat1 <- summary(REG3)$coef[2,2]
N <- length(residuals(REG3))

# Find the critical t-distribution values... same as before

AL <- 0.05
df <- N-2
tcrit <- qt(AL/2,df,lower.tail = FALSE)

# Use the formula... same as before

(LEFT <- Bhat1 - tcrit * SBhat1)
```

```
## [1] 0.1167203
```

```
(RIGHT <- Bhat1 + tcrit * SBhat1)
```

```
## [1] 0.1637017
```

$$Pr(0.1167 \leq \beta_1 \leq 0.1637) = 0.95$$

This states that while an increase in house size by one square foot will increase the house price by \$140 ($\hat{\beta}_1$) on average in the sample, we can also state that an increase in house size by one square foot will increase the house price *in the population* somewhere between \$116.70 and \$163.70 with 95% confidence.

While the code above showed you how to calculate a confidence interval from scratch as we did before, there is an easier (one-line) way in R:

```
confint(REG3)
```

```
##                   2.5 %     97.5 %
## (Intercept) -37.9825309 60.3908210
## sqrft         0.1167203  0.1637017
```

7.6.2 Hypothesis Tests

We are able to conduct hypothesis tests regarding the values of the population regression coefficients. For example:

$$H_0 : \beta_1 = 0 \quad vs. \quad H_1 : \beta_1 \neq 0$$

In the context of our house price application, this null hypothesis states that the population slope between house price and size is zero… meaning that there is *no* relationship between the two variables.

Given the null hypothesis above, we follow the remaining steps laid out previously: we calculate a test statistic under the null, calculate a p-value, and conclude.

The test statistic under the null is given by

$$t = \frac{\hat{\beta}_1 - \beta_1}{S_{\beta_1}}$$

and this test statistic is drawn from a t distribution with $n - 2$ degrees of freedom. Concluding this test is no more difficult that what we've done previously.

```
B1 = 0
(tstat <- (Bhat1 - B1)/SBhat1)
```

```
## [1] 11.86555
```

```
(Pval <- pt(tstat,N-2,lower.tail=FALSE)*2)
```

```
## [1] 8.423405e-20
```

```
(1-Pval)
```

```
## [1] 1
```

Our results state that we can reject this null hypothesis with approximately 100% confidence, meaning that there is a statistically significant relationship between house prices and house sizes.

As with the confidence interval exercise above, we actually do not need to conduct hypothesis tests where the null sets the population parameter to zero because R does this automatically. If you look again at columns to the right of the estimated coefficient $\hat{\beta}_1$, you will see a t value that is exactly what we calculated above and a p value that is essentially zero. This implies that a test with the null hypothesis set to zero is always done for you.

```
summary(REG3)
```

```
##
## Call:
## lm(formula = price ~ sqrft, data = hprice1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -117.112  -36.348   -6.503   31.701  235.253
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.20415   24.74261   0.453    0.652
## sqrft        0.14021    0.01182  11.866   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 63.62 on 86 degrees of freedom
## Multiple R-squared:  0.6208, Adjusted R-squared:  0.6164
## F-statistic: 140.8 on 1 and 86 DF,  p-value: < 2.2e-16
```

This isn't to say that all hypothesis tests are automatically done for you. Suppose a realtor believes that homes sell for $150 per square foot. This delivers the following hypotheses, followed by a test statistic, p-value, and conclusion.

$$H_0 : \beta_1 = 0.150 \quad vs. \quad H_1 : \beta_1 \neq 0.150$$

```
B1 = 0.150
(tstat <- (Bhat1 - B1)/SBhat1)
```

```
## [1] -0.8284098
```

```
(Pval <- pt(tstat,N-2)*2)
```

```
## [1] 0.4097316
```

```
(1-Pval)
```

```
## [1] 0.5902684
```

Our p-value of 0.41 implies that there is a 41% chance of being wrong if we reject the null hypothesis. We therefore do not have evidence that the population slope is different from 0.150… so we do not reject.

One-sided tests are also like before. Suppose a realtor believes that homes sell *more than* $160 per square foot. This delivers the following hypotheses, followed by a test statistic, p-value, and conclusion.

$$H_0 : \beta_1 \leq 0.160 \quad vs. \quad H_1 : \beta_1 > 0.160$$

```
B1 = 0.160
(tstat <- (Bhat1 - B1)/SBhat1)
```

```
## [1] -1.674674
```

```
(Pval <- pt(tstat,N-2))
```

```
## [1] 0.04881561
```

```
(1-Pval)
```

```
## [1] 0.9511844
```

Our test concludes that we can reject the null with at most 95.11% confidence.

7.6.3 Confidence Intervals (around forecasts)

A regression can also build confidence intervals around the conditional expectations (i.e., forecasts) of the dependent variable.

Suppose you want to use our model to predict the price of a 1000 square foot house. The conditional expectation is calculated by using our regression coefficients, a value of house size of 1000, and setting our forecast error to zero.

```
X = 1000
Bhat0 = summary(REG3)$coef[1,1]
Bhat1 = summary(REG3)$coef[2,1]

(Yhat = Bhat0 + Bhat1 * X)
```

```
## [1] 151.4151
```

Another way to calculate this forecast is using the predict command in R. This command creates a new data frame that includes only the value for the independent variable you want to make a prediction with. The rest is done for you.

```
predict(REG3,data.frame(sqrft = 1000))
```

```
##        1
## 151.4151
```

Our model predicts that a 1,000 square foot house will sell for \$151,415 on average. While this is an expected value based on the sample, what is the prediction in the population? We are able to build a confidence interval around this forecast in a number of ways.

- A confidence interval for the mean response
- A confidence interval for an individual response

The mean response: a confidence interval

Suppose you want to build a confidence interval around the mean price for a 1000 square foot house in the population. This is a conditional mean. In other words, we want the average house price but *only* for homes with a particular size. This conditional mean is generally given by $\mu_{Y|X=X_i}$ and in this case by $\mu_{Y|X=1000}$. Building a confidence interval for the mean response is given by

$$\hat{Y}_{X=X_i} \pm t_{(\frac{\alpha}{2},df=n-2)} S_{YX} \sqrt{h_i}$$

or

$$\hat{Y}_{X=X_i} - t_{(\frac{\alpha}{2},df=n-2)} S_{YX} \sqrt{h_i} \leq \mu_{Y|X=X_i} \leq \hat{Y}_{X=X_i} + t_{(\frac{\alpha}{2},df=n-2)} S_{YX} \sqrt{h_i}$$

where

- $\hat{Y}_{X=X_i}$ is the expectation of the dependent variable conditional on the desired value of X_i.
- S_{YX} is the standard error of the estimate (calculated previously)