Introduction to Statistics for Researchers

Derek Sonderegger

April 24, 2015

These notes were originally written for an introductory statistics course for grad students in biological sciences.

The problem with most introductory statistics courses is that they don't prepare the student for the use of advanced statistics. Rote hand calculation is easy to test, easy to grade, and easy for students to learn to do, but is useless for actually understanding how to apply statistics. Because students pursuing a Ph.D. will likely be using statistics for the rest of their professional careers, I feel that this sort of course should attempt to steer away from a "cookbook" undergraduate pedagogy, and give the student enough theoretical background to continue their statistical studies at a high level while staying away from the painful mathematical details that statisticians must work through.

Recent pedagogical changes have been made at the undergraduate level to introduce sampling distributions via permutation and bootstrap procedures. Because those are extremely useful tools in their own right and because of the ability to think about statistical inference from the very start of the course is invaluable, I've attempted to duplicate this approach. I am grateful to the ICOTS 9 organizers and presenters for their expertise, perspective, and motivation for making such a large shift in my teaching.

Statistical software has progressed by leaps and bounds over the last decades. Scientists need access to reliable software that is flexible enough to handle new problems, with minimal headaches. R has become a widely used, and extremely robust Open Source platform for statistical computing and most new methodologies will appear in R before being incorporated into commercial software. Second, data exploration is the first step of any analysis and a user friendly yet powerful mechanism for graphing is a critical component in a researchers toolbox. R succeeds in this area as R has the most flexible graphing library of any statistical software I know of and the basic plots can created quickly and easily. The only downside is that there is a substantial learning curve to learning a scripting language, particularly for students without any programming background. The R package `mosaic` attempts to overcome this difficulty by providing a minimal set of tools for doing introductory statistics that all follow the same syntactical formula. I've made every attempt to use this package to minimize the amount of R necessary.

Because the mathematical and statistical background of typical students varies widely, the course seems to have a split-personality disorder. We wish to talk about using calculus to maximize the likelihood function and define the expectation of a continuous random variable, but also must spend time defining how to calculate the a mean. I attempt to address both audiences, but recognize that it is not ideal.

As these notes are in a continual state of being re-written, I endeavor to keep the latest version available on my website http://oak.ucc.nau.edu/dls354/Home/. In general, I recommend printing the chapter we are currently covering in class.

I encourage instructors to use these notes for their own classes and appreciate notification of the use to encourage me to keep tweaking the content and presentation. Finally, I hope these notes useful to a broad range of students.

Derek Sonderegger, Ph.D.
Department of Mathematics and Statistics
Northern Arizona University

Contents

Chapter 1

Summary Statistics and Graphing

When confronted with a large amount of data, we seek to summarize the data into statistics that somehow capture the essence of the data with as few numbers as possible. Graphing the data has a similar goal... to reduce the data to an image that represents all the key aspects of the raw data. In short, we seek to simplify the data in order to understand the trends while not obscuring important structure.

For this chapter, we will consider data from a the 2005 Cherry Blossom 10 mile run that occurs in Washington DC. This data set has 8636 observations that includes the runners `state` of residence, official `time` (gun to finish, in seconds), `net` time (start line to finish, in seconds), `age`, and `gender` of the runners.

```
library(mosaic)     # library of user friendly functions we'll use
library(mosaicData) # library of datasets we'll use
head(TenMileRace)   # examine the first few rows of the data
```

```
##   state time  net age sex
## 1    VA 6060 5978  12   M
## 2    MD 4515 4457  13   M
## 3    VA 5026 4928  13   M
## 4    MD 4229 4229  14   M
## 5    MD 5293 5076  14   M
## 6    VA 6234 5968  14   M
```

In general, I often need to make a distinction between two types of data.

- Discrete (also called Categorical) data is data that can only take a small set of particular values. For example a college student's grade can be either A, B, C, D, or F. A person's sex can be only Male or Female.[1] Discrete data could also be numeric, for example a bird could lay 1, 2, 3, ... eggs in a breeding season.
- Continuous data is data that can take on an infinite number of numerical values. For example a person's height could be 68 inches, 68.2 inches, 68.23212 inches.

To decided if a data attribute is discrete or continuous, I often as "Does a fraction of a value make sense?" If so, then the data is continuous.

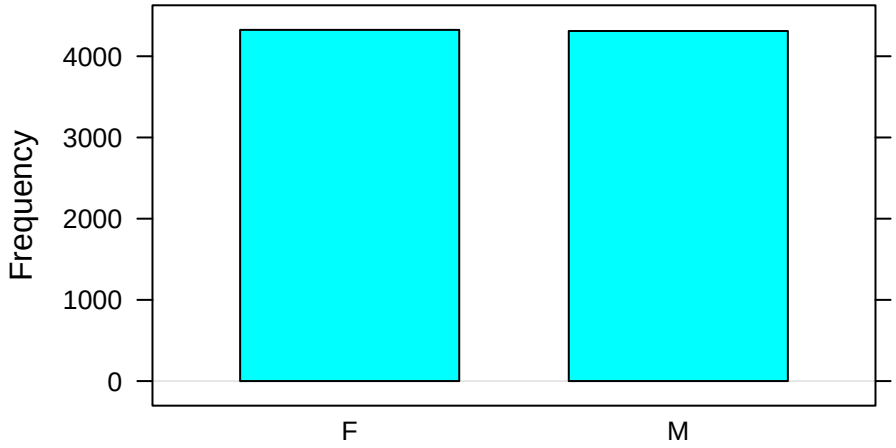
[1] Actually this isn't true as both gender and sex are far more complex. However from a statistical point of view it is often useful to simplify our model of the world. George Box famously said, "All models are wrong, but some are useful."

1.1 Graphical summaries of data

1.1.1 Univariate - Categorical

If we have univariate data about a number of groups, often the best way to display it is using barplots. They have the advantage over pie-charts that groups are easily compared.

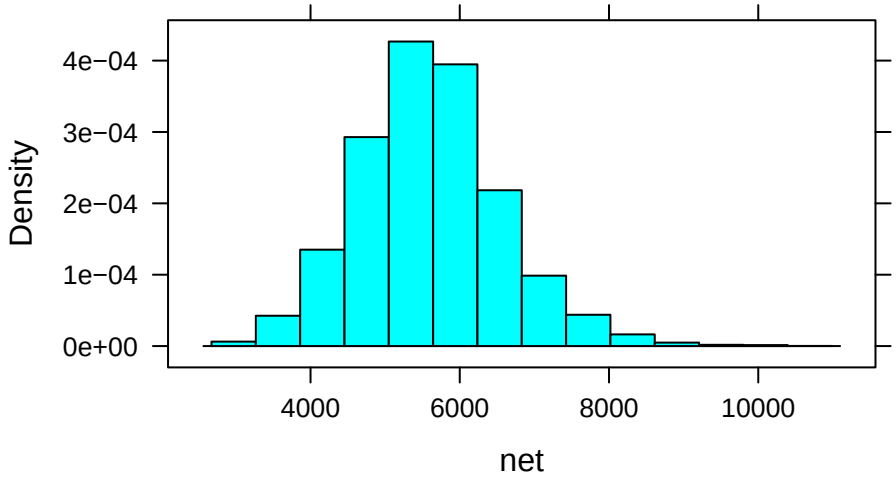
```
bargraph( ~sex, data=TenMileRace)
```



1.1.2 Univariate - Continuous

A histogram looks very similar to a bar plot, but is used to represent continuous data instead of categorical and therefore the bars will actually be touching.

```
histogram( ~net, data=TenMileRace )
```



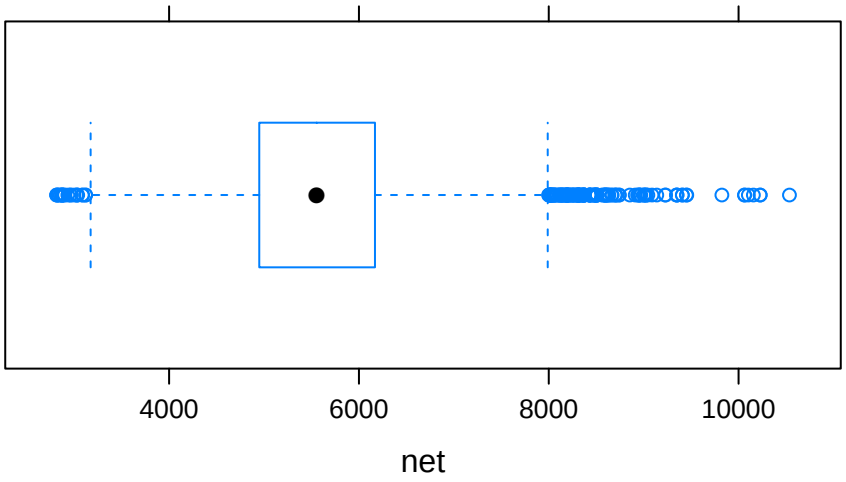
Often when a histogram is presented, the y-axis is labeled as "frequency" which is the number of observations that fall within a particular bin. However, it is often desirable to scale the percent so

that if we were to sum up the area (height * width) then the area would sum to 1. The rescaling that accomplishes this is

$$density = \frac{\#\ observations\ in\ bin}{total\ number\ observations} \cdot \frac{1}{bin\ width}$$

A second way to look at this data is to use a box-and-whisker plot.

```
bwplot( ~net, data=TenMileRace )
```



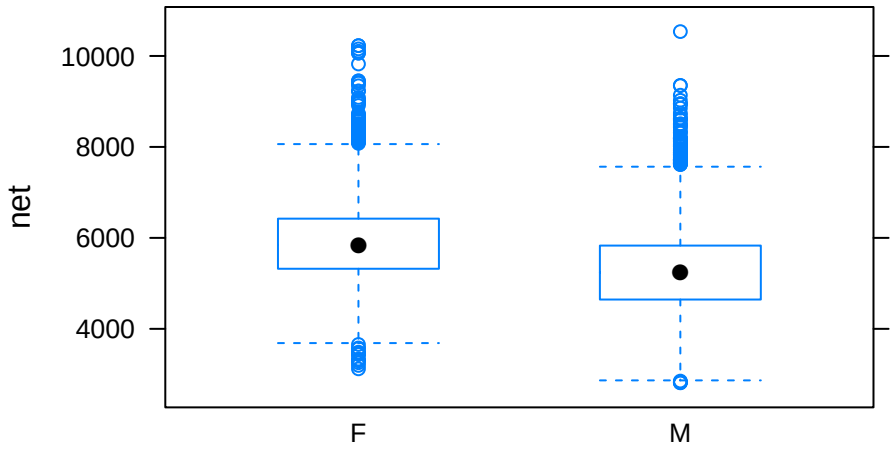
In this graph, the edges of the box are defined by the 25% and 75% quantiles. That is to say, 25% of the data is to the left of the box, 50% of the data is in the box, and the final 25% of the data is to the right of the box. The dots are data points that traditionally considered outliers.[2]

1.1.3 Bivariate - Categorical vs Continuous

We often wish to compare response levels from two or more groups of interest. To do this, we often use side-by-side boxplots. Notice that each observation is associated with a continuous response value and a categorical value.

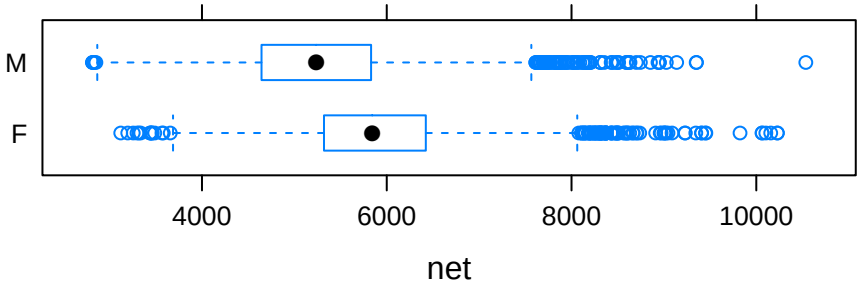
[2]Define the Inter-Quartile Range (IQR) as the length of the box. Then any observation more than 1.5*IQR from the box is considered an outlier.

```
bwplot( net ~ sex, data=TenMileRace )
```



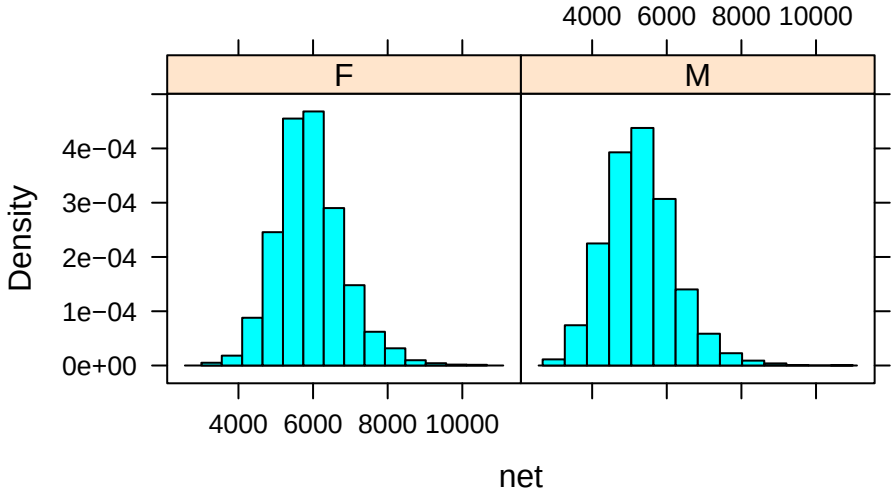
We could change the orientation of the graph just by changing the formula.

```
bwplot( sex ~ net, data=TenMileRace )
```



Sometimes I think that box-and-whisker plot obscures too much of the details of the data and we should look at the side-by-side histograms instead.

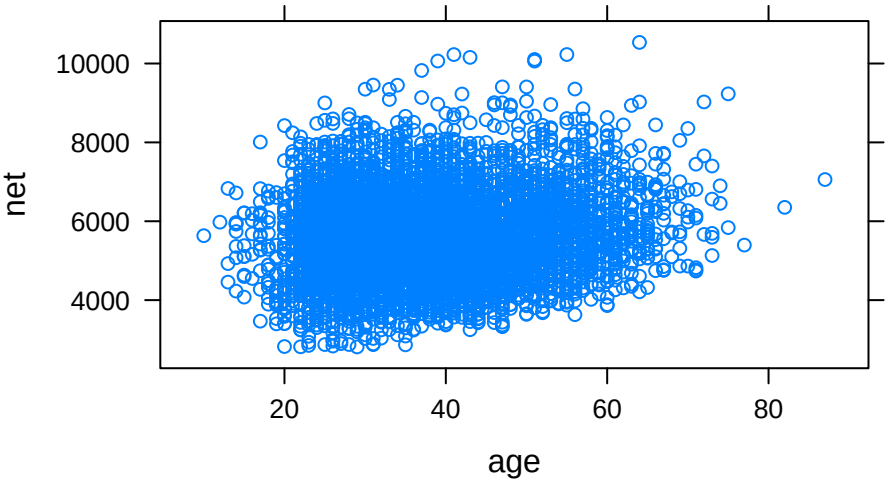
```
histogram( ~ net | sex, data=TenMileRace )
```



1.1.4 Bivariate - Continuous vs Continuous

Finally we might want to examine the relationship between two continuous random variables.

```
xyplot(net ~ age, data=TenMileRace )
```



1.2 Measures of Centrality

The most basic question to ask of any dataset is 'What is the typical value?' There are several ways to answer that question and they should be familiar to most students.

Mean

Often called the average, or arithmetic mean, we will denote this special statistic with a bar. We define

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\left(x_1 + x_2 + \cdots + x_n\right)$$

If we want to find the mean of five numbers $\{3, 6, 4, 8, 2\}$ the calculation is

$$\begin{aligned} \bar{x} &= \frac{1}{5}(3+6+4+8+2) \\ &= \frac{1}{5}(23) \\ &= 23/5 \\ &= 4.6 \end{aligned}$$

This can easily be calculated in R by using the function `mean()`.

```
mean( ~ net, data=TenMileRace )

## [1] 5599.065
```

Median

If the data were to be ordered, the median would be the middle most observation (or, in the case that n is even, the mean of the two middle most values).

In our simple case of five observations $\{3, 6, 4, 8, 2\}$, we first sort the data into $\{2, 3, 4, 6, 8\}$ and then the middle observation is clearly 4.

In R the median is easily calculated by the function `median()`.

```
mean( ~net, data=TenMileRace )

## [1] 5599.065
```

Mode

This is the observation value with the most number of occurrences.

Examples

- If my father were to become bored with retirement and enroll in my STA 570 course, how would that affect the mean and median age of my 570 students?
  - The mean would move much more than the median. Suppose the class has 5 people right now, ages 21, 22, 23, 23, 24 and therefore the median is 23. When my father joins, the ages will be 21, 22, 23, 23, 24, 72 and the median will remain 23. However, the mean would move because we add in such a large outlier. Whenever we are dealing with skewed data, the mean is pulled toward the outlying observations.
- In 2010, the median NFL player salary was $770,000 while the mean salary was $1.9 million. Why the difference?
  - Because salary data is *skewed* superstar players that make huge salaries (in excess of 20 million) while the minimum salary for a rookie is $375,000. Financial data often reflects a highly skewed distribution and the median is often a better measure of centrality in these cases.

1.3 Measures of Variation

The second question to ask of a dataset is 'How much variability is there?' Again there are several ways to measure that.

Range

Range is the distance from the largest to the smallest value in the dataset.

```
max(~net, data=TenMileRace) - min(~net, data=TenMileRace)

## [1] 7722
```

Inter-Quartile Range

The **p-th** percentile is the observation (or observations) that has at most p percent of the observations below it and $(1-p)$ above it, where p is between 0 and 100. The median is the 50th percentile. Often we are interested in splitting the data into four equal sections using the 25th, 50th, and 75th percentiles (which, because it splits the data into four sections, we often call these the 1st, 2nd, and 3rd quartiles).

In general I could be interested in dividing my data up into an arbitrary number of sections, and refer to those as *quantiles* of my data.

```
# notice the formate of this command is inconsisent. Not all functions we'll use
# have been "standardized" by the mosaic package.
quantile( TenMileRace$net )

##    0%   25%   50%   75%  100%
##  2814  4950  5555  6169 10536
```

The inter-quartile range (IQR) is defined as the distance from the 3rd quartile to the 1st.

```
IQR( ~net, data=TenMileRace )

## [1] 1219
```

Notice that we've defined IQR before when we looked at box-and-whisker plots and this is exactly the length of the box.

Variance

One way to measure the spread of a distribution is to ask "what is the average distance of an observation to the mean?" We could define the ith **deviate** as $e_i = x_i - \bar{x}$ and then ask what is the average deviate? The problem with this approach is that the sum (and thus the average) of all deviates is *always* 0.

$$\begin{aligned}
\sum_{i=1}^{n}(x_i - \bar{x}) &= \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} \\
&= n\frac{1}{n}\sum_{i=1}^{n} x_i - n\bar{x} \\
&= n\bar{x} - n\bar{x} \\
&= 0
\end{aligned}$$

The big problem is that about half the deviates are negative and the others are positive. What we really care is the distance from the mean, not the sign. So we could either take the absolute value, or square it.

Absolute values are a gigantic pain to deal with. So we square the deviates and then find the average deviate size (approximately) and call that the **sample variance**.

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

Why do we divide by $n-1$ instead of n?

1. If I divide by n, then on average, we would tend to underestimate the population variance σ^2.
2. The reason is because we are using the same set of data to estimate σ^2 as we did to estimate the population mean (μ). If I could use $\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$ as my estimator, we would be fine. But since I have to replace μ with $\bar{x}$ we have to pay a price.
3. Because the estimation of σ^2 requires the estimation of one other quantity, and using using that quantity, you only need $n-1$ data points and can then figure out the last one, we have used one *degree of freedom* on estimating the mean and we need to adjust the formula accordingly.

In later chapters we'll give this quantity a different name, so we'll introduce the necessary vocabulary here. Let $e_i = x_i - \bar{x}$ be the *error* left after fitting the sample mean. This is the deviation from the observed value to the "expected value" $\bar{x}$. We can then define the Sum of Squared Error as

$$SSE = \sum_{i=1}^{n} e_i^2$$

and the Mean Squared Error as

$$MSE = \frac{SSE}{df} = \frac{SSE}{n-1} = s^2$$

where $df = n-1$ is the appropriate degrees of freedom.

Calculating the variance of our small sample of five observations $\{3, 6, 4, 8, 2\}$, recall that the sample mean was $\bar{x} = 4.6$

x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
3	-1.6	2.56
6	1.4	1.96
4	-0.6	0.36
8	3.4	11.56
2	-2.6	6.76
sum		23.2

and so the sample variance is $23.2/(n-1) = 23.2/4 = 5.8$

Clearly this calculation would get very tedious to do by hand and computers will be much more accurate in these calculations. In R, the sample variance is easily calculated by the function `var()`.

```
var( ~net, data=TenMileRace )

## [1] 940233.5
```

Standard Deviation

The biggest problem with the sample variance statistic is that the units are in the original units-*squared*. That means if you are looking at data about car fuel efficiency, then the values would be in mpg^2 which are units that I can't really understand. The solution is to take the positive square root, which we will call the sample standard deviation.

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

But why do we take the jog through through variance? Mathematically the variance is more useful and most distributions (such as the normal) are defined by the variance term. Practically though, standard deviation is easier to think about.

The sample standard deviation is important enough for R to have function that will calculate it for you.

```
sd( ~net, data=TenMileRace )

## [1] 969.6564
```

Coefficient of Variation

Suppose we had a group of animals and the sample standard deviation of the animals lengths was 15 cm. If the animals were elephants, you would be amazed at their uniformity in size, but if they were insects, you would be astounded at the variability. To account for that, the **coefficient of variation** takes the sample standard deviation and divides by the absolute value of the sample mean (to keep everything positive)

$$CV = \frac{s}{|\bar{x}|}$$

Empirical Rule of Thumb

For any mound-shaped sample of data the following is a reasonable rule of thumb:

Interval	Approximate percent of measurements
$\bar{x} \pm s$	68%
$\bar{x} \pm 2s$	95%
$\bar{x} \pm 3s$	99.7%

Chapter 2

Hypothesis Tests Using Simulation

2.1 Hypotheses and errors

2.1.1 Null and alternative hypotheses

In elementary school most students are taught the scientific method follows the following steps:

1. Ask a question of interest.
2. Construct a hypothesis.
3. Design and conduct an experiment that challenges the hypothesis.
4. Depending on how consistent the data is with the hypothesis:
   (a) If the observed data is inconsistent with the hypothesis, then we have proven it wrong and we should consider competing hypotheses.
   (b) If the observed data is consistent with the hypothesis, design a more rigorous experiment to continue testing the hypothesis.

Through the iterative process of testing ideas and refining them under the ever growing body of evidence, we continually improve our understanding of how our universe works. The heart of the scientific method is the falsification of hypothesis and statistics is the tool we'll use to assess the consistency of our data with a hypothesis.

Science is done by examining competing ideas for how the world works and throwing evidence at them. Each time a hypothesis is removed, the remaining hypotheses appear to be more credible. This doesn't mean the remaining hypotheses are correct, only that they are consistent with the available data.

1. In approximately 300 BC, Eratosthenes[1] showed that the world was not flat[2] by measuring the different lengths of shadows of identical sticks in two cities that were 580 miles apart but lay on the same meridian (Alexandria is directly north of Aswan). His proposed alternative was that the Earth was a sphere. While his alternative is not technically true (it is actually an oblate spheroid that bulges at the equator), it was substantially better than the flat world hypothesis.

[1] For more about Eratosthenes, start at his wikipedia page. http://en.wikipedia.org/wiki/Eratosthenes
[2] Carl Sagan has an excellent episode of *Cosmos* on this topic. https://www.youtube.com/watch?v=G8cbIWMv0rI

2. At one point it was believed that plants "ate" the soil and turned it into plant mass. A experiment to test this hypothesis was performed by Johannes Baptista van Helmont in 1648 in which he put exactly 200 pounds of soil in a pot and then grew a willow tree out of it for five years. At the end of the experiment, the pot contained 199.875 pounds of soil and 168 pounds of willow tree. He correctly concluded that the plant matter was not substantially taken from the soil but incorrectly jumped to the conclusion that the mass must of have come from the water that was used to irrigate the willow.

It is helpful to our understanding to label the different hypothesis, both the ones being tested and the different alternatives. We'll label the hypothesis being tested as H_0 which we often refer to as the "**null hypothesis**." The **alternative hypothesis**, which we'll denote H_a, should be the opposite of the null hypothesis. Had Eratosthenes known about modern scientific methods, he would have correctly considered H_0 : *the world is flat* verses H_a: *the world is not flat* and not incorrectly concluded that the world is a sphere[3]. Likewise Helmont should have considered the hypotheses H_0 : *plants only consume soil* versus the alternative H_a : *plants consume something besides soil.*

In both of cases, the observed data was compared to what would have been expected if the null hypothesis was true. If the null was true Eratosthenes would have seen the same length shadow in both cities and Helmont would have seen 168 pounds of willow tree and $200 - 168 = 32$ pounds of soil remaining.

2.1.2 Error

Unfortunately the world is not a simple place and experiments rarely can isolate exactly the hypothesis being tested. We can repeat an experiment and get slightly different results each time due to variation in weather, temperature, or diligence of the researcher. If we are testing the effectiveness of a new drug to treat a particular disease, we don't trust the results of a single patient, instead we wish to examine many patients (some that receive the new drug and some the receive the old) to average out the noise between the patients. The questions about how many patients do we need to have and how large of a difference between the treatments is large enough to conclude the new drug is better are the heart of modern statistics.

Suppose we consider the population of all US men aged 40-60 with high blood pressure (there might be about 20 million people in this population). We want to know if exercise and ACE inhibitors lower systolic blood pressure better than exercise alone for these people. We'll consider the null hypothesis that *exercise is equivalent to exercise and ACE inhibitors* versus *exercise is different than exercise and ACE inhibitors.* If we could take every single member of the population and expose them to exercise or exercise with ACE inhibitors, we would know for certain how the population reacts to the different treatments. Unfortunately this is too expensive and ethically dubious.

Instead of testing the entire population we'll take a sample of n men from the population and treat half of them with exercise alone and half of them with exercise and ACE inhibitors. What might our data look like if there is a difference between the two treatments at different samples sizes compared to if there is no difference? At small sample sizes it is difficult to distinguish the effect of the treatment when it is masked by individual variation. At high sample sizes, the individual variation is smoothed out and the difference between the treatments can be readily seen.

[3]Amusingly Eratosthenes' data wasn't inconsistent with the hypothesis that the world was shaped like a donut, but he thought the sphere to be more likely.

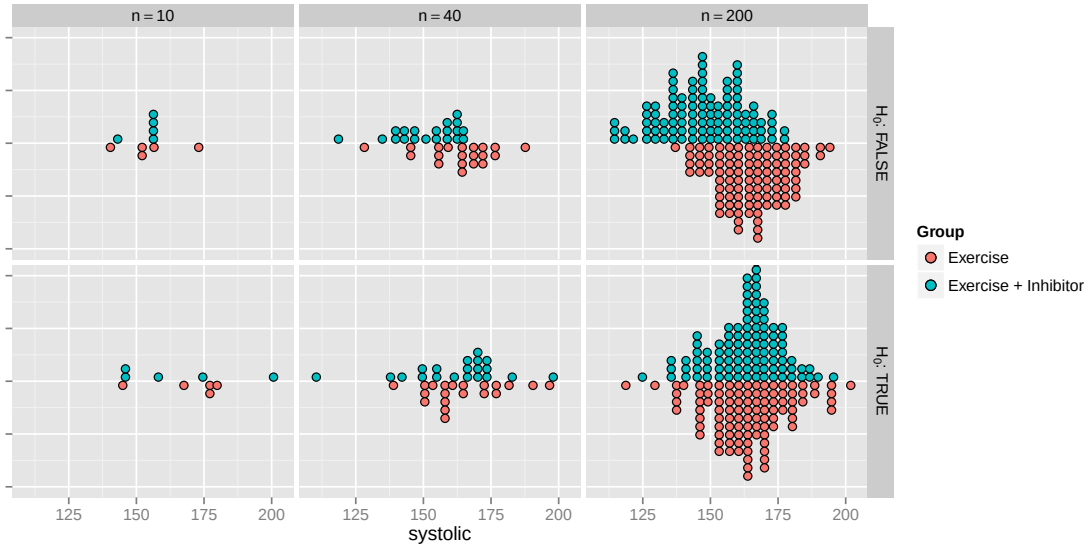


Figure 2.1.1: Comparing possible data assuming there is a difference between treatments versus no difference. In the top row of graphs, there is a difference between the *Exercise* and the *Exercise + Inhibitor* treatments. However, at small sample sizes, we can't tell if the observed difference is due to the difference in treatment or just random variation in the data. In the second row, there is no difference between the treatments.

When the sample size is large it is easy to see if the treatments differ in their effect on systolic blood pressure, but at medium or small sample sizes, the question is much harder. It is important to recognize that the core of the problem is still *"is the observed data consistent with the null hypothesis?"* but we now have to consider an addition variability term that is unrelated to the research hypothesis of interest. In the above example, the small sample data is consistent with the null hypothesis even when the null hypothesis is false!

To clarify this idea, it is useful to clarify our notation to distinguish between the population parameters (numbers the describe the population) versus the sample statistics (numbers that describe the sample).

- Population Parameters - These quantities will be denoted with Greek letters such as μ for means, π for proportions, σ^2 for variances. These are the quantities that we are interested in, but do not know the actual values (unless we perform a complete census on the population).
- Sample Statistics - These are the quantities that are calculated from the sample and will be denoted with Roman letters, $\bar{x}$ for means, p for proportions, s^2 for variances. We only care about these values in the sense that they are proxies for the population parameters.

For a sample to be representative of the population, we want it to be selected in a manner that avoid bias. For example, a political phone survey should avoid just calling retired (and generally more conservative) individuals but try to sample both young and old people in roughly the same percentage as in the population. If we are performing a study on 50-60 year old males with heart disease, we should strive to get an ethnic diversity that reflects the population, or recognize the sample is actually representative of a particular subpopulation.

2.2 Sampling distribution of a proportion

In order to conclude the null hypothesis is incorrect, we must compare the data we observed to what we could have seen if the null hypothesis was correct. Just as Eratosthenes expected to see equal

length shadows if the earth was flat, we need to consider what data we could have observed if the null hypothesis was correct.

To assess if the our data is consistent with the null hypothesis, it is useful to reduce the observed data into a summary statistic that is related to the hypothesis of interest. Sometimes the summary statistic is obvious, but sometimes not. We then will compare the observed sample statistic to the distribution of statistic values that we could have seen. We will call this distribution of possible sample statistics (assuming H_0 is true) the sampling distribution of the statistic.

There are several different methods we'll use to produce the sampling distribution and are often specific to a particular to the design of the study but the goal of producing possible realizations of the statistic that are consistent with the null hypothesis and the study design will be a constant theme.

2.2.1 $H_0 : \pi = \frac{1}{2}$

We will first consider the case where we wish to test the null hypothesis that some proportion is equal to $\frac{1}{2}$. The method will then be slightly modified to address testing for proportions other than $\frac{1}{2}$.

While the human sex ratio tends to be very close to $\frac{1}{2}$ female and $\frac{1}{2}$ male, this isn't necessarily true for the gender[4] ratio of graduate students. It is reasonable to suppose a null and alternative hypotheses of

$$\begin{aligned} H_0 &: \quad \pi = \tfrac{1}{2} \\ H_a &: \quad \pi \neq \tfrac{1}{2} \end{aligned}$$

where π represents the true proportion of graduate students that are female. In our class (Spring 2015) the proportion of females is $p = \frac{16}{38} \approx 0.421$. If we consider this class as a sample of NAU grad students (a dubious assumption), is our observed data consistent with the null hypothesis?

What set of values should we expect to see if the probability that any randomly selected student is female is $\frac{1}{2}$? In later chapters we'll use the rules of probability answer the question, but for now we'll turn to simulation. We could use fair coins and have every student flip a coin separately and count the number of heads, and denote the proportion of heads as p_1. We can repeat this procedure 10 times and graph the observed proportions $p_1, p_2, \ldots, p_{10}$.

[4] Broadly categorizing sex and gender (they are different!) into just Male/Female is a broad simplification of a complex continuum.

```
library(mosaic)                    # load a library of functions useful for teaching
rflip(38)                          # flip a fair coin 38 times

##
## Flipping 38 coins [ Prob(Heads) = 0.5 ] ...
##
## T H H T T H H H H H T T T T T T T H T T H H T H T H H H H H T H T T H H T T
##
## Number of Heads: 19 [Proportion Heads: 0.5]

coin.flips <- do(10) * rflip(38) # repeat this process 10 times
coin.flips

##     n heads tails      prop
## 1  38    15    23 0.3947368
## 2  38    21    17 0.5526316
## 3  38    21    17 0.5526316
## 4  38    21    17 0.5526316
## 5  38    14    24 0.3684211
## 6  38    12    26 0.3157895
## 7  38    20    18 0.5263158
## 8  38    18    20 0.4736842
## 9  38    20    18 0.5263158
## 10 38    22    16 0.5789474
```

```
# width tells it not to round distinct .54 and .51 into one group
dotPlot( ~prop, data=coin.flips, width=.01)
```

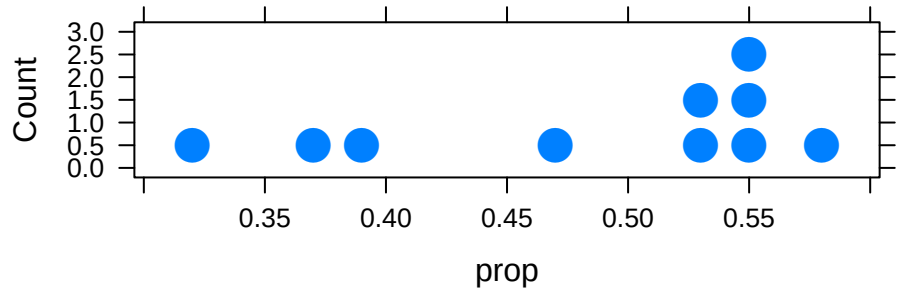


Figure 2.2.1: Each dot represents the outcome of a simulation where a simulation consists of calculating the proportion of 38 coin flips that were heads assuming $\pi = 0.5$.

This gives some idea of what possible values the sample proportion could take on, but we should generate more, say 300 times.

```
coin.flips <- do(300) * rflip(38) # repeat this process 300 times
dotPlot( ~prop, data=coin.flips, width=0.01 )
```

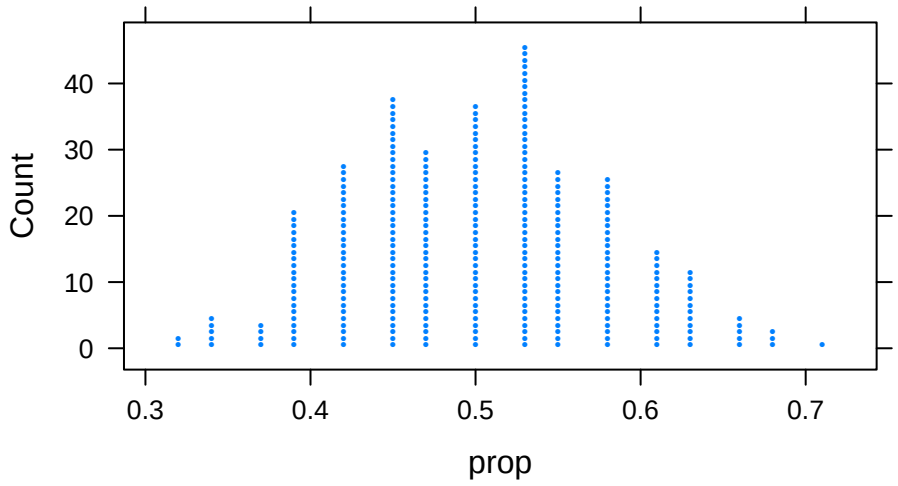


Figure 2.2.2: Each dot represents the proportion of 38 coin flips that were heads assuming $\pi = 0.5$. This is the sampling distribution of the sample proportion of 38 coin flips assuming $\pi = 0.5$.

With this number of observations, a dotplot isn't the greatest plot and we'll switch to a `histogram` with bins narrow enough so that the each bin represents a particular number of heads.

```
histogram( ~prop, data=coin.flips, width=1/38 )
```

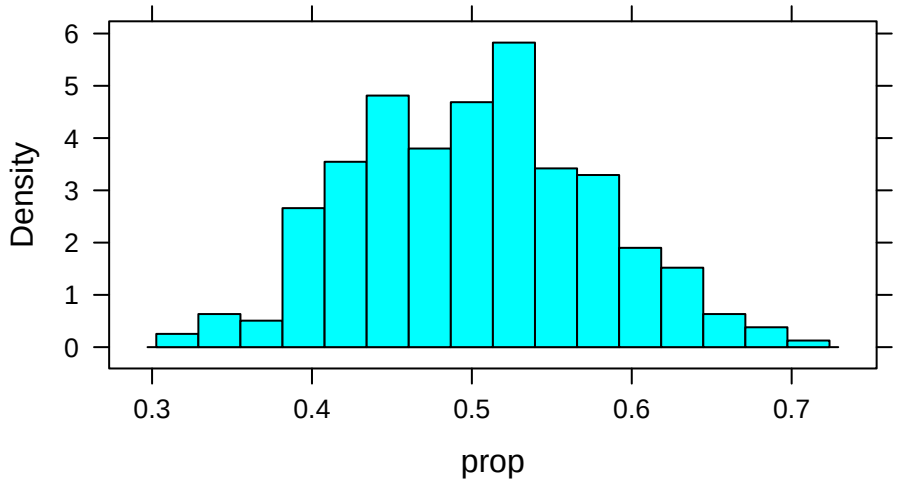


Figure 2.2.3: Each column's area represents the percent of 300 simulations of the percent heads 38 coin flips, assuming the probability of heads is $\pi = 0.5$. This is the sampling distribution of the sample proportion of 38 coin flips assuming $\pi = 0.5$.

Given this graph, the observed percentage of female students in class $p \approx .421$ doesn't appear

inconsistent with the hypothesis that there are an equal number of female grad students as male grad students. We want to make it more obvious where our observed data is in the graph, so we'll color the histogram bars using the `groups` argument.

```
histogram( ~prop, data=coin.flips,
           width=1/38, groups= (prop <= 16/38) )
```

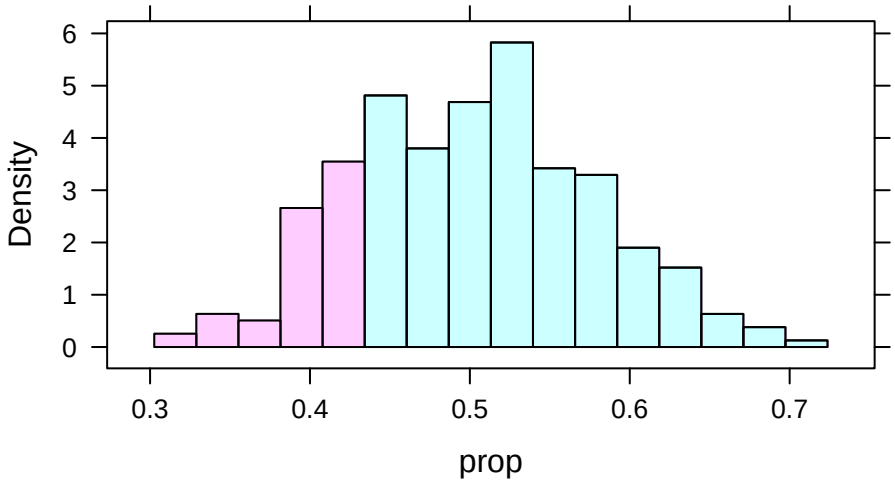


Figure 2.2.4: Each column's area represents the percent of 300 simulations of the percent heads 38 coin flips, assuming the probability of heads is $\pi = 0.5$. This is the sampling distribution of the sample proportion of 38 coin flips assuming $\pi = 0.5$. We have colored the simulations that resulted in a sample proportion $\leq \frac{16}{38}$.

We can summarize our simulation using the `tally` command which counts the number of observations in each category

```
# How many simulations had 10 heads, how many had 11, etc...
tally(~heads, data=coin.flips, format='count')

##
## 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27
##  2  5  4 21 28 38 30 37 46 27 26 15 12  5  3  1

# How many simulations had a proportion less than 16/38?
tally( ~ (prop <= 16/38), data=coin.flips, format='count' )

##
##  TRUE FALSE
##    60   240
```

```
# What proportion of simulations had a sample proportion less than 16/38?
tally( ~ (prop <= 16/38), data=coin.flips, format='proportion' )

##
##  TRUE FALSE
##   0.2   0.8
```

and we see that 60 of the 300 flips (0.2 percent) of the simulated sample proportions are less than our observed classroom proportion of $p = \frac{16}{38} \approx 0.421$.

We will formalize this logic by defining the *p-value*.

p-value: The p-value is the probability of seeing the observed test statistic, or something more extreme, given the null hypothesis is true.

Our simulated distribution was created assuming the null hypothesis is true so we merely have to consider "What outcomes are more extreme than our test statistic?" In the case of our null hypothesis, any sample proportion farther than $0.5 - 0.421 = 0.079$ from the null hypothesis of 0.5 would be more extreme. This includes clearly anything less than our observed sample proportion of 0.421, but also anything greater than 0.579! In terms of numbers of heads, we want to find the proportion that is $\leq \frac{16}{38}$ or $\geq \frac{22}{38}$.

```
histogram( ~prop, data=coin.flips,
           width=1/38, groups= (prop <= 16/38 | 22/38 <= prop ) )
```

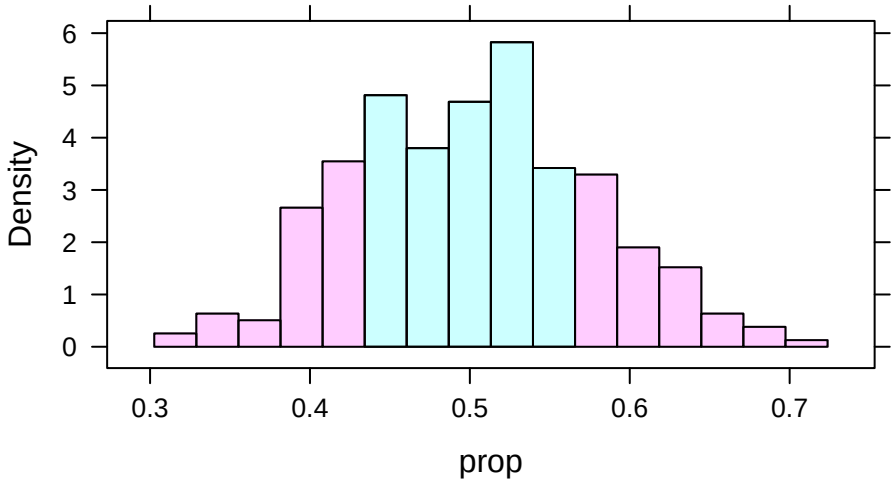


Figure 2.2.5: Each column's area represents the percent of 300 simulations of the percent heads 38 coin flips, assuming the probability of heads is $\pi = 0.5$. This is the sampling distribution of the sample proportion of 38 coin flips assuming $\pi = 0.5$. We have colored the simulations that resulted in a sample proportion more extreme than the observed $p = \frac{16}{38}$.

We first notice that while we observed 60 simulations with $p \leq \frac{16}{38}$, there were an additional 62 with $p \geq \frac{22}{38}$ for a total of 122 simulations that were "at least as extreme as what we actually observed."

```
tally( ~ (prop <= 16/38 | 22/38 <= prop), data=coin.flips, format='count' )

##
##  TRUE FALSE
##   122   178
```

The resulting p-value can be calculated via:

```
# p-value: The TRUE column is the "extreme" simulation percentage
tally( ~ (prop <= 16/38 | 22/38 <= prop), data=coin.flips, format='proportion' )

##
##      TRUE     FALSE
## 0.4066667 0.5933333
```

2.2.2 General case: $H_0 : \pi = \pi_0$

We might be interested in a phenomena that where the null hypothesis is something other than $H_0 : \pi = \frac{1}{2}$. For example, in Gregor Mendel's classic pea experiments, the proportion of observations with the recessive phenotype should be $\frac{1}{4}$. Similarly, it is known that the percent of high school students (in 2012) that smoke is 14% and we wish to test if students exposed to some treatment have smoking rates that are different than the national rate.

Example: The the British television show *The Man Lab*, James May performed an experiment to test the Monte Hall in Series 3 Episode 3[5]. The mathematical theory suggests that the probability of winning given that James switches doors is 2/3. James tested this by playing 100 games, but James only won $p = \frac{60}{100} = 0.6$ of the games. Is the experimental data consistent with the null hypothesis of $H_0 : \pi = \frac{2}{3}$? Is there result possible due to random chance or is this evidence that the procedure they used had some sort of bias or that the mathematicians are wrong?

Again we need to consider what is meant by "more extreme?" Because the null hypothesis is that the true proportion is $\pi = \frac{2}{3}$ then an experiment where we saw too few successes (as James did) would be evidence against the null hypothesis as would an experiment where we saw too many. So in this case we are interested in any simulation case where the observed sample proportion is more than $\left|\frac{2}{3} - \frac{60}{100}\right| = 0.06\bar{6}$ from the hypothesized value of $\frac{2}{3}$. That is we want to find the percentage of simulations that satisfy

$$|p^* - \pi_0| \geq |p - \pi_0|$$

where p^* are my simulated sample proportions, p is my observed sample proportion, and π_0 is my null hypothesis value (in this case $\frac{2}{3}$)

[5] https://www.youtube.com/watch?v=tvODuUMLLgM

```
coin.flips <- do(1000) * rflip(100, prob=2/3) # repeat the simulation 1000 times
histogram( ~prop, data=coin.flips,
           width=1/100, groups=(abs(prop - 2/3) >= abs(0.6 - 2/3) ),
          main='Sampling distribution of p*', xlab='p*')
```

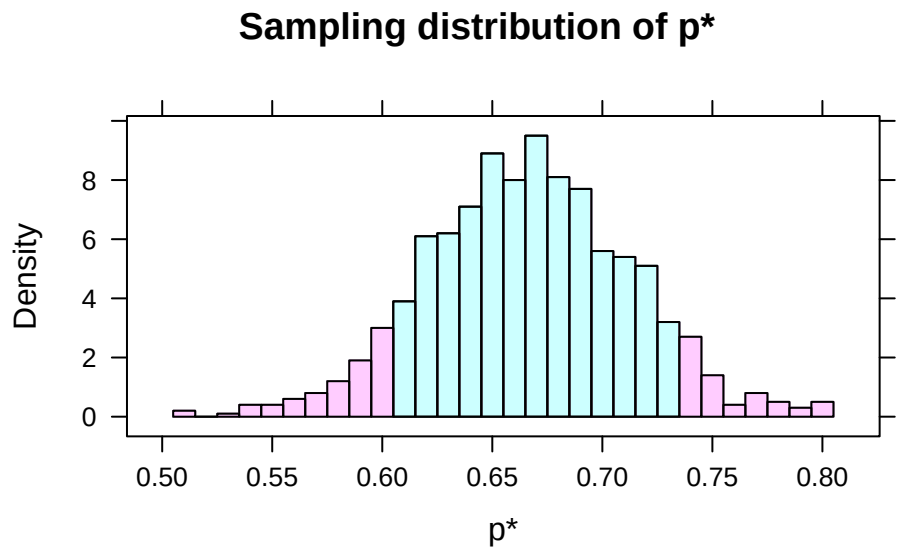


Figure 2.2.6: The histogram shows the sampling distribution of p^*, the proportion won of $n = 100$ games where each game has a true probability of winning of $\pi = \frac{2}{3}$. The colored tails denote the proportion of simulations with successes more inconsistent with $\pi = \frac{2}{3}$ than James May's 60% success proportion.

Summarizing the simulations, we see the proportion of simulations with a more extreme sample proportion as

```
# what percent of simulations were more extreme than 60/100?
p.value <- tally( ~ (abs(prop - 2/3) >= abs(0.6 - 2/3)),
                  data=coin.flips,
                  format='proportion' )
p.value

##
##  TRUE FALSE
## 0.152 0.848
```

we see that 15.2% of the simulations resulted in sample proportions more extreme than 60% so the *Man Lab*'s observed data is not particularly inconsistent with the mathematical theory that we should see $\frac{2}{3}$ of trials be successes.

2.3 Experimental assignment to groups

There are two broad classifications of types of research, *observational studies* and *designed experiments*. These two types of research differ in the way that the researcher interacts with the subjects being observed. In an observational study, the researcher doesn't force a subject into some behavior or treatment, but merely observes the subject (making measurements but not changing behaviors). In contrast, in an experiment, the researcher imposes different treatments onto the subjects and the pairing between the subject and treatment group happens at random.

Example: For many years hormone (Estrogen and Progestin) replacement therapy's primary use for post-menopausal woman was to reduce the uncomfortable side-effects of menopause but it was thought to also reduced the rate of rate of breast cancer in post-menopausal women. This belief was the result of many observational studies where women who chose to take hormone replacement therapy also had reduced rates of breast cancer. The *lurking*[6] *variable* thing that the observational studies missed was that hormone therapy is relatively expensive and was taken by predominately women of a high socio- economic status. Those women tended to be more health conscious, lived in areas with less pollution, and were generally at a lower risk for developing breast cancer. Even when researchers realized that socio-economic status was *confounded*[7] with the therapy, they couldn't be sure which was the cause of the reduced breast cancer rates. To correctly test this, nearly 17,000 women underwent an experiment in which each women was randomly assigned to take either the treatment (E+P) or a placebo. The Women's Health Initiative (WHI) Estrogen plus Progestin Study[8] (E+P) was stopped on July 7, 2002 (after an average 5.6 years of follow-up) because of increased risks of cardiovascular disease and breast cancer in women taking active study pills, compared with those on placebo (inactive pills). The study showed that the overall risks exceeded the benefits, with women taking E+P at higher risk for heart disease, blood clots, stroke, and breast cancer, but at lower risk for fracture and colon cancer. Lurking variables such as income levels and education are correlated to overall health behaviors and with an increased use of hormone replacement therapy. By randomly assigning each woman to a treatment, the unidentified lurking variables were evenly spread across treatments and the dangers of hormone replacement therapy were revealed.

There is a fundamental difference between imposing treatments onto subjects versus taking a random sample from a population and observing relationships between variables. In general, designed experiments allow us to determine cause-and-effect relationships while observational studies can only determine if variables are correlated. This difference in how the data is generated will result in different methods for generating a sampling distribution for a statistic of interest. In this chapter we will focus on experimental designs.

2.3.1 Two groups, continuous response variable

Often researchers will obtain a group of subjects and them into two groups, provide different treatments to each, and the observe some response. The goal is to see if the two groups are different in some fashion.

The first thing to consider is that the group of subjects in our sample is should be representative of a population of interest. Because we cannot impose an experiment on an entire population, we often are forced to examine a small sample and we hope that the sample statistics (the sample mean $\bar{x}$, and sample standard deviation s) are good estimates of the population parameters (the population mean μ, and population standard deviation σ) First recognize that these are a sample and we generally think of them to be representative of some population.

Second, the way that the two groups are different could be if the mean of group 1 is greater than the mean of the other. This is the most common difference to be interested in and we are usually interested in the difference between the mean responses. There is nothing to say that we couldn't look at the difference in variance, using the same method we'll cover next.

Example: Finger Tapping and Caffeine

The effects of caffeine on the body have been well studied. In one experiment,[9] a group of male college students were trained in a particular tapping movement and to tap at a rapid rate. They

[6] A *lurking variable* is a variable the researcher hasn't considered but affects the response variable. In observational studies a researcher will try to measure all the variables that might affect the response but will undoubtable miss something.

[7] Two variables are said to be *confounded* if the design of a given experiment or study cannot distinguish the effect of one variable from the other.

[8] Risks and Benefits of Estrogen Plus Progestin in Healthy Postmenopausal Women: Principal Results From the Women's Health Initiative Randomized Controlled Trial. JAMA. 2002;288(3):321-333. doi:10.1001/jama.288.3.321.

[9] Hand, A.J., Daly, F., Lund, A.D., McConway, K.J. and Ostrowski, I., *Handbook of Small Data Sets*, Chapman and Hall, London, 1994, p. 40.

were randomly divided into caffeine and non-caffeine groups and given approximately two cups of coffee (with either 200 mg of caffeine or none). After a 2-hour period, the students tapping rate was measured.

The population that we are trying to learn about is male college-aged students and we the most likely question of interest is if the mean tap rate of the caffeinated group is different than the non-caffeinated group. Notice that we don't particularly care about these 20 students, but rather the population of male college-aged students so the hypotheses we are interested in are

$$\begin{aligned} H_0 &: \quad \mu_c = \mu_{nc} \\ H_a &: \quad \mu_c \neq \mu_{nc} \end{aligned}$$

where μ_c is the mean tap rate of the caffeinated group and μ_{nc} is the mean tap rate of the non-caffeinated group. We could equivalently express these hypotheses via

$$\begin{aligned} H_0 &: \quad \mu_{nc} - \mu_c = 0 \\ H_a &: \quad \mu_{nc} - \mu_c \neq 0 \end{aligned}$$

Or we could let $\delta = \mu_{nc} - \mu_c$ and write the hypotheses as

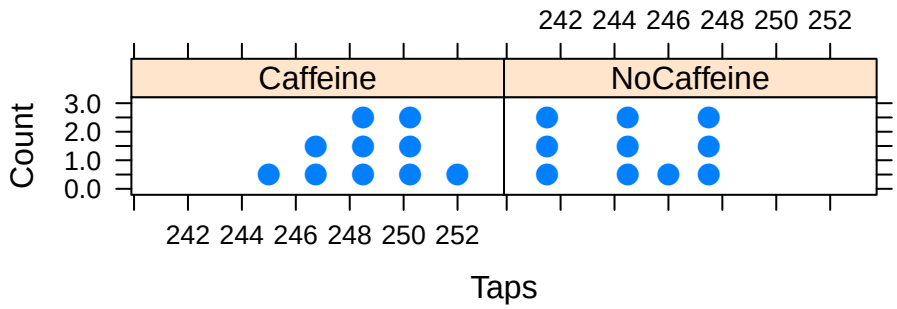
$$\begin{aligned} H_0 &: \delta \quad = \quad 0 \\ H_a &: \delta \quad \neq \quad 0 \end{aligned}$$

The data are available in many different formats at http://www.lock5stat.com/datapage.html

```
# the gdata package allows a user to load Excel files. Most R folks prefer
# comma separated files ".csv" files because they are not dependent on MS.
library(gdata)
CaffeineTaps <- read.xls('http://www.lock5stat.com/datasets/CaffeineTaps.xls')
CaffeineTaps <- read.csv('http://www.lock5stat.com/datasets/CaffeineTaps.csv')

# gdata and mosaic fight... best to unload gdata after we have our data
detach('package:gdata', unload=TRUE)

str(CaffeineTaps)

## 'data.frame': 20 obs. of  2 variables:
##  $ Taps : int  246 248 250 252 248 250 246 248 245 250 ...
##  $ Group: Factor w/ 2 levels "Caffeine","NoCaffeine": 1 1 1 1 1 1 1 1 1 1 ...
```

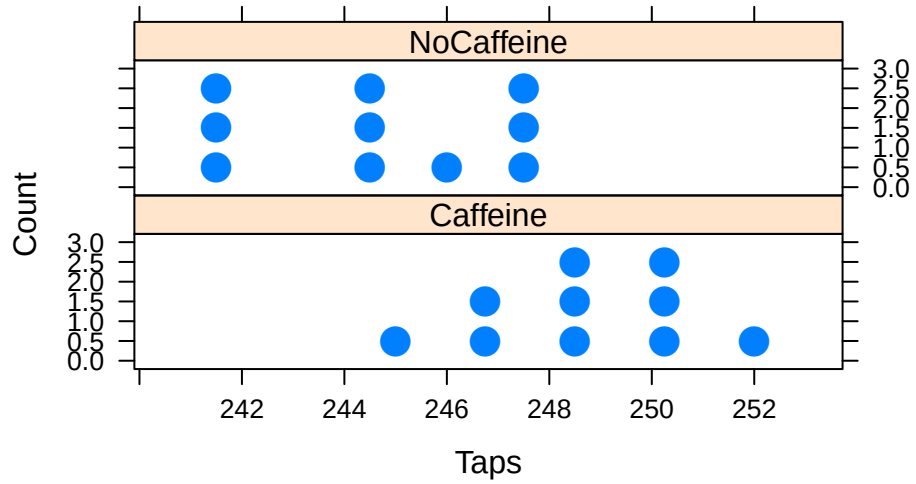
The first thing we should do is, as always, is graph the data. Because there are only 10 individuals in each group, a dotplot is an appropriate choice.

```
dotPlot(~ Taps | Group, data=CaffeineTaps)
```



The default orientation isn't as helpful as we'd like. To stack the two plots, we'll use the `layout` argument.

```
dotPlot(~ Taps | Group, data=CaffeineTaps, layout=c(1,2) )
```



From this view, it looks like the caffeine group has a higher tapping rate. It will be helpful to summarize the difference between these two groups with a single statistic by calculating the mean for each group and then calculate the difference between the group means.

```
# same result regardless of how we ask for it
mean(~ Taps | Group, data=CaffeineTaps)

##   Caffeine NoCaffeine
##      248.3      244.8

mean(  Taps ~ Group, data=CaffeineTaps)

##   Caffeine NoCaffeine
##      248.3      244.8

mean( ~ Taps, group = Group, data=CaffeineTaps )

##   Caffeine NoCaffeine
##      248.3      244.8
```

```
# No Caffeine  -  Caffeine
244.8 - 248.3

## [1] -3.5

d <- diffmean(Taps ~ Group, data=CaffeineTaps)
d

## diffmean
##     -3.5
```

Notationally, lets call this statistic $d = \bar{x}_{nc} - \bar{x}_c = -3.5$. We are interested in testing if this observed difference might be due to just random chance and we just happened to assigned more of

the fast tappers to the caffeine group. How could we test the null hypothesis that the mean of the caffeinated group is different than the non-caffeinated?

The key idea is *"How could the data have turned out if the null hypothesis is true?"* If the null hypothesis is true, then the caffeinated/non-caffeinated group treatment had no effect on the tap rate and it was just random chance that the caffeinated group got a larger percentage of fast tappers. That is to say the group variable has no relationship to tap rate. I could have just as easily assigned the fast tappers to the non-caffeinated group purely by random chance. So our simulation technique is to **shuffle the group labels** and then calculate a difference between the group means!

We can perform this shuffling with the following code:

```
# mutate: creates a new column
# shuffle: takes an input column and reorders it randomly
ShuffledCaffeine <- mutate(CaffeineTaps, ShuffledGroup = shuffle(Group))
ShuffledCaffeine
```

```
##    Taps      Group ShuffledGroup
## 1   246   Caffeine    NoCaffeine
## 2   248   Caffeine      Caffeine
## 3   250   Caffeine      Caffeine
## 4   252   Caffeine    NoCaffeine
## 5   248   Caffeine      Caffeine
## 6   250   Caffeine      Caffeine
## 7   246   Caffeine    NoCaffeine
## 8   248   Caffeine    NoCaffeine
## 9   245   Caffeine      Caffeine
## 10  250   Caffeine    NoCaffeine
## 11  242 NoCaffeine      Caffeine
## 12  245 NoCaffeine    NoCaffeine
## 13  244 NoCaffeine    NoCaffeine
## 14  248 NoCaffeine    NoCaffeine
## 15  247 NoCaffeine      Caffeine
## 16  248 NoCaffeine    NoCaffeine
## 17  242 NoCaffeine      Caffeine
## 18  244 NoCaffeine      Caffeine
## 19  246 NoCaffeine      Caffeine
## 20  242 NoCaffeine    NoCaffeine
```

We can then calculate the mean difference but this time using the randomly generated groups, and now the non-caffeinated group just happens to have a slightly higher mean tap rate just by the random sorting into two groups.

```
mean(    Taps ~ ShuffledGroup, data=ShuffledCaffeine)
```

```
##   Caffeine NoCaffeine
##      246.2      246.9
```

```
diffmean(Taps ~ ShuffledGroup, data=ShuffledCaffeine)
```

```
## diffmean
##      0.7
```

We could repeat this shuffling several times and see the possible values we might have seen if the null hypothesis is correct and the treatment group doesn't matter at all.

```
do(5) * diffmean(Taps ~ shuffle(Group), data=CaffeineTaps)

##   diffmean
## 1      0.1
## 2      2.1
## 3     -0.3
## 4      0.5
## 5      0.5
```

Of course, five times isn't sufficient to understand the sampling distribution of the mean difference under the null hypothesis, we should do more.

```
SamplingDist <- do(10000) * diffmean(Taps ~ shuffle(Group), data=CaffeineTaps)
histogram( ~ diffmean, data=SamplingDist, groups=( abs(diffmean) >= 3.5 ),
           main='Sampling distribution of d*', xlab='d*')
```

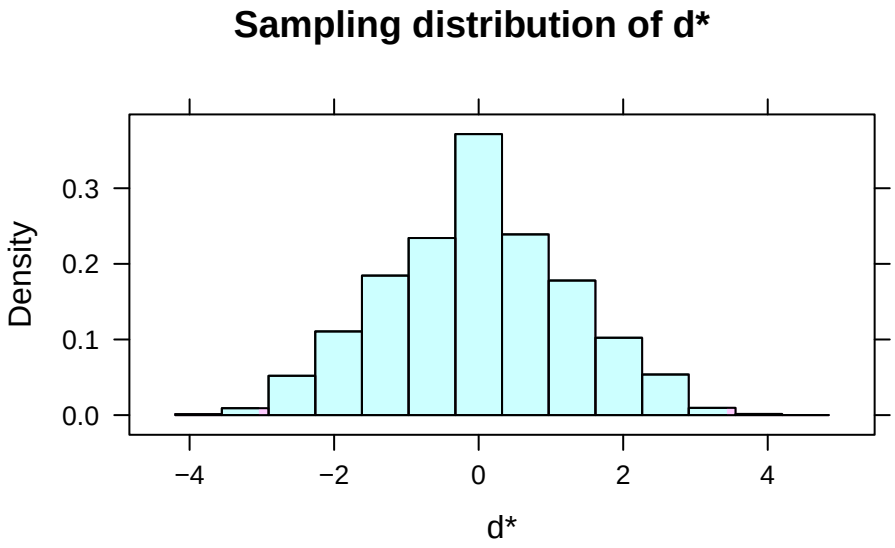


Figure 2.3.1: The histogram shows the sampling distribution of the statistic d^*, the difference between group means when groups are randomly assigned.

We have almost no cases where the randomly assigned groups produced a difference as extreme as the actual groups. We can calculate the percentage of the sampling distribution of the difference in means that is farther from zero

```
p.value <- tally( ~ (abs(diffmean) >= 3.5),
                  data=SamplingDist,
                  format='proportion'     )
p.value

##
##   TRUE  FALSE
## 0.0041 0.9959
```

We see that 0.41% of the simulations were as or more extreme than our observed difference in sample means so we can reject the null hypothesis $H_0 : \mu_{nc} - \mu_c = 0$ in favor of the alternative $H_a : \mu_{nc} - \mu_c \neq 0$.

Everything we know about the biological effects of ingesting caffeine suggests that we should have expected the caffeinated group to tap faster, so we might want to set up our experiment so only faster tapping represents "extreme" data compared to the null hypothesis. In this case we want an alternative of $H_a : \mu_{nc} - \mu_c < 0$? Therefore the null and alternative hypothesis are

$$\begin{aligned} H_0 : \mu_{nc} - \mu_c &\geq 0 \\ H_a : \mu_{nc} - \mu_c &< 0 \end{aligned}$$

or using the parameter $\delta = \mu_{nc} - \mu_c$ the null and alternative are

$$\begin{aligned} H_0 : \delta &\geq 0 \\ H_a : \delta &< 0 \end{aligned}$$

The creation of the sampling distribution of the mean difference d^* is identical to our previous technique because if our observed difference d is so negative that it is incompatible with the hypothesis that $\delta = 0$ then it *must* also be incompatible with any positive value of δ, so we evaluate the consistency of our data with the value of δ that is closest to the observed d while still being true to the null hypothesis. Thus for either the the one-sided (i.e. $\delta < 0$) or the two-sided case (i.e. $\delta \neq 0$), we generate the sampling distribution of d^* in the same way. The only difference in the analysis is at the end when we calculate the p-value and don't consider the positive tail. That is, the p-value is the percent of simulations where $d^* < d$.

```
p.value <- tally( ~ (diffmean >= 3.5), data=SamplingDist, format='proportion' )
p.value

##
##   TRUE  FALSE
## 0.0021 0.9979
```

and we see that the p-value is approximately cut in half by ignoring the upper tail, which makes sense considering the observed symmetry in the sampling distribution of d^*.

In general, we prefer to use a two-sided test because if the two-sided test leads us to reject the null hypothesis then so would the appropriate one-sided hypothesis[10]. Second, by using a two-sample test, it prevents us from from "tricking" ourselves when we don't know the which group should have a higher mean going into the experiment, but after seeing the data, thinking we should have known and using the less stringent test. Some statisticians go so far as to say that using a 1-sided test is outright fraudulent. Generally, we'll concentrate on two-sided tests as they are the most widely acceptable.

Example:

In places in the country substantial mosquito populations, the question of whether drinking beer causes the drinker to be more attractive to the mosquitoes than drinking something else has plagued campers. To answer such a question, researchers[11] conducted a study to determine if drinking beer attracts more mosquitoes than drinking water. Of $n = 43$ subjects, $n_b = 25$ drank a liter beer and $n_w = 18$ drank a liter of water and mosquitoes were caught in traps as they approached the different subjects. The critical part of this study is that the treatment (beer or water) was randomly assigned to each subject.

For this study, we want to test

$$H_0 : \delta = 0 \quad \text{vs} \quad H_a : \delta \neq 0$$

where we define $\delta = \mu_w - \mu_b$ and μ_b is the mean number of mosquitoes attracted to a beer drinker and μ_w is the mean number attracted to a water drinker. As usual we begin our analysis by plotting

[10] Except in the case where the alternative was chosen before the data was collected and the observed data was in the other tail. For example: the alternative was $H_a : \delta > 0$ but the observed difference was actually negative.

[11] Lefvre, T., et. al., "Beer Consumption Increases Human Attractiveness to Malaria Mosquitoes, " PLoS ONE, 2010; 5(3): e9546

the data

```
# I can't find this dataset on-line so I'll just type it in.
Mosquitoes <- data.frame(
  Number = c(27,19,20,20,23,17,21,24,31,26,28,20,27,
             19,25,31,24,28,24,29,21,21,18,27,20,
             21,19,13,22,15,22,15,22,20,
             12,24,24,21,19,18,16,23,20),
  Group = c( rep('Beer', 25), rep('Water',18) ) )

# Plot the data
dotPlot(~ Number | Group, data=Mosquitoes, layout=c(1,2) )
```

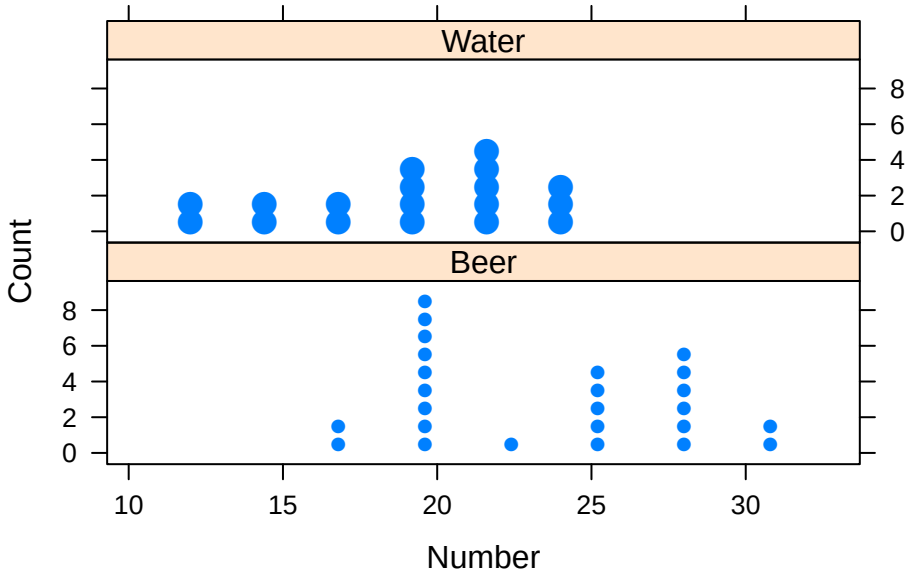


Figure 2.3.2: Raw data for the number of mosquitoes attracted to each subject, divided by treatment type (beer or water).

and calculating a summary statistic that captures the difference we are trying to understand

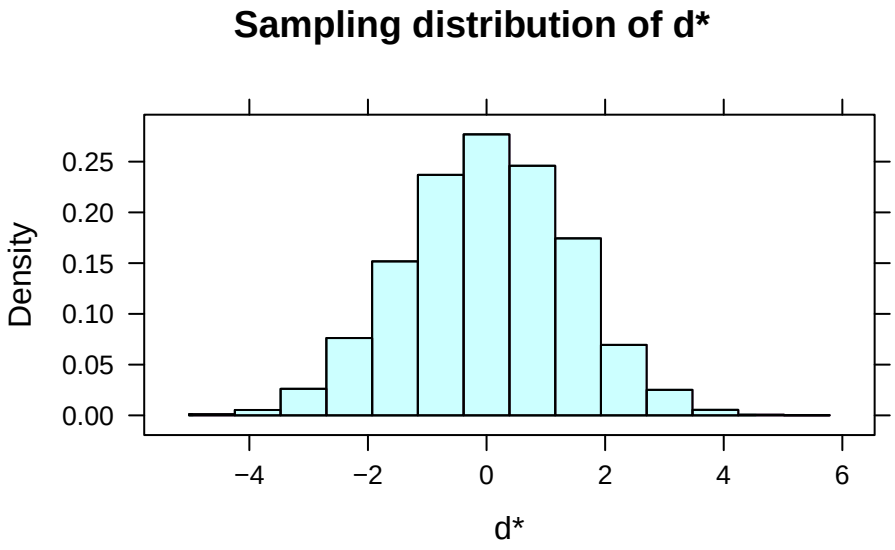
$$d = \bar{x}_w - \bar{x}_b$$

where $\bar{x}_w$ is the sample mean number of mosquitoes attracted by the water group and $\bar{x}_b$ is the sample mean number of mosquitoes attracted by the beer group.

```
# Observed difference in number of mosquitoes
d <- diffmean(Number ~ Group, data=Mosquitoes)
d

##  diffmean
## -4.377778
```

The calculation of the sampling distribution of d^*, the distribution of the difference in group means assuming that the null hypothesis is true and the beer/water grouping doesn't matter, is done by repeatedly shuffling the group labels and calculating differences.

```
SamplingDist <- do(10000) * diffmean(Number ~ shuffle(Group), data=Mosquitoes)
histogram( ~ diffmean, data=SamplingDist, groups=( abs(diffmean) >= abs(d) ),
           main='Sampling distribution of d*', xlab='d*')
```




```
p.value <- tally( ~ (abs(diffmean) >= abs(d)),
                  data=SamplingDist,
                  format='proportion'        )
p.value

##
##   TRUE  FALSE
## 0.0013 0.9987
```

The calculated p-value is extremely small (p = 0.0013) and we can conclude that the choice of drink does cause a change in attractiveness to mosquitoes.

Example:

Caffeine is often used to increase alertness in the United States. A study[12] compared the effect of a brief nap versus caffeine on the ability to recall memorized information. Using the summary statistics provided by the article, the plausible data were generated for a textbook.[13] The question is if the number of words a subject can recall is related to the randomly assigned treatment received before the trial (nap or caffeine). The simulated data are available on the book's website.

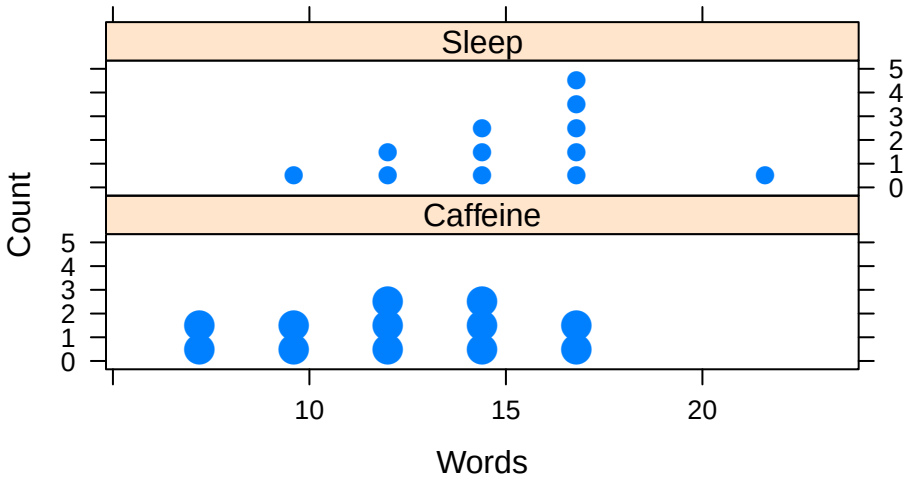
```
# If we wanted to read an excel file...
# library(gdata)
# SleepCaffeine <- read.xls('http://www.lock5stat.com/datasets/SleepCaffeine.xls')

# Just using the standard .csv file...
SleepCaffeine <- read.csv('http://www.lock5stat.com/datasets/SleepCaffeine.csv')
```

As always we plot the data first, and observe that we have 12 subjects assigned to the sleep group and 12 assigned to the caffeine group.

[12] Mednick, S., Cai, D., Kanady, J., and Drummond, S., "Comparing the Benefits of Caffeine, Naps, and Placebo on Verbal Motor and Perceptual Memory", *Behavioral Brain Research* 2008; 193: 79-86.

[13] Problem 4.82 of *Statistics: Unlocking the power of data* by Lock, Lock, Lock, Lock, and Lock.

```
dotPlot( ~ Words | Group, data=SleepCaffeine, layout=c(1,2))
```

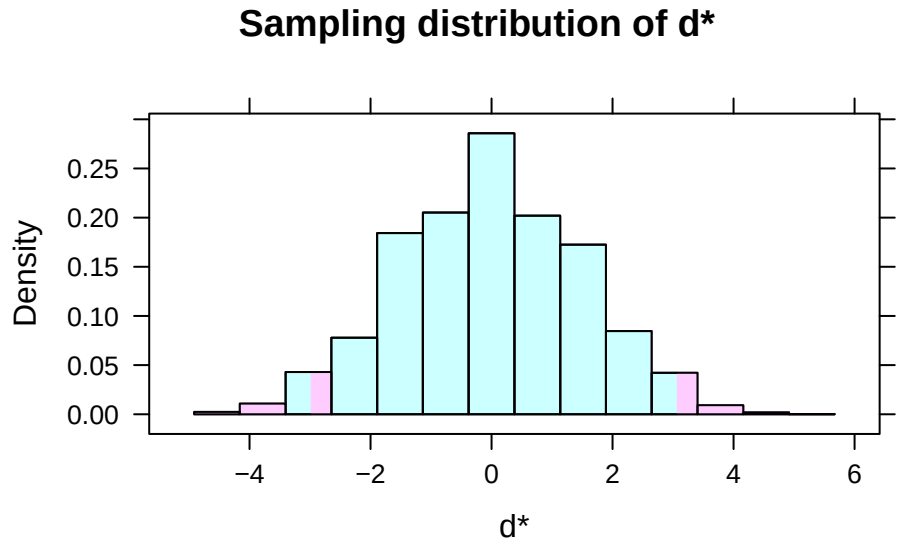


We generate the observed difference in the means $d = \bar{x}_C - \bar{x}_S$

```
d <- diffmean(Words ~ Group, data=SleepCaffeine)
d

## diffmean
##        3
```

We then generate the sampling distribution of d^*, the possible differences assuming that there is no difference in effect of the two treatments.

```
SamplingDist <- do(10000) * diffmean(Words ~ shuffle(Group), data=SleepCaffeine)
histogram( ~ diffmean, data=SamplingDist, groups=( abs(diffmean) >= abs(d) ),
           main='Sampling distribution of d*', xlab='d*')
```



We can then calculate the p-value as the proportion of simulations with $|d^*| > d = 3$

```
p.value <- tally( ~ (abs(diffmean) >= abs(d)),
                  data=SamplingDist,
                  format='proportion'         )
p.value

##
##   TRUE  FALSE
## 0.0462 0.9538
```

The p-value is very near the common cut-off value of 0.05 for deciding if the result is significant. I would run more simulations if I were submitting this to a journal, but we should recognize that this p-value indicates we have moderately strong evidence to reject that naps are equivalent to caffeine, but a larger study would necessary for a definitive result.

2.3.2 Two groups, binary response

We are often interested in experiments where the response variable is binary in nature. For example, perhaps we treat a bunch of plots with two types of insecticides and after 2 weeks observe if the plots are infested or not infested with some insect. Our goal would be to decide if the proportion of plots invested is different amongst the two treatment types.

We will have two questions:

1. What statistic could be calculated from the observed data to measure how far the observed data is from the null hypothesis?

2. Given the statistic in part 1, how should it vary from sample to sample assuming the null hypothesis (no difference in treatments) is true?

Example:

We will examine these questions in the context of a study where researchers suspected that attack of a plant by one organism induced resistance to subsequent attack by a different organism. Individually potted cotton plants were randomly allocated to two groups: infestation by spider mites or no infestation. After two weeks the mites were dutifully removed by a conscientious research assistant, and both groups were inoculated with Verticillium, a fungus that causes Wilt disease.

```
library(mosaicData)  # where the data lives
data(Mites)
str(Mites)

## 'data.frame': 47 obs. of  2 variables:
##  $ treatment: Factor w/ 2 levels "mites","no mites": 1 1 1 1 1 1 1 1 1 1 ...
##  $ outcome  : Factor w/ 2 levels "no wilt","wilt": 2 2 2 2 2 2 2 2 2 2 ...
```

We will summarize the data into a *contingency table* that counts the number of plants in each treatment/wilt category[14].

[14] If the code below doesn't give the counts, it might be because the packages `mosaic` and `dplyr` are fighting over which gets to define the function `tally()`. You can force your code to use the mosaic version of the function by using `mosaic::tally( outcome ~ treatment, data=mites )` where the key part is to give the package name and then the function. Within `mosaic`'s `tally()` function there is an option `format=` option that allows you to specify if you want the raw `counts`, the `proportion` in each column, or as `percent` in each column.

```
tally(outcome ~ treatment, data=Mites, # table of outcome by treatment
     format='count'                    # give the raw counts, not percentages
)

##          treatment
## outcome   mites no mites
##   no wilt    15        4
##   wilt       11       17
```

From this table we can see that of the $n = 47$ plants, 28 of them wilted. Furthermore we see that the mites were applied to $n_m = 26$ of the plants. Is this data indicative of mites inferring a disease resistance? More formally we are interested in testing

$$H_0: \quad \pi_w = \pi_{w|m}$$
$$H_0: \quad \pi_w \neq \pi_{w|m}$$

where the relevant parameters are π_w, the probability that a plant will wilt, and $\pi_{w|m}$, the probability that a plant will wilt given that it has been treated with mites.

What would you expect to see if there was absolutely no effect of the mite treatment? The wilting plants should be equally dispersed between the mite and non-mite treatments, but we also have to account for the fact that we have more mite treatments. Let $p_w = \frac{28}{47}$ be the proportion all the plants that wilted, and $n_m = 26$ be the number of plants receiving the mite treatment. If the treatment has no effect on wilting, then we expect $p_w \cdot n_m = 15.49$ of the mite treated plants to wilt and $(1 - p_w)\, n_m = 10.51$ to not wilt. A similar calculation for the non-mite treatment shows $p_w \cdot (n - n_m) = 12.51$ should wilt and $(1 - p_w)\,(n - n_m) = 8.49$ should not.

		Treatment		
		Mites	No Mites	
Outcome	No Wilt	$O_{nw,m} = 15$ $E_{nw,m} = (10.51)$	$O_{nw,nm} = 4$ $E_{nw,m} = (8.49)$	$n_{nw} = 19$
	Wilt	$O_{w,m} = 11$ $E_{nw,m} = (15.49)$	$O_{w,nm} = 17$ $E_{nw,m} = (12.51)$	$n_w = 28$
		$n_m = 26$	$n_{nm} = 21$	$n = 47$

In general the expected count for a cell can be calculated as

$$E_{i,j} = \frac{n_i n_j}{n}$$

where n_i is the row sum, and n_j is the column sum. For example $E_{nw,m} = 15.49 = (28 * 26)/47$.

This is the first case where our test statistic will not be just plugging in the sample statistic into the null hypothesis. Instead we will consider a test statistic that is more flexible and will handle more general cases (say 3 or more response or treatment groups) Our statistic for assessing how far our observed data is from what we expect under the null hypothesis involves the difference between the observed and the expected for each of the cells, but again we don't want to just sum the differences, instead will make the differences positive by squaring the differences. Second, a difference of 10 between the observed and expected cell count is very different if the number expected is 1000 than if it is 10, so we will scale the observed difference by dividing by the expected cell count.

We define

$$\begin{aligned} X^2 &= \sum_{\text{all ij cells}} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \\ &= \frac{(15-10.51)^2}{10.51} + \frac{(4-8.49)^2}{8.49} + \frac{(11-15.49)^2}{15.49} + \frac{(17-12.51)^2}{12.51} \\ &= 1.92 + 2.37 + 1.30 + 1.61 \\ &= 7.20 \end{aligned}$$

If the null hypothesis is true, then this statistic should be small, and a large value of the statistic is indicative of the null hypothesis being incorrect. But how large must the statistic be before we reject the null hypothesis? Again, we'll randomly shuffle the treatment assignments and recalculate the statistic many times and examine the sampling distribution of X^2.

To do this efficiently, we'll need a way of easily calculating this test statistic. In a traditional course I would introduce this test by the name of "Pearson's Chi-squared test" and we can obtain the test statistic using the following code:

```
# function is chisq.test() and we need to tell it not to do the Yates continuity
# correction and just calculate the test statistic as we've described
chisq.test(                                                  # do a Chi-sq test
  tally(outcome ~ treatment, data=Mites, format='count'),    # on this table
  correct=FALSE                                              # Don't do the correction
)

##
##  Pearson's Chi-squared test
##
## data:  tally(outcome ~ treatment, data = Mites, format = "count")
## X-squared = 7.2037, df = 1, p-value = 0.007275
```

R is performing the traditional Pearson's Chi-Squared test which assumes our sample sizes are large enough for several approximations to be good. Fortunately, we don't care about this approximation to the p-value and will use simulation methods which will be more accurate. In order to use the `chisq.test()` function to do our calculations, we need to extract the test-statistic from the output of the function.

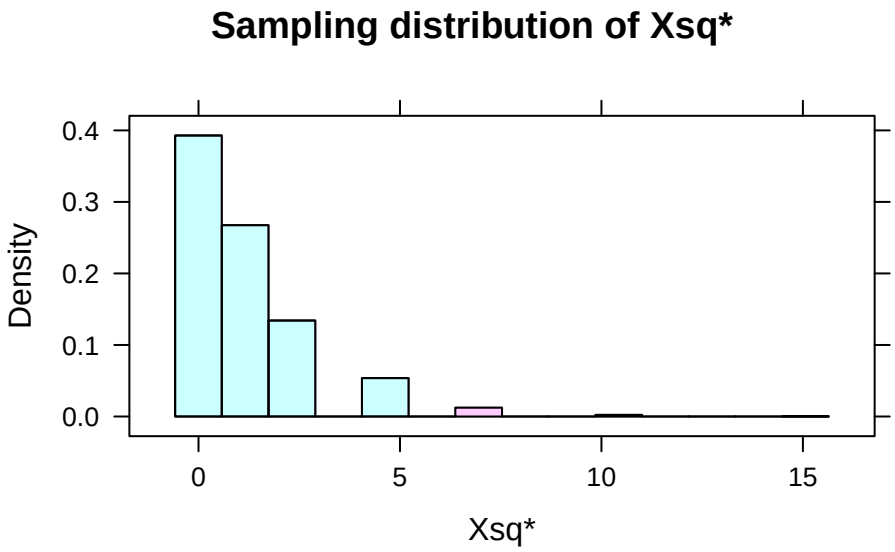
```
# extract the X^2 test statistic from the output
Xsq <- chisq.test(                                           # do a Chi-sq test
  tally(outcome ~ treatment, data=Mites, format='count'),    # on this table
  correct=FALSE                                              # Don't do the correction
)$statistic                                                  # grab only the test statistic

Xsq

## X-squared
##  7.203748
```

Next we wish to repeat our shuffling trick of the treatment labels to calculate the sampling distribution of X^{2*}, which is the distribution of X^2 when the null hypothesis of no difference between treatments is true.

```
SamplingDist <- do(3)*
  chisq.test(
    tally(outcome ~ shuffle(treatment), data=Mites, format='count'),
    correct=FALSE
  )$statistic

SamplingDist

##    X.squared
## 1 0.79284750
## 2 4.40515781
## 3 0.09320003
```

We see that this code is creating a data frame with a single column called `X.squared` and next we simulate a large number of times and display the sampling distribution of X^{2*}.

```
SamplingDist <- do(10000)*
  chisq.test(
    tally(outcome ~ shuffle(treatment), data=Mites, format='count'),
    correct=FALSE
  )$statistic

histogram( ~ X.squared, data=SamplingDist, groups=( X.squared >= Xsq ),
           main='Sampling distribution of Xsq*', xlab='Xsq*')
```



At first glance this seems wrong because it is not a nice looking distribution. However there are only a small number of ways to allocate the treatments labels to the two possible outcomes. Second, for the test statistic we have chosen only the right hand side of the distribution (large values of X^*) would be evidence against the null hypothesis, so we only look at $X^{2*} > 7.20$.

```
p.value <- tally( ~ (X.squared >= Xsq), data=SamplingDist, format='proportion' )
p.value

##
##  TRUE FALSE
## 0.017 0.983
```

We see that the p-value is 0.017 and conclude that there is strong evidence to reject the null hypothesis that the mite treatment does not affect the probability of wilting. That is to say, the probability of observing data as extreme as ours is unlikely to occur by random chance when the null hypothesis is true.

Example:

In a study to investigate possible treatments for human infertility, researchers[15] performed a double-blind study and randomly divided 58 patients into two groups. The treatment group ($n_t = 30$) received 100 mg per day of Doxycycline and the placebo group ($n_p = 28$) received a placebo but were unaware that it was a placebo. Within 5 months, the treatment group had 5 pregnancies, while the placebo group had 4.

		Treatment		
		Doxycycline	Placebo	
Outcome	Conceived	$O_{c,t} = 5$ $E_{c,t} = \left(\frac{9 \cdot 30}{58}\right) = 4.66$	$O_{c,p} = 4$ $E_{c,p} = \left(\frac{9 \cdot 28}{58}\right) = 4.34$	$n_c = 9$
	Not Conceived	$O_{nc,t} = 25$ $E_{nc,t} = \left(\frac{49 \cdot 30}{58}\right) = 25.34$	$O_{nc,p} = 24$ $E_{nc,p} = \left(\frac{49 \cdot 28}{58}\right) = 23.66$	$n_{nc} = 49$
		$n_t = 30$	$n_p = 28$	$n = 58$

Just looking at the observed vs expected there doesn't seem to be much difference between the treatments. To confirm this we do a similar test as before.

```
Conceived <-  data.frame(
  CoupleID=1:58,
  Treatment=c(rep('Doxycyline',30), rep('Placebo',28)),
  Outcome=c(rep('Conceived',5), rep('Not Conceived',25),
            rep('Conceived',4), rep('Not Conceived',24)))

chisq.test(
  tally(Outcome ~ Treatment, data=Conceived, format='count'),
  correct=FALSE )

## Warning in chisq.test(tally(Outcome ~ Treatment, data = Conceived, format =
"count"), :  Chi-squared approximation may be incorrect

##
##  Pearson's Chi-squared test
##
## data:  tally(Outcome ~ Treatment, data = Conceived, format = "count")
## X-squared = 0.0626, df = 1, p-value = 0.8024
```

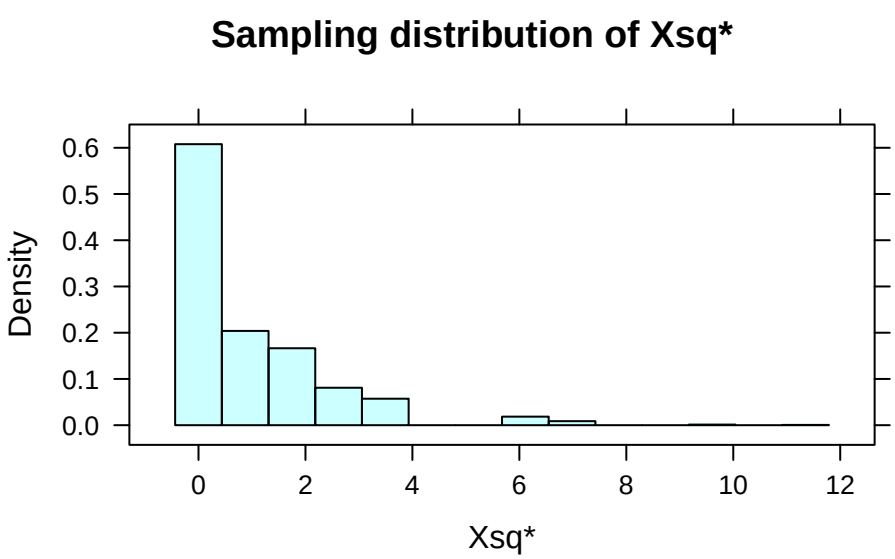
Notice that the `chisq.test()` function is warning us that it doesn't think the asymptotic (large sample) approximations it must make to calculate a p-value are appropriate. Fortunately we have an alternative methodology of simulation to rely on and all we want is to save the observed X^2 test statistic from this analysis.

[15] Harrison, R. F., Blades, M., De Louvois, J., & Hurley, R. (1975). Doxycycline treatment and human infertility. *The Lancet*, 305(7907), 605-607.

```
Xsq <- chisq.test(
  tally(Outcome ~ Treatment, data=Conceived, format='count'),
  correct=FALSE
)$statistic

Xsq

##  X-squared
## 0.06262823
```

As usual, we create the sampling distribution of X^{2*} by shuffling the treatment labels and re-calculating the statistic of interest.

```
SamplingDist <- do(10000)*
  chisq.test(
    tally(Outcome ~ shuffle(Treatment), data=Conceived, format='count'),
    correct=FALSE
  )$statistic

histogram( ~ X.squared, data=SamplingDist, groups=( X.squared >= Xsq ),
           main='Sampling distribution of Xsq*', xlab='Xsq*')
```



Note that the the entire distribution is "more extreme" than what our actual data was because our there is no way to perfectly partition the 9 pregnancies into the $n_t = 30$ and $n_p = 28$ groups and our observed data was as close as possible to an even split as possible.

```
p.value <- tally( ~ (X.squared >= Xsq), data=SamplingDist, format='proportion' )
p.value

##
##  TRUE FALSE
##     1     0
```

We see that the p-value is 1 and conclude that there is no evidence to reject the null hypothesis that Doxycycline affects the probability of conceiving.

2.4 Summary

The scientific method relies on being able to decide if a given set of data is compatible or consistent with a particular hypothesis (which we called the null hypothesis). We answer that question by first simplifying the data to a single statistic that concisely measures the hypothesis of interest, and then examining the distribution of that statistic assuming the null hypothesis is correct. We called that distribution the *sampling distribution*.

In the case of a sample proportion, we were able to simulate the sampling distribution directly and all that was needed was the null hypothesis value for π. With repeated draws from the sampling distribution, we were able to infer how probable our observed data would be if the null hypothesis was true. We called that probability the *p-value*.

Designed experiments are the "gold standard" of scientific studies because the random assignment to treatment groups controls for unknown lurking variables. Due to the random assignment, each level of the lurking variables end up in all the treatment groups. This random assignment is absolutely critical in the study and our simulated sampling distribution is consistent with design by incorporating this randomness by the shuffling of treatment labels. Assuming the null hypothesis of "no effect" is true, then the assignment to treatment groups is not related to the observed response and our simulation method respects the experimental design.

The methodology we used isn't restricted to these particular statistics. If we wanted to know if the median values of two treatment groups are different, we would use

$$d = \tilde{x}_1 - \tilde{x}_2$$

as our sample statistic where $\tilde{x}_1$and $\tilde{x}_2$ are the sample medians of the groups and then do usual trick of repeatedly calculating that same statistic on the shuffled data sets. The comparison of the observed statistic to the sampling distribution when H_0 is true would remain the same.

Chapter 3

Confidence Intervals Using Bootstrapping

Often our research goal isn't to compare our data to some specific hypothesized value of a parameter, but rather to take the data we've observed and ask "What values of the parameter are consistent with the data?"

3.1 Observational Studies

Unfortunately it is not always possible to perform a designed experiment. It might be unethical (randomly assigning people to receive a dangerous dose of radiation) logistically difficult (assigning a heating treatment to hectare level landscapes), or just too expensive or time consuming.

Instead we often settle for a inferior study method of taking a random sample from the population of interest and observing the relationships between the variables of interest. In the designed experiment, the random assignment to treatment groups was critical to our inference. In an observational study, the random sample from the population is critical trusting that the observed sample data is representative of the population of interest.

As with experimental data, we will want to test whether the observed data (and test statistic d) is consistent with some null hypothesis H_0. However we will have to modify our process slightly to account for the difference between just observing randomly sampled data versus the random assignment of treatments.

Suppose that we had a population of interest and we wish to estimate the mean of that population (the population mean we'll denote as μ). We can't observe every member of the population (which would be prohibitively expensive) so instead we take a random sample and from that sample calculate a sample mean (which we'll denote $\bar{x}$). We believe that $\bar{x}$ will be a good estimator of μ, but it will vary from sample to sample and won't be exactly equal to μ.

Next suppose we wish to ask if a particular value for μ, say μ_0, is consistent with our observed data? We know that $\bar{x}$ will vary from sample to sample, but we have no idea *how much it will vary* between samples. However, if we could understand how much $\bar{x}$ varied sample to sample, we could answer the question. For example, suppose that $\bar{x} = 5$ and we know that $\bar{x}$ varied about ± 2 from sample to sample. Then I'd say that possible values of μ_0 in the interval 3 to 7 (5 ± 2) are reasonable values for μ and anything outside that interval is not reasonable.

Therefore, if we could take many, many repeated samples from the population and calculate our test statistic $\bar{x}$ for each sample, we could rule out possible values of μ. Unfortunately we don't have the time or money to repeatedly sample from the actual population, but we could sample from our best approximation to what the population is like.

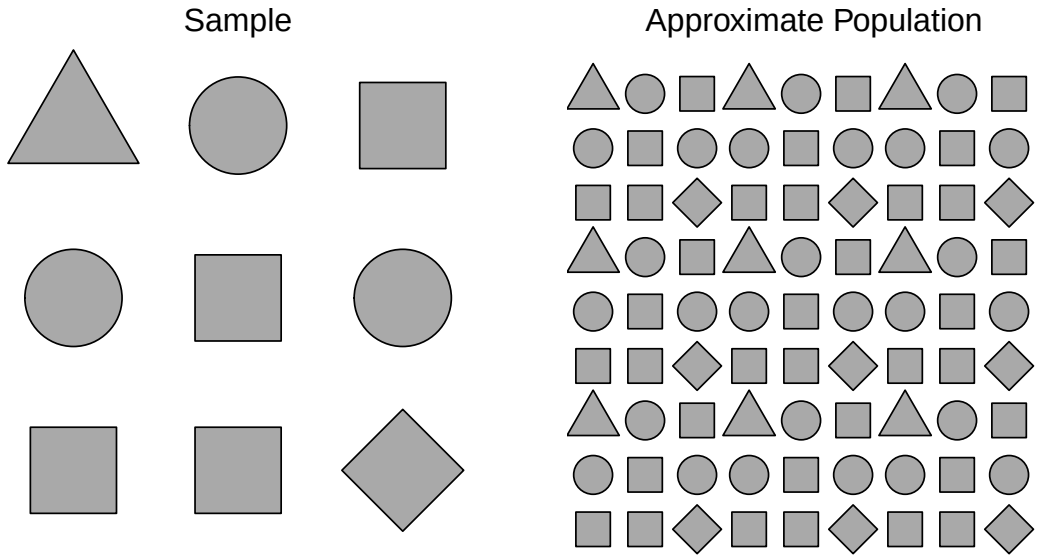


Figure 3.1.1: A possible sample from a population of shapes. Because 4/9 of our sample were squares, our best estimate is that the population is also approximately 4/9 squares. We can think of the approximated population as just many many copies of the observed sample data.

Suppose we were to sample from a population of shapes, and we observed 4/9 of the sample were squares, 3/9 were circles, and a triangle and a diamond. Then our best guess of what the population that we sampled from was a population with 4/9 squares, 3/9 circles, and 1/9 of triangles and diamonds.

Using this approximated population (which is just many many copies of our sample data), we can repeated sample $\bar{x}^*$ values to create the sampling distribution of $\bar{x}$.

Because our approximate population is just an infinite number of copies of our sample data, then sampling from the approximate population is equivalent to sampling *with replacement* from our sample data. If I take n samples from n distinct objects with replacement, then the process can be thought of as mixing the n objects in a bowl and taking an object at random, noting which it is, replace it into the bowl, and then draw the next sample. Practically, this means some objects will be selected more than once and some will not be chosen at all. To sample our observed data with replacement, we'll use the `resample()` function in the `mosaic` package. We see that some rows will be selected multiple times, and some will not be selected at all.

```
Testing.Data <- data.frame(
  name=c('Alison','Brandon','Chelsea','Derek','Elise'))
Testing.Data
```

```
##      name
## 1  Alison
## 2 Brandon
## 3 Chelsea
## 4   Derek
## 5   Elise
```

```
# Sample rows from the Testing Data (with replacement)
resample(Testing.Data)

##        name orig.ids
## 1    Alison        1
## 4     Derek        4
## 3   Chelsea        3
## 1.1  Alison        1
## 5     Elise        5
```

Notice `Alison` has selected twice, while `Brandon` has not been selected at all. We can use the `resample()` function similarly as we did the `shuffle()` function.

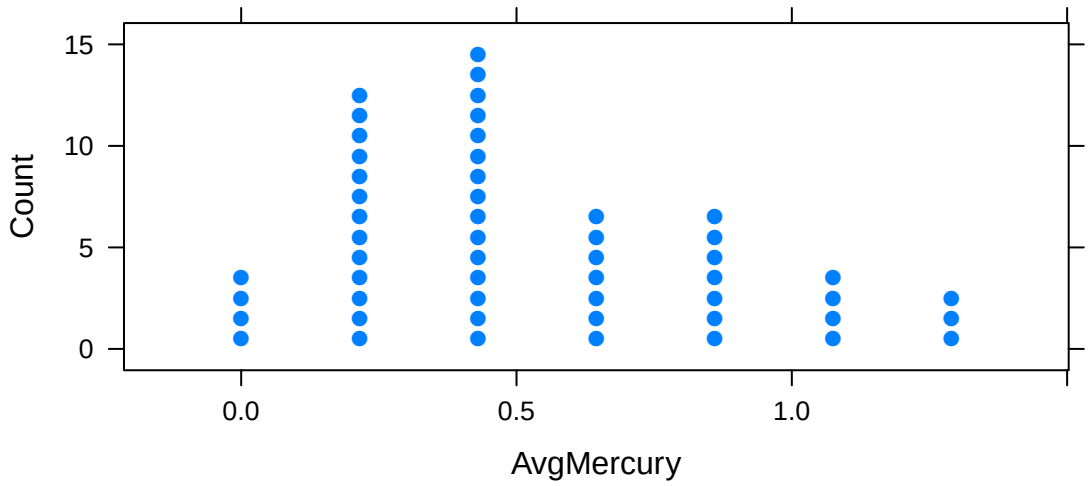
The sampling from the estimated population via sampling from the observed data is called *bootstrapping* because we are making no distributional assumptions about where the data came from, and the idiom "Pulling yourself up by your bootstraps" seemed appropriate.

Example: Mercury Levels in Fish from Florida Lakes

A data set provided by the Lock[5] textbook looks at the mercury levels in fish harvested from lakes in Florida. There are approximately 7,700 lakes in Florida that are larger than 10 acres. As part of a study to assess the average mercury contamination in these lakes, a random sample of $n = 53$ lakes, an unspecified number of fish were harvested and the average mercury level (in ppm) was calculated for fish in each lake. The goal of the study was to assess if the average mercury concentration was greater than the 1969 EPA "legally actionable level" of 0.5 ppm.

```
# as always, our first step is to load the mosaic package
library(mosaic)

# read the Lakes data set
Lakes <- read.csv('http://www.lock5stat.com/datasets/FloridaLakes.csv')
```

```
# make a nice picture
dotPlot( ~ AvgMercury, data=Lakes)
```



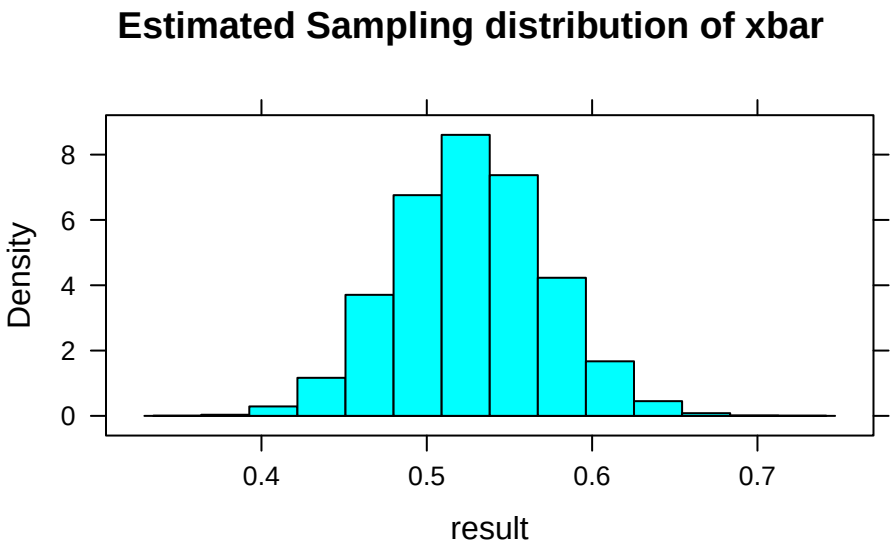
We can calculate mean average mercury level for the $n = 53$ lakes

```
mean( ~ AvgMercury, data=Lakes )

## [1] 0.5271698
```

The sample mean is greater than 0.5 but not by too much. Is a true population mean concentration μ_{Hg} that is 0.5 or less incompatible with our observed data? Is our data sufficient evidence to conclude that the average mercury content is greater than 0.5? Perhaps the true average mercury content is less than (or equal to) 0.5 and we just happened to get a random sample that with a mean greater than 0.5?

The first step in answering these questions is to create the sampling distribution of $\bar{x}_{Hg}$. To do this, we will sample from the approximate population of lakes, which is just many many replicated copies of our sample data.

```
# create the sampling distribution of xbar
SamplingDist <- do(10000) * mean( ~ AvgMercury, data=resample(Lakes) )

# what columns does the data frame "SamplingDist" have?
str(SamplingDist)

## Classes 'do.data.frame' and 'data.frame': 10000 obs. of  1 variable:
##  $ result: num  0.568 0.5 0.527 0.576 0.479 ...

# show a histogram of the sampling distribution of xbar
histogram( ~result, data=SamplingDist, main='Estimated Sampling distribution of xbar' )
```

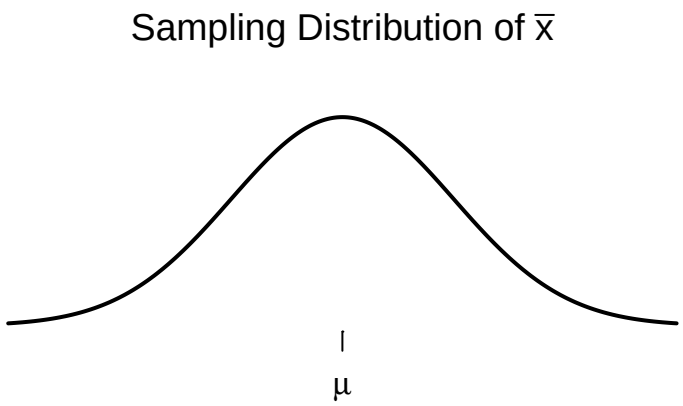


3.2 Using Quantiles of the Estimated Sampling Distributions to create a Confidence Interval

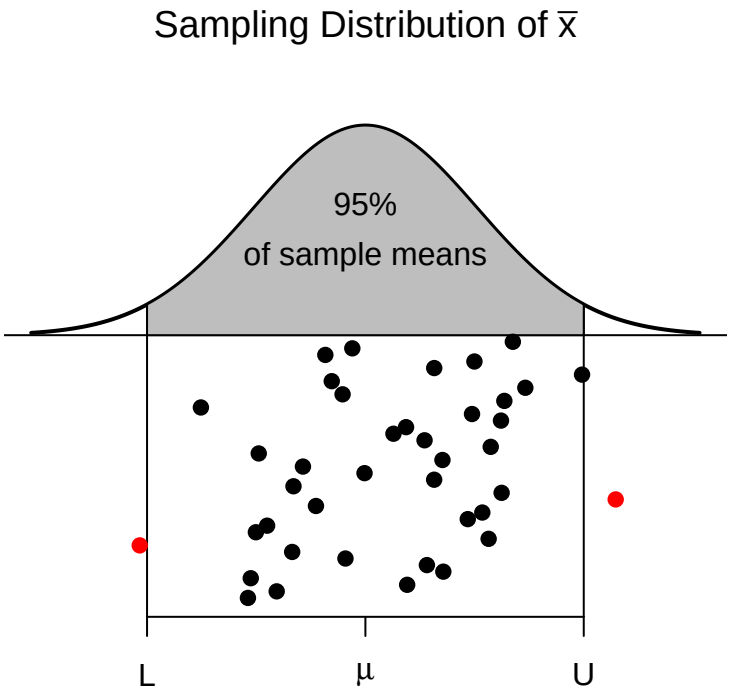
In many cases we have seen, the sampling distribution of a statistic is centered on the parameter we are interested in estimating and is symmetric about that parameter[1]. For example, we expect that the sample mean $\bar{x}$ should be a good estimate of the population mean μ and the sampling

[1]There are actually several ways to create a confidence interval from the estimated sampling distribution. The method presented here is called the "percentile" method and works when the sampling distribution is symmetric and the estimator we are using is unbiased.

distribution of $\bar{x}$ should look something like the following.

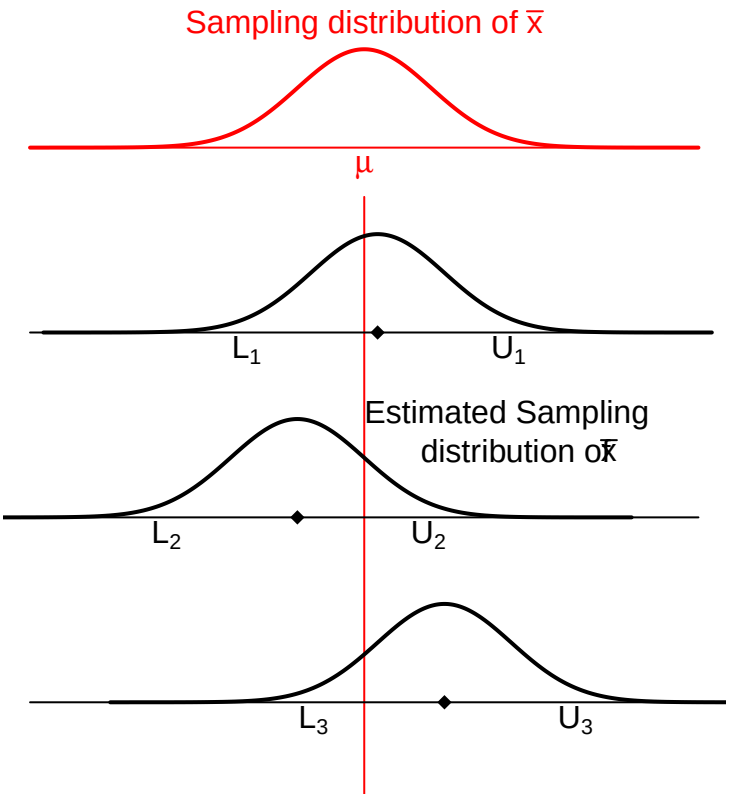


There are two points, (call them L and U) where for our given sample size and population we are sampling from, where we expect that 95% of the sample means to fall within. That is to say, L and U capture the middle 95% of the sampling distribution of $\bar{x}$.

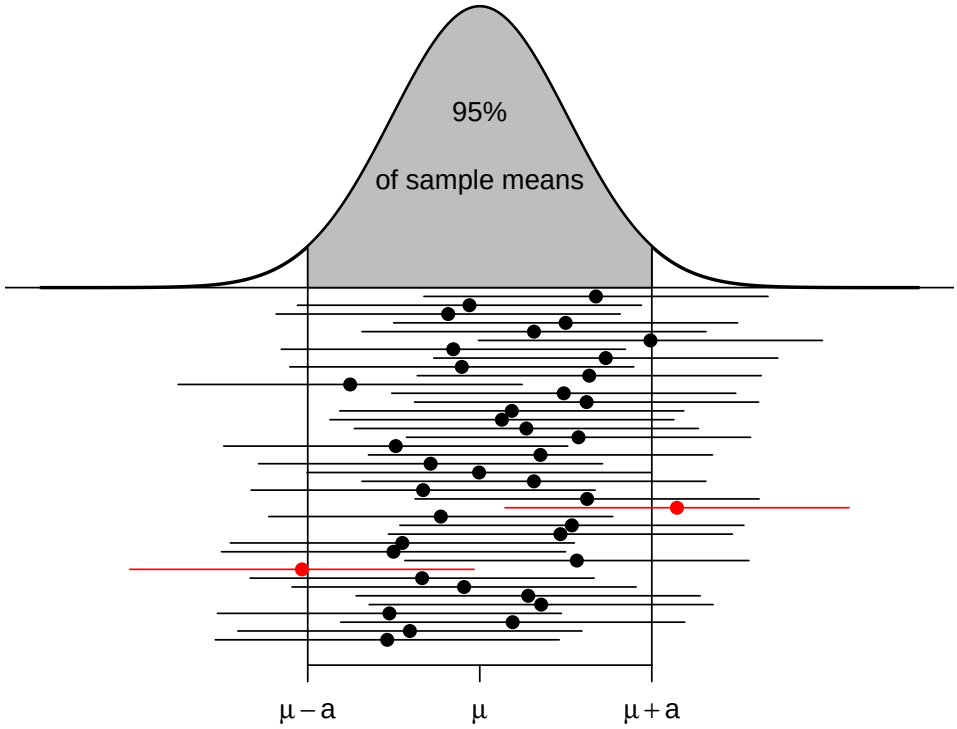


These sample means are randomly distributed about the population mean μ. Given our sample data and sample mean $\bar{x}$, we can examine how our *simulated* values of $\bar{x}^*$ vary about $\bar{x}$. I expect that these simulated sample means $\bar{x}^*$ should vary about $\bar{x}$ in the same way that $\bar{x}$ values vary around μ.

Below are three estimated sampling distributions that we might obtain from three different samples and their associated sample means.



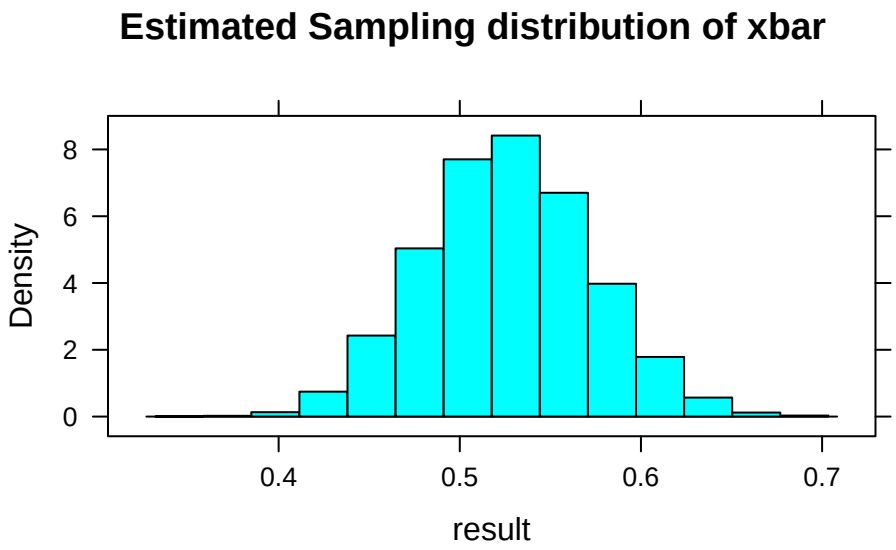
For each possible sample, we could consider creating the estimated sampling distribution of $\bar{x}$ and calculating the L and U values that capture the middle 95% of the estimated sampling distribution. Below are twenty samples, where we've calculated this interval for each sample.



Most of these intervals contain the true parameter μ, that we are trying to estimate. In practice, I will only take one sample and therefore will only calculate one sample mean and one interval, but I want to recognize that the method I used to produce the interval (i.e. take a random sample, calculate the mean and then the interval) will result in intervals where only 95% of those intervals will contain the mean μ. Therefore, I will refer to the interval as a 95% *confidence interval.*

After the sample is taken and the interval is calculated, the numbers lower and upper bounds of the confidence interval are fixed. Because μ is a constant value and the confidence interval is fixed, nothing is changing. To distinguish between a future random event and the fixed (but unknown) outcome of if I ended up with an interval that contains μ and we use the term confidence interval instead of probability interval.

```
# create the sampling distribution of xbar
SamplingDist <- do(10000) * mean( ~ AvgMercury, data=resample(Lakes) )

# what columns does the data frame "SamplingDist" have?
str(SamplingDist)

## Classes 'do.data.frame' and 'data.frame': 10000 obs. of  1 variable:
##  $ result: num  0.55 0.598 0.502 0.507 0.542 ...

# show a histogram of the sampling distribution of xbar
histogram( ~result, data=SamplingDist, main='Estimated Sampling distribution of xbar' )
```




```
# calculate the 95% confidence interval using middle 95%
quantile( SamplingDist$result, probs=c(.025, .975) )

##      2.5%     97.5%
## 0.4388679 0.6184953
```

There are several ways to interpret this interval.

1. The process used to calculate this interval (take a random sample, calculate a statistic, repeatedly resample, and take the middle 95%) is a process that results in an interval that contains the parameter of interest on 95% of the samples we could have collected, however we don't know if the particular sample we collected and its resulting interval of (0.44, 0.62) is one of the intervals containing μ.

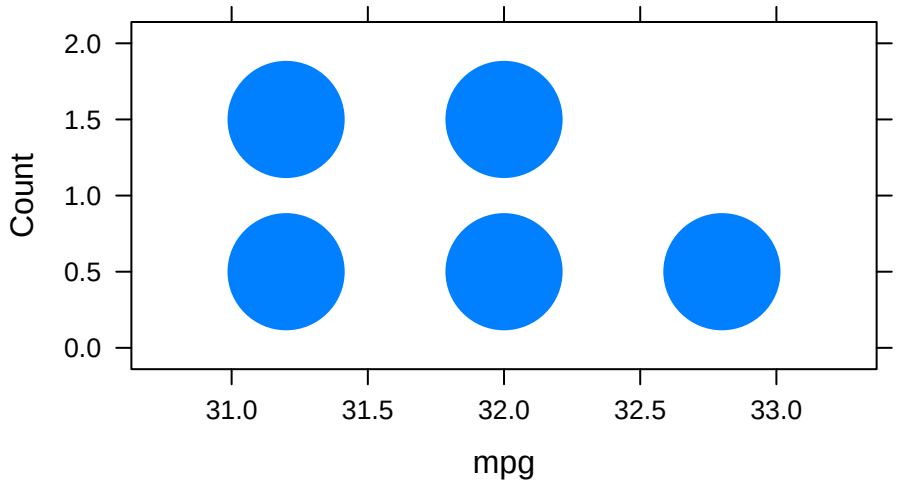
2. We are 95% confident that μ is in the interval (0.44, 0.62). This is delightfully vague and should be interpreted as a shorter version of the previous interpretation.

3. The interval (0.44, 0.62) is the set of values of μ that are consistent with the observed data at the 0.05 threshold of statistical significance for a two-sided hypothesis test.

Example:

Suppose we have data regarding fuel economy of 5 new vehicles of the same make and model and we wish to test if the observed fuel economy is consistent with the advertised 31 mpg at highway speeds. We the data are

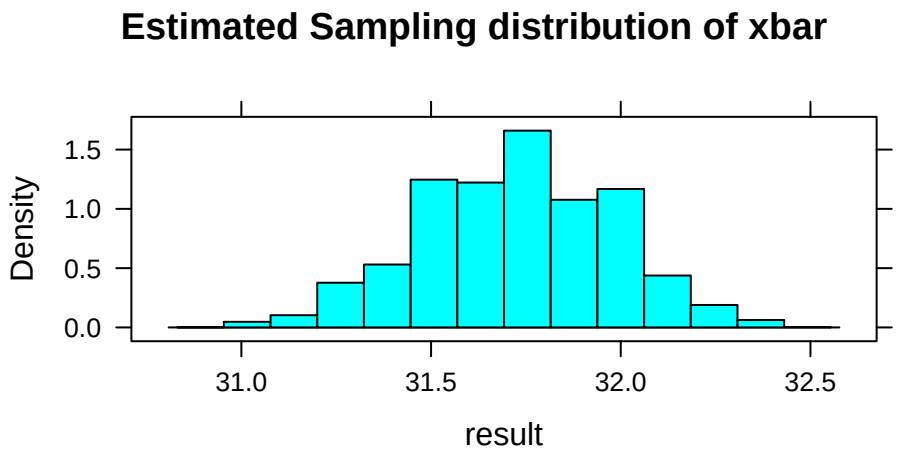
```
CarMPG <- data.frame( ID=1:5, mpg = c(31.8, 32.1, 32.5, 30.9, 31.3) )
dotPlot( ~ mpg, data=CarMPG )
```




```
mean( ~ mpg, data=CarMPG )

## [1] 31.72
```

We will use the sample mean to assess if the sample fuel efficiency is consistent with the advertised number. Because these cars could be considered a random sample of all new cars of this make, we will create the estimated sampling distribution using the bootstrap resampling of the data.

```
SamplingDist <- do(10000) * mean( ~ mpg, data=resample(CarMPG) )
histogram(~result, data=SamplingDist,
          main='Estimated Sampling distribution of xbar' )
```




```
quantile( SamplingDist$result, probs=c(.025, .975) )

##  2.5% 97.5%
## 31.22 32.20
```

We see that the 95% confidence interval is (31.2, 32.2) and does not actually contain the advertised 31 mpg. However, I don't think that in this case we would object to a car manufacturer sell us a car that is *better* than was advertised.

Example

Recall the pollution ratio data from homework 1 (O&L 3.21). The ratio of DDE (related to DDT) to PCB concentrations in bird eggs has been shown to have had a number of biological implications. The ratio is used as an indication of the movement of contamination through the food chain. The paper "The ratio of DDE to PCB concentrations in Great Lakes herring gull eggs and its us in interpreting contaminants data" reports the following ratios for eggs collected at 13 study sites from the five Great Lakes. The eggs were collected from both terrestrial and aquatic feeding birds.

	DDE to PCB Ratio										
Terrestrial	76.50	6.03	3.51	9.96	4.24	7.74	9.54	41.70	1.84	2.5	1.54
Aquatic	0.27	0.61	0.54	0.14	0.63	0.23	0.56	0.48	0.16	0.18	

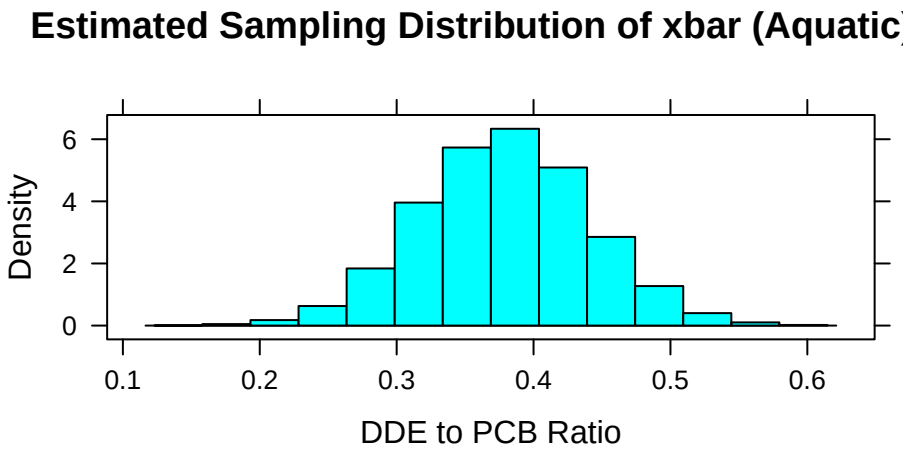
Suppose that the eggs were collected at random and we observe both the ratio and the feeding type. That is to say, we didn't decide to sample 11 Terrestrial birds and 10 Aquatic, that was just how it ended up. To create confidence intervals in that case, we should resample from the 21 eggs.

```
# write a data frame as before
PollutionRatios <- data.frame(
  Ratio = c(76.50, 6.03, 3.51, 9.96, 4.24, 7.74, 9.54, 41.70, 1.84, 2.5, 1.54,
            0.27, 0.61, 0.54, 0.14, 0.63, 0.23, 0.56,  0.48, 0.16, 0.18       ),
  Type  = c( rep('Terrestrial',11), rep('Aquatic',10) ) )

# what happens when I calculate multiple means...
do(3) * mean( Ratio ~ Type, data=resample(PollutionRatios) )

##     Aquatic Terrestrial
## 1 0.3840000    11.32545
## 2 0.3685714    23.47500
## 3 0.3623077     3.48250
```

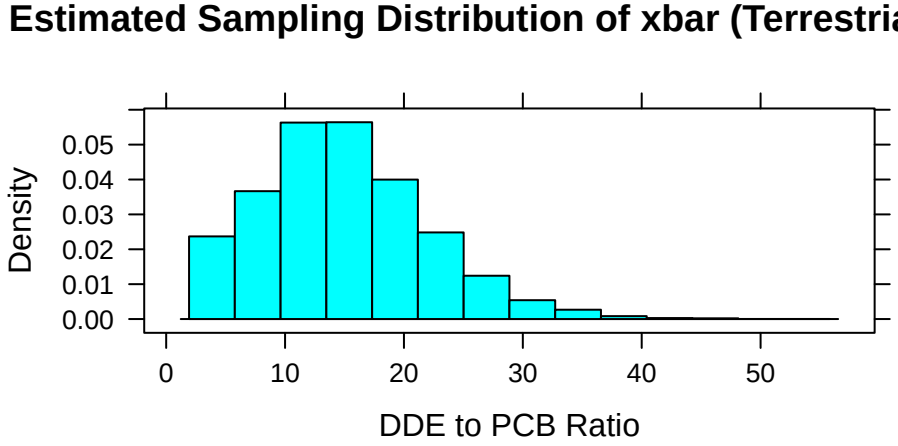
```
SamplingDist <- do(10000) * mean( Ratio ~ Type, data=resample(PollutionRatios) )
```

As always we can look at the estimated sampling distributions of the means

```
histogram( ~ Aquatic, data=SamplingDist,
           main='Estimated Sampling Distribution of xbar (Aquatic)',
           xlab='DDE to PCB Ratio')
```




```
histogram( ~ Terrestrial, data=SamplingDist,
           main='Estimated Sampling Distribution of xbar (Terrestrial)',
           xlab='DDE to PCB Ratio')
```

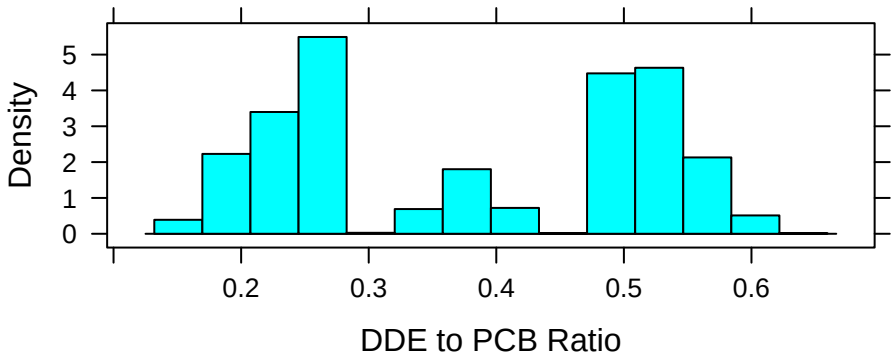



```
# Calculate confidence intervals
quantile( SamplingDist$Aquatic,     probs=c(.025, .975) )

##      2.5%     97.5%
## 0.2585714 0.5016750

quantile( SamplingDist$Terrestrial, probs=c(.025, .975) )

##      2.5%     97.5%
##  4.363306 30.573899
```

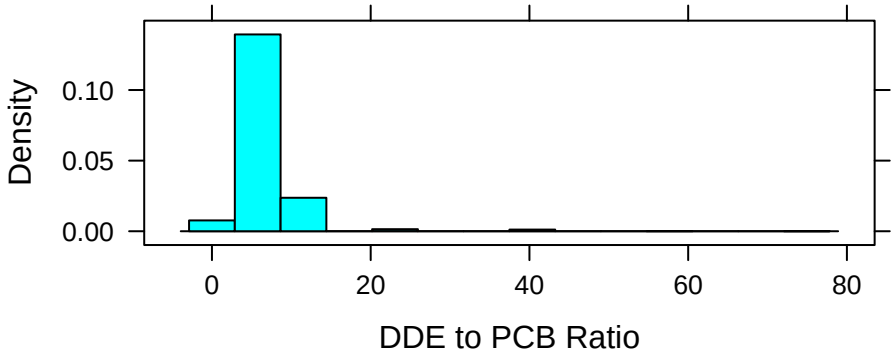
Because the terrestrial eggs had those large outliers we had recommend that the median might be a better measure of the "center" of the terrestrial observations. With the resampling method of calculating confidence intervals, it is easy to create a 95% confidence interval for the medians.

```
SamplingDist <- do(10000) * median( Ratio ~ Type, data=resample(PollutionRatios) )
```

```
histogram( ~ Aquatic, data=SamplingDist,
           main='Est. Sampling Dist. of sample median (Aquatic)',
           xlab='DDE to PCB Ratio')
```




```
histogram( ~ Terrestrial, data=SamplingDist,
           main='Est. Sampling Dist. of sample median (Terrestrial)',
           xlab='DDE to PCB Ratio')
```




```
# Calculate confidence intervals
quantile( SamplingDist$Aquatic,     probs=c(.025, .975) )

##  2.5% 97.5%
##  0.18  0.56

quantile( SamplingDist$Terrestrial, probs=c(.025, .975) )

##  2.5% 97.5%
##  2.50  9.96
```

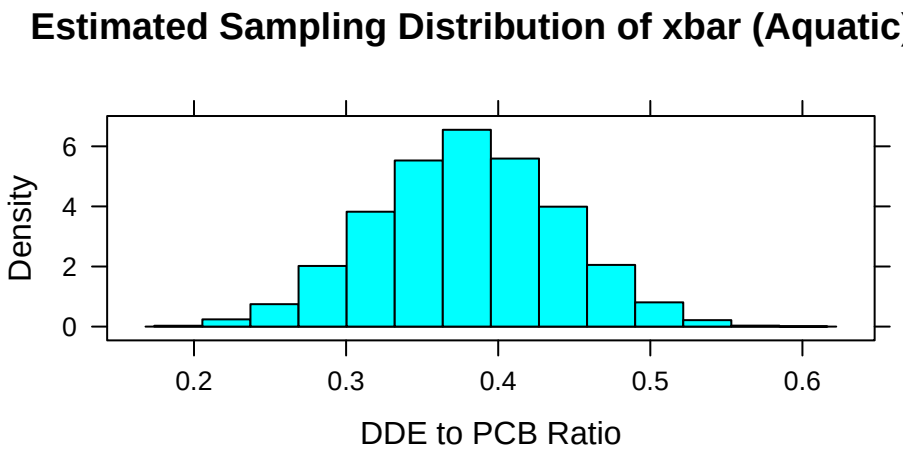
Suppose that the researchers had *deliberately* chosen to sample 11 terrestrial birds and 10 aquatic. Then our resampling method should respect that choice and always produce datasets with 11 ter-

restrial and 10 aquatic eggs. This can be done by the `groups=` argument to the resample command.

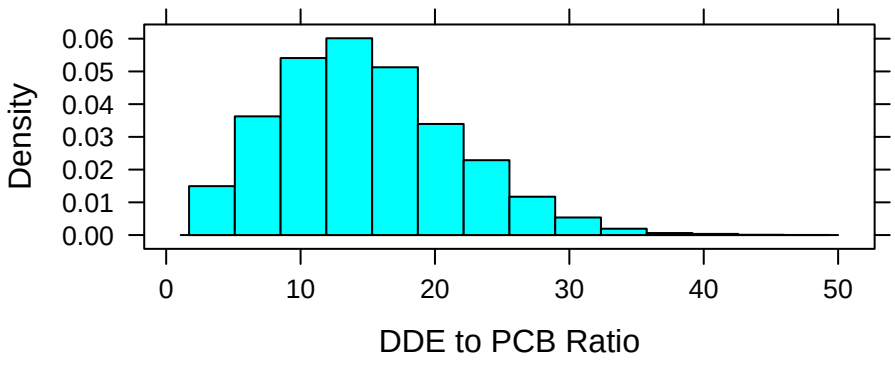
```
SamplingDist <- do(10000) *
   mean( Ratio ~ Type,
         data = resample(PollutionRatios, groups=Type) )
```

As always we can look at the estimated sampling distributions of the means

```
histogram( ~ Aquatic, data=SamplingDist,
           main='Estimated Sampling Distribution of xbar (Aquatic)',
           xlab='DDE to PCB Ratio')
```




```
histogram( ~ Terrestrial, data=SamplingDist,
           main='Estimated Sampling Distribution of xbar (Terrestrial)',
           xlab='DDE to PCB Ratio')
```



```
# Calculate confidence intervals
quantile( SamplingDist$Aquatic,     probs=c(.025, .975) )

##  2.5% 97.5%
## 0.263 0.497

quantile( SamplingDist$Terrestrial, probs=c(.025, .975) )

##      2.5%     97.5%
##  4.509023 30.135909
```

In this case, the difference between requiring the bootstrap datasets to conform to the 11 Terrestrial and 10 Aquatic didn't make much of a difference because that is what would happen on average sampling from the full 21 observations, but it could make a difference if the two groups were not as balanced.

Chapter 4

Probability

We need to work out the mathematics of what we mean by probability. To begin with we first define an *outcome*. An outcome is one observation from a random process or event. For example we might be interested in a single roll of a six-side die. Alternatively we might be interested in selecting one NAU student at random from the entire population of NAU students.

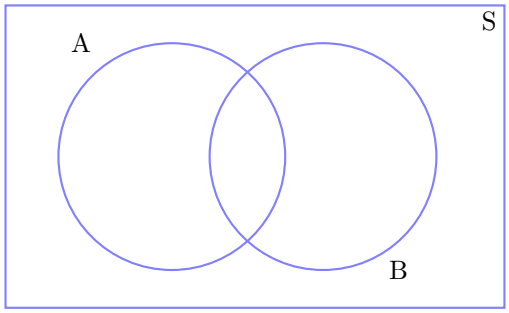
4.1 Introduction to Set Theory

Before we jump into probability, it is useful to review a little bit of set theory.

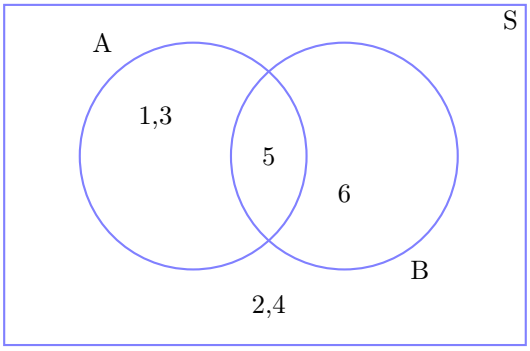
Events are properties of a particular outcome. For a coin flip, the event "Heads" would be the event that a heads was flipped. For the single roll of a six-sided die, a possible event might be that the result is even. For the NAU student, we might be interested in the event that the student is a biology student. A second event of interest might be if the student is an undergraduate.

4.1.1 Venn Diagrams

Let S be the set of all outcomes of my random trial. Suppose I am interested in two events A and B. The traditional way of representing these events is using a *Venn diagram.*



For example, suppose that my random experiment is rolling a fair 6-sided die once. The possible outcomes are $S = \{1, 2, 3, 4, 5, \text{ or } 6\}$. Suppose I then define events $A =$ *roll is odd* and $B =$ *roll is 5 or greater*. In this case our picture is:



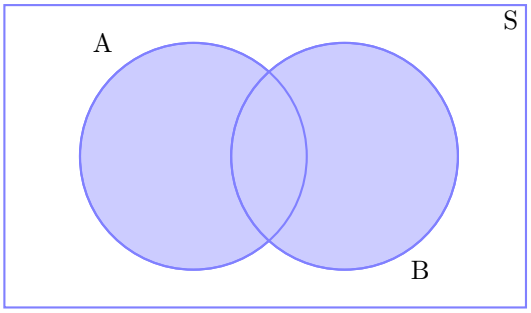
All of our possible events are present, and distributed amongst our possible events.

4.1.2 Composition of events

I am often interested in discussing the composition of two events and we give the common set operations below.

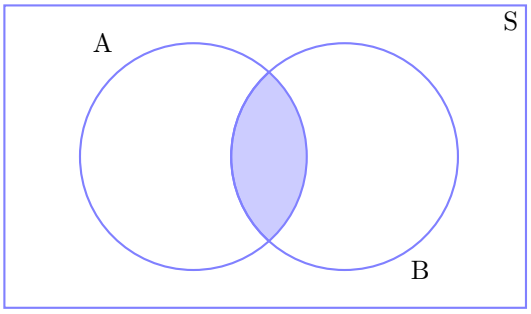
- Union: Denote the event that either A or B occurs as $A \cup B$.

$A \cup B$



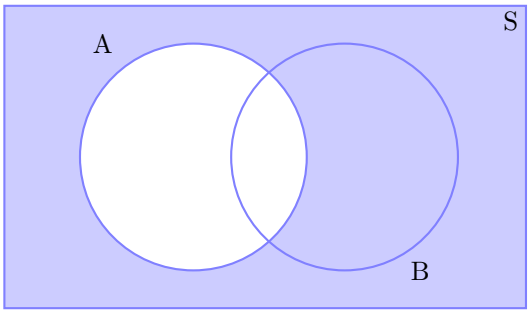
- Denote the event that both A and B occur as $A \cap B$

$A \cap B$



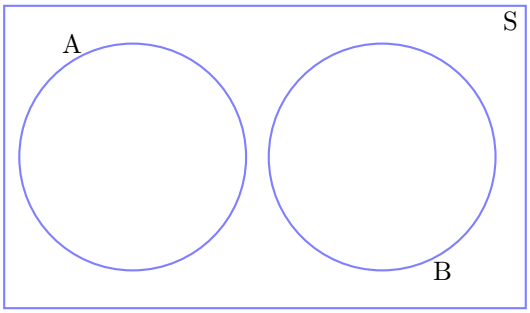
- Denote the event that A does not occur as $\bar{A}$ or A^C (different people use different notations)

$\bar{A}$ or A^C



Definition 1. Two events A and B are said to be mutually exclusive (or disjoint) if the occurrence of one event precludes the occurrence of the other. For example, on a single roll of a die, a two and a five cannot both come up. For a second example, define A to be the event that the die is even, and B to be the event that the die comes up as a 5.

A and B disjoint



4.2 Probability Rules

4.2.1 Simple Rules

We now take our Venn diagrams and use them to understand the rules of probability. The underlying idea that we will use is the the probability of an event is the *area* in the Venn diagram.

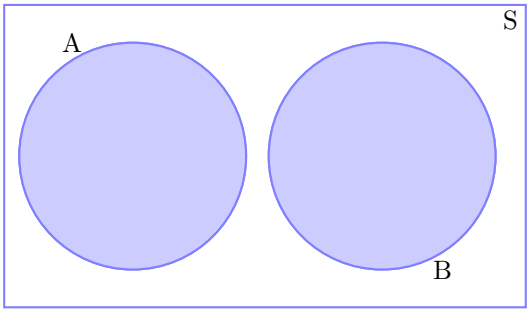
Definition 2. Probability is the proportion of times an event occurs in many repeated trials of a random phenomenon. In other words, probability is the long-term relative frequency.

Fact. *For any event A the probability of the event $P(A)$ satisfies $0 \leq P(A) \leq 1$ since proportions always lie in $[0, 1]$*

Because S is the set of all events that might occur, the area of our bounding rectangle will be 1 and the probability of event A occurring will be the area in the circle A.

Fact. *If two events are mutually exclusive, then $P(A \cup B) = P(A) + P(B)$*

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



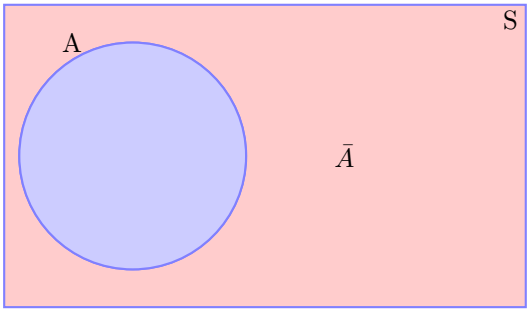
Example. Let S be the sum of two different colored dice. Suppose we are interested in $P(S \leq 4)$. Notice that the pair of dice can fall 36 different ways (6 ways for the first die and six for the second results in 6x6 possible outcomes, and each way has equal probability 1/36. Since the dice cannot simultaneously sum to 2 *and* to 3, we could write

$$\begin{aligned} P(S \leq 4) &= P(S = 2) + P(S = 3) + P(S = 4) \\ &= P(\{1,1\}) + P(\{1,2\} \text{ or } \{2,1\}) + P(\{1,3\} \text{ or } \{2,2\} \text{ or } \{3,1\}) \\ &= \frac{1}{36} + \frac{2}{36} + \frac{3}{36} \\ &= \frac{6}{36} \\ &= \frac{1}{6} \end{aligned}$$

Fact. $P(A) + P(\bar{A}) = 1$.

The above statement is true because the probability of whole space S is one (remember S is all possible outcomes), then either we get an outcome in which A occurs or we get an outcome in which A does not occur.

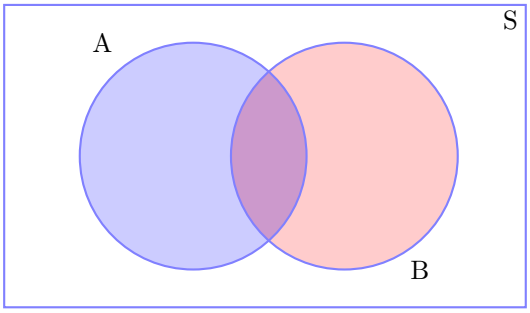
$$P(A) + P(\bar{A}) = 1$$



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

The reason behind this fact is that if there is if A and B are not disjoint, then some area is added *twice* when I calculate $P(A) + P(B)$. To account for this, I simply subtract off the area that was double counted.

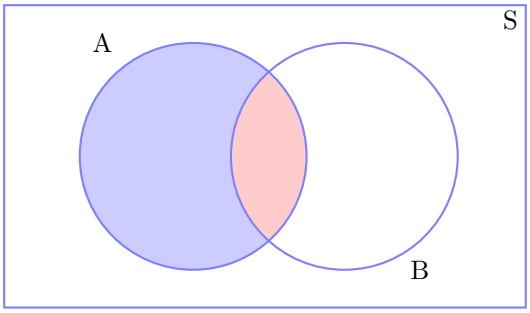
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$



Fact 3. $P(A) = P(A \cap B) + P(A \cap \bar{B})$

This identity is just breaking the event A into two disjoint pieces.

$$P(A) = P(A \cap \bar{B}) + P(A \cap B)$$



4.2.2 Conditional Probability

We are given the following data about insurance claims. Notice that the data is given as $P($ *Category* $\cap$ *PolicyType* $)$ which is apparent because the sum of all the elements in the table is 100%:

	Type of Policy (%)		
Category	Fire	Auto	Other
Fraudulent	6%	1%	3%
Non-fraudulent	14%	29%	47%

Summing across the rows and columns, we can find the probabilities of for each category and policy type.

	Type of Policy (%)			
Category	Fire	Auto	Other	Total %
Fraudulent	6%	1%	3%	10%
Non-fraudulent	14%	29%	47%	90%
Total	20%	30%	50%	100%

It is clear that fire claims are more likely fraudulent than auto or other claims. In fact, the proportion of fraudulent claims, given that the claim is against a fire policy is

$$\begin{aligned} P(\text{ Fraud} \mid \text{FirePolicy }) &= \frac{\text{proportion of claims that are fire policies and are fraudulent}}{\text{proportion of fire claims}} \\ &= \frac{6\%}{20\%} \\ &= 0.3 \end{aligned}$$

In general we define conditional probability (assuming $P(B) \neq 0$) as

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

which can also be rearranged to show

$$\begin{aligned} P(A \cap B) &= P(A \,|\, B)\, P(B) \\ &= P(B \,|\, A)\, P(A) \end{aligned}$$

since the order doesn't matter and $P\,(A \cap B) = P\,(B \cap A)$.

Using this rule, we might calculate the probability that a claim is an Auto policy given that it is not fraudulent.

$$\begin{aligned} P\,(\,Auto \,|\, NotFraud\,) &= \frac{P\,(\,Auto \,\cap\, NotFraud)}{P\,(\,NotFraud\,)} \\ &= \frac{0.29}{0.9} \\ &= 0.3\bar{2} \end{aligned}$$

Definition 4. Two events A and B are said to be **independent** if

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

What independence is saying that knowing the outcome of event A doesn't give you any information about the outcome of event B.

- In simple random sampling, we assume that any two samples are independent.
- In cluster sampling, we assume that samples within a cluster are not independent, but clusters are independent of each other.

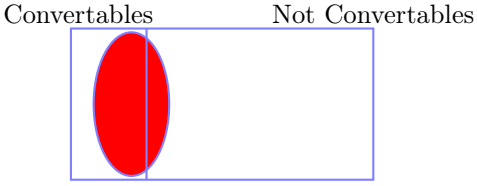
Fact 5. *If A and B are independent events, then*

$$\begin{aligned} P(A \cap B) &= P(A|B)P(B) \\ &= P(A)P(B) \end{aligned}$$

Example 6. Suppose that we are interested in the relationship between the color and the type of car. Specifically I will divide the car world into convertibles and non-convertibles and the colors into red and non-red.

Suppose that convertibles make up just 10% of the domestic automobile market. This is to say $P(\,Convertable\,) = 0.10$. Of the non-convertibles, red is not unheard of but it isn't common either. So suppose $P(\,Red \mid NonConvertable\,) = 0.15$. However red is an extremely popular color for convertibles so let $P(\,Red \mid Convertable\,) = 0.60$.

We can visualize this information via another Venn diagram:



Given the above information, we can create the following table:

	Convertable	non-Convertable	
Red			
Not Red			
	0.10	0.90	

We can fill in some of the table using our the definition of conditional probability. For example:

$$\begin{aligned} P\left(Red \cap Convertable\right) &= P\left(Red \,|\, Convertable\right) P\left(Convertable\right) \\ &= 0.60 * 0.10 \\ &= 0.06 \end{aligned}$$

Lets think about what this conditional probability means. Of the 90% of cars that are not convertibles, 15% those non-convertibles are red and therefore the proportion of cars that are red non-convertibles is $0.90 * 0.15 = 0.135$. Of the 10% of cars that are convertibles, 60% of those are red and therefore proportion of cars that are red convertibles is $0.10 * 0.60 = 0.06$. Thus the total percentage of red cars is actually

$$\begin{aligned} P\left(\, Red\,\right) &= P\left(\, Red \,\cap\, Convertible\,\right) + P\left(\, Red \cap NonConvertible\,\right) \\ &= P\left(\, Red \,|\, Convertable\,\right) P\left(\, Convertible\,\right) + P\left(\, Red \,|\, NonConvertible\,\right) P\left(\, NonConvertible\,\right) \\ &= 0.60 * 0.10 + 0.15 * 0.90 \\ &= 0.06 + 0.135 \\ &= 0.195 \end{aligned}$$

So when I ask for $P($ *red* $|$ *convertable* $)$, I am narrowing my space of cars to consider only convertibles. While there percentage of cars that are red and convertible is just 6% of all cars, when I restrict myself to convertibles, we see that the percentage of this smaller set of cars that are red is 60%.

Notice that because $P\left(Red\right) = 0.195 \neq 0.60 = P\left(Red \,|\, Convertable\right)$ then the events *Red* and *Convertable* are not independent.

4.2.3 Summary of Probability Rules

$$0 \leq P\left(A\right) \leq 1$$

$$P\left(A\right) + P\left(\bar{A}\right) = 1$$

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

$$P\left(A \cap B\right) = \begin{cases} P\left(A \,|\, B\right) P\left(B\right) \\ P\left(B \,|\, A\right) P\left(A\right) \\ P(A)P(B) \qquad \text{if A,B are independent} \end{cases}$$

$$P\left(A \,|\, B\right) = \frac{P\left(A \cap B\right)}{P\left(B\right)}$$

4.3 Discrete Random Variables

The different types of probability distributions (and therefore your analysis method) can be divided into two general classes:

1. Continuous Random Variables - the variable takes on numerical values and could, in principle, take any of an uncountable number of values. In practical terms, if fractions or decimal points in the number make sense, it is usually continuous.

2. Discrete Random Variables - the variable takes on one of small set of values (or only a countable number of outcomes). In practical terms, if fractions or decimals points don't make sense, it is usually discrete.

Examples:

1. Presence or Absence of wolves in a State?

2. Number of Speeding Tickets received?

3. Tree girth (in cm)?

4. Photosynthesis rate?

4.3.1 Introduction to Discrete Random Variables

The following facts hold for discrete random variables:

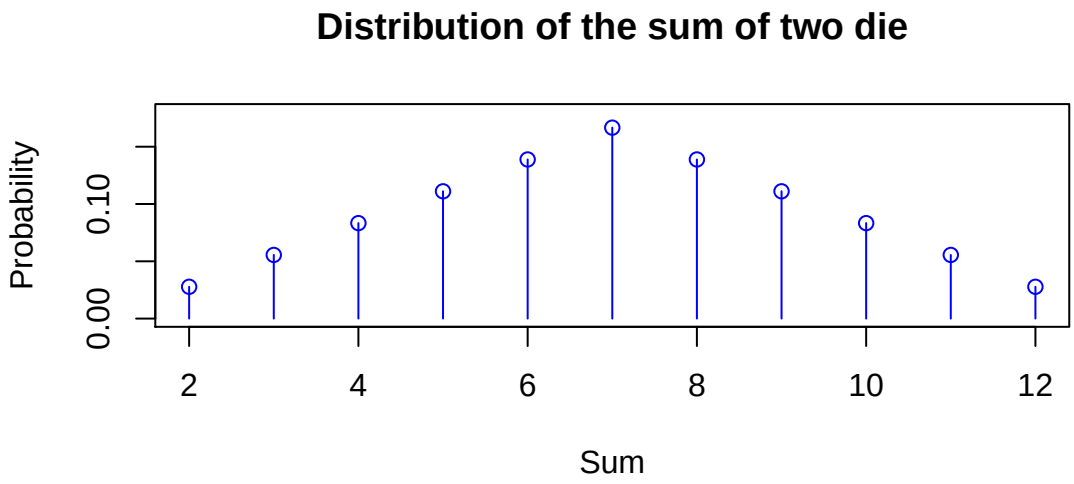
1. The probability associated with every value lies between 0 and 1

2. The sum of all probabilities for all values is equal to 1

3. Probabilities for discrete RVs are additive. i.e., $P(3 \text{ or } 4) = P(3) + P(4)$

Expected Value

Example: Consider the discrete random variable S, the sum of two fair dice.



We often want to ask 'What is expected value of this distribution?' You might think about taking a really, really large number of samples from this distribution. We define the expected value (often denoted by μ) as *a weighted average of the possible values and the weights are the proportions*

with which those values occur.

$$\begin{aligned}
\mu = E[S] &= \sum_{\text{possible } s} s \cdot P(S = s) \\
&= \sum_{s=2}^{12} s \cdot P(S = s) \\
&= 2 \cdot P(S = 2) + 3 \cdot P(S = 3) + \cdots + 11 \cdot P(S = 11) + 12 \cdot P(S = 12) \\
&= 2\left(\frac{1}{36}\right) + 3\left(\frac{2}{36}\right) + \cdots + 11\left(\frac{2}{36}\right) + 12\left(\frac{1}{36}\right) \\
&= 7
\end{aligned}$$

Variance

Similarly we could define the variance of S (which we often denote σ^2) as *a weighted average of the squared-deviations that could occur.*

$$\begin{aligned}
\sigma^2 = V[S] &= \sum_{s=2}^{12} (s - \mu)^2 P(S = s) \\
&= (2-7)^2\left(\frac{1}{36}\right) + (3-7)^2\left(\frac{2}{36}\right) + \cdots + (12-7)^2\left(\frac{1}{36}\right) \\
&= \frac{35}{6} = 5.8\bar{3}
\end{aligned}$$

We could interpret the expectation as the sample mean of an infinitely large sample, and the variance as the sample variance of the same infinitely large sample. These are two very important numbers that describe the distribution.

Example 7. My wife is a massage therapist and over the last year, the number of clients she sees per work day (denoted Y) varied according the following table:

Number of Clients	0	1	2	3	4
Frequency/Probability	0.3	0.35	0.20	0.10	0.05

Because this is the long term relative frequency of the number of clients (over 200 working days!), it is appropriate to interpret these frequencies as probabilities. This table is often called a *probability mass function (pmf)* because it lists how the probability is spread across the possible values of the random variable. We might next ask ourselves what is the average number of clients per day? It looks like it ought to be between 1 and 2 clients per day.

$$\begin{aligned}
E(Y) &= \sum_{\text{possible } y} y\, P(Y = y) \\
&= \sum_{y=0}^{4} y\, P(Y = y) \\
&= 0\, P(Y = 0) + 1\, P(Y = 1) + 2\, P(Y = 2) + 3\, P(Y = 3) + 4\, P(Y = 4) \\
&= 0\,(0.3) + 1\,(0.35) + 2\,(0.20) + 3\,(0.10) + 4\,(0.05) \\
&= 1.25
\end{aligned}$$

Assuming that successive days are independent (which might be a bad assumption) what is the probability she has two days in a row with no clients?

$$\begin{aligned}
P(0 \text{ on day1 } \textit{and } 0 \text{ on day2}) &= P(0 \text{ on day 1})\, P(0 \text{ on day 2}) \\
&= (0.3)\,(0.3) \\
&= 0.09
\end{aligned}$$

What is the variance of this distribution?

$$\begin{aligned} V(S) &= \sum_{\text{possible y}} (y-\mu)^2 \, P(Y=y) \\ &= \sum_{y=0}^{4} (y-\mu)^2 \, P(Y=y) \\ &= (0-1.25)^2(0.3) + (1-1.25)^2(0.35) + (2-1.25)^2(0.20) + (3-1.25)^2(0.10) + (4-1.25)^2(0.05) \\ &= 1.2875 \end{aligned}$$

Note on Notation: There is a difference between the upper and lower case letters we have been using to denote a random variable. In general, we let the upper case denote the random variable and the lower case as a value that the the variable could possibly take on. So in the massage example, the number of clients seen per day Y could take on values $y = 0, 1, 2, 3$, or 4.

4.4 Common Discrete Distributions

4.4.1 Binomial Distribution

Example: Suppose we are trapping small mammals in the desert and we spread out three traps. Assume that the traps are far enough apart that having one being filled doesn't affect the probability of the others being filled and that all three traps have the same probability of being filled in an evening. Denote the event that a trap is filled as F_i and if it is empty E_i (note I could have used $\bar{F}_i$). Denote the probability that a trap is filled by $\pi = 0.8$. (This sort of random variable is often referred to as a Bernoulli RV.)

The possible outcomes are

Outcome
$E_1E_2E_3$
$F_1E_2E_3$
$E_1F_2E_3$
$E_1E_2F_3$
$E_1F_2F_3$
$F_1E_2F_3$
$F_1F_3E_3$
$F_1F_2F_3$

Because these are far apart enough in space that the outcome of Trap1 is independent of Trap2 and Trap3, the

$$\begin{aligned} P(E_1 \cap F_2 \cap E_3) &= P(E_1)P(F_2)P(E_3) \\ &= (1-0.8)0.8(1-0.8) \\ &= 0.032 \end{aligned}$$

Notice how important the assumption of independence is!!! Similarly we could calculate the probabilities for the rest of the table.

Outcome	Probability	S outcome	Probability
$E_1E_2E_3$	0.008	$S = 0$	0.008
$F_1E_2E_3$	0.032	$S = 1$	3(0.032)
$E_1F_2E_3$	0.032		
$E_1E_2F_3$	0.032		
$E_1F_2F_3$	0.128	$S = 2$	3(0.128)
$F_1E_2F_3$	0.128		
$F_1F_3E_3$	0.128		
$F_1F_2F_3$	0.512	$S = 3$	0.512

Next we are interested in the random variable S, the number of traps that were filled:

Outcome	Probability
$S = 0$	$1(0.008) = 0.008$
$S = 1$	$3(0.032) = 0.096$
$S = 2$	$3(0.128) = 0.384$
$S = 3$	$1(0.512) = 0.512$

S is an example of a **Binomial Random Variable.** A binomial experiment is one that:

1. Experiment consists of n identical trials
2. Each trial results in one of two outcomes (Heads/Tails, presence/absence). One will be labeled a success and the other a failure.
3. The probability of success on a single trial is equal to π and remains the same from trial to trial.
4. The trials are independent (this is implied from property 3)
5. The random variable Y is the number of successes observed during n trials

Recall that the probability mass function (pmf) describes how the probability is spread across the possible outcomes, and in this case, I can describe this via a nice formula. The pmf of a a binomial random variable Y taken from n trials each with probability of success π is

$$P(Y = y) = \underbrace{\frac{n!}{y!(n-y)!}}_{orderings} \underbrace{\pi^y}_{y\ successes} \underbrace{(1-\pi)^{n-y}}_{n-y\ failures}$$

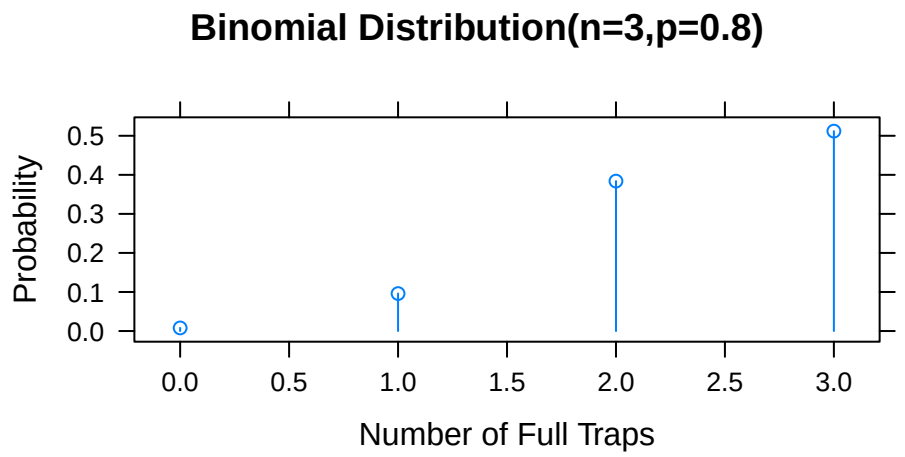
where we define $n! = n(n-1)\dots(2)(1)$ and further define $0! = 1$. Often the ordering term is written more compactly as $\binom{n}{y} = \frac{n!}{y!(n-y)!}$.

There will be a variety of distributions we'll be interested in and R refers to them using the following abbreviations

Distribution	R stem	parameters (and defaults)	parameter interpretation
Binomial	`binom`	`size` `prob`	number of trials probability of success
Poisson	`pois`	`lambda`	mean
Exponential	`exp`	`rate` or `lambda`	λ represents the mean, while rate is $\frac{1}{\lambda}$.
Normal	`norm`	`mean=0` `sd=1`	mean standard deviation
Uniform	`unif`	`min=0` `max=1`	lower bound upper bound

For our small mammal example we can create a graph that shows the binomial distribution with the following R code:

```
library(mosaic)
plotDist('binom', size=3, prob=0.8,
         main='Binomial Distribution(n=3,p=0.8)',
         ylab='Probability', xlab='Number of Full Traps')
```



To calculate the height of any of these bars, we can evaluate the pmf at the desired point. For example, to calculate the probability the number of full traps is 2, we calculate the following

$$\begin{aligned} P(S=2) &= \binom{3}{2}(0.8)^2(1-0.8)^{3-2} \\ &= \frac{3!}{2!(3-2)!}(0.8)^2(0.2)^{3-2} \\ &= \frac{3\cdot 2\cdot 1}{(2\cdot 1)1}(0.8)^2(0.2) \\ &= 3(0.128) \\ &= 0.384 \end{aligned}$$

You can use R to calculate these probabilities. In general, for any distribution, the "d-function" gives the distribution function (pmf or pdf). So to get R to do the preceding calculation we use:

```
# P( Y = 2 | n=3, pi=0.8 )
dbinom(2, size=3, prob=0.8)

## [1] 0.384
```

The expectation of this distribution can be shown to be

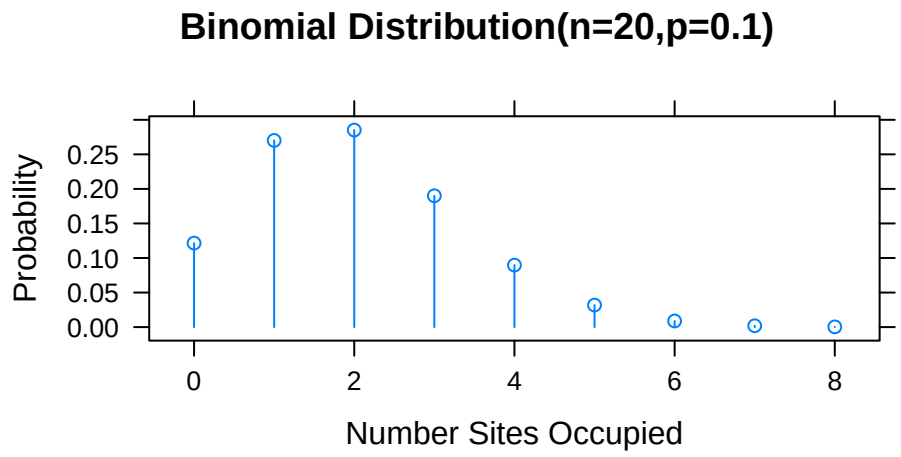
$$\begin{aligned} E[Y] &= \sum_{y=0}^{n} y\, P(Y=y) \\ &= \sum_{y=0}^{n} y\, \frac{n!}{y!\,(n-y)!}\pi^y\,(1-\pi)^{n-y} \\ &= \vdots \\ &= n\pi \end{aligned}$$

and the variance can be similarly calculated

$$\begin{aligned} V[Y] &= \sum_{y=0}^{n} (y - E\,[Y])^2\; P\,(Y=y|n,\pi) \\ &= \sum_{y=0}^{n} (y - E\,[Y])^2\; \frac{n!}{y!\,(n-y)!}\pi^y\,(1-\pi)^{n-y} \\ &= \vdots \\ &= n\pi(1-\pi) \end{aligned}$$

Example 8. Suppose a bird survey only captures the presence or absence of a particular bird (say the mountain chickadee). Assuming the true presence proportion at national forest sites around Flagstaff $\pi = 0.1$, then for $n = 20$ randomly chosen sites, the number of sites in which the bird was observed would have the distribution

```
plotDist('binom', size=20, prob=0.1,
         main='Binomial Distribution(n=20,p=0.1)',
         ylab='Probability', xlab='Number Sites Occupied')
```



Often we are interested in questions such as $P(Y \leq 2)$ which is the probability that we see 2 or fewer of the sites being occupied by mountain chickadee. These calculations can be tedious to calculate by hand but R will calculate these cumulative distribution function values for you using the "p-function". This cumulative distribution function gives the sum of all values up to and including the number given.

```
# P(Y=0) + P(Y=1) + P(Y=2)
sum <- dbinom(0, size=20, prob=0.1) +
       dbinom(1, size=20, prob=0.1) +
       dbinom(2, size=20, prob=0.1)
sum

## [1] 0.6769268

# P(Y <= 2)
pbinom(2, size=20, prob=0.1)

## [1] 0.6769268
```

In general we will be interested in asking four different questions about a distribution.

1. What is the height of the probability mass function (or probability density function). For discrete variable Y this is $P(Y = y)$ for whatever value of y we want.
2. What is the probability of observing a value less than or equal to y? In other words, to calculate $P(Y \leq y)$.
3. What is a particular quantile of a distribution? For example, what value separates the lower 25% from the upper 75%?
4. Generate a random sample of values from a specified distribution.

All the probability distributions available in R are accessed in exactly the same way, using a `d-function`, `p-function`, `q-function`, and `r-function`. For the rest of this section suppose that X is a random variable from the distribution of interest and x is some possible value that X could take on.

Function	Result	Example

`d-function(x)`	Discrete: $P(X = x)$
	Continuous: the height of the density function at the given point
		`dbinom(0, size=20, prob=.1)`
		`dnorm(0, mean=0, sd=1)`
		is the height ≈ 0.40

`p-function(x)`	$P(X \leq x)$	`pnorm(-1.96, mean=0, sd=1)` is $P(Z < 1.96) = 0.025$
`q-function(q)`	x such that $P(X \leq x) = q$	`qnorm(0.05, mean=0, sd=1)` is z such that $P(Z \leq z) = 0.05$ which is $z = -1.645$
`r-function(n)`	n random observations from the distribution	`rnorm(n=10, mean=0, sd=1)` generates 10 observations from a N(0,1) distribution

4.4.2 Poisson Distribution

A commonly used distribution for count data is the Poisson.

1. Number of customers arriving over a 5 minute interval
2. Number of birds observed during a 10 minute listening period
3. Number of prairie dog towns per 1000 hectares
4. Number of alga clumps per cubic meter of lake water

For a RV is a Poisson RV if the following conditions apply:

1. Two or more events do not occur at precisely the same time or in the same space

2. The occurrence of an event in a given period of time or region of space is independent of the occurrence of the event in a non overlapping period or region.

3. The expected number of events during one period or region, λ, is the same in all periods or regions of the same size.

Assuming that these conditions hold for some count variable Y, the the probability mass function is given by

$$P(Y = y) = \frac{\lambda^y e^{-\lambda}}{y!}$$

where λ is the expected number of events over 1 unit of time or space and e is the constant 2.718281828.

$$\begin{aligned} E[Y] &= \lambda \\ Var[Y] &= \lambda \end{aligned}$$

Example 9. Suppose we are interested in the population size of small mammals in a region. Let Y be the number of small mammals caught in a large trap (multiple traps in the same location?) in a 12 hour period. Finally, suppose that $Y \sim Poi(\lambda = 2.3)$. What is the probability of finding exactly 4 critters in our trap?

$$\begin{aligned} P(Y = 4) &= \frac{2.3^4\, e^{-2.3}}{4!} \\ &= 0.1169 \end{aligned}$$

What about the probability of finding at most 4?

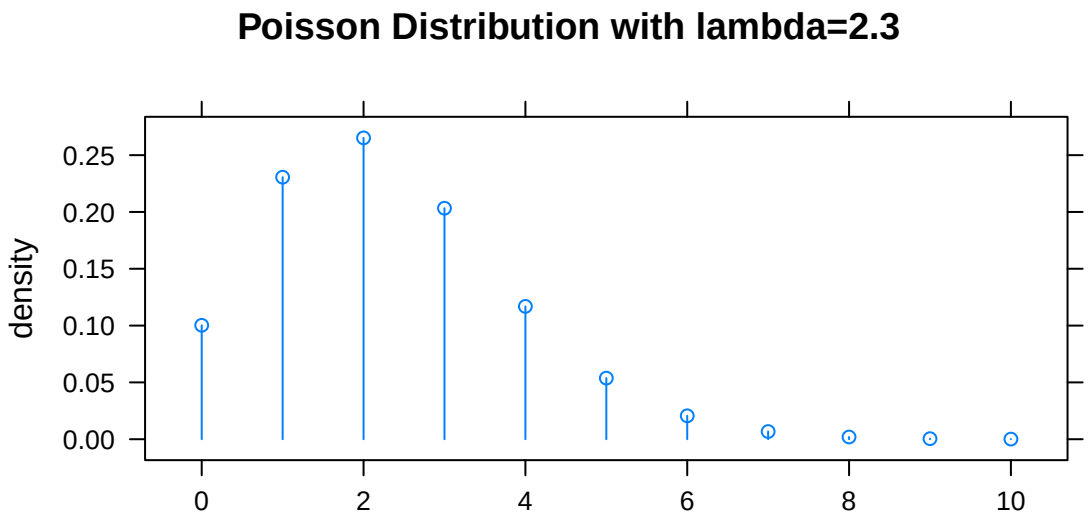
$$\begin{aligned} P(Y \leq 4) &= P(Y = 0) + P(Y = 1) + P(Y = 2) + P(Y = 3) + P(Y = 4) \\ &= 0.1003 + 0.2306 + 0.2652 + 0.2033 + 0.1169 \\ &= 0.9163 \end{aligned}$$

What about the probability of finding 5 or more?

$$\begin{aligned} P(Y \geq 5) &= 1 - P(Y \leq 4) \\ &= 1 - 0.9163 \\ &= 0.0837 \end{aligned}$$

These calculations can be done using the distribution function (`d-function`) for the poisson and the cumulative distribution function (`p-function`).

```
plotDist('pois', lambda=2.3,
         main='Poisson Distribution with lambda=2.3', ylab='density')
```




```
# P( Y = 4)
dpois(4, lambda=2.3)

## [1] 0.1169022

# P( Y <= 4)
ppois(4, lambda=2.3)

## [1] 0.9162493

# 1-P(Y <= 4)  ==  P( Y > 4)  ==  P( Y >= 5)
1-ppois(4, 2.3)

## [1] 0.08375072
```

4.5 Continuous Random Variables

Finding the area under the curve of a particular density function $f(x)$ requires the use of calculus, but since this isn't a calculus course, we will resort to using R or tables of calculated values.

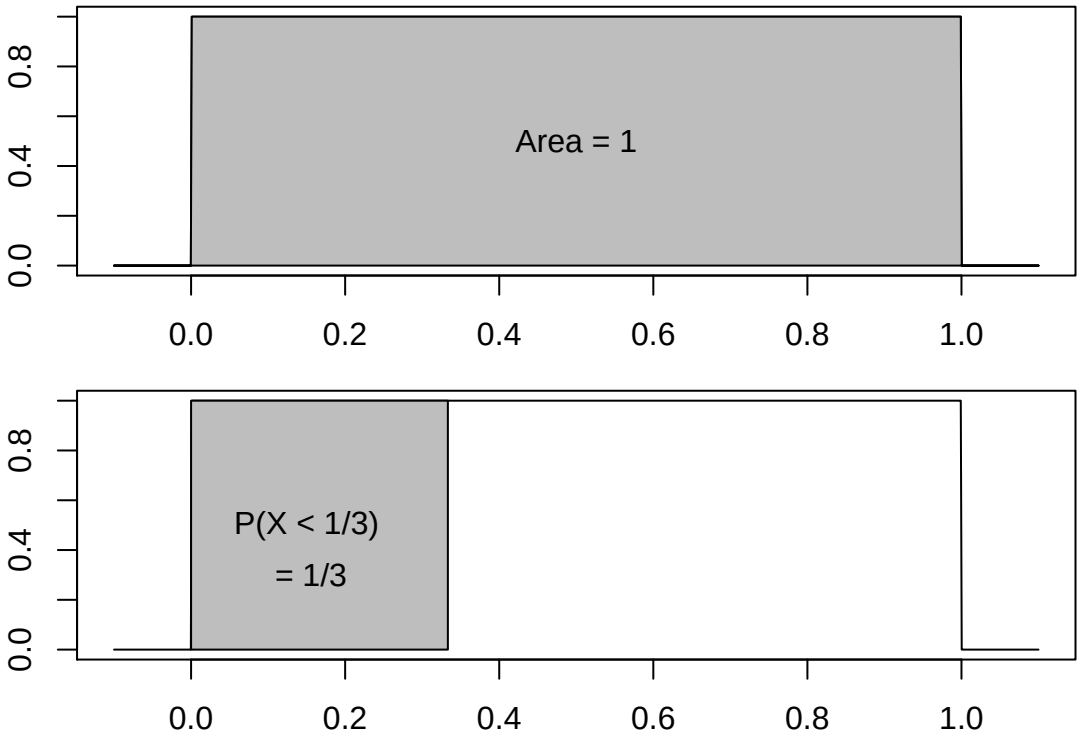
4.5.1 Uniform(0,1) Distribution

Suppose you wish to draw a random number number between 0 and 1 and each number should have an equal chance of being selected. This random variable is said to have a *Uniform(0,1)* distribution.

Because there are an infinite number of rational numbers between 0 and 1, the probability of any particular number being selected is $1/\infty = 0$. But even though each number has 0 probability of being selected, some number must end up being selected. Because of this conundrum, probability theory doesn't look at the probability of a single number, but rather focuses on a *region of numbers*.

To make this distinction, we will define the distribution using a *probability density function* instead of the probability mass function. In the discrete case, we had to constrain the probability

mass function to sum to 1. In the continuous case, we have to constrain the probability density function to integrate to 1.

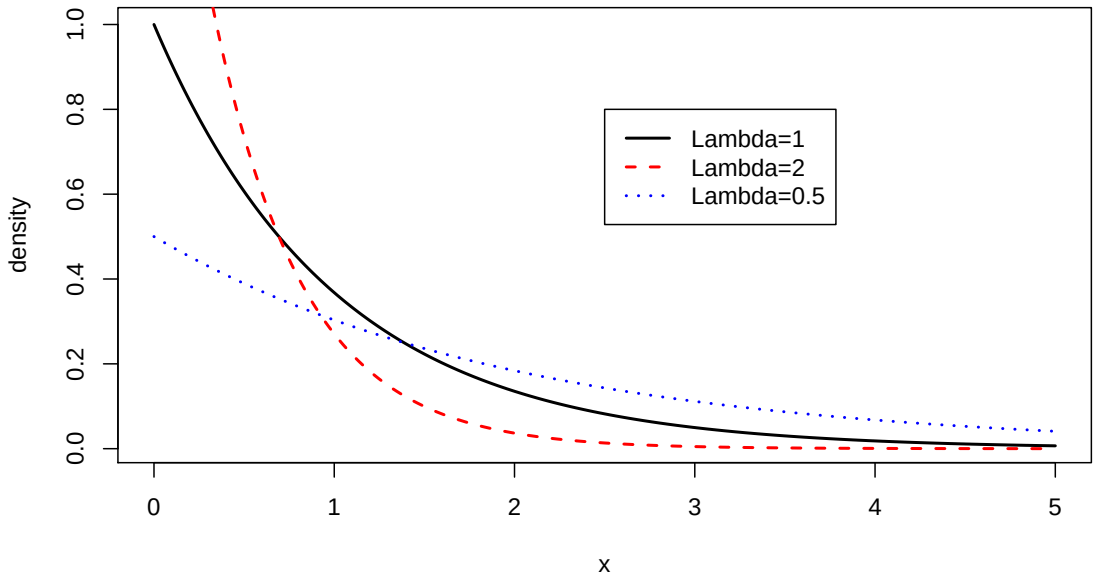


Finding the area under the curve of a particular density function $f(x)$ usually requires the use of calculus, but since this isn't a calculus course, we will resort to using R or tables of calculated values.

4.5.2 Exponential Distribution

The exponential distribution is the continuous analog of the Poisson distribution and is often used to model the time between occurrence of successive events. Perhaps we are modeling time between transmissions on a network, or the time between feeding events or prey capture. If the random variable X has an Exponential distribution, its distribution function is

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \text{ and } \lambda > 0 \\ 0 & \text{otherwise} \end{cases}$$



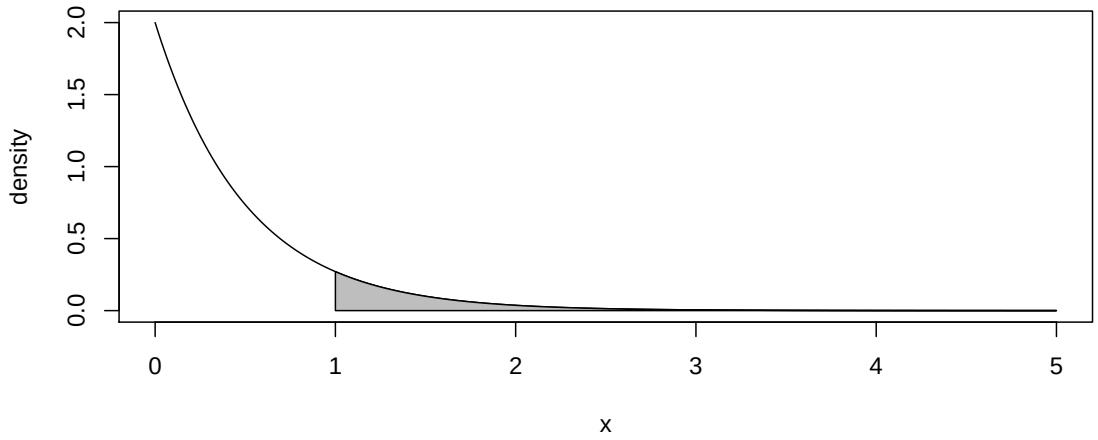
Analogous to the discrete distributions, we can define the Expectation and Variance of these distributions by replacing the summation with an integral

$$\begin{aligned} E[X] &= \int_0^\infty x\, f(x)\, dx \\ &= \vdots \\ &= \frac{1}{\lambda} \\ Var[X] &= \int_0^\infty (x - E[X])^2\, f(x)\, dx \\ &= \vdots \\ &= \frac{1}{\lambda^2} \end{aligned}$$

Since the exponential distribution is defined by the rate of occurrence of an event, increasing that rate *decreases* the time between events. Furthermore since the rate of occurrence cannot be negative, we restrict $\lambda > 0$

Example 10. Suppose the time between insect captures X during a summer evening for a species of bat follows a exponential distribution with capture rate of $\lambda = 2$ insects per minute and therefore the expected waiting time between captures is $1/\lambda = 1/2$ minute. Suppose that we are interested in the probability that it takes a bat more than 1 minute to capture its next insect.

$P(X > 1) =$



We now must resort to calculus to find this area. Or use tables of pre-calculated values. Or use R (remembering that `p-functions` give the area under the curve *to the left of the given value.*

```
# P(X > 1)  == 1 - P(X <= 1)
1 - pexp(1, rate=2)

## [1] 0.1353353
```

4.5.3 Normal Distribution

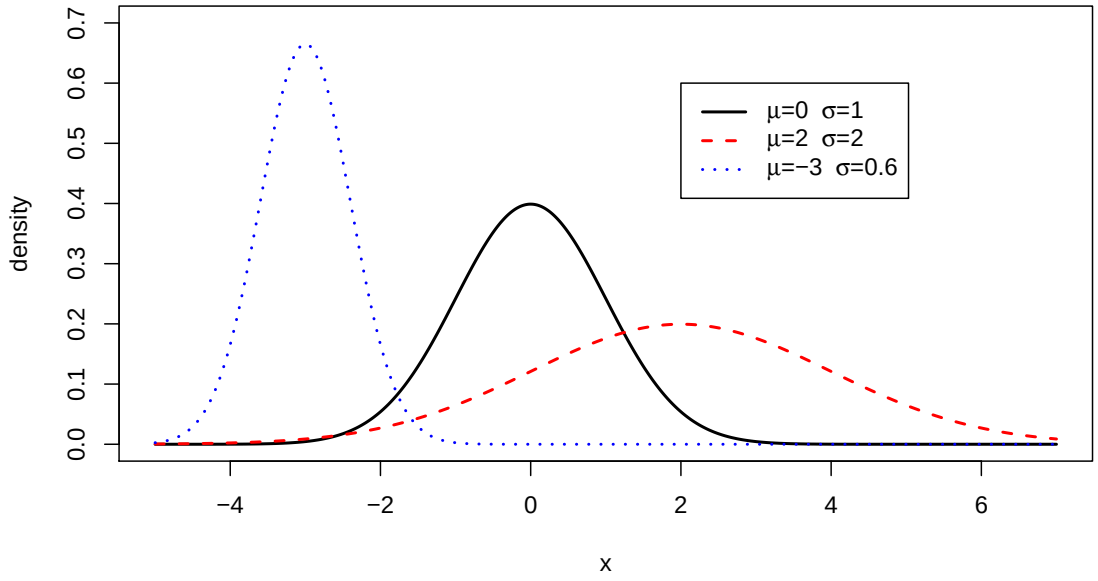
Undoubtably the most important distribution in statistics is the normal distribution. If my RV X is normally distributed with mean μ and standard deviation σ, its probability distribution function is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right]$$

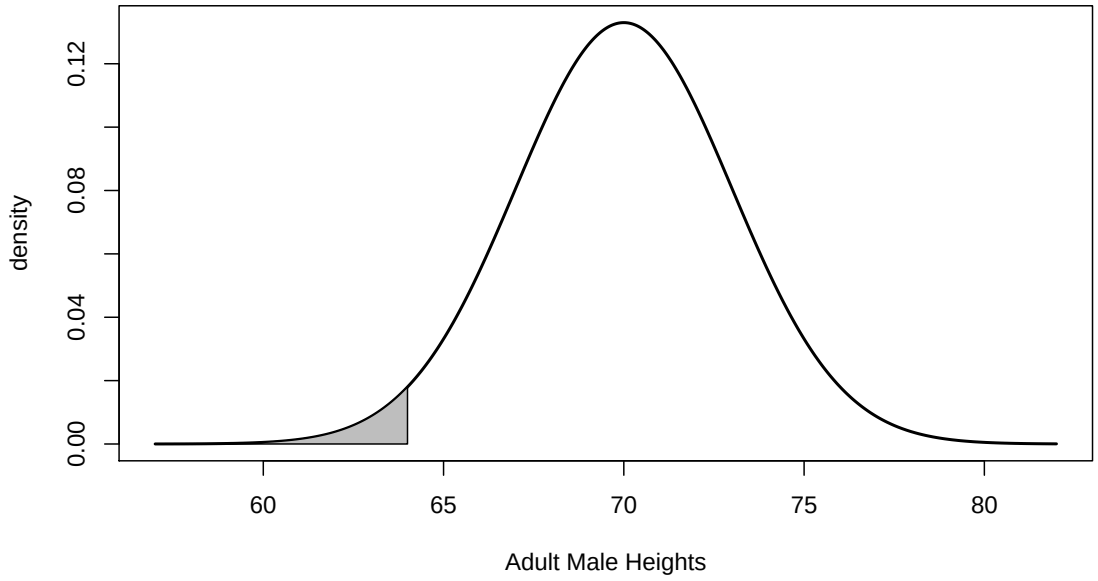
where $\exp[y]$ is the exponential function e^y. We could slightly rearrange the function to

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

and see this distribution is defined by its expectation $E[X] = \mu$ and its variance $Var[X] = \sigma^2$. Notice I could define it using the standard deviation σ, and different software packages will expect it to be defined by one or the other. R defines the normal distribution using the standard deviation.



Example 11. It is known that the heights of adult males in the US is approximately normal with a mean of 5 feet 10 inches ($\mu = 70$ inches) and a standard deviation of $\sigma = 3$ inches. Your instructor is a mere 5 feet 4 inches (64 inches). What proportion of the population is shorter than your professor?



Using R you can easily find this

```
# P( Y <= 64 )
pnorm(64, mean=70, sd=3)

## [1] 0.02275013
```

but unfortunately your professor may need to ask similar questions on an exam and so we have to talk about how to do the same calculation using a table. First we notice that the normal distribution is defined by the number of standard deviations from the mean (which we denote as z)

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

and we note that he is -2 standard deviations from the mean because

$$\begin{aligned} z &= \frac{64 - 70}{3} \\ &= \frac{-6}{3} \\ &= -2 \end{aligned}$$

Next we look at the table in the front of the book for $z = -2.00$. To do this we go down to the -2.0 row and over to the .00 column and find 0.0228. Only slightly over 2% of the adult male population is shorter!

How tall must a male be to be taller than 80% of the rest of the male population? To answer that we must use the table in reverse and look for the 0.8 value. We find the closest value possible (0.7995) and the z value associated with it is $z = 0.84$. Next we solve the standardizing equation for x

$$\begin{aligned} z &= \frac{x - \mu}{\sigma} \\ 0.84 &= \frac{x - 70}{3} \\ x &= 3(0.84) + 70 \\ &= 72.49 \text{ inches} \end{aligned}$$

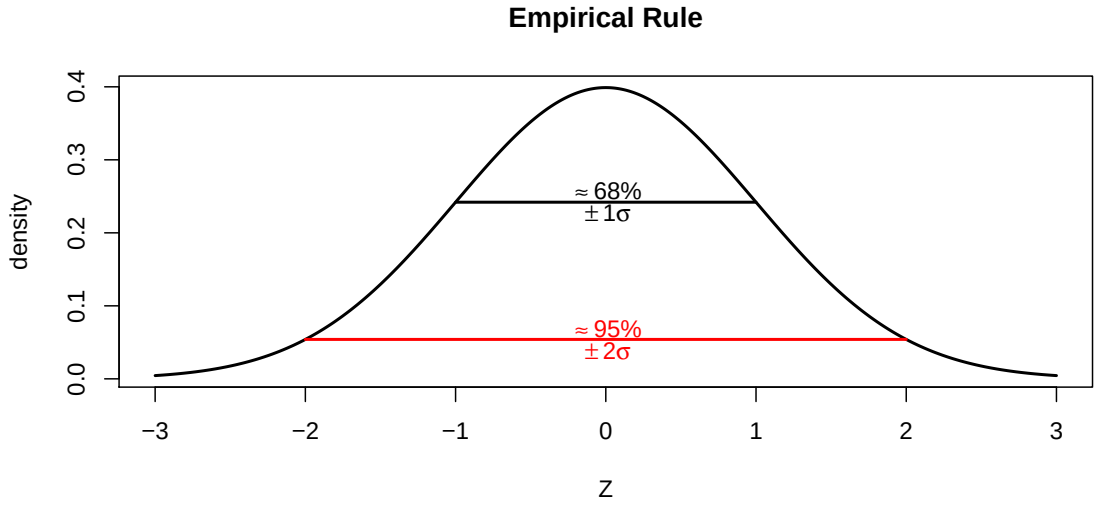
Alternatively we could use the quantile function for the normal distribution `(q-function)` in R and avoid the imprecision of using a table.

```
qnorm(.8, mean=0, sd=1)

## [1] 0.8416212
```

$$\begin{aligned} x &= 3(0.8416) + 70 \\ &= 72.52 \text{ inches} \end{aligned}$$

Empirical Rule - It is from the normal distribution that the empirical rule from chapter 3 is derived. If $X \sim N(\mu, \sigma^2)$ then

$$\begin{aligned} P(\mu - \sigma \leq X \leq \mu + \sigma) &= P(-1 \leq Z \leq 1) \\ &= P(Z \leq 1) - P(Z \leq -1) \\ &\approx 0.8413 - 0.1587 \\ &= 0.6826 \end{aligned}$$


Empirical Rule
density
0.0
0.1
0.2
0.3
0.4
≈ 68%
±1σ
≈ 95%
±2σ
−3
−2
−1
0
1
2
3
Z

Chapter 5

Sampling Distributions

Claim: For random variables X and Y and constant a the following statements hold:

$$\begin{aligned} E(aX) &= aE(X) \\ Var(aX) &= a^2 Var(X) \\ E(X+Y) &= E(X)+E(Y) \\ E(X-Y) &= E(X)-E(Y) \\ Var(X \pm Y) &= Var(X)+Var(Y) \text{ if X,Y are independent} \end{aligned}$$

Example for a Discrete case: Suppose that the number of cavities (X) that are detected during a trip to the dentist can be modeled via a Poisson with $\lambda = 1$. This dentist charges \$50 for filling each cavity, and we are interested in calculating the estimated cost $C = \$50X$. Lets walk through this:

Num Cavities	0	1	2	3	4	5	6	7	8	...
Cost	0	50	100	150	200	250	300	350	400	...
Probability	0.3679	0.3679	0.1839	0.0613	0.0153	0.0031	0.0005	0.0001	0.0000	...

Recall that we calculated the expectation of a Poisson random variable as

$$\begin{aligned} E[X] &= \sum_{x=0}^{\infty} x\, P(X=x) \\ &= 0(0.3679)+1(0.3679)+2(0.1839)+\ldots \\ &= 1 \\ &= \lambda \end{aligned}$$

Now doing the same calculation for my cost random variable,

$$\begin{aligned} E[C] &= \sum_{costs} c\, P(C=c) \\ &= \sum_{x=0}^{\infty} 50\, x\, P(X=x) \\ &= 50 \sum_{x=0}^{\infty} x\, P(X=x) \\ &= 50\, E[X] \end{aligned}$$

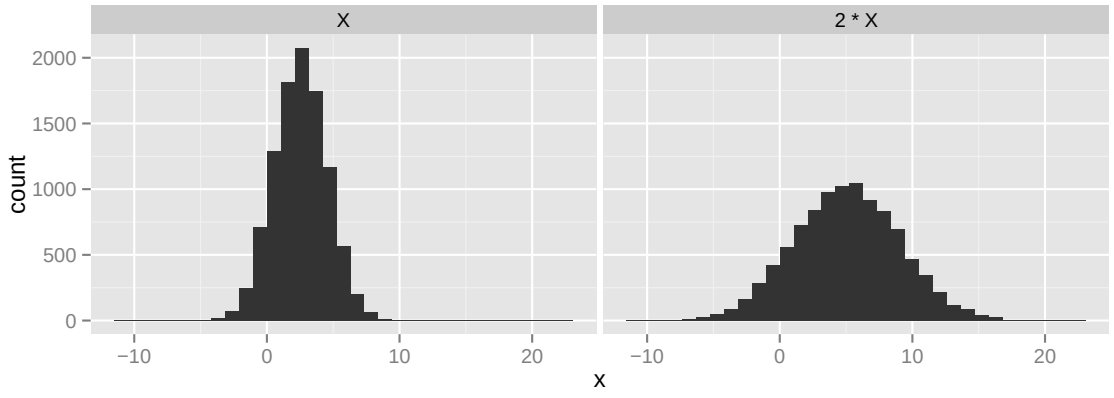
A similar calculation for variance can be done.

$$\begin{aligned} Var[C] &= \sum_{costs} (c - E[C])^2 P(C = c) \\ &= \sum_{x=0}^{\infty} (50x - 50E[X])^2 \ P(X = x) \\ &= \sum_{x=0}^{\infty} 50^2 (x - E[X])^2 \ P(X = x) \\ &= 50^2 \sum_{x=0}^{\infty} (x - E[X])^2 \ P(X = x) \\ &= 50^2 Var(X) \end{aligned}$$

Qualitative support: Recalling that we can think of the expectation and variance of a distribution as the sample mean and variance of an infinitely large sample, lets run some simulations of really large samples.

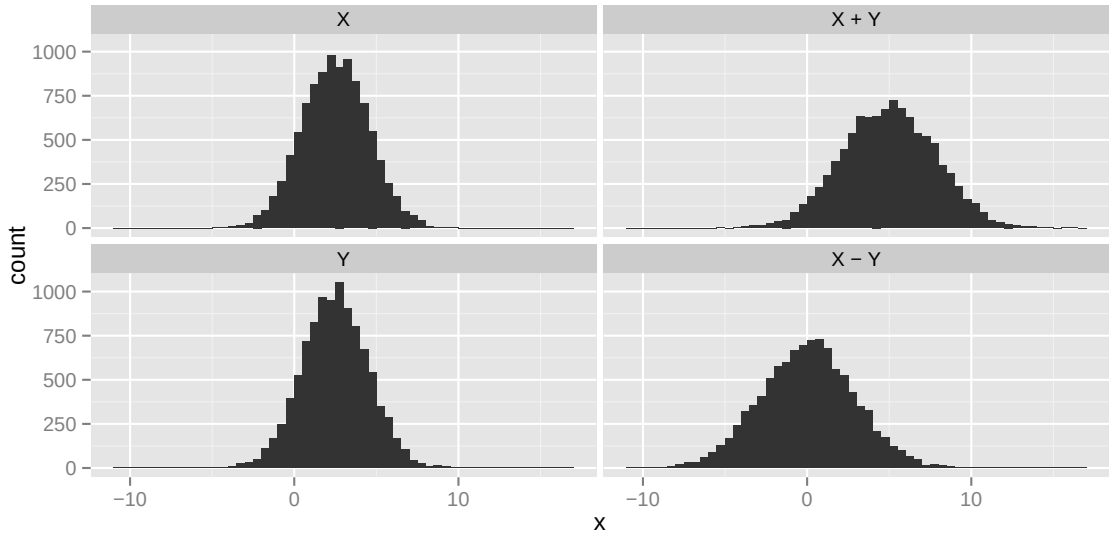
```
n <- 10000
x <- rnorm(n, mean=2.5, sd=2)
mean(x)

## [1] 2.518048

mean(2*x)

## [1] 5.036096

var(x)

## [1] 3.974362

var(2*x)

## [1] 15.89745
```

Why is this the case? Multiplying by a constant only rescales the distribution (a value of 9 is now 18, etc) and the mean is rescaled along with all the rest of the values. However since the variance is defined by the squared distances from the mean, the variance is multiplied by the constant *squared.*



```
x <- rnorm(n, mean=2.5, sd=2)
y <- rnorm(n, mean=2.5, sd=2)
mean( x+y )

## [1] 4.976937

mean( x-y )

## [1] 0.01956081

var( x+y )

## [1] 7.983336

var( x-y )

## [1] 7.948916
```

Adding two independent random variables will result in a new random variable whose expectation is the sum of the other two. However the standard deviations do not add together but the variances do. This is why statisticians prefer to work with variances instead of standard deviations.



Notice that the standard deviation of the sums is $\sqrt{8} \approx 2.8$ which is bigger than the two original distributions, but not twice as big.

These calculations can be done using *any* distributions and the results will still hold. Try it at home!

5.1 Mean and Variance of the Sample Mean

We have been talking about random variables drawn from a known distribution and being able to derive their expected values and variances. We now turn to the mean of a collection of random variables. Because sample values are random, any function of them is also random. So even though the act of calculating a mean is not a random process, the numbers that are feed into the algorithm *are random.* Thus the sample mean will change from sample to sample and we are interested in how it varies.

Using the rules we have just confirmed, it is easy to calculate the expectation and variance of the sample mean. Given a sample $X_1, X_2, \ldots, X_n$ of observations where all the observations

are independent of each other and all the observations have expectation $E[X_i] = \mu$ and variance $Var[X_i] = \sigma^2$ then

$$\begin{aligned} E\left[\bar{X}\right] &= E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \\ &= \frac{1}{n} E\left[\sum_{i=1}^{n} X_i\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} E\left[X_i\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} \mu \\ &= \frac{1}{n}\, n\mu \\ &= \mu \end{aligned}$$

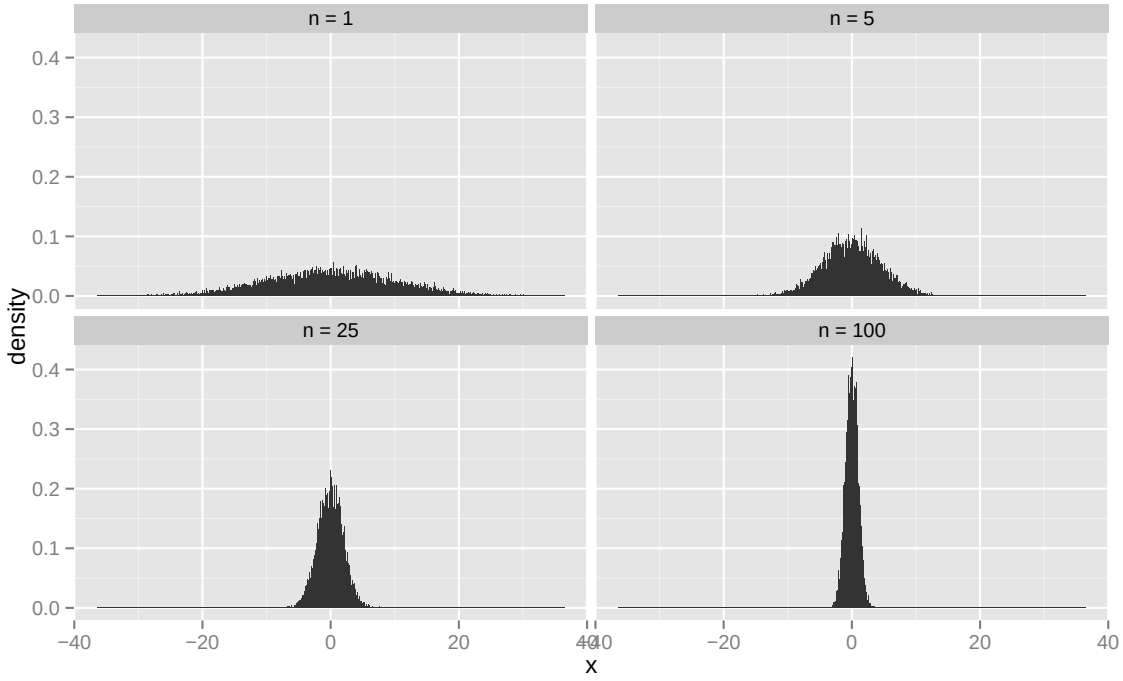
and

$$\begin{aligned} Var\left[\bar{X}\right] &= Var\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \\ &= \frac{1}{n^2} Var\left[\sum_{i=1}^{n} X_i\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n} Var\left[X_i\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 \\ &= \frac{1}{n^2}\, n\sigma^2 \\ &= \frac{\sigma^2}{n} \end{aligned}$$

Notice that the sample mean has the same expectation as the original distribution that the samples were pulled from, *but it has a smaller variance!* So the sample mean is an unbiased estimator of the population mean μ and the average distance of the sample mean to the population mean decreases as the sample size becomes larger. We can also explore this phenomena by simulation.

```
Num.Sims <- 10000
n <- 5
samples <- rep(0, Num.Sims)
for( i in 1:Num.Sims ){
  samples[i] <- mean( rnorm(n, mean=0, sd=10) )
}
mean(samples)

## [1] -0.005462296

var(samples)

## [1] 19.72994
```

We can look at how different sample sizes affect the variance by looking at $n = 1, 5, 25, 100$. Notice that $n = 1$ is just averaging 1 observation which is not averaging at all and is just the original random variable.



5.2 Distribution of $\bar{X}$ if the samples were drawn from a normal distribution

Looking at the graphs in the previous section, it should not be surprising that if $X_i \overset{iid}{\sim} N\left(\mu, \sigma^2\right)$ then $\bar{X}$ is also normally distributed with a mean and variance that were already calculated. That is

$$\bar{X} \sim N\left(\mu_{\bar{X}} = \mu,\ \sigma^2_{\bar{X}} = \frac{\sigma^2}{n}\right)$$

Notation: Since the expectations of X and $\bar{X}$ are the same, I could drop the subscript for the expectation of $\bar{X}$ but it is sometimes helpful to be precise. Since the variances are different we will use $\sigma_{\bar{X}}$ to denote the standard deviation of $\bar{X}$ and $\sigma^2_{\bar{X}}$ to denote variance of $\bar{X}$. If there is no subscript, we are referring to the population parameter of the distribution from which we taking the sample from.

Exercise: A researcher measures the wingspan of a captured Mountain Plover three times. Assume that each of these X_i measurements comes from a $N\left(\mu = 6 \text{ inches}, \sigma^2 = 1^2 \text{ inch}\right)$ distribution.

1. What is the probability that the first observation is greater than 7?

$$\begin{aligned} P\left(X \geq 7\right) &= P\left(\frac{X-\mu}{\sigma} \geq \frac{7-6}{1}\right) \\ &= P\left(Z \geq 1\right) \\ &= 0.1587 \end{aligned}$$

2. What is the distribution of the sample mean?

$$\bar{X} \sim N\left(\mu = 6,\ \frac{1^2}{3}\right)$$

3. What is the probability that the sample mean is greater than 7?

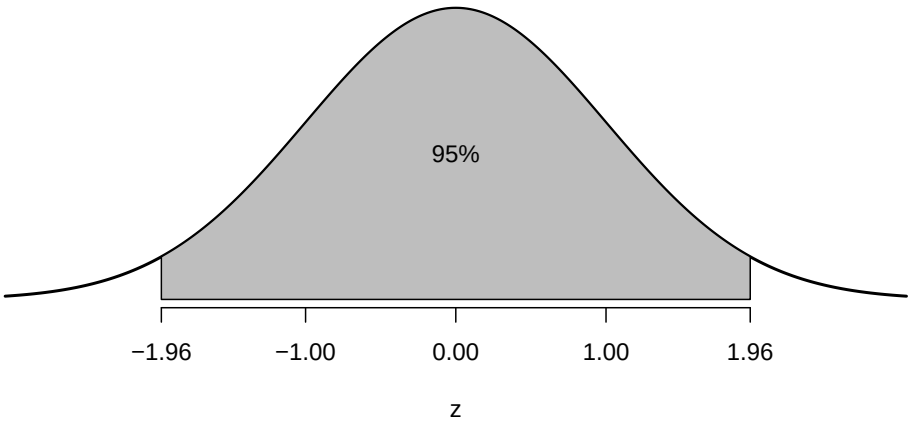
$$\begin{aligned} P\left(\bar{X} \geq 7\right) &= P\left(\frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} \geq \frac{7-6}{\sqrt{\frac{1}{3}}}\right) \\ &= P\left(Z \geq \sqrt{3}\right) \\ &= P\left(Z \geq 1.73\right) \\ &= 0.0418 \end{aligned}$$

Example: Suppose that the weight of an adult black bear is normally distributed with standard deviation $\sigma = 50$ pounds. How large a sample do I need to take to be 95% certain that my sample mean is within 10 pounds of the true mean μ?

So we want $\left|\bar{X} - \mu\right| \leq 10$ which we rewrite as

$$\begin{aligned} -10 &\leq \bar{X} - \mu_{\bar{X}} \leq 10 \\ \frac{-10}{\left(\frac{50}{\sqrt{n}}\right)} &\leq \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} \leq \frac{10}{\left(\frac{50}{\sqrt{n}}\right)} \\ \frac{-10}{\left(\frac{50}{\sqrt{n}}\right)} &\leq Z \leq \frac{10}{\left(\frac{50}{\sqrt{n}}\right)} \end{aligned}$$

Next we look in our standard normal table to find a z-value such that $P\left(-z \leq Z \leq z\right) = 0.95$ and that value is $z = 1.96$.



So all we need to do is solve the following equation for n

$$\begin{aligned} 1.96 &= \frac{10}{\frac{50}{\sqrt{n}}} \\ \frac{1.96}{10}\left(50\right) &= \sqrt{n} \\ 96 &\approx n \end{aligned}$$

Central Limit Theorem

> I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the "Law of Frequency of Error". The law would have been

> personified by the Greeks and deified, if they had known of it. It reigns with serenity and in complete self-effacement, amidst the wildest confusion. The huger the mob, and the greater the apparent anarchy, the more perfect is its sway. It is the supreme law of Unreason. Whenever a large sample of chaotic elements are taken in hand and marshaled in the order of their magnitude, an unsuspected and most beautiful form of regularity proves to have been latent all along. - Sir Francis Galton (1822-1911)

It was not surprising that the average of a number of normal random variables is also a normal random variable. Since the average of a number of binomial random variables cannot be binomial since the average could be something besides a 0 or 1 and the average of Poisson random variables does not have to be an integer.

The question arises, what can we say he distribution of the sample mean if the data comes from a non-normal distribution? The answer is quite a lot, but provided the original distribution has a non-infinite variance and we have a sufficient sample size.

[Central Limit Theorem]

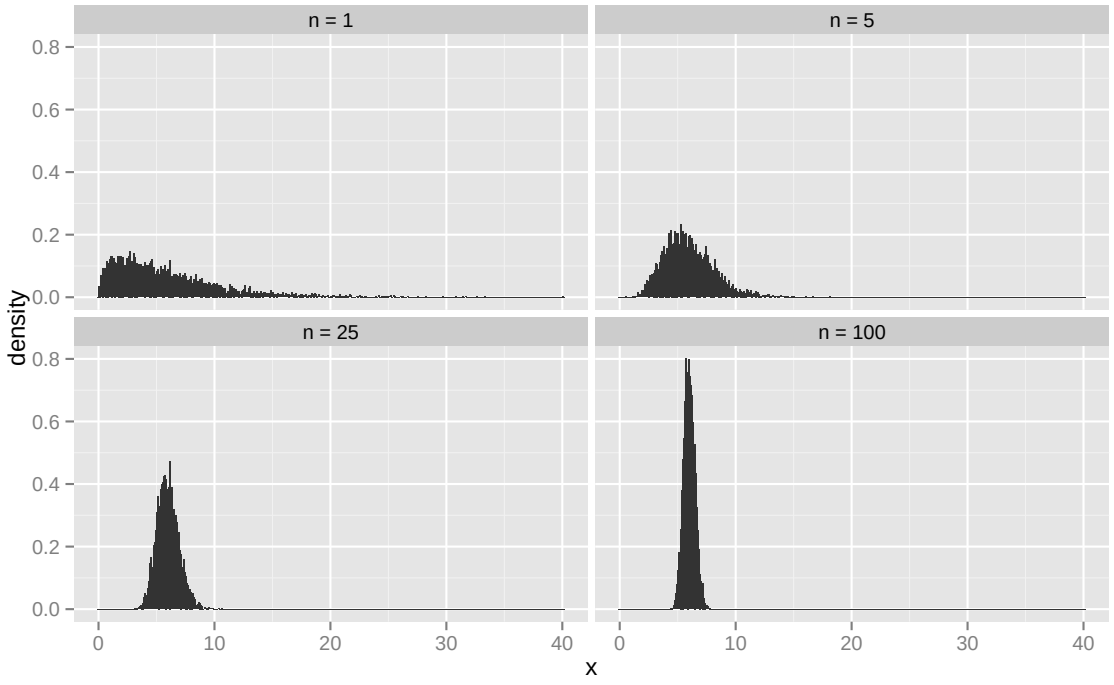
Let $X_1, \ldots X_n$ be independent observations collected from a distribution with expectation μ and variance σ^2. Then the distribution of $\bar{X}$ converges to a normal distribution with expectation μ and variance σ^2/n as $n \to \infty$.

In practice this means that if n is large (usually $n >$30 is sufficient), then

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

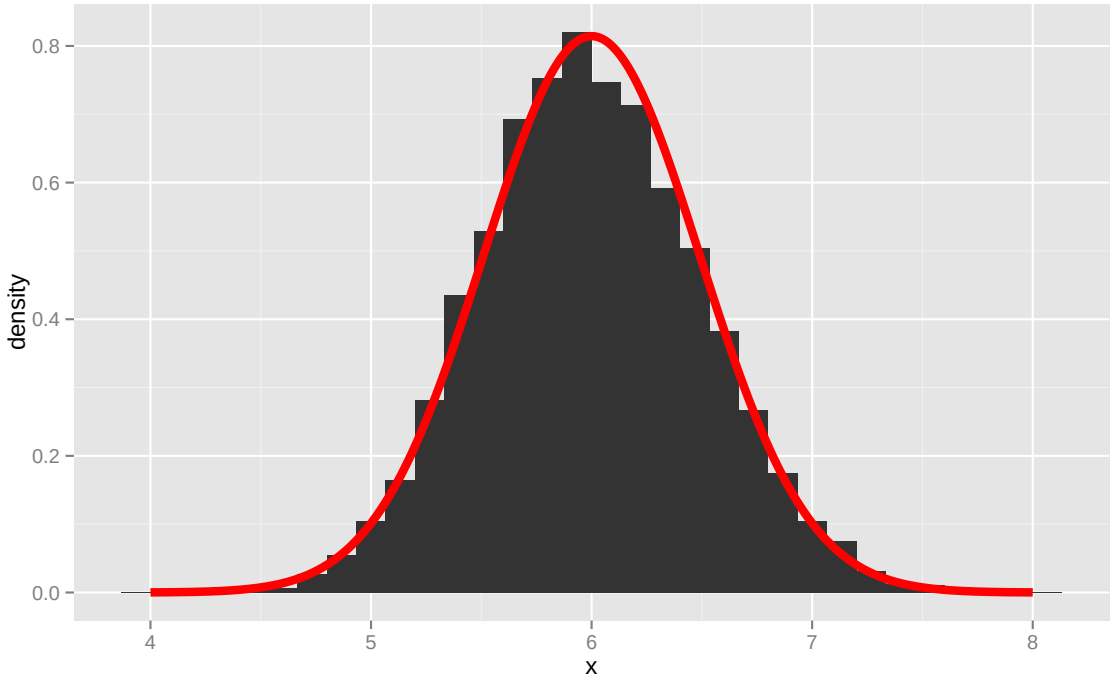
Evidence:

Again we turn to simulations. We take samples from a Gamma(1.5,4) distribution which has expectation $\mu = 1.5 * 4 = 6$ and variance $\sigma^2 = 1.5 * 4^2 = 24$ look at histograms of 2,000 sample means for each sample size, $n \in \{1, 5, 25, 100\}$.



By the time $n = 25$ the distribution of $\bar{X}$ is starting to take on the familiar mound shape of the normal. The case $n = 100$ should be approximately $N(6, 24/100)$ and to demonstrate that, we zoom in on just the $n = 100$ and super-impose the approximate normal density.

```
data.n100 <- samples[which(samples$n == 'n = 100'),]
x.grid=seq(4,8,length=1001)
normal.curve <- data.frame(x=x.grid,
                           y=dnorm(x.grid, mean=6, sd=sqrt(24/100)))
ggplot(data.n100, aes(x=x)) +
  geom_histogram(aes(y=..density..)) +
  geom_line( data=normal.curve, aes(y=y), size=2, color='red' )
```



So what does this mean?

1. Variables that are the sum or average of a bunch of other random variables will be close to normal. Example: human height is determined by genetics, pre-natal nutrition, food abundance during adolescence, etc. Similar reasoning explains why the normal distribution shows up surprisingly often in natural science.

2. With sufficient data, the sample mean will have a known distribution and we can proceed as if the sample mean came from a normal distribution.

Example: Suppose the waiting time from order to delivery at a fast-food restaurant is a exponential random variable with rate $\lambda = 1/2$ minutes and so the expected wait time is 2 minutes and the variance is 4 minutes. What is the approximate probability that we observe a sample of size $n = 40$ with a mean time greater than 2.5 minutes?

$$\begin{aligned} P\left(\bar{X} \geq 2.5\right) &= P\left(\frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} \geq \frac{2.5 - \mu_{\bar{X}}}{\sigma_{\bar{X}}}\right) \\ &\approx P\left(Z \geq \frac{2.5 - 2}{\frac{2}{\sqrt{40}}}\right) \\ &= P\left(Z \geq 1.58\right) \\ &= 0.0571 \end{aligned}$$

5.3 Summary

- If we have sampled n elements $Y_1, Y_2, \ldots, Y_n$ independently and $E(Y_i) = \mu$ and $Var(Y_i) = \sigma^2$ then we want to understand the distribution of the sample mean, that is we want to understand how the sample mean varies from sample to sample.
  - $E\left(\bar{Y}\right) = \mu$. That states that the distribution of the sample mean will centered at μ. We expect to sometimes take samples where the sample mean is higher than μ and sometimes less than μ, but the average underestimate is the same magnitude as the average overestimate.
  - $Var\left(\bar{Y}\right) = \frac{\sigma^2}{n}$. That states that as our sample size increases, we trust the sample mean to be close to μ. The larger the sample size, the greater our expectation that the $\bar{Y}$ will be close to μ.
- If $Y_1, Y_2, \ldots, Y_n$ were sampled from a $N\left(\mu, \sigma^2\right)$ distribution then $\bar{Y}$ is normally distributed.

$$\bar{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

- If $Y_1, Y_2, \ldots, Y_n$ were sampled from a distribution that is not normal but has mean μ and variance σ^2, and our sample size is large, then $\bar{Y}$ is *approximately* normally distributed.

$$\bar{Y} \dot{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$$

Chapter 6

Confidence Intervals and T-tests

We have seen that the sample mean is a random variable and will vary from sample to sample. So even though I will use my observed sample mean $\bar{x}$ as an estimate of μ I don't believe that μ is exactly equal to $\bar{x}$. In fact, for a continuous distribution, $P(\bar{X} = \mu)$=0. So what we really know is that sample means tend to land near μ and so I want to define a region near $\bar{x}$ that has a high likelihood of containing μ.

We know a great deal about the how the distribution of the sample mean is related to the distribution from which the data was sampled.

- If we have sampled n elements $Y_1, Y_2, \ldots, Y_n$ independently and $E(Y_i) = \mu$ and $Var(Y_i) = \sigma^2$ then we want to understand the distribution of the sample mean, that is we want to understand how the sample mean varies from sample to sample.
  - $E(\bar{Y}) = \mu$. That states that the distribution of the sample mean will centered at μ. We expect to sometimes take samples where the sample mean is higher than μ and sometimes less than μ, but the average underestimate is the same magnitude as the average overestimate.
  - $Var(\bar{Y}) = \frac{\sigma^2}{n}$. That states that as our sample size increases, we trust the sample mean to be close to μ. The larger the sample size, the greater our expectation that the $\bar{Y}$ will be close to μ.
- If $Y_1, Y_2, \ldots, Y_n$ were sampled from a $N(\mu, \sigma^2)$ distribution then $\bar{Y}$ is normally distributed.

$$\bar{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

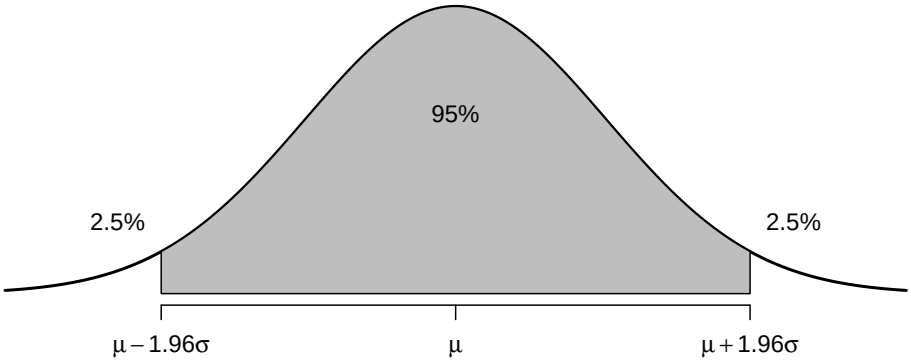
- If $Y_1, Y_2, \ldots, Y_n$ were sampled from a distribution that is not normal but has mean μ and variance σ^2, and our sample size is large, then $\bar{Y}$ is *approximately* normally distributed.

$$\bar{Y} \dot{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$$

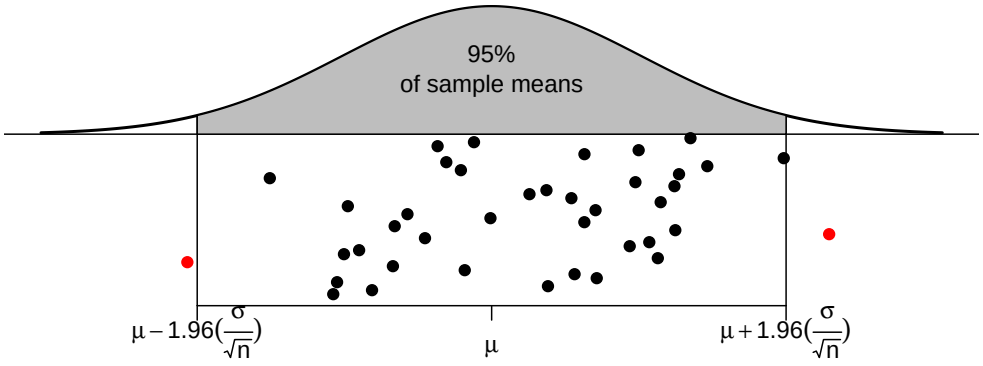
6.1 Confidence Intervals assuming σ is known

Recall that if $Z \sim N(\mu = 0,\ \sigma = 1)$ then we can find the middle $1-\alpha$ percent of the distribution by finding the $1-\alpha/2$ quantile. For example, I might be interested in the middle 95% of the distribution and thus $\alpha = 0.05$ the 97.5^{th} quantile of the standard normal distribution is $z_{0.975} = 1.96$.

Therefore I could find the middle 95% of any normal distribution by using $\mu \pm z_{0.975}\sigma$:

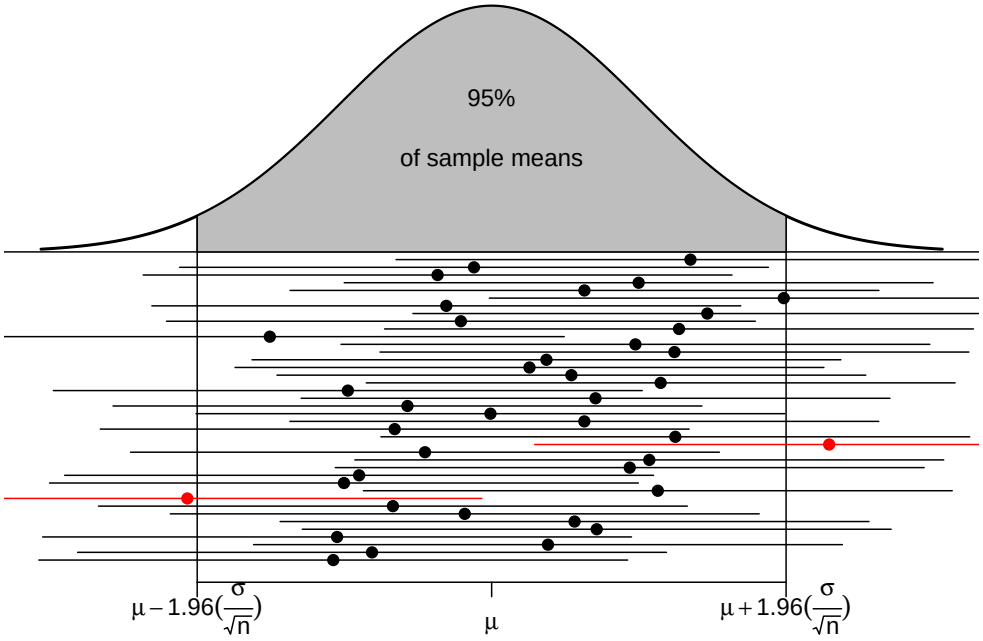


Generalizing this statement to its logical conclusion, the middle $1-\alpha$ of any normal distribution is found by $\mu \pm z_{1-\alpha/2}\ \sigma$.Therefore we know that 95% of sample means will lie in the interval $\mu \pm z_{0.975}\left(\frac{\sigma}{\sqrt{n}}\right)$. Or more generally we could say for $\alpha \in [0,1]$ we have $100\,(1-\alpha)\,\%$ of sample means lie between $\mu \pm z_{1-\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$. Notice that we use the $1-\frac{\alpha}{2}$ quantile of the standard normal because we wish to evenly divide the α probability equally to the left and right tails.



If 95% of the time the sample mean $\bar{x}$ lies within $z_{0.975}\,(\sigma/\sqrt{n})$ of the μ, then it is also true that μ lies within $z_{0.975}\,(\sigma/\sqrt{n})$ of 95% of those sample means.

Therefore 95% of the intervals $\bar{X} \pm z_{0.975}\,(\sigma/\sqrt{n})$ will contain μ.



In practice, I will only take one sample and therefore will only calculate one sample mean and one interval, but I want to recognize that the method I used to produce the interval (i.e. take a random sample, calculate the mean and then the interval) will result in intervals, but only 95% of those intervals will contain the mean μ. Therefore, I will refer to the interval as a 95% *confidence*

interval.

The general formula for a $100(1-\alpha)\%$ confidence interval is for μ is

$$\bar{x} \pm z_{1-\alpha/2}\left(\sigma/\sqrt{n}\right)$$

Notice in this formula I have denoted the sample mean with a lower case letter denoting that it is *a realization of a random variable.* Since I will only take one sample, I want to emphasize that the after the sample is taken and the data are fixed, the sample mean is not random.

For future reference we note z-values for some commonly used confidence levels:

Confidence Level	α	$z_{1-\alpha/2}$
90%	0.10	1.645
95%	0.05	1.96
99%	0.01	2.575

The interpretation of a confidence interval is that over repeated sampling, $100(1-\alpha)\%$ of the resulting intervals will contain the population mean μ but we don't know if the interval we have actually observed is one of the good intervals that contains the mean μ or not. Since this is quite the mouthful, we will say "we are $100\,(1-\alpha)\,\%$ confident that the observed interval contains the mean μ."

Example: Suppose a bottling facility has a machine that supposedly fills bottles to 300 milliliters (ml) and is known to have a standard deviation of $\sigma = 3$ ml. However, the machine occasionally gets out of calibration and might be consistently overfilling or under-filling bottles. To discover if the machine is calibrated correctly, we take a random sample of $n = 40$ bottles and observe the mean amount filled was $\bar{x} = 299$ ml. We calculate a 95% confidence interval (CI) to be

$$\begin{aligned} \bar{x} &\pm z_{1-\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \\ 299 &\pm 1.96\left(\frac{3}{\sqrt{40}}\right) \\ 299 &\pm 0.93 \end{aligned}$$

and conclude that we are 95% confident that the that the true mean fill amount is in $[298.07, 299.93]$ and that the machine has likely drifted off calibration.

6.1.1 Sample Size Selection

Often a researcher is in the position of asking how many sample observations are necessary to achieve a specific width of confidence interval. Let the *margin of error,* which we denote ME, be the half-width desired (so the confidence interval would be $\bar{x} \pm ME$). So given the desired confidence level, and if we know σ, then we can calculate the necessary number of samples to achieve a particular ME. To do this calculation, we must also have some estimate of the population standard deviation σ.

$$ME = z_{1-\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

and therefore

$$n \approx \left[z_{1-\alpha/2}\left(\frac{\sigma}{ME}\right)\right]^2$$

Notice that because $n \propto \left[\frac{1}{ME}\right]^2$ then if we want a margin of error that is twice as precise (i.e. the CI is half as wide) then we need to quadruple our sample size! Second, this result requires having some knowledge of σ. We could acquire an estimate through: 1) a literature search, 2) a pilot study, or 3) expert opinion.

Example. A researcher is interested in estimating the mean weight of an adult elk in Yellowstone's northern herd after the winter and wants to obtain a 90% confidence interval with a half-width $E = 10$ pounds. Using prior collection data from the fall harvest (road side checks by game wardens), the researcher believes that $\sigma = 60$ lbs is a reasonable standard deviation number to use.

$$\begin{aligned} n &\approx \left[z_{0.95}\left(\frac{\sigma}{E}\right)\right]^2 \\ &= \left[1.645\left(\frac{60}{10}\right)\right]^2 \\ &= 97.41 \end{aligned}$$

6.2 Confidence interval for μ assuming σ is unknown

6.2.1 t-distributions

It is unrealistic to expect that we know the population variance σ^2 but do not know the population mean μ. So in calculations that involve σ, we want to use the sample standard deviation S instead.

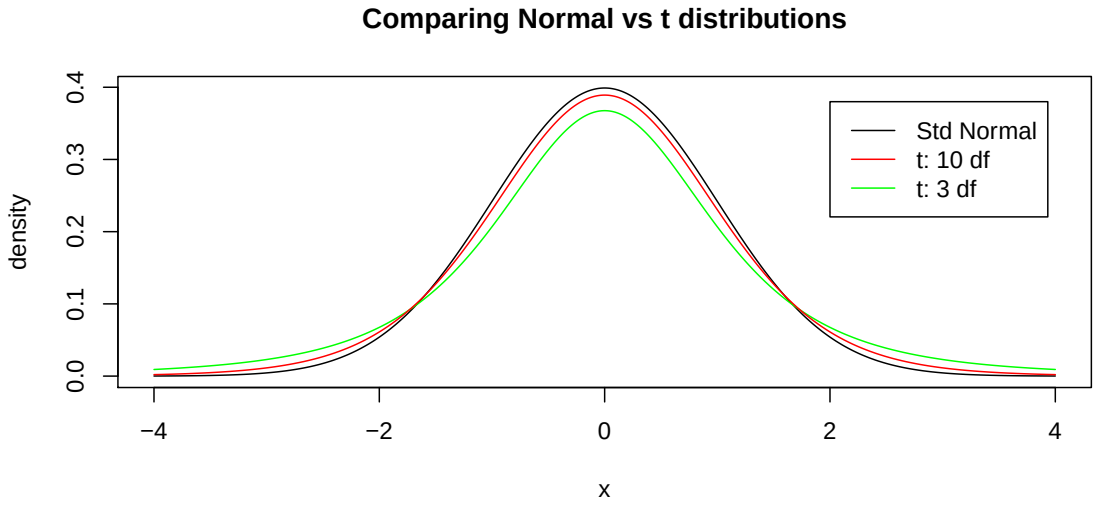
Our previous results about confidence intervals assumed that $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ (or is approximately so) and therefore

$$\frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0, 1).$$

I want to just replace σ^2 with S^2 but the sample variance S^2 is also a random variable and incorporating it into the standardization function might affect the distribution.

$$\frac{\bar{X} - \mu}{\sqrt{\frac{S^2}{n}}} \sim ???$$

Unfortunately this substitution of S^2 for σ^2 comes with a cost and this quantity is not normally distributed. Instead it has a t-distribution with $n - 1$ degrees of freedom. However as the sample size increases and S^2 becomes a more reliable estimator of σ^2, this penalty should become smaller.



The t-distribution is named after William Gosset who worked at Guinness Brewing and did work with small sample sizes in both the brewery and at the farms that supplied the barley. Because Guinness prevented its employees from publishing any of their work, he published under the pseudonym *Student*.

Notice that as the sample size increases, the t-distribution gets closer and closer to the normal distribution. From here on out, we will use the following standardization formula:

$$\frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$$

and emphasize that this formula is valid if the sample observations came from a population with a normal distribution or if the sample size is large enough for the Central Limit Theorem to imply that $\bar{X}$ is approximately normally distributed.

Substituting the sample standard deviation into the confidence interval formula, we also substitute a t-quantile for the standard normal quantile. We will denote $t_{n-1}^{1-\alpha/2}$ as the $1-\alpha/2$ quantile of a t-distribution with $n-1$ degrees of freedom. Therefore we will use the following formula for the calculation of $100\,(1-\alpha)\,\%$ confidence intervals for the mean μ:

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

Notation: We will be calculating confidence intervals for the rest of the course and it is useful to recognize the skeleton of a confidence interval formula. The basic form is always the same

$$Estimate \ \pm \ t_{df}^{1-\alpha/2} \ Standard\,Error\,(\,Estimate\,)$$

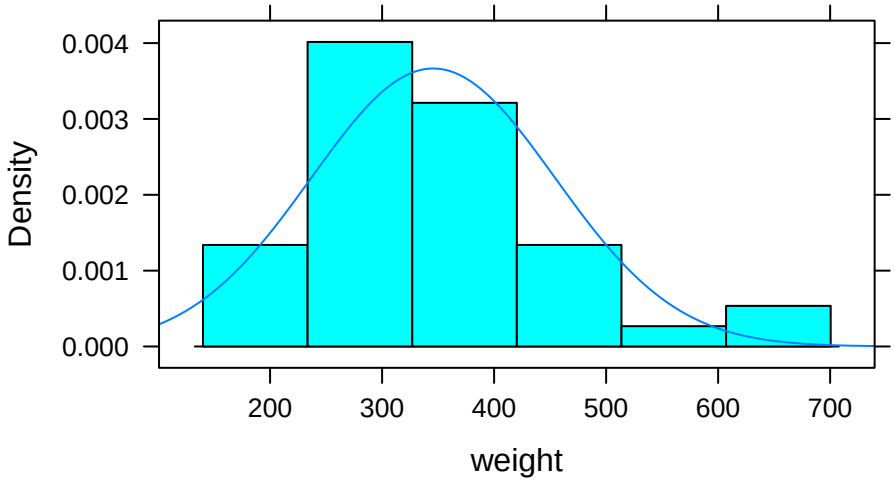
In our current problem, $\bar{x}$ is our estimate of μ and the estimated standard deviation (which is commonly called the *standard error*) is $s/\sqrt{n}$ and the appropriate degrees of freedom are $df = n-1$.

Example. Suppose we are interested in calculating a 95% confidence interval for the mean weight of adult black bears. We collect a random sample of 40 individuals (large enough for the CLT to kick in) and observe the following data:

```
library(mosaic)
bears <- data.frame(weight =
   c(306, 446, 276, 235, 295, 302, 374, 339, 624, 266,
     497, 384,  429, 497, 224, 157, 248, 349, 388, 391,
     266, 230, 621, 314, 344,  413, 267, 380, 225, 418,
     257, 466, 230, 548, 277, 354, 271, 369,  275, 272))
mean( ~ weight, data=bears)

## [1] 345.6

sd(   ~ weight, data=bears)

## [1] 108.8527
```

Notice that the data do not appear to come from a normal distribution, but a slightly heavier right tail.

```
histogram( ~ weight, data=bears)
plotDist( 'norm',                              # plot a normal
          mean = mean(~weight,data=bears),     # with same mean as my data
          sd=sd(~weight,data=bears),           # with same sd as my data
          add=TRUE )                           # add it to the current histogram
```



The observed sample mean is $\bar{x} = 345.6$ pounds and a sample standard deviation $s = 108.9$ pounds. Because we want a 95% confidence interval $\alpha = 0.05$. Using the t-tables in your book or the following R code

```
qt(.975, df=39)

## [1] 2.022691
```

we find that $t_{n-1}^{1-\alpha/2} = 2.02$. Therefore the 95% confidence interval is

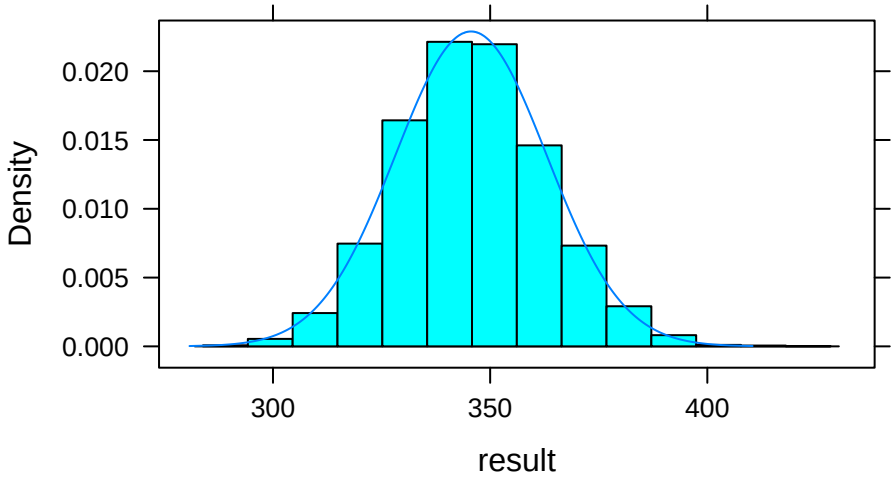
$$\begin{aligned} \bar{x} &\pm t_{n-1}^{1-\alpha/2}\left(\frac{s}{\sqrt{n}}\right) \\ 345.6 &\pm 2.02\left(\frac{108.9}{\sqrt{40}}\right) \\ 345.6 &\pm 34.8 \end{aligned}$$

or $(310.8, 380.4)$ which is interpreted as "We are 95% confident that the true mean μ is in this interval" which is shorthand for "The process that resulted in this interval (taking a random sample, and then calculating an interval using the algorithm presented) will result in intervals such that 95% of them contain the mean μ, but we cannot know of this particular interval is one of the good ones or not."

We can wonder how well this interval[1] matches up to the sampling distribution we would have gotten from a bootstrap estimation of the sampling distribution of $\bar{x}$. In this case, where the sample size n is relatively large, the Central Limit Theorem is certainly working and the distribution of the sample mean certainly looks fairly normal.

[1] I'm going to use the normal distribution instead of the t because it is harder to use the `mosaic` package to easily plot a scaled and shifted t distribution.

```
SampDist <- do(10000) * mean( ~ weight, data=resample(bears))
histogram(~result, data=SampDist)
plotDist('norm',                                    # make a normal distribution
         mean = mean(~weight,data=bears),           # with same mean as our data
         sd   = sd(~weight,data=bears) / sqrt( nrow(bears)-1 ),  # sd as theory suggests
         add=TRUE )                                 # add to current plot
```



Grabbing the appropriate quantiles from the bootstrap estimate of the sampling distribution, we see that the bootstrap 95% confidence interval matches up will with the confidence interval we obtained from asymptotic theory.

```
quantile( SampDist$result, probs=c(0.025, 0.975) )

##     2.5%    97.5%
## 312.8487 381.0756
```

Example. Assume that the percent of alcohol in casks of whisky is normally distributed. From the last batch of casks produced, the brewer samples $n = 5$ casks and wants to calculate a 90% confidence interval for the mean percent alcohol in the latest batch produced. The sample mean was $\bar{x} = 55$ percent and the sample standard deviation was $s = 4$ percent.

$$\begin{aligned} \bar{x} &\pm t_{n-1}^{1-\alpha/2}\left(\frac{s}{\sqrt{n}}\right) \\ 55 &\pm 2.13\left(\frac{4}{\sqrt{5}}\right) \\ 55 &\pm 3.8 \end{aligned}$$

Question: If we wanted a 95% confidence interval, would it have been wider or narrower?
Question: If this interval is too wide to be useful, what could we do to make it smaller?

6.2.2 Simulation study comparing asymptotic vs bootstrap confidence intervals

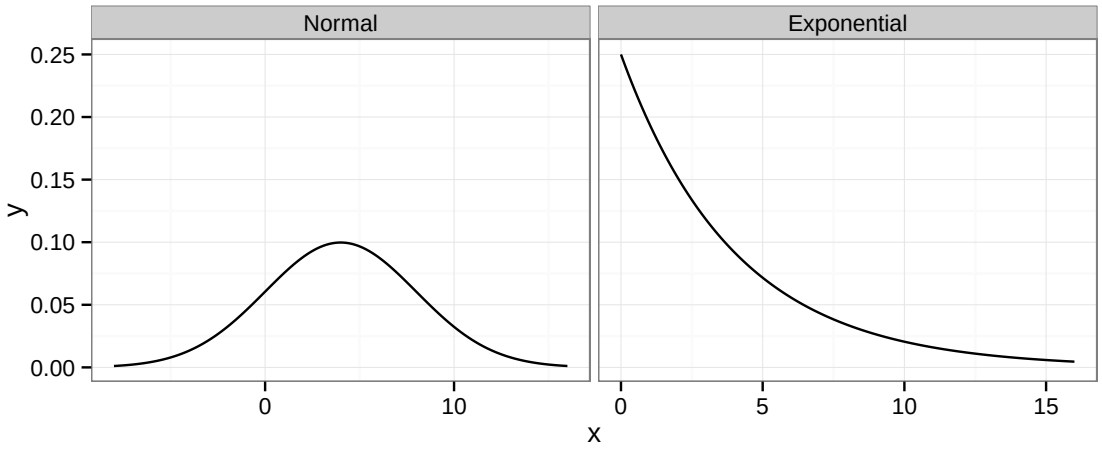
We would like to understand the benefits and limitations of the two methods to create confidence intervals. Both methods of creating confidence intervals rely on some critical assumptions.

1. Asymptotic intervals
   (a) One of the following applies
      i. The original population we sampled from was normally distributed and therefore the sample mean has a normal distribution.
      ii. The sample size is large enough that the Central Limit Theorem applies and the sample mean has an approximate normal distribution.
   (b) The observed sample data is representative of the population.
2. Bootstrap intervals (calculated using the percentile method, i.e. middle $(1-\alpha)*100\%$ of the distribution)
   (a) The observed sample data is representative of the population.
   (b) The sampling distribution is symmetric[2].

In particular, the bootstrap more heavily relies on the assumption that the data is representative, while the asymptotic interval assumes that the data is also "nicely behaved".

It is useful to investigate how well these two methods perform when sampling from different distributions. We will consider two distributions for the population, both of which have a mean $\mu = 4$ and a standard deviation of 4.

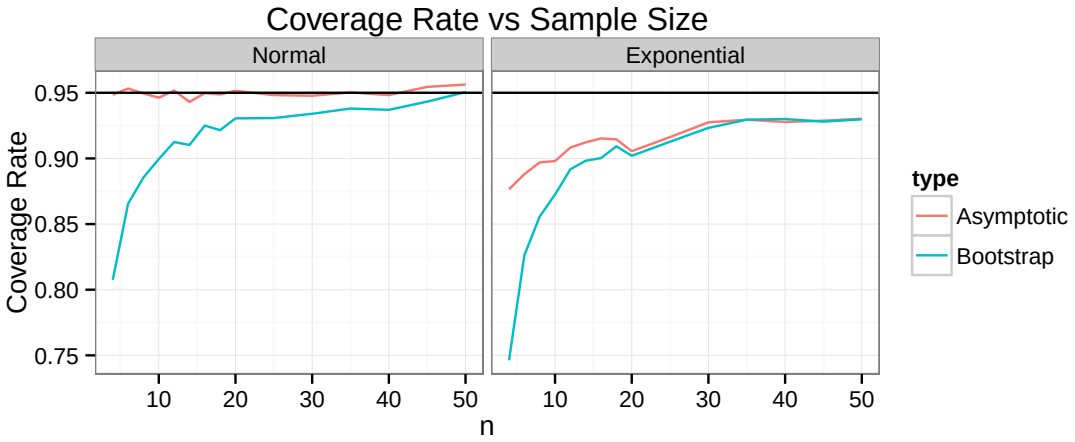
1. Normal($\mu = 4$, $\sigma = 4$) - Under this scenario, the Asymptotic interval should be optimal
2. Exponential - We'll use an exponential distribution with mean 4 and standard deviation of 4. This distribution has a longer tail than the normal distribution, but isn't too extreme.



For this simulation, we will generate a sample of size n from one of the above distributions and then create a 95% confidence interval for the mean $\mu = 4$ using both the Asymptotic method and the Bootstrap method (using 1000 bootstrap replicates). We will repeat this process 4000 times for each sample size $n \in \{4, 6, \ldots, 20, 25, 30, \ldots, 50\}$.

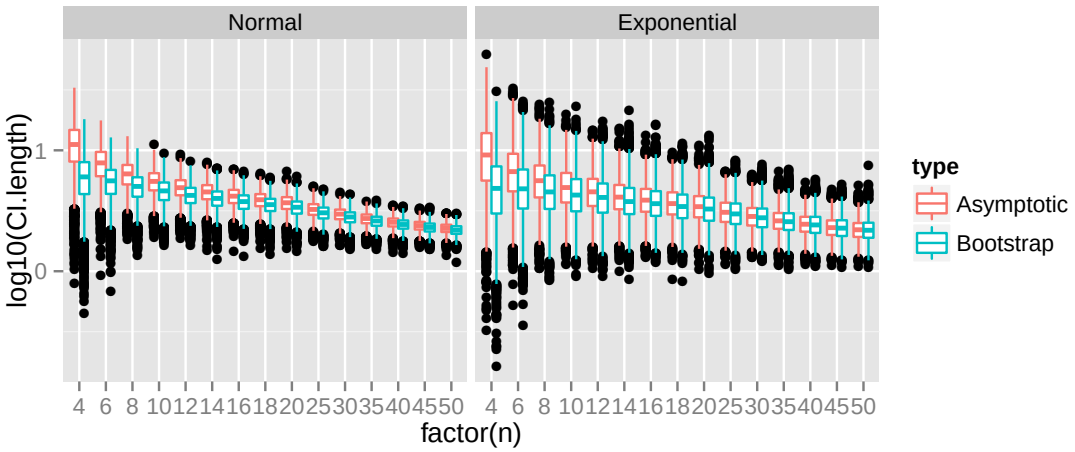
The first figure that we'll examine is the proportion of sample who's resulting 95% CI actually contained the true value of $\mu = 4$. Ideally, this proportion should be exactly 95%; if the method captures μ less than that, our method is being "too liberal" and the intervals are not long enough, if the method captures μ in more than the desired 95% of samples, then our confidence intervals are too wide and we could shorten them up and still be accurate 95% CIs.

[2]The symmetric assumption can be relaxed by choosing a different method for creating confidence intervals. In particular, the "BCa" method handles non-symmetric bootstrap distributions quite well. In the following simulation, I used this better method.



When the data was normally distributed, then regardless of sample size, our 95% CI calculated using the Asymptotic results is doing quite well and in about 95% of the simulations the Asymptotic CI contains μ. When the data came from the exponential distribution, the Asymptotic CI is liberal and is capturing the mean μ in too few of the simulations. In contrast, the Bootstrap 95% CI is failing to capture μ often enough in both situations, but in the exponential distribution case, the two methods appear to capture μ with equal probability at moderate sample sizes.

Next we consider the lengths of the confidence intervals. If both methods capture the mean μ at the same rate, then I prefer the method that generally produces shorter CIs.



This graph shows that, on average, the Bootstrap CIs are shorter than the corresponding Asymptotic CIs, even in the case where the data was drawn from the exponential distribution and the coverage rates were similar.

The major points of this simulation are:

1. If the normality assumption is met, the Asymptotic CI is better.

2. If the normality assumption is not met, both methods are liberal (e.g. are actually 92% CIs when they advertise as 95% CIs), but the bootstrap method produces *slightly* shorter intervals.

3. The Asymptotic approach is relatively robust to departures from normality of the population (due to the Central Limit Theorem coercing the sample mean to be normally distributed).

Because the bootstrap methodology is very consistent across a wide variety of situations, it is applicable to a wide range of problems. In particular, it is useful when no good asymptotic result

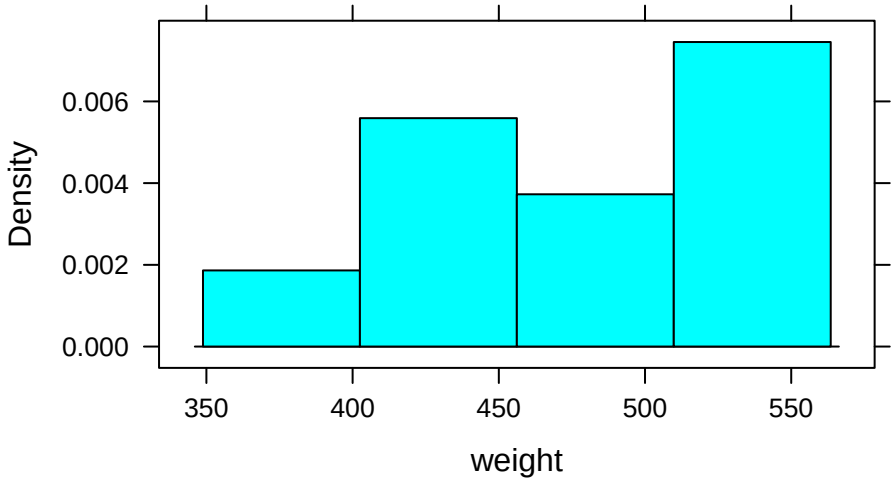
is known, or when the assumptions of the asymptotic test are strongly violated. In this course, we will continue to develop asymptotic results, but also show the corresponding bootstrap analysis.

6.3 Hypothesis Testing

Science is done by observing how the world works, making a conjecture (or hypothesis) about the mechanism and then performing experiments to see if real data agrees or disagrees with the proposed hypothesis.

Example. Suppose a rancher in Texas (my brother-in-law Bryan) wants to buy cattle from another rancher. This rancher claims that the average weight of his steers is 500 pounds. My brother-in-law likes the cows and buys 10. A few days later he starts looking at the cows and begins to wonder if the average really is 500 pounds. He weighs his 10 cows and the sample mean is $\bar{x} = 475$ and the sample standard deviation is $s = 50$. Below are the data

```
cows <- data.frame(
  weight = c(553, 466, 451, 421, 523,
             517, 451, 510, 392, 466) )
histogram( ~weight, data=cows )
```



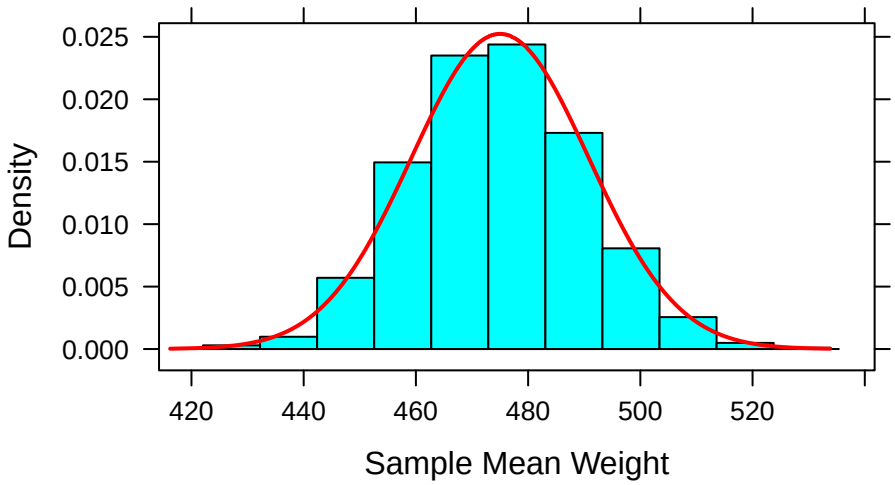
There are two possibilities. Either Bryan was just unlucky the random selection of his 10 cows from the heard, or the true average weight within the herd is less than 500.

$$\begin{aligned} H_0 : \mu &= 500 \\ H_a : \mu &< 500 \end{aligned}$$

Assuming[3] the true mean is 500, how likely is it to get a sample mean of 475? What should the sampling distribution of $\bar{x}$ look like? Because we know that distribution of $\bar{x}$ is normally distributed, the bootstrap sampling distribution is a $\bar{X} \sim N\left(\mu = 475, \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\right)$.

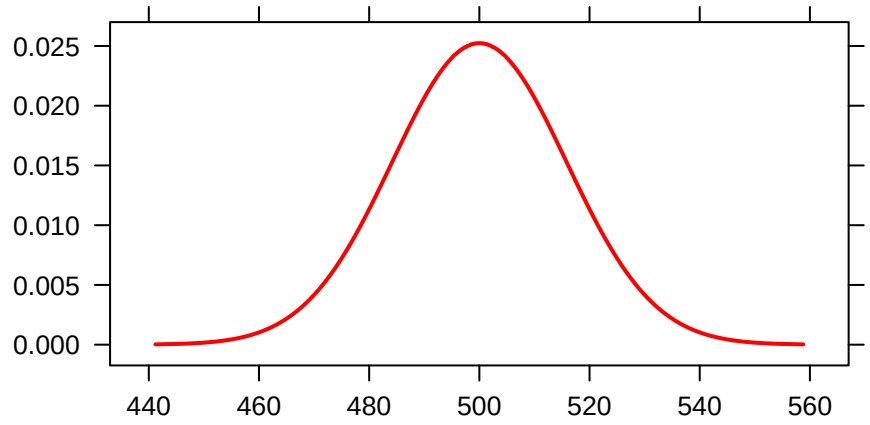
[3] For this calculation we'll assume the weight of a steer is normally distributed $N(\mu, \sigma)$, and therefore $\bar{X}$ is normally distributed $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.

```
SampDist <- do(1000)*mean(~weight, data=resample(cows))
histogram( ~result, data=SampDist, xlab='Sample Mean Weight' )
# cannot easily plot the appropriate shifted and scaled t-distribution, so do a normal
plotDist('norm',
         mean=mean(~weight,data=cows),
         sd=sd(~weight,data=cows)/sqrt(nrow(cows)),
         col='red', lwd=2, add=TRUE)
```



But this distribution is centered at the sample mean $\bar{x} = 475$. If the null hypothesis is true, then this distribution should be centered at $\mu = 500$. So we should *shift* the whole distribution to be centered at the hypothesized value of $\mu = 500$.

```
# cannot easily plot the appropriate t-distribution, so do a normal
plotDist('norm',
         mean=500,
         sd=sd(~weight,data=cows)/sqrt(nrow(cows)),
         col='red', lwd=2,
         main='Estimated Sampling Distribution of Xbar')
```



Finally we we can calculate how far into the tail our observed sample mean $\bar{x} = 475$ is by measuring the area of the distribution that is farther into the tail than the observed value.

$$\begin{aligned} P\left(\bar{X} \leq 475\right) &= P\left(\frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \leq \frac{475 - 500}{\frac{50}{\sqrt{10}}}\right) \\ &= P\left(T_9 \leq -1.58\right) \\ &= 0.074 \end{aligned}$$

We see that the observed $\bar{X}$ is in the tail of the distribution and tends to not support H_0.

P-value is the probability of seeing the observed data *or something more extreme* given the null hypothesis is true. By "something more extreme", we mean samples that would be more evidence for the alternative hypothesis.

$$\text{p-value} = P(T_9 < -1.58) = 0.074$$

The above value is the actual value calculated using R

```
pt(-1.58, df=9)

## [1] 0.07428219
```

but using the table in your book, the most precise thing you would be able to say is

$$0.05 \leq \text{p-value} \leq 0.10$$

So there is a small chance that my brother-in-law just got unlucky with his ten cows. While the data isn't entirely supportive of H_0, we don't have strong enough data to out right reject H_0. So we will say that *we fail to reject* H_0. Notice that we aren't saying that we accept the null hypothesis, only that there is insufficient evidence to call-out the neighbor as a liar.

6.3.1 Writing Hypotheses

Perhaps the hardest part about conducting a hypothesis test is figuring out what the null and alternative hypothesis should be. The null hypothesis is a statement about a population parameter.

$$H_0: \text{ population parameter} = \text{hypothesized value}$$

and the alternative will be one of

$$\begin{aligned} H_a: &\quad \text{population parameter} < \text{hypothesized value} \\ H_a: &\quad \text{population parameter} > \text{hypothesized value} \\ H_a: &\quad \text{population parameter} \neq \text{hypothesized value} \end{aligned}$$

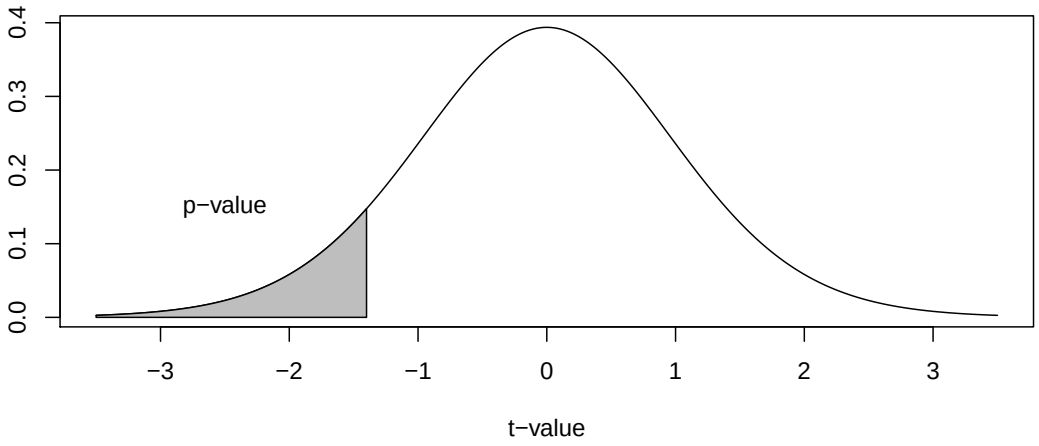
The hard part is figuring which of the possible alternatives we should examine. The alternative hypothesis is what the researcher believes is true. By showing that the complement of H_a (that is H_0) can not be true, we support the alternative which we believe to be true.

H_0 is often a statement of no effect, or no difference between the claimed and observed.

Example A light bulb company advertises that their bulbs last for 1000 hours. Consumers will be unhappy if the bulbs last less time, but will not mind if the bulbs last longer. Therefore we would test

$$\begin{aligned} H_0: \mu &= 1000 \\ H_a: \mu &< 1000 \end{aligned}$$

Suppose we perform an experiment and get a test statistics of $t_{19} = -1.4$. Then the p-value would be



and we calculate

$$p - value = P\left(T_{19} < -1.4\right) = 0.0888$$

using R

```
pt(-1.4, df=19)
```

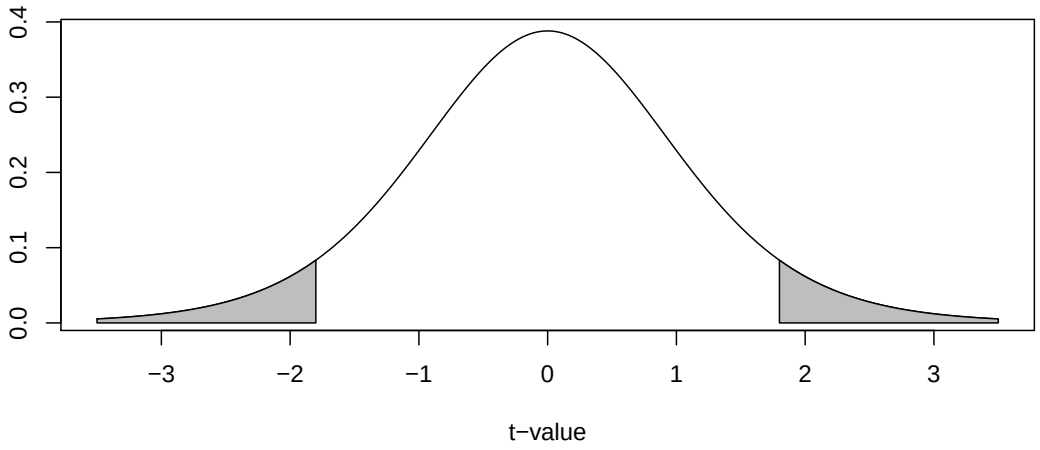
```
## [1] 0.08881538
```

Example A computer company is buying resistors from another company. The resistors are supposed to have a resistance of 2 Ohms and too much or too little resistance is bad. Here we would

be testing

$$\begin{aligned} H_0: \ \mu &= 2 \\ H_a: \ \mu &\neq 2 \end{aligned}$$

Suppose we perform a test of a random sample of resistors and obtain a test statistics of $t_9 = 1.8$. Because the p-value is "the probability of your data or something more extreme" and in this case more extreme implies extreme values in both tails then



and we calculate

$$p - value = P\left(|T_9| > 1.8\right) = 2P\left(T_9 < -1.8\right) = 2\left(0.0527\right) = 0.105$$

using the R commands

```
2 * pt(-1.8, df=9)

## [1] 0.1053907
```

Why should hypotheses use μ and not $\bar{x}$?

There is no need to make a statistical test of the form

$$\begin{aligned} H_0: \ \bar{x} &= 3 \\ H_a: \ \bar{x} &\neq 3 \end{aligned}$$

because we *know the value of* $\bar{x}$*;* we calculated the value there is no uncertainty to what it is. However I want to use the sample mean $\bar{x}$ as an estimate of the population mean μ and because I don't know what μ is but know that it should be somewhere near $\bar{x}$, my hypothesis test is a question about μ and if it is near the value stated in the null hypothesis.

Hypotheses are *always* statements about population parameters such as μ or σ and *never* about sample statistic values such as $\bar{x}$ or s.

Examples

1. A potato chip manufacturer advertises that it sells 16 ounces of chips per bag. A consumer advocacy group wants to test this claim. They take a sample of $n = 18$ bags and carefully weights the contents of each bag and calculate a sample mean $\bar{x} = 15.8$ oz and a sample standard deviation of $s = 0.2$.

(a) State an appropriate null and alternative hypothesis.

$$\begin{aligned} H_0 : \mu &= 16 \text{ oz} \\ H_a : \mu &< 16 \text{ oz} \end{aligned}$$

(b) Calculate an appropriate test statistic given the sample data.

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{15.8 - 16}{\frac{.2}{\sqrt{18}}} = -4.24$$

(c) Calculate the p-value.

$$\text{p-value } = P(T_{17} < -4.24) = 0.000276$$

(d) Do you reject or fail to reject the null hypothesis at the $\alpha = 0.05$ level?
Since the p-value is less than $\alpha = 0.05$ we will reject the null hypothesis.

(e) State your conclusion in terms of the problem.
There is statistically significant evidence to conclude that the mean weight of chips is less than 16 oz.

2. A pharmaceutical company has developed an improved pain reliever and believes that it acts faster than the leading brand. It is well known that the leading brand takes 25 minutes to act. They perform an experiment on 16 people with pain and record the time until the patient notices pain relief. The sample mean is $\bar{x} = 23$ minutes, and the sample standard deviation was $s = 10$ minutes.

(a) State an appropriate null and alternative hypothesis.

$$\begin{aligned} H_0 : \mu &= 25 \text{ minutes} \\ H_a : \mu &< 25 \text{ minutes} \end{aligned}$$

(b) Calculate an appropriate test statistic given the sample data.

$$t_{15} = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{23 - 25}{\frac{10}{\sqrt{16}}} = -0.8$$

(c) Calculate the p-value.

$$\text{p-value } = P(T_{15} < -0.8) = 0.218$$

(d) Do you reject or fail to reject the null hypothesis at the $\alpha = .10$ level?
Since the p-value is larger than my α-level, I will fail to reject the null hypothesis.

(e) State your conclusion in terms of the problem.
These data do not provide statistically significant evidence to conclude that this new pain reliever acts faster than the leading brand.

3. Consider the case of SAT test preparation course. They claim that their students perform better than the national average of 1019. We wish to perform a test to discover whether or not that is true.

$$\begin{aligned} H_0 : \mu &= 1019 \\ H_a : \mu &> 1019 \end{aligned}$$

They take a sample of size $n = 10$ and the sample mean is $\bar{x} = 1020$, with a sample standard deviation $s = 50$. The test statistic is

$$t_9 = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{1}{\frac{50}{\sqrt{10}}} = .06$$

So the p-value is

$$\text{p-value } = P(T_9 > .06) \approx 0.5$$

So we fail to reject the null hypothesis. However, what if they had performed this experiment with $n = 20000$ students and gotten the same results?

$$t_{19999} = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{1}{\frac{50}{\sqrt{20000}}} = 2.83$$

and thus

$$\text{p-value } = P(T_{19999} > 2.83) = 0.0023$$

At $\alpha = .05$, we will reject the null hypothesis and conclude that there is statistically significant evidence that the students who take the course perform better than the national average.

So what just happened and what does "statistically significant" mean? It appears that there is *very* slight difference between the students who take the course versus those that don't. With a small sample size we can not detect that difference, but by taking a large sample size, I can detect the difference of even 1 SAT point. So here I would say that there is a statistical difference between the students who take the course versus those that don't because given such a large sample, we are *very* unlikely to see a sample mean of $\bar{x} = 1020$ if the true mean is $\mu = 1020$. So statistically significant really means "unlikely to occur by random chance".

But is there a practical difference in 1 SAT point? Not really. Since SAT scores are measured in multiple of 5 (you can score 1015, or 1020, but not 1019), there isn't any practical value of raising a students score by 1 point. By taking a sample so large, I have been able to detect a completely worthless difference.

Thus we have an example of a statistically significant difference, but it is not a practical difference.

6.3.2 Calculating p-values

Students often get confused by looking up probabilities in tables and don't know which tail of the distribution supports the alternative hypothesis. This is further exacerbated by tables sometimes giving area to the left, sometimes area to the right, and R only giving area to the left. In general, your best approach to calculating p-values correctly is to draw the picture of the distribution of the test statistic (usually a t-distribution) and decide which tail(s) supports the alternative and figuring out the area farther out in the tail(s) than your test statistic. However, since some students need a more algorithmic set of instructions, the following will work:

1. If your alternative has a $\neq$ sign
   (a) Look up the value of your test statistic in whatever table you are going to use and get some probability... which I'll call p^*.
   (b) Is p^* greater than 0.5? If so, you just looked up the area in the wrong tail. To fix your error, subtract from one... that is $p^* \leftarrow 1 - p^*$
   (c) Because this is a two sided test, multiply p^* by two and that is your p-value. p-value $= 2\,(p^*)$
   (d) A p-value is a probability and therefore must be in the range$[0, 1]$. If what you've calculated is outside that range, you've made a mistake.
2. If your alternative is $<$ (or $>$) then the p-value is the area to the left (to the right for the greater than case) of your test statistic.
   (a) Look up the value of your test statistic in whatever table you are using and get the probability... which again I'll call p^*

(b) If p^* is greater than 0.5, you have most likely screwed up and looked up the area for the wrong tail.[4] Most of the time you'll subtract from one $p^* = 1 - p^*$.

(c) After possibly adjusting for looking up the wrong tail, your p-value is p^* with no multiplication necessary.

6.3.3 Calculating p-values vs cutoff values

We have been calculating p-values and then comparing those values to the desired alpha level. It is possible, however, to use the alpha level to back-calculate a cutoff level for the test statistic, or even original sample mean. Often these cutoff values are referred to as *critical values.* Neither approach is wrong, but is generally a matter of preference, although knowing both techniques can be useful.

Example. We return to the pharmaceutical company that has developed a new pain reliever. Recall null and alternative hypothesis was

$$\begin{aligned} H_0 : \mu &= 25 \text{ minutes} \\ H_a : \mu &< 25 \text{ minutes} \end{aligned}$$

and we had observed a test statistic

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{23 - 25}{\frac{10}{\sqrt{16}}} = -0.8$$

with 15 degrees of freedom. Using an $\alpha = 0.10$ level of significance, if this test statistic is smaller than the 0.10th quantile of a t-distribution with 15 degrees of freedom, then we will reject the null hypothesis. This cutoff value is $t_{crit} = -1.341$ and can be using either R or the t-table in your book. Bcause the observed test statistic is less extreme than the cutoff value, we failed to reject the null hypothesis.

We can push this idea even farther and calculate a critical value on the original scale of $\bar{x}$ by solving

$$\begin{aligned} t_{crit} &= \frac{\bar{x}_{crit} - \mu_0}{\frac{s}{\sqrt{n}}} \\ -1.341 &= \frac{\bar{x}_{crit} - 25}{\frac{10}{\sqrt{16}}} \\ -1.341\left(\frac{10}{\sqrt{16}}\right) + 25 &= \bar{x}_{crit} \\ 21.65 &= \bar{x}_{crit} \end{aligned}$$

So if we observe a sample mean $\bar{x} < 21.65$ then we would reject the null hypothesis. Here we actually observed $\bar{x} = 23$ so this comparison still fails to reject the null hypothesis and concludes there is insufficient evidence to reject that the new pain reliever has the same time till relief as the old medicine.

In general, I prefer to calculate and report p-values because they already account for any ambiguity in if we are dealing with a 1 sided or 2 sided test and how many degrees of freedom there are.

6.3.4 t-tests in R

While it is possible to do t-tests by hand, most people will use a software package to perform these calculations. Here we will use the R function `t.test()`. This function expects a vector of data (so that it can calculate $\bar{x}$ and s) and a hypothesized value of μ.

[4] Be careful here, because if your alternative is "greater than" and your test statistic is negative, then the p-value really is greater than 0.5. This situation is rare and 9 times out of 10, the student has just used the table incorrectly.

Example. Suppose we have data regarding fuel economy of 5 vehicles of the same make and model and we wish to test if the observed fuel economy is consistent with the advertised 31 mpg at highway speeds. Assuming the fuel economy varies normally amongst cars of the same make and model, we test

$$\begin{aligned} H_0 : \mu &= 31 \\ H_a : \mu &\neq 31 \end{aligned}$$

and calculate

```
cars <- data.frame(mpg = c(31.8, 32.1, 32.5, 30.9, 31.3))
mean( ~ mpg, data=cars)

## [1] 31.72

sd( ~ mpg, data=cars )

## [1] 0.6340347
```

The test statistic is:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{31.72 - 31}{\frac{0.634}{\sqrt{5}}} = 2.54$$

The p-value is

$$p - value = 2 \cdot P(T_4 > 2.54) = 0.064$$

and a 95% confidence interval is

$$\begin{aligned} \bar{x} &\pm t_{n-1}^{1-\alpha/2} \left(\frac{s}{\sqrt{n}} \right) \\ 31.72 &\pm 2.776445 \left(\frac{0.63403}{\sqrt{5}} \right) \\ 31.72 &\pm 0.7872 \\ &[30.93, 32.51] \end{aligned}$$

```
# Mosaic style call to the t.test function
t.test( ~ mpg, data=cars, mu=31, alternative='two.sided' )

##
##  One Sample t-test
##
## data:  data$mpg
## t = 2.5392, df = 4, p-value = 0.06403
## alternative hypothesis: true mean is not equal to 31
## 95 percent confidence interval:
##  30.93274 32.50726
## sample estimates:
## mean of x
##     31.72

# Base R call to the t.test returns the same analysis
# t.test( cars$mpg, mu=31, alternative='two.sided' )
```

The `t.test()` function supports testing one-sided alternatives and more information can be found in the R help system using `help(t.test)`.

6.4 Type I and Type II Errors

We can think of the p-value as measuring how much evidence we have for the null hypothesis. If the p-value is small, the evidence for the null hypothesis is small. Conversely if the p-value is large, then the data is supporting the null hypothesis.

There is an important philosophical debate about how much evidence do we need in order to reject the null hypothesis. My brother-in-law would have to have extremely strong evidence before he stated the other rancher was wrong. Likewise, researchers needed solid evidence before concluding that Newton's Laws of Motion were incorrect.

Since the p-value is a measure of evidence for the null hypothesis, if the p-value drops below a specified threshold (call it α), I will chose to reject the null hypothesis. Different scientific disciplines have different levels of rigor. Therefore they set commonly used α levels differently. For example physicists demand a high degree of accuracy and consistency, thus might use $\alpha = 0.01$, while ecologists deal with very messy data and might use an $\alpha = 0.10$.

The most commonly used α-level is $\alpha = 0.05$, which is traditional due to an off-hand comment by R.A. Fisher. There is nothing that fundamentally forces us to use $\alpha = 0.05$ other than tradition. However, when sociologists do experiments presenting subjects with unlikely events, it is usually when the events have a probability around 0.05 that the subjects begin to suspect they are being duped.

People who demand rigor might want to set α as low as possible, but there is a trade off. Consider the following possibilities, where the "True State of Nature" is along the top, and the decision is along the side.

		True State of Nature	
		H_0 True	H_0 False
Decision	Fail to Reject H_0	Correct	Type II error
	Reject H_0	Type I error	Correct

There are two ways to make a mistake. The type I error is to reject H_0 when it is true. This error is controlled by α. We can think of α as the probability of rejecting H_0 when we shouldn't. However there is a trade off. If α is very small then we will fail to reject H_0 in cases where H_0 is not true. This is called a type II error and we will define β as the probability of failing to reject H_0 when it is false.

This trade off between type I and type II errors can be seen by examining our legal system. A person is presumed innocent until proven guilty. So the hypothesis being tested in the court of law are

$$\begin{aligned} H_0 &: \text{ defendent is innocent} \\ H_a &: \text{ defendent is guilty} \end{aligned}$$

Our legal system theoretically operates under the rule that it is better to let 10 guilty people go free, than wrongly convict 1 innocent. In other words, it is worse to make a type I mistake (concluding guilty when innocent), than to make a type II mistake (concluding not guilty when guilty). Critically, when a jury finds a person "not guilty" they are not saying that defense team has proven that the defendant is innocent, but rather that the prosecution has not proven the defendant guilty.

This same idea manifests itself in science with the α-level. Typically we decide that it is better to make a type II mistake. An experiment that results in a large p-value does not prove that H_0 is true, but that there is insufficient evidence to conclude H_a.

If we still suspect that H_a is true, then we must repeat the experiment with a larger samples size. A larger sample size makes it possible to detect smaller differences.

6.4.1 Power and Sample Size Selection

Just as we calculated the necessary sample size to achieve a confidence interval of a specified width, we are also often interested in calculating the necessary sample size to to find a significant difference from the hypothesized mean μ_0. Just as in the confidence interval case where we had to specify the half-width E and some estimate of the population standard deviation $\hat{\sigma}$, we now must specify a difference we want to be able to detect δ and an estimate of the population standard deviation $\hat{\sigma}$.

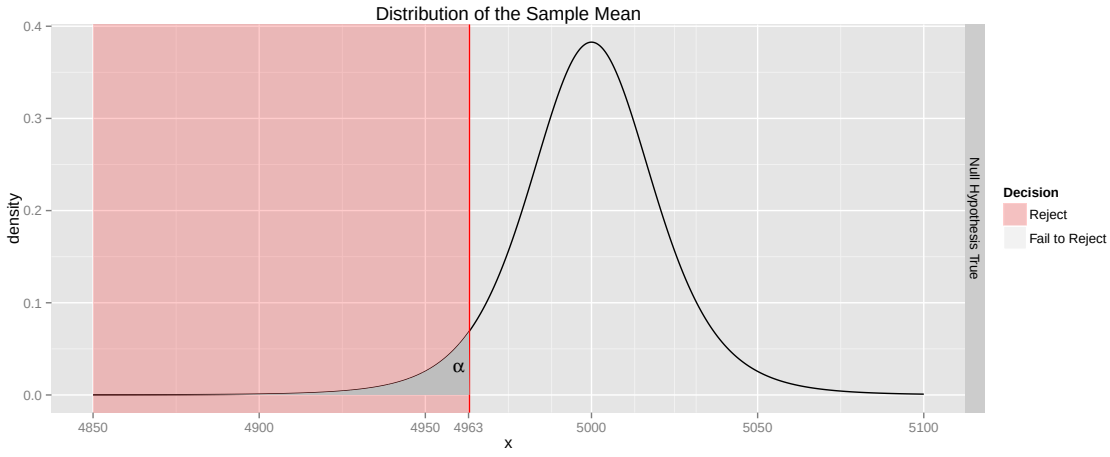
Example. Suppose that I work in Quality Control for a company that manufactures a type of rope. This rope is supposed to have a mean breaking strength of 5000 pounds and long experience with the process suggests that the standard deviation is approximately $s = 50$. As with many manufacturing processes, sometimes the machines that create the rope get out of calibration. So each morning we take a random sample of $n = 7$ pieces of rope and using $\alpha = 0.05$, test the hypothesis

$$\begin{aligned} H_0 : \mu &= 5000 \\ H_a : \mu &< 5000 \end{aligned}$$

Notice that I will reject the null hypothesis if $\bar{x}$ is less than some cut-off value (which we denote $\bar{x}_{crit}$), which we calculate by first recognizing that the critical t-value is $t_{crit} = t^{\alpha}_{n-1} = -1.943$ and then solving the following equation for $\bar{x}_{crit}$

$$\begin{aligned} t_{crit} &= \frac{\bar{x}_{crit} - \mu_0}{\frac{s}{\sqrt{n}}} \\ t_{crit}\left(\frac{s}{\sqrt{n}}\right) + \mu_0 &= \bar{x}_{crit} \\ -1.943\left(\frac{50}{\sqrt{7}}\right) + 5000 &= \bar{x}_{crit} \\ 4963 &= \bar{x}_{crit} \end{aligned}$$

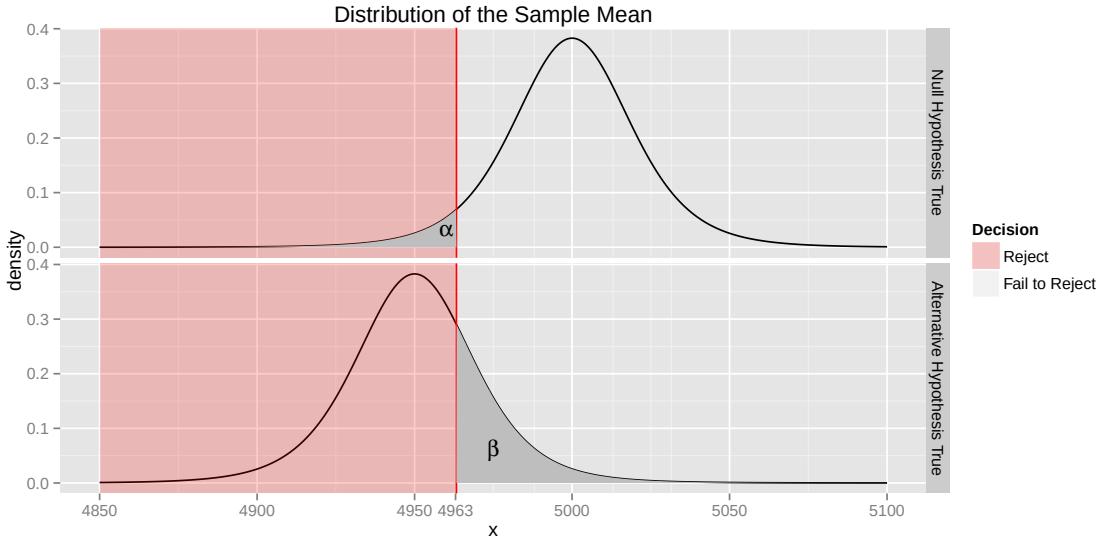
There is a trade off between the Type I and Type II errors. By making a Type I error, I will reject the null hypothesis when the null hypothesis is true. Here I would stop manufacturing for the day while recalibrating the machine. Clearly a Type I error is not good. The probability of making a Type I error is denoted α.



A type II error occurs when I fail to reject the null hypothesis when the alternative is true. This would mean that we would be selling ropes that have a breaking point less than the advertised amount. This opens the company up to a lawsuit. We denote the probability of making a Type II error is denoted as β and define **Power** $= 1 - \beta$. But consider that I don't want to be shutting down the plant when the breaking point is just a few pounds from the true mean. The head of

engineering tells me that if the average breaking point is more than 50 pounds less than 5000, we have a problem, but less than 50 pounds is acceptable.

So I want to be able to detect if the true mean is less than 4950 pounds. Consider the following where we assume $\mu = 4950$.

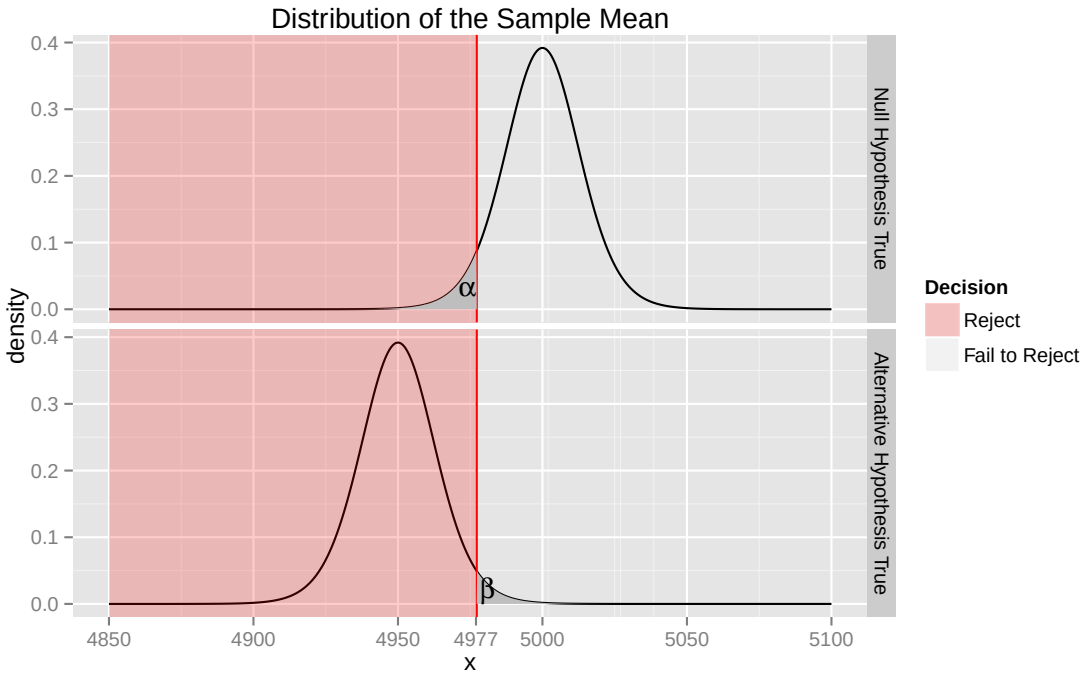


The the probability of a type II error is

$$\begin{aligned} \beta &= P\left(\bar{X} > 4963.3 \,|\, \mu = 4950\right) \\ &= P\left(\frac{\bar{X} - 4950}{50/\sqrt{7}} > \frac{4963.3 - 4950}{50/\sqrt{7}}\right) \\ &= P\left(T_6 > 0.703\right) \\ &= 0.254 \end{aligned}$$

and therefore my power for detecting a mean breaking strength less than or equal to 4950 is $1 - \beta = 0.7457$ which is very close to what any statistical package will calculate for us.[5] This power is rather low and I would prefer to have the power be near 0.95. We can improve our power by using a larger sample size. We'll repeat these calculations using $n = 15$.

[5]The power calculation should done using a t-distribution with non-centrality parameter instead of just shifting the distribution. The difference is slight, but is enough to cause our calculation to be slightly off.



Power calculations are relatively tedious to do by hand, but fortunately there are several very good resources for exploring how power and sample size interact. My favorite is a Java Applet web page maintained by Dr. Russ Lenth at http://www.stat.uiowa.edu/~rlenth/Power/. It will provide you a list of analysis to do the calculations for and the user is responsible for knowing that we are doing a one-sample t-test with a one-sided alternative.

Alternatively, we can do these calculations in R using the function `power.t.test()`.

Fundamentally there are five values that can be used and all power calculators will allow a user to input four of them and the calculator will calculate the fifth.

1. The difference δ from the hypothesized mean μ_0 that we wish to detect
2. The population standard deviation σ.
3. The significance level of the test α.
4. The power of the test $1 - \beta$.
5. The sample size n.

```
power.t.test(delta=50, sd=50, sig.level=0.05, n=7,
             type="one.sample", alternative="one.sided")

##
##      One-sample t test power calculation
##
##               n = 7
##           delta = 50
##              sd = 50
##       sig.level = 0.05
##           power = 0.7543959
##     alternative = one.sided
```

```
power.t.test(delta=50, sd=50, sig.level=0.05, power=0.95,
             type="one.sample", alternative="one.sided")

##
##      One-sample t test power calculation
##
##               n = 12.32052
##           delta = 50
##              sd = 50
##       sig.level = 0.05
##           power = 0.95
##     alternative = one.sided
```

The general process for selecting a sample size is to

1. Pick a α-level. Usually this is easy and people use $\alpha = 0.05$.
2. Come up with an estimate for the standard deviation σ. If you don't have an estimate, then a pilot study should be undertaken to get a rough idea what the variability is. Often this is the only good data that comes out of the first field season in a dissertation.
3. Decide how large of an effect is scientifically interesting.
4. Plug the results of steps 1-3 into a power calculator and see how large a study you need to achieve a power of 90% or 95%.

6.5 Variations of the t-test: Comparing two population means

It is very common to want to compare the means of two different distributions. Suppose I am interested in NAU students and wish examine whether the mean GPA of men is different from the mean GPA of women. Another example, researchers working for a pharmaceutical company might wish to compare the mean time to relief of their drug versus the mean time of relief from a competing drug. Finally a third example might be comparing trees with a certain morphological trait to "normal" trees.

In general, we can consider the problem of comparing the means of two populations and testing the hypothesis that the means are the same.

$$H_0 : \mu_1 = \mu_2$$

versus one of the following alternative hypothesis

$$H_a : \mu_1 \neq \mu_2$$
$$H_a : \mu_1 > \mu_2$$
$$H_a : \mu_1 < \mu_2$$

I could also re-write these hypothesis in terms of the difference between the two hypothesis

$$H_0 : \mu_1 - \mu_2 = 0$$

versus on the following

$$H_a : \mu_1 - \mu_2 \neq 0$$
$$H_a : \mu_1 - \mu_2 > 0$$
$$H_a : \mu_1 - \mu_2 < 0$$

There are two ways to do these tests. The first method, a paired test is generally more powerful, but is only applicable in very specific instances. The more general two sample t-test is easy to do and is more applicable.

6.5.1 Paired t-Tests

If the context of the problem (or data) is such that we can logically pair an observation from the first population to a particular observation in the second, then we can perform what is called a *Paired Test*. In a paired test, we will take each set of paired observations, calculate the difference, and then perform a regular hypothesis test on the *differences*.

Example. Cross country skiers use ski poles to propel themselves across the snow. As such, the ergonomics of the connection between the hand and the pole might be important. It is common for serious cross country racers to use a specialized type of grip that we will call a "racing grip". Suppose a researcher is interested in comparing whether an expensive "racing grip" provides more power transfer on cross country ski poles to the standard type of grip. The researcher rigs a pressure sensor on two sets of poles, one with a standard grip, and one with the racing grip. He then gets a group of $n = 20$ cross country ski racers to use both sets of poles. Data from this experiment might look like this...

Skier	Standard Grip	Racing Grip	Difference (R-S)
Bob	19.2 lbs	21.1 lbs	1.9 lbs
Jeff	18.6 lbs	19.7 lbs	1.1 lbs
⋮	⋮	⋮	⋮

Here I chose to look at the difference of $Racing - Standard$, and we will now test the hypothesis

$$H_0 : \mu_{diff} = 0$$

$$H_a : \mu_{diff} > 0$$

This hypothesis test will be carried out exactly as we have before, but the only difference is that I will be using the average of the differences. Suppose we took our sample and got a sample mean $\bar{x}_{diff} = 1.5$ lbs, and a standard deviation of the differences $s_{diff} = 3$ lbs. Assuming that these differences come from an approximately normal distribution, our test statistic is

$$t_{19} = \frac{\bar{x}_{diff} - \mu_0}{\frac{s_{diff}}{\sqrt{n}}} = \frac{1.5 - 0}{\frac{3}{\sqrt{20}}} = 2.23$$

The p-value is $P(T_{19} > 2.23) = 0.019$, so at an $\alpha = 0.05$ level, I reject the null hypothesis and conclude that the racing grip does transfer more power than the standard grip.

The important thing to notice, is that for each observation that I have for the racing grip, there is a particular observation using the standard grip. The reason that this test is so powerful, is that everything else is constant between those two observations. With the same skier, same skis, same snow, etc, we are able to effectively isolate the impact of the grip.

As a practical point, notice that we should randomly choose whether a skier uses the racing grip first or second to control for possible effects of order on the power. Perhaps being suitably warmed up helps, or perhaps the skier has become tired after the first test. Either way, the researcher should control for this possibility.

6.5.2 Two Sample t-test

Unfortunately there is not always a logical way to pair observations. Fortunately the solution is not too difficult.

Suppose we are interested in examining the heights of men and women and wish to test the seeming obvious proposition that the average height of men is taller than the average height of women. Here we will examine the hypotheses

$$H_0 : \mu_m - \mu_w = 0$$

$$H_a : \mu_m - \mu_w > 0$$

The idea here will be to calculate the mean of a sample of men, the mean of a sample of women and then compare the two.

Theory. In principle I wish to examine the distribution of $\bar{X}_m - \bar{X}_w$. This is a function of two random variables, so this difference is also a random variable, which I'll denote as D. Then D has a distribution with mean $\mu_m - \mu_w$ as might be expected, but the standard deviation is more tricky. Recall that earlier in class we said that variance was easier to use mathematically. Here is a case of that.

$$Var(D) = Var(\bar{X}_m) + Var(\bar{X}_w)$$

$$StdDev(D) = \sqrt{Var(\bar{X}_m) + Var(\bar{X}_w)}$$

Therefore my two sample t-test statistic will be

$$t = \frac{(\bar{x}_m - \bar{x}_w) - 0}{\sqrt{\frac{s_m^2}{n_m} + \frac{s_w^2}{n_w}}}$$

Suppose I take a sample of $n_m = 30$ men and calculate $\bar{x}_m = 69.5$ inches, and sample standard deviation $s_m = 3.2$ inches. For the women, I take a sample of $n_w = 25$ and calculate $\bar{x}_w = 64.0$ inches, and a sample standard deviation $s_w = 2.2$ inches. So our test statistic will be

$$t_{???} = \frac{(\bar{x}_m - \bar{x}_w) - 0}{\sqrt{\frac{s_m^2}{n_m} + \frac{s_w^2}{n_w}}} = \frac{69.5 - 64.0}{\sqrt{\frac{3.2^2}{30} + \frac{2.2^2}{25}}} = \frac{5.5}{\sqrt{0.3413 + 0.1936}} = 7.52$$

But now we must deal with the question of what is the appropriate degrees of freedom? The men have a sample of $n_m = 30$, but the women only have $n_w = 25$. You might guess that the degrees of freedom ought to be somewhere between $min(n_m, n_w)$ and $n_m + n_w$. However there is an efficient way to approximate what the degrees of freedom are called *Satterthwaite's Approximation.*

$$df = \frac{(V_m + V_w)^2}{\frac{V_m^2}{n_m - 1} + \frac{V_w^2}{n_w - 1}}$$

where

$$V_m = \frac{s_m^2}{n_m} \text{ and } V_w = \frac{s_w^2}{n_w}$$

So in our case we have

$$V_m = \frac{3.2^2}{30} = 0.3413 \qquad \text{and} \qquad V_w = \frac{2.2^2}{25} = 0.1936$$

$$df = \frac{(0.3413 + 0.1936)^2}{\frac{0.3413^2}{29} + \frac{0.1936^2}{24}} = \frac{0.2861}{.005578} = 51.29$$

Since we degrees of freedom must be an integer, we will always round down to the next integer, in this case, 51. This should make sense, because there are a total of $n = 55$ observations, but we should take a penalty because they are not coming from the same distribution.

Now that we have our degrees of freedom we can calculate the p-value $= P(T_{51} > 7.52) \approx 0$. So we reject the null hypothesis and conclude that there is statistically significant evidence that the average height of men is larger than the average height of women.

Example. Suppose we have data from an experiment which compared the productivity of desert plants under elevated CO_2 versus ambient conditions. Suppose that 40 plants were grown from seedlings with half subjected to CO_2 levels of 550 ppm and the other half subjected to ambient levels. Both groups were grown in green houses under identical conditions and at the end of the

experiment all plants were kiln dried and weighed. We denote Ambient as population 1, and Elevated as population 2. We wish to test the hypotheses

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

The data for ambient was $\bar{x}_1 = 601.1$, $s_1 = 36.60$, $n_1 = 20$ while the data for the elevated plants was $\bar{x}_2 = 646.85$, $s_2 = 32.92$, $n_2 = 20$. The sample statistic is therefore

$$\begin{aligned} t_{???} &= \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \\ &= \frac{(601.1 - 646.85)}{\sqrt{\frac{36.60^2}{20} + \frac{32.92^2}{20}}} \\ &= -4.15 \end{aligned}$$

and the degrees of freedom are

$$\begin{aligned} V_1 &= \frac{s_1^2}{n_1} = \frac{36.60^2}{20} = 66.98 \\ V_2 &= \frac{s_2^2}{n_2} = \frac{32.92^2}{20} = 54.19 \\ df &= \frac{(V_1 + V_2)^2}{\frac{V_1^2}{n_1 - 1} + \frac{V_2^2}{n_2 - 1}} = \frac{(66.98 + 54.19)^2}{\frac{66.98^2}{19} + \frac{54.19^2}{19}} = 37.58 \end{aligned}$$

and so the p-value is

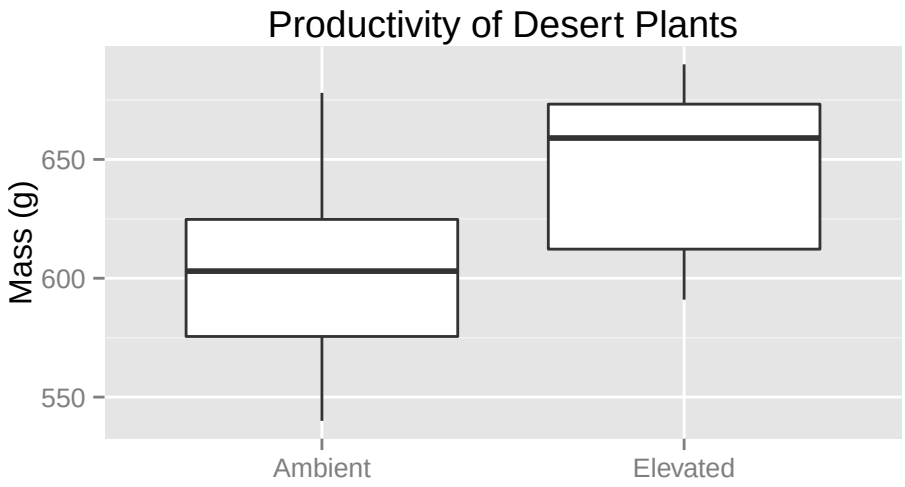
$$\begin{aligned} p - value &= 2 \cdot P\left(T_{37} < -4.15\right) \\ &= 0.000187 \end{aligned}$$

The equivalent analysis in R is as follows:

```r
CO2 <- data.frame(
    mass = c(540, 634, 620, 606, 598, 627, 593, 541, 577, 638,
             571, 649, 678, 604, 559, 624, 553, 614, 602, 594,
             685, 677, 610, 601, 682, 659, 638, 687, 609, 607,
             690, 591, 613, 647, 672, 664, 659, 618, 669, 659),
    trt  = c( rep('Ambient', 20), rep('Elevated',20) ) )

# Mosaic style call to t.test
t.test(mass ~ trt, data=CO2, var.equal=FALSE)

##
##  Welch Two Sample t-test
##
## data:  mass by trt
## t = -4.1565, df = 37.582, p-value = 0.0001797
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -68.04061 -23.45939
## sample estimates:
##  mean in group Ambient mean in group Elevated
##                 601.10                 646.85

# Base R styel call to t.test
# amb <- CO2$mass[  1:20 ]
# ele <- CO2$mass[ 21:40 ]
# t.test(amb, ele, data=CO2, var.equal=FALSE)
```

6.5.3 Two sample t-test using a pooled variance estimator

In some instances, it is possible that the two populations have the same variance parameter despite having different means. Consider the following graph of the CO_2 data.



It isn't unreasonable to think that these two groups have variances that are similar. Let's assume that the variance of the ambient group is equal to that of the elevated group. That is, we assume $\sigma_1 = \sigma_2 = \sigma$ and we will see how our two-sample t-test would change.

First, we need to calculate a pooled variance estimate of σ. First recall the formula for the sample variance for one group was

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

In the case with two samples, want a similar formula but it should take into account data from both sample groups. Define the notation x_{1i} to be the ith observation of group 1, and x_{2j} to be the jth observation of group 2. We want to subtract each observation from the its appropriate sample mean and that since we had to estimate two means, we need to subtract two degrees of freedom from the denominator.

$$\begin{aligned} s^2_{pooled} &= \frac{1}{n_1+n_2-2}\left[\sum_{i=1}^{n_1}(x_{1i}-\bar{x}_1)^2 + \sum_{j=1}^{n_2}(x_{2j}-\bar{x}_2)^2\right] \\ &= \frac{1}{n_1+n_2-2}\left[(n_1-1)\,s_1^2 + (n_2-1)\,s_2^2\right] \end{aligned}$$

where $\bar{x}_1$ is the sample mean of the first group and s_1^2 is the sample variance of the first group and similarly for $\bar{x}_2$ and s_2^2. Finally we notice that this pooled estimate of the variance term σ^2 has n_1+n_2-2 degrees of freedom. One of the biggest benefits of the pooled procedure is that we don't have to mess with the Satterthwaite's approximate degrees of freedom.

Recall our test statistic in the unequal variance case was

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

but in the equal variance case, we will use the pooled estimate of the variance term s^2_{pooled} instead of s_1^2 and s_2^2. So our test statistic becomes

$$\begin{aligned} t_{df=n_1+n_2-2} &= \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s^2_{pool}}{n_1} + \frac{s^2_{pool}}{n_2}}} \\ &= \frac{(\bar{x}_1 - \bar{x}_2) - 0}{s_{pool}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \end{aligned}$$

Example. If we had decided to pool the variance in the elevated CO_2 example we would have

$$\begin{aligned} s^2_{pooled} &= \frac{1}{n_1+n_2-2}\left[(n_1-1)\,s_1^2 + (n_2-1)\,s_2^2\right] \\ &= \frac{1}{38}\left[(19)\,36.6^2 + (19)\,32.92^2\right] \\ &= 1211.643 \\ s_{pooled} &= 34.81 \end{aligned}$$

and the test statistic would be

$$\begin{aligned} t_{n_1+n_2-2} &= \frac{(\bar{x}_1 - \bar{x}_2) - 0}{s_{pooled}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \\ t_{38} &= \frac{(601.1 - 646.85)}{34.81\sqrt{\frac{1}{20} + \frac{1}{20}}} \\ &= -4.1561 \end{aligned}$$

and the p-value is

$$\begin{aligned} p - value &= 2 \cdot P\left(T_{38} < -4.1561\right) \\ &= 0.000177 \end{aligned}$$

Again we present the same analysis in R to confirm our calculations.

```
t.test(mass ~ trt, data=CO2, var.equal=TRUE)

##
##  Two Sample t-test
##
## data:  mass by trt
## t = -4.1565, df = 38, p-value = 0.000177
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -68.03246 -23.46754
## sample estimates:
##  mean in group Ambient mean in group Elevated
##                 601.10                 646.85
```

Chapter 7

Testing Model Assumptions

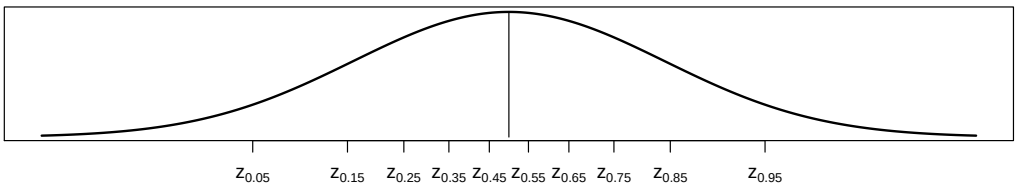
Performing a t-test requires that the data was drawn from a normal distribution or that the sample size is large enough that the Central Limit Theorem will guarantee that the sample means are approximately normally distributed. However, how do you decide if the data were drawn from a normal distribution, say if your sample size is between 10 and 20? If we are using a model that assumes equal variance between groups, how should we test if that assumption is true?

7.1 Testing Normality

7.1.1 Visual Inspection - QQplots

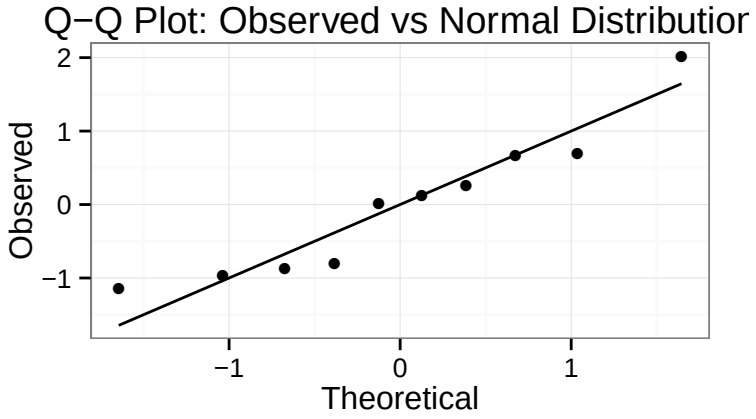
If we are taking a sample of size $n = 10$ from a standard normal distribution, then I should expect that the smallest observation will be negative. Intuitively, you would expect the smallest observation to be near the 10th percentile of the standard normal, and likewise the second smallest should be near the 20th percentile.

This idea needs a little modification because the largest observation cannot be near the 100th percentile (because that is ∞). So we'll adjust the estimates to still be spaced at $(1/n)$ quantile increments, but starting at the $0.5/n$ quantile instead of the $1/n$ quantile. So the smallest observation should be near the 0.05 quantile, the second smallest should be near the 0.15 quantile, and the largest observation should be near the 0.95 quantile. I will refer to these as the *theoretical quantiles*.



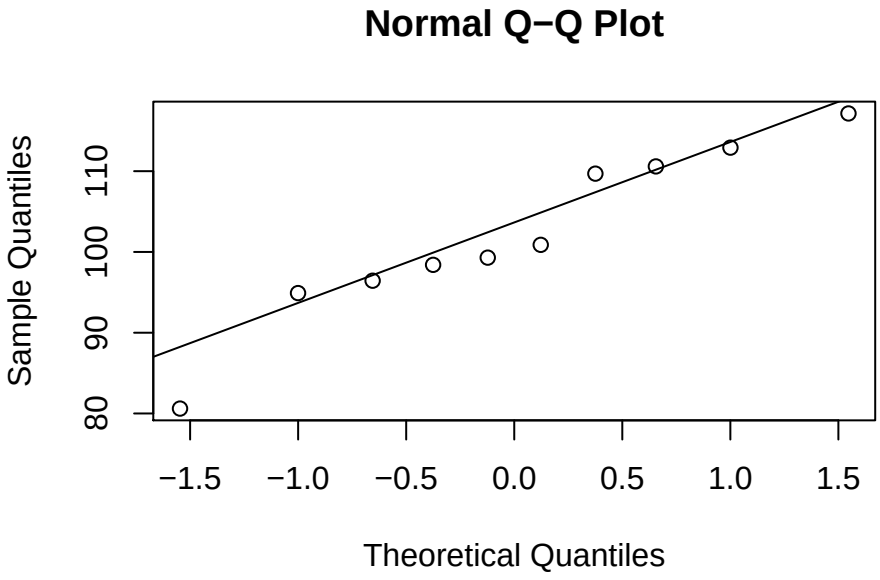
I can then graph the theoretical quantiles vs my observed values and if they lie on the 1-to-1 line, then my data comes from a standard normal distribution.

```
n <- 10
data <- data.frame( observed    = sort( rnorm(n, mean=0, sd=1) ),
                    theoretical = qnorm( (1:n -.5)/n, mean=0, sd=1 ) )
library(ggplot2)
ggplot(data) +
  geom_point( aes(x=theoretical, y=observed) ) +
  geom_line(  aes(x=theoretical, y=theoretical) ) +
  labs(x='Theoretical', y='Observed', title="Q-Q Plot: Observed vs Normal Distribution') +
  theme_bw()
```

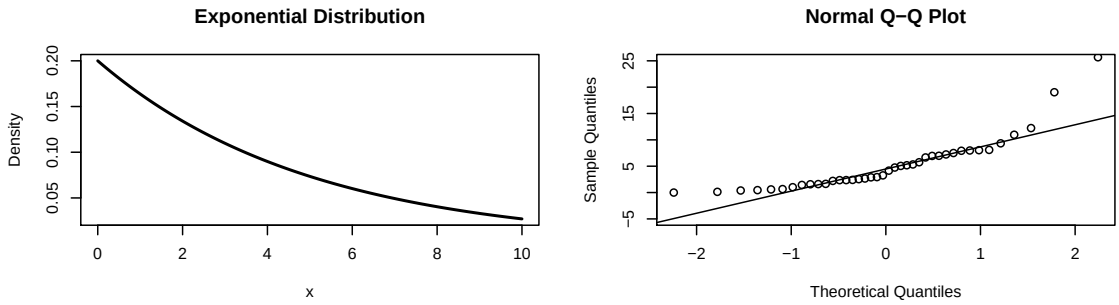


If I think my data are normal, but with some mean μ and standard deviation σ, we still make the same graph, but the 1-to-1 line will be moved to pass through the 1st and 3rd quartiles. Again, the data points should be near the line. This is common enough that R has built in functions to make this graph:

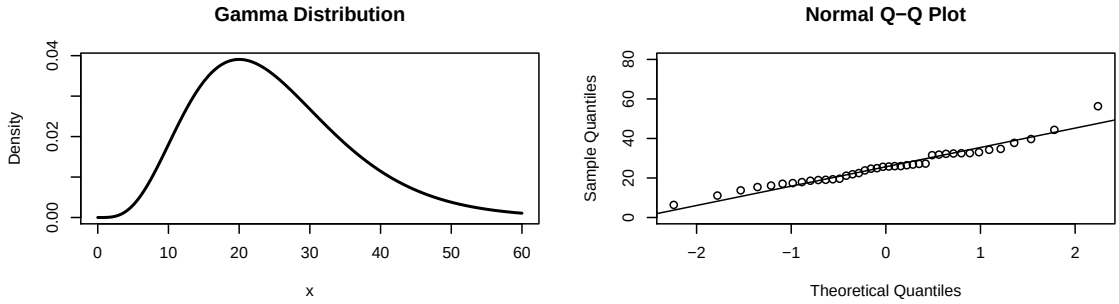
```
n <- 10
x <- rnorm(n, mean=100, sd=10)
qqnorm(x)
qqline(x)
```



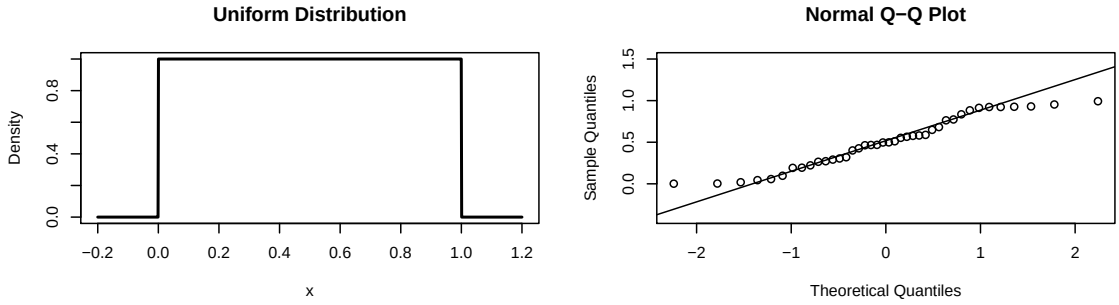
We now will examine a sample of $n = 40$ from a bunch of different distributions that are not normal and see what the normal QQ plot looks like. In the following graphs, pay particular attention to the tails. Notice the the T-distribution has significantly heavier tails than the normal distribution and that is reflected in the dots being lower than the line on the left and higher on the right. Likewise the logNormal distribution, which is defined by $\log(X) \sim Normal$ has too light of a tail on the left (because logNormal variables must be greater than 0) and too heavy on the right. The uniform distribution, which is cut off at 0 and 1, has too light of tails in both directions.


Exponential Distribution
Density
x

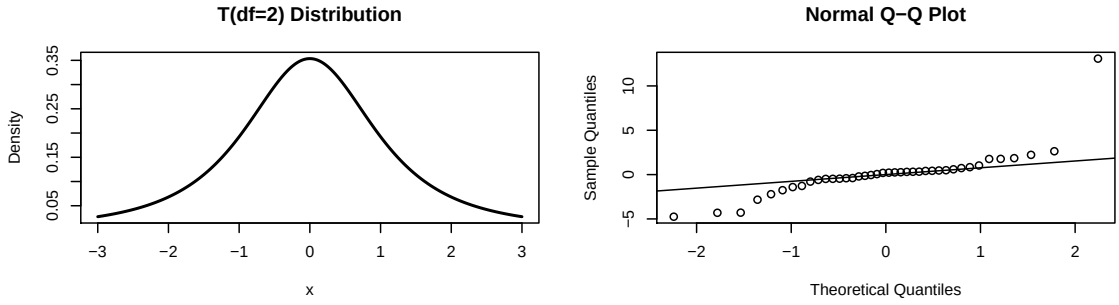
Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles


Gamma Distribution
Density
x

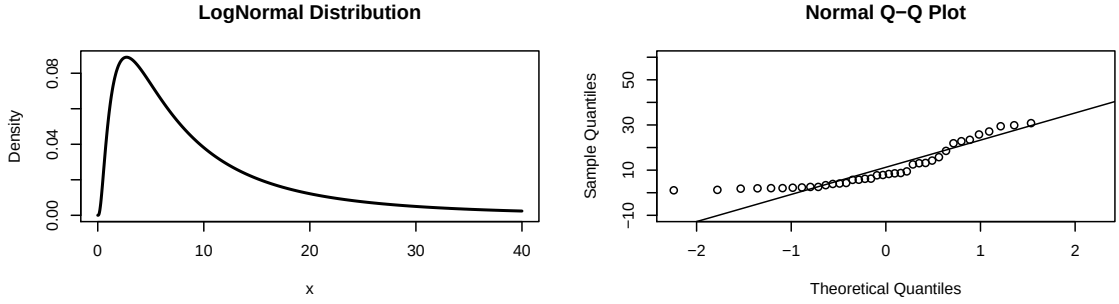
Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles


Uniform Distribution
Density
x

Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles


T(df=2) Distribution
Density
x

Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles


LogNormal Distribution
Density
x

Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles

7.1.2 Tests for Normality

It seems logical that there should be some sort of statistical test for if a sample is obviously non-normal. Two common ones are the Shapiro-Wilks test and the Anderson-Darling test. The Shapiro-Wilks test is available in the base installation of R with the function `shapiro.test()`. The Anderson-Darling test is available in the package `nortest`. Here we will not focus on the theory of these tests, but instead their use. In both tests the null hypothesis is that the data are normally distributed.

$$\begin{aligned} H_0: &\quad \text{data are normally distributed} \\ H_a: &\quad \text{data are not normally distributed} \end{aligned}$$

Therefore a small *p-value* is evidence against normality.

Often we want to know if our data comes from a normal distribution because our sample size is too small to rely on the Central Limit Theorem to guarantee that the sampling distribution of the sample mean is Normal. So how well do these tests detect non-normality in a small sample size case?

```
x <- rlnorm(10, meanlog=2, sdlog=2)
shapiro.test(x)

##
##  Shapiro-Wilk normality test
##
## data:  x
## W = 0.3954, p-value = 2.207e-07
```

So the Shapiro-Wilks test detects the non-normality in the extreme case of a logNormal distribution, but what about something closer to normal like the gamma distribution?

```
x <- rgamma(10, shape=5, rate=1/5)
shapiro.test(x)

##
##  Shapiro-Wilk normality test
##
## data:  x
## W = 0.927, p-value = 0.4193
```

Here the Shapiro test fails to detect the non-normality due to the small sample size. Unfortunately, the small sample size case is exactly when we need a good test. So what do we do?

My advise is to look at the histograms of your data, normal QQ plots, and to use the Shapiro-Wilks test to find extreme non-normality, but recognize that in the small sample case, we have very little power and can only detect extreme departures from normality. If I cannot detect non-normality and my sample size is moderate (15-30), I won't worry too much since the data isn't too far from normal and the CLT will help normalize the sample means but for smaller sample sizes, I will use nonparametric methods (such as the bootstrap) that do not make distributional assumptions.

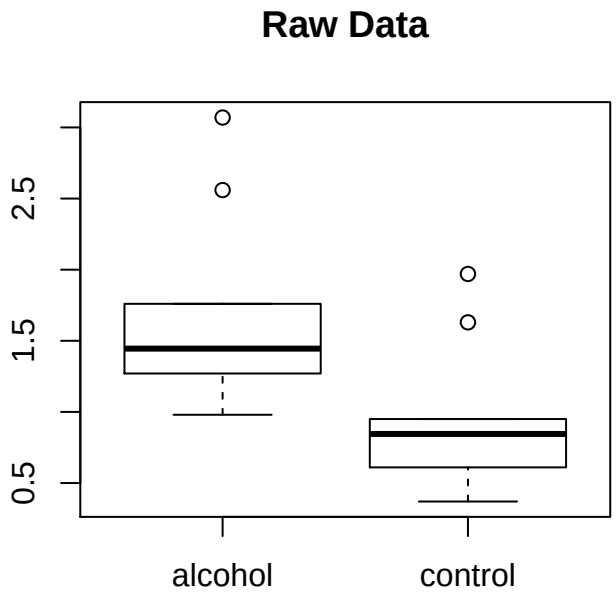
7.2 Testing Equal Variance

7.2.1 Visual Inspection

Often a test procedure assumes equal variances amongst groups or constant variance along a prediction gradient. The most effect way of checking to see if that assumption is met is to visually inspect the data. For the case of t-tests, boxplots are an excellent visual check. If the lengths of the boxes are not substantially different, then the equal variance assumption is acceptable.

Consider an experiment where we measure the speed of reaction to a stimulus. The subjects are told to press a button as soon as they hear a noise. Between 2 and 30 seconds later an extremely loud noise is made. Of primary interest is how inebriation affects the reaction speed. Since we can't surprise subjects twice, only one measurement per subject is possible and a paired test is not possible. Subjects were randomly assigned to a control or alcohol group[1]

```
Alcohol <- data.frame(
  time=c( 0.90, 0.37, 1.63, 0.83, 0.95, 0.78, 0.86, 0.61, 0.38, 1.97,
          1.46, 1.45, 1.76, 1.44, 1.11, 3.07, 0.98, 1.27, 2.56, 1.32 ),
  trt = rep(c("control","alcohol"),each=10))
boxplot(time ~ trt, data=Alcohol, main='Raw Data')
```



7.2.2 Tests for Equal Variance

Consider having samples drawn from normal distributions

$$X_{ij} = \mu_i + \epsilon_{ij} \qquad \text{where } \epsilon_{ij} \sim N\left(0,\, \sigma_i^2\right)$$

where the i subscript denotes which population the observation was drawn from and the j subscript denotes the individual observation and from the ith population we observe n_i samples. In general I might be interested in evaluating if $\sigma_i^2 = \sigma_j^2$.

Let's consider the simplest case of two populations and consider the null and alternative hypotheses:

$$\begin{aligned} H_0 : \sigma_1^2 &= \sigma_2^2 \\ H_a : \sigma_1^2 &\neq \sigma_2^2 \end{aligned}$$

If the null hypothesis is true, then the ratio s_1^2/s_2^2 should be approximately one. It can be shown

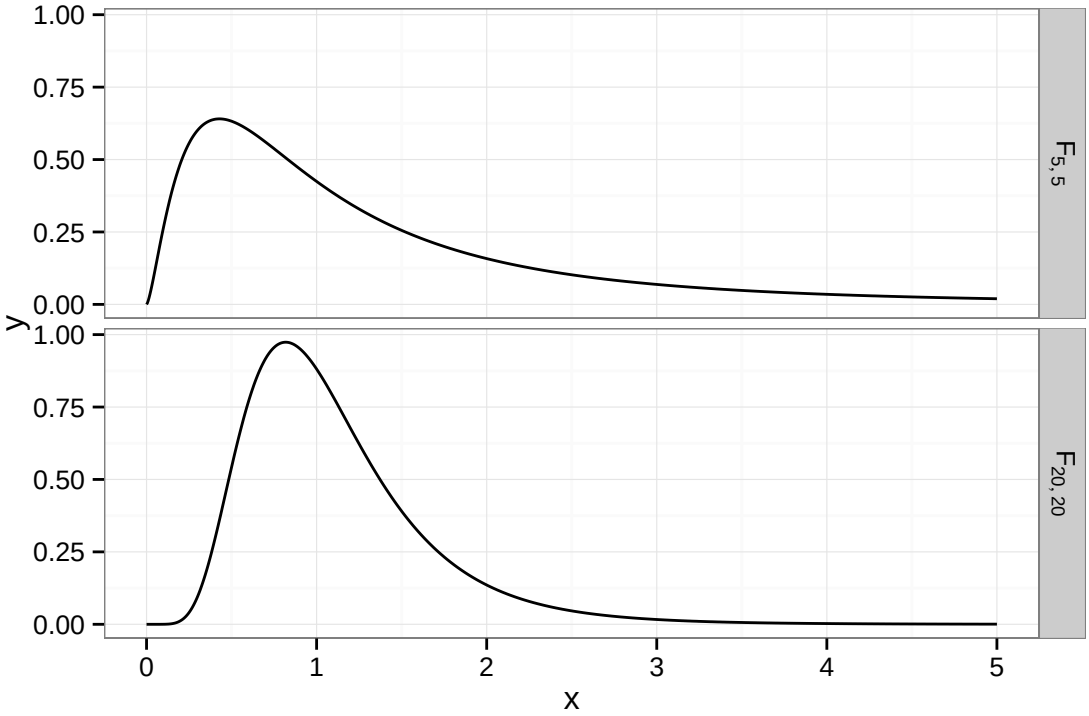
[1] This study was long enough ago that review boards let this sort of thing be done.

that under the null hypothesis,

$$f = \frac{s_1^2}{s_2^2} \sim F_{df_1, df_2}$$

where df_1 and df_2 are the associated degrees of freedom for s_1^2 and s_2^2. The order of these is traditionally given with the degrees of freedom of the top term first and the degrees of freedom of the bottom term second.

Variables that follow a F distribution must be non-negative and two F distributions are shown below. The F distribution is centered at $E\left(F_{df_1,df_2}\right) = \frac{df_2}{df_2-2} \approx 1$ for large values of df_2. The variance of this distribution goes to 0 as df_1 and df_2 get large.



If the value of my test statistic $f = s_1^2/s_2^2$ is too large or too small, then we will reject the null hypothesis. If we preform an F-test with an $\alpha = 0.05$ level of significance then we'll reject H_0 if $f < F_{0.025,n_1-1,n_2-1}$ or if $f > F_{0.975,n_1-1,n_2-1}$.

Example. Suppose we have two samples, the first has $n_1 = 7$ observations and a sample variance of $s_1^2 = 25$ and the second sample has $n_2 = 10$ and $s_2^2 = 64$. Then

$$f_{6,9} = \frac{25}{64} = 0.391$$

which is in between the lower and upper cut-off values

```
qf(0.025, 6, 9)

## [1] 0.1810477

qf(0.975, 6, 9)

## [1] 4.319722
```

so we will fail to reject the null hypothesis. Just for good measure, we can calculate the p-value as

$$\begin{aligned} p-value &= 2 \cdot P(F_{n_1-1,n_2-1} < 0.391) \\ &= 2 \cdot P\left(F_{6,9} < 0.391\right) \end{aligned}$$

```
2*pf(0.391, 6, 9)

## [1] 0.2654714
```

We calculate the p-value by finding the area to the left and multiplying by two because my test statistic was less than 1 (the expected value of f if H_0 is true). If my test statistic was greater than 1, we would have found the area to the *right* of f and multiplied by two.

Symmetry of the F-distribution

When testing

$$\begin{aligned} H_0: &\quad \sigma_1^2 = \sigma_2^2 \\ H_a: &\quad \sigma_1^2 \neq \sigma_2^2 \end{aligned}$$

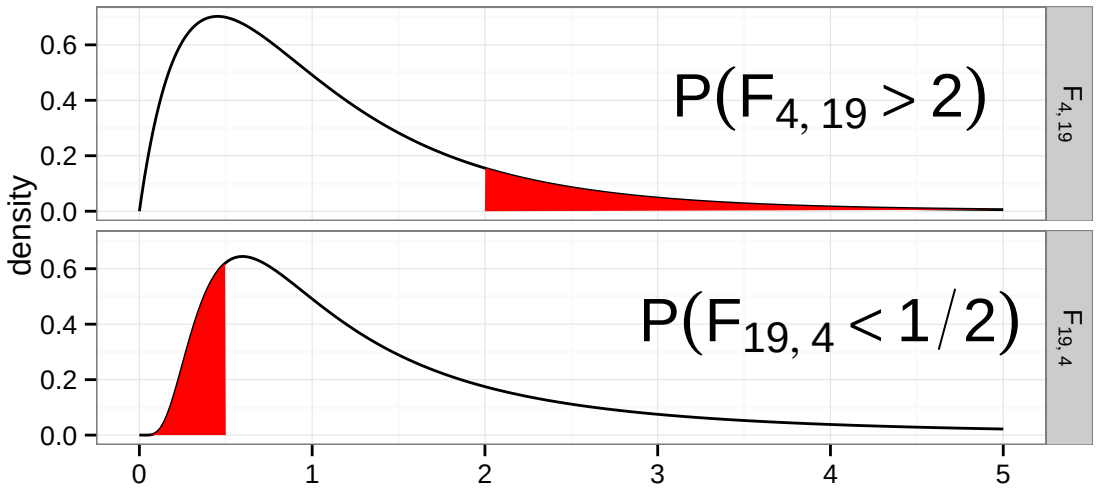
The labeling of group 1 and group 2 is completely arbitrary and I should view $f = s_1^2/s_2^2$ as the same evidence against null as $f^* = s_2^2/s_1^2$. Therefore we have

$$P\left(F_{df_1,\,df_2} > \frac{s_1^2}{s_2^2}\right) = P\left(F_{df_2,\,df_1} < \frac{s_2^2}{s_1^2}\right)$$

For example, suppose that $n_1 = 5$ and $n_2 = 20$ and $s_1^2 = 6$ and $s_2^2 = 3$ then

$$P\left(F_{4,\,19} > \frac{6}{3}\right) = P\left(F_{19,\,4} < \frac{3}{6}\right)$$

```
1 - pf(6/3, 4, 19)

## [1] 0.1354182

pf(3/6, 19, 4)

## [1] 0.1354182
```



Power of the F-test

But how well does this test work? To find out we'll take samples from different normal distributions and test them.

```
sigma1 <- 1
sigma2 <- 2
n1 <- 10
n2 <- 10
v1 <- var(rnorm(n1, mean=0, sd=sigma1))
v2 <- var(rnorm(n2, mean=0, sd=sigma2))
f <- v1/v2
if( f < 1 ){
  p.value <- 2 *      pf( f, df1 = n1-1, df2 = n2-1 )
}else{
  p.value <- 2 * (1 - pf( f, df1 = n1-1, df2 = n2-1))
}
p.value

## [1] 0.1142902
```

So even though the standard deviation in the second sample was twice as large as the first, we were unable to detect it do to the small sample sizes. What happens when we take a larger sample size?

```
sigma1 <- 1
sigma2 <- 2
n1 <- 30
n2 <- 30
v1 <- var(rnorm(n1, mean=0, sd=sigma1))
v2 <- var(rnorm(n2, mean=0, sd=sigma2))
f <- v1/v2
if( f < 1 ){
  p.value <- 2 *      pf( f, df1 = n1-1, df2 = n2-1 )
}else{
  p.value <- 2 * (1 - pf( f, df1 = n1-1, df2 = n2-1))
}
p.value

## [1] 4.276443e-06
```

What this tells us is that just like every other statistical test, *sample size effects the power of the test*. In small sample situations, you cannot rely on a statistical test to tell you if your samples have unequal variance. Instead you need to think about if the assumption is scientifically valid or if you can use a test that does not rely on the equal variance assumption.

Theoretical distribution vs bootstrap

Returning to the research example with the alcohol and control group, an F-test for different variances results in a p-value of

```
# Calculating everything by hand
library(dplyr)
F <- Alcohol %>%
  group_by(trt) %>%                # for each trt group,
  summarise( s2 = var(time)) %>%   # calculate variance.
  summarise( F = s2[1] / s2[2] )   # and then take the ratio
F

## Source: local data frame [1 x 1]
##
##          F
## 1 1.704753
```

```
obs.F <- as.numeric( F )           # Convert 1-by-1 data frame to simple number
pvalue <- 2* (1-pf( obs.F, 9,9 ))
pvalue

## [1] 0.4390223
```

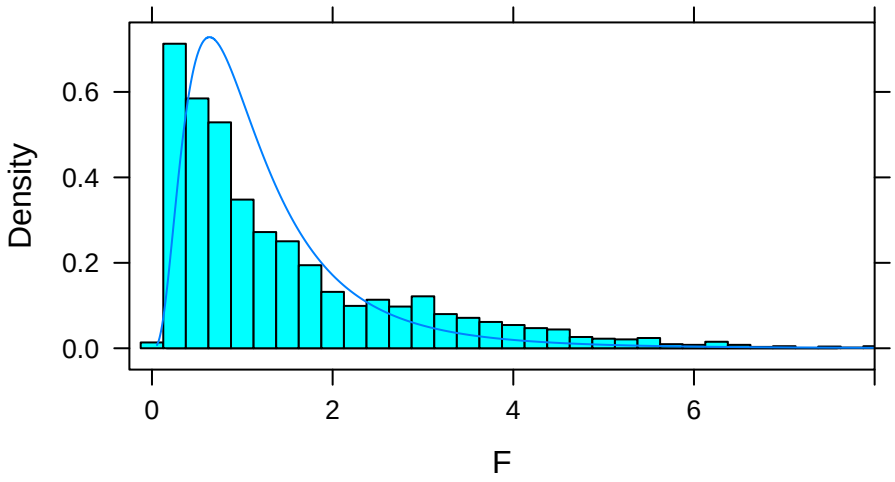
```
# Using Rs built in function
var.test( time ~ trt, data=Alcohol )

##
##  F test to compare two variances
##
## data:  time by trt
## F = 1.7048, num df = 9, denom df = 9, p-value = 0.439
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.4234365 6.8633246
## sample estimates:
## ratio of variances
##           1.704753
```

We can wonder how well the theoretical estimate of the sampling distribution ($F_{9,9}$) compares to the simulation based[2] estimate of the sampling distribution.

```
library(mosaic)
# estimate the sampling distribution of Observed F via simulation
SampDist <- do(5000) *
  var.test(time ~ shuffle(trt), data=Alcohol)$statistic

# Figure which parts of the distribution are more extreme than my observed F
SampDist <- SampDist %>%
  mutate( extreme = F > obs.F | F < 1/obs.F )
```

[2]It is easiest to create F-values under the null distribution of equal variances by shuffling the group labels, but we could do this via bootstrapping by resampling the residuals. We'll see this approach later in the semester.

```
# Make a histogram of the boostrap sampling distribution and theoretical
histogram( ~ F, data=SampDist,
  xlim  = c(-.25,8),  # Don't bother showing huge values
  width = .25)       # Bin Widths of 1/4
plotDist('f', df1=9, df2=9, add=TRUE)  # add the theoretical distiribution on top
```



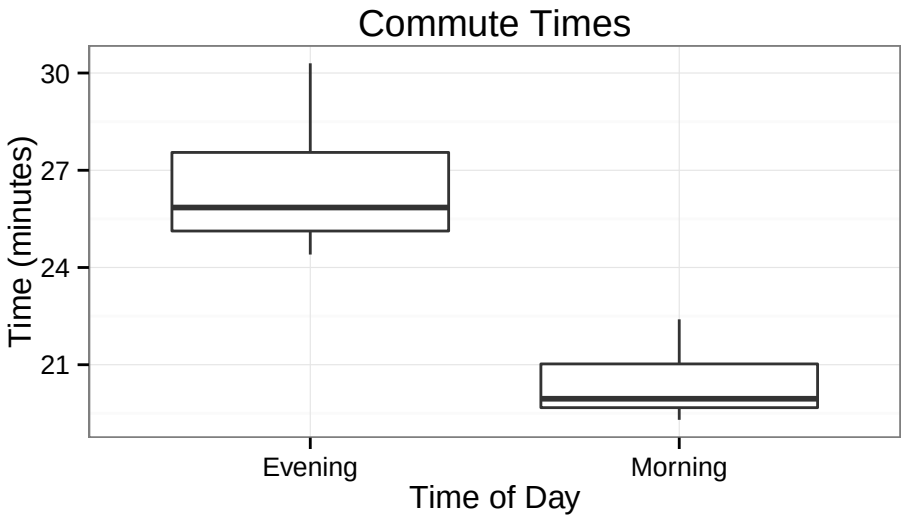

```
# Bootstrap p-value... what percent is more extreme than what I observed?
mosaic::tally( ~ extreme, data=SampDist, format='proportion')

##
##   TRUE  FALSE
## 0.6102 0.3898
```

The theoretical sampling distribution is more concentrated near 1 than the simulation estimate. As a result, the p-value is a bit larger, but in both cases, we cannot reject equal variances.

Example: Lets consider a case where we have two groups of moderate sample sizes where there is a difference in variance. Suppose we consider the set of times in takes me to bike to work in the morning versus biking home[3].

[3]Surprisingly often on the way home I run into other cyclists I know and we stop and chat or we end up riding someplace neither of us has to go.

```
Commute <- data.frame(
  time = c(21.0, 22.1, 19.3, 22.4, 19.6, 19.8,
           19.6, 20.4, 21.1, 19.7, 19.9, 20.0,
           25.0, 27.8, 25.2, 25.1, 25.4, 25.9,
           30.3, 29.5, 25.1, 26.4, 24.4, 27.7,
           25.8, 27.1),
  type = c( rep('Morning',12), rep('Evening',14)))

ggplot(Commute, aes(x=type, y=time)) +
  geom_boxplot() +
  labs(title='Commute Times', y='Time (minutes)', x='Time of Day') +
  theme_bw()
```



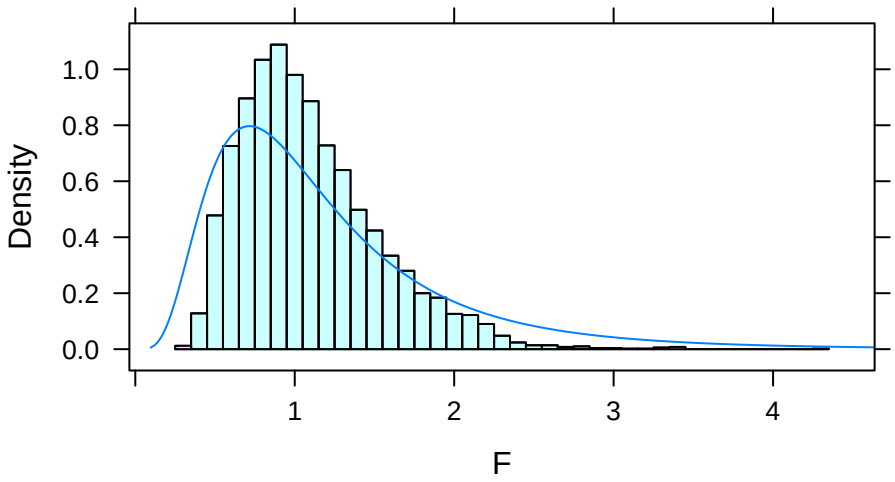
We now test to see if there is a significant difference between the variances of these two groups. If we feel comfortable with assuming that these data come from normal distributions, then the theoretical method is appropriate

```
var.test( time ~ type, data=Commute )

##
##  F test to compare two variances
##
## data:  time by type
## F = 3.039, num df = 13, denom df = 11, p-value = 0.07301
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.8959971 9.7171219
## sample estimates:
## ratio of variances
##           3.038978
```

But if we are uncomfortable with the normality assumption (the Shapiro-Wilks test indicates moderate evidence to reject normality for both samples due to the positive skew in both) we could compare our observed F-statistic to the simulation based estimate of the sampling distribution.

```
# obs.F = 3.04
obs.F <- var.test(time ~ type, data=Commute)$statistic

# estimate the sampling distribution of Observed F via simulation
SampDist <- do(5000) *
  var.test(time ~ shuffle(type), data=Commute)$statistic

# Figure which parts of the distribution are more extreme than my observed F
SampDist <- SampDist %>%
  mutate( extreme = F > obs.F | F < 1/obs.F )   # F > 3.04 or F < 1/3.04
```

```
# Make a histogram of the boostrap sampling distribution and theoretical
histogram( ~ F, data=SampDist,
  width = .1,         # Bin Widths of 1/10
  groups = (extreme==TRUE) )
plotDist('f', df1=13, df2=11, add=TRUE)   # add the theoretical distiribution on top
```




```
# Bootstrap p-value... what proportion is more extreme than what I observed?
mosaic::tally( ~ extreme, data=SampDist, format='proportion')

##
##  TRUE FALSE
## 0.003 0.997
```

We again see that with this small of a data set, our simulation based p-value is different from the theoretical based p-value. This is primarily due to the non-normality of our data along with the small sample sizes. In general as our sample sizes increase the simulation based and theoretical based distributions should give similar inference and p-values.

Chapter 8

Analysis of Variance

Introduction

We are now moving into a different realm of statistics. We have covered enough probability and the basic ideas of hypothesis tests and p-values to move onto the type of inference that you took this class to learn. The heart of science is comparing and evaluating which hypothesis is better supported by the data.

To evaluate a hypothesis, scientists will write a grant, hire grad students (or under-grads), collect the data, and then analyze the data using some sort of model that reflects the hypothesis under consideration. It could be as simple as "What is the relationship between iris species and petal width?" or as complex as "What is the temporal variation in growing season length in response to elevated CO_2 in desert ecosystems?"

At the heart of the question is which predictors should be included in my model of the response variable. Given twenty different predictors, I want to pare them down to just the predictors that matter. I want to make my model as simple as possible, but still retain as much explanatory power as I can.

Our attention now turns to building models of our observed data in a fashion that allows us to ask if a predictor is useful in the model or if we can remove it. Our model building procedure will be consistent:

1. Write two models, one that is perhaps overly simple and another that is a complication of the simple model.
2. Verify that the assumptions that are made in both models are satisfied.
3. Evaluate if the complex model explains significantly more of the variability in the data than the simple model.

Our goal here isn't to find "the right model" because no model is right. Instead our goal is to find a model that is *useful* and helps me to understand the science.

We will start by developing a test that helps me evaluate if a model that has a categorical predictor variable for a continuous response should have a mean value for each group or just one overall mean.

8.1 Model

The two-sample t-test provided a convenient way to compare the means from two different populations and test if they were equal. We wish to generalize this test to more than two different populations.[1]

[1] Later when we have more tools in our statistical tool box, it is useful to notice that ANOVA uses a categorical variable (which group) to predict a continuous response.

Suppose that my data can be written as

$$Y_{ij} = \mu_i + \epsilon_{ij} \quad \text{where } \epsilon_{ij} \overset{iid}{\sim} N(0, \sigma)$$

and μ_i is the mean of group i and ϵ_{ij} are the deviations from the group means. Let the first subscript denote which group the observation is from $i \in \{1, \dots k\}$ and the second subscript is the observation number within that sample. Each group has its own mean μ_i and we might allow the number of observations in each group n_i to be of different across the populations.

Assumptions:

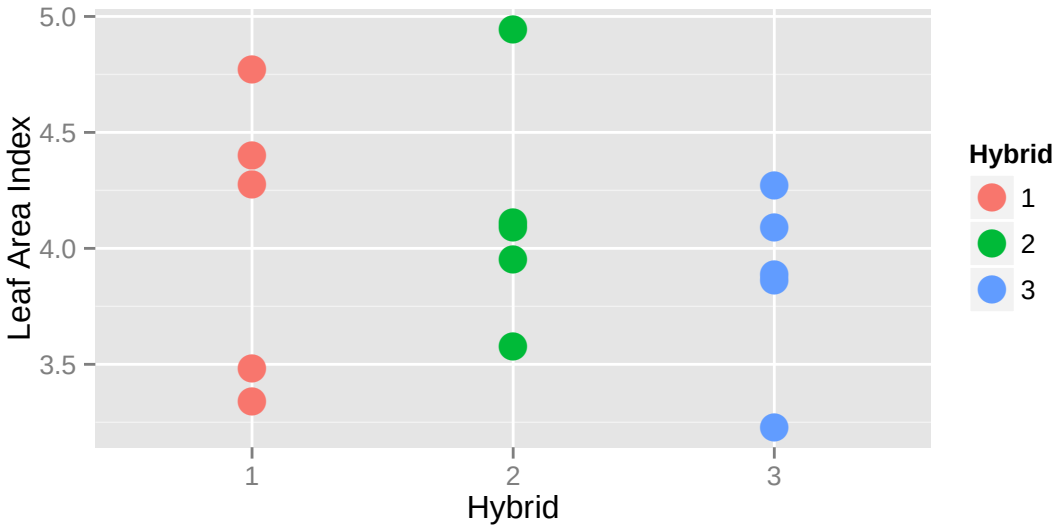
1. *The error terms come from a normal distribution*
2. *The variance of each group is the same*
3. *The observations are independent*
4. *The observations are representative of the population of interest*

In general I want to test the hypotheses

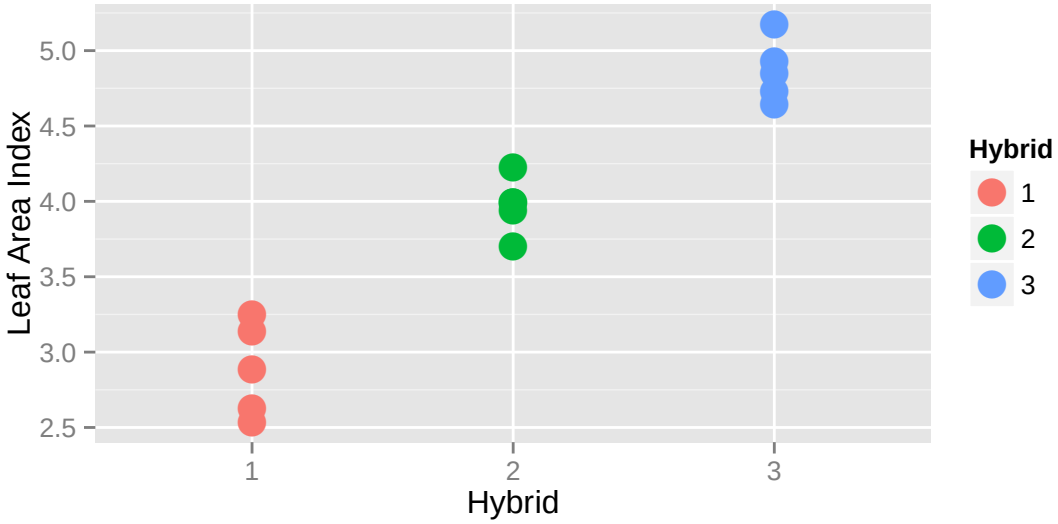
$$\begin{aligned} H_0: &\quad \mu_1 = \mu_2 = \cdots = \mu_k \\ H_a: &\quad \text{at least on mean is different than the others} \end{aligned}$$

Example 12. Suppose that we have three hybrids of a particular plant and we measure the leaf area for each hybrid.

In the following graph, there does not appear to be a difference between the hybrid means:



However, in this case, it looks like there is a difference in the means of each hybrid:



What is the difference between these two?

1. If the variance *between* hybrids is small compared the variance *within* a hybrid variance is huge compared, then I would fail to reject the null hypothesis of equal means (this would be the first case). In this case, the additional model complexity doesn't result in more accurate model, so Occam's Razor would lead us to prefer the simpler model where each group has the same mean.
2. If there is a large variance *between* hybrids compared to the variance *within* a hybrid then I'd conclude there is a difference (this would be the first case). In this case, I prefer the more complicated model with each group having separate means.

8.2 Theory

Notation:

1. $n = n_1 + n_2 + \cdots + n_k$ as the total number of observations
2. $\bar{y}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$ as the sample mean from the ith group
3. $\bar{y}_{\cdot\cdot}$ be the mean of all the observations.

Regardless of if the null hypothesis is true, the following is an estimate of σ^2. We could use a pooled variance estimate similar to the estimator in the pooled two-sample t-test. We will denote this first estimator as the *within-group* estimate because the sums in the numerator are all measuring the variability within a group.

$$\begin{aligned} s_W^2 &= \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_k}\left(y_{ij} - \bar{y}_{i\cdot}\right)^2}{n-k} \\ &= \frac{\sum_{j=1}^{n_1}\left(y_{1j} - \bar{y}_{1\cdot}\right)^2 + \sum_{j=1}^{n_2}\left(y_{2j} - \bar{y}_{2\cdot}\right)^2 + \cdots + \sum_{j=1}^{n_k}\left(y_{kj} - \bar{y}_{k\cdot}\right)^2}{(n_1-1) + (n_2-1) + \cdots + (n_k-1)} \\ &= \frac{(n_1-1)\,s_1^2 + (n_2-1)\,s_2^2 + \cdots + (n_k-1)\,s_k^2}{n-k} \end{aligned}$$

If the null hypothesis is true and $\mu_1 = \cdots = \mu_k$, then a second way that I could estimate the σ^2 is using the sample means. If H_0 is true then each sample mean has sampling distribution

$\bar{Y}_{i\cdot} \sim N\left(\mu, \frac{\sigma^2}{n_i}\right)$ and therefore the k sample means could be used to estimate σ^2.

$$s_B^2 = \frac{1}{k-1}\sum_{i=1}^{k} n_i \left(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}\right)^2$$

Under the null hypothesis, these two estimates are both estimating σ^2 and should be similar and the ratio s_B^2/s_W^2 follows an F-distribution with numerator degrees of freedom $k-1$ and denominator degrees of freedom $n-k$ degrees of freedom. We define our test statistic as

$$f = \frac{s_B^2}{s_W^2}$$

In the case that the null hypothesis is false, s_B^2 should be much larger than s_W^2 and our test statistic f will be very large and so we will reject the null hypothesis if f is greater than the $1-\alpha$ quantile from the F-distribution with $k-1$ and $n_t - k$ degrees of freedom. If s_B^2 is small, then the difference between the group means and the overall means is small and we shouldn't reject the null hypothesis. So this F-test will always be a one sided test, rejecting only if f is large.

$$p - value = P\left(F_{k-1,\, n_t - k} > f\right)$$

8.2.1 Anova Table

There are several sources of variability that we are dealing with.

SSW: Sum of Squares Within - This is the variability within sample groups. It has an associated $df_W = n - k$

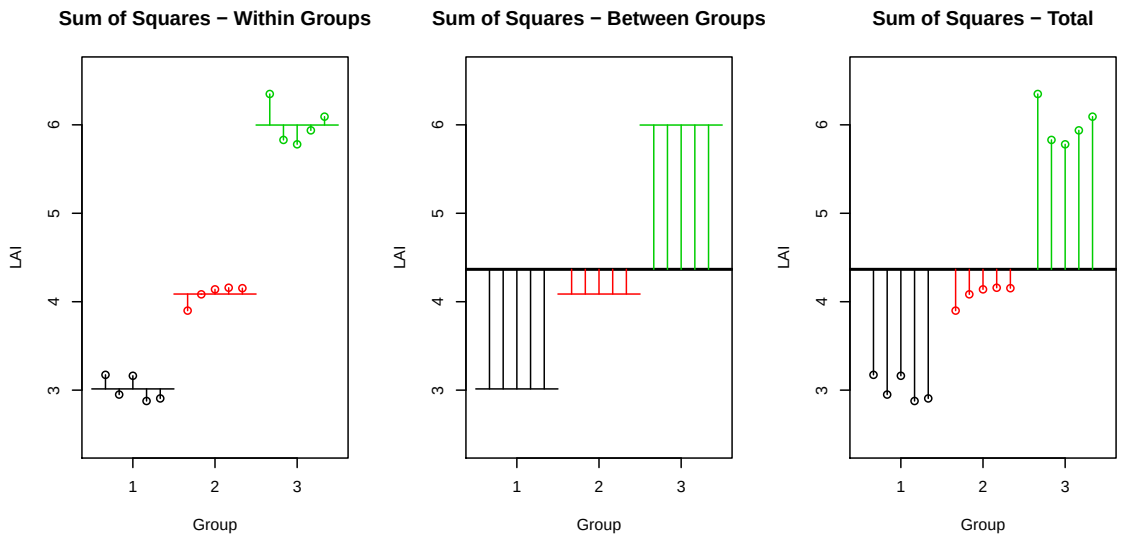
$$SSW = \sum_{i=1}^{k}\sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_{i\cdot}\right)^2$$

SSB: Sum of Squares Between - This is the variability between sample groups. It has an associated $df_B = k - 1$

$$SSB = \sum_{i=1}^{k} n_i \left(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}\right)^2$$

SST: Sum of Squares Total - This is the total variability in the data set. It has an associated $df = n - 1$ because under the null hypothesis there is only one mean μ.

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_j} \left(y_{ij} - \bar{y}_{\cdot\cdot}\right)^2$$



An anova table is usually set up the in the following way (although the total row is sometimes removed):

Source	df	Sum of Squares	Mean Squares	F-stat	P-value
Between Samples	$k-1$	SSB	$s_B^2 = SSB/(k-1)$	$f = s_B^2/s_W^2$	$P(F_{k-1,n-k} > f)$
Within Samples	$n-k$	SSW	$s_W^2 = SSW/(n_t-k)$		
Total	$n-1$	SST			

It can be shown that

$$SST = SSB + SSW$$

and we can think about what these sums actually mean by returning to our idea about simple vs complex models.

8.2.2 ANOVA using Simple vs Complex models.[2]

The problem under consideration boils down to how complicated of a model should we fit.

Simple

The simple model is

$$Y_{ij} = \mu + \epsilon_{ij} \quad \text{where } \epsilon_{ij} \overset{iid}{\sim} N\left(0, \sigma^2\right)$$

has each observation having the same expectation μ. Thus we use the overall mean of the data $\bar{y}_{\cdot\cdot}$ as the estimate of μ and therefore our error terms are

$$e_{ij} = y_{ij} - \bar{y}_{\cdot\cdot}$$

[2]Upon the second reading of these notes, the student is likely asking why we even bothered introducing the ANOVA table using SST, SSW, SSB. The answer is that these notations are common in the ANOVA literature and that we can't justify using an F-test without variance estimates. Both interpretations are valid, but the Simple/Complex models are a better paradigm as we move forward.

The sum of squared error associated with the simple model is thus

$$\begin{aligned} SSE_{simple} &= \sum_{i=1}^{k}\sum_{j=1}^{n_i} e_{ij}^2 \\ &= \sum_{i=1}^{k}\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2 \\ &= SST \end{aligned}$$

Complex

The more complicated model

$$Y_{ij} = \mu_i + \epsilon_{ij} \quad \text{where } \epsilon_{ij} \overset{iid}{\sim} N\left(0, \sigma^2\right)$$

has each observation having the expectation of it's group mean μ_i. We'll use the group means $\bar{y}_{i.}$ as estimates for μ_i and thus the error terms are

$$e_{ij} = y_{ij} - \bar{y}_{i.}$$

and the sum of squared error associated with the complex model is thus

$$\begin{aligned} SSE_{complex} &= \sum_{i=1}^{k}\sum_{j=1}^{n_i} e_{ij}^2 \\ &= \sum_{i=1}^{k}\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2 \\ &= SSW \end{aligned}$$

Difference

The difference between the simple and complex sums of squared error is denoted SSE_{diff} and we see

$$\begin{aligned} SSE_{diff} &= SSE_{simple} - SSE_{complex} \\ &= SST - SSW \\ &= SSB \end{aligned}$$

Note that SSE_{diff} can be interpreted as the amount of variability that is explained by the more complicated model vs the simple. If this SSE_{diff} is large, then we should use the complex model. Our only question becomes "How large is large?"

First we must account for the number of additional parameters we have added. If we added five parameters, I should expect to account for more variability that if I added one parameter, so first we will divide SSE_{diff} by the number of added parameters to get MSE_{diff} which is the amount of variability explained by each additional parameter. If that amount is large compared to the leftover from the complex model, then we should use the complex model.

These calculations are preformed in the ANOVA table, and the following table is identical to the previous ANOVA table, and we have only changed the names given to the various quantities.

Source	df	Sum of Squares	Mean Squares	F-stat	P-value
Difference	$k-1$	SSE_{diff}	$MSE_{diff} = \frac{SSE_{diff}}{k-1}$	$f = \frac{MSE_{diff}}{MSE_{complex}}$	$P\left(F_{k-1,n-k} > f\right)$
Complex	$n-k$	$SSE_{complex}$	$MSE_{complex} = \frac{SSE_{complex}}{n-k}$		
Simple	$n-1$	SSE_{simple}			

8.2.3 Parameter Estimates and Confidence Intervals

As usual, the sample mean $\bar{y}_{i\cdot}$ is a good estimator for the mean of group μ_i.

But what about σ^2? If we conclude that we should use the complex model, and since one of our assumptions is that each group has equal variance, then I should use all of the residual terms $e_{ij} = y_{ij} - \bar{y}_{i\cdot}$ in my estimation of σ. In this case we will use

$$\hat{\sigma}^2 = s_W^2 = MSE_{complex} = \frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2$$

as the estimate of σ^2. Notice that this is analogous to the pooled estimate of the variance in a two-sample t-test with the assumption of equal variance.

Therefore an appropriate confidence interval for μ_i is

$$\bar{y}_{i\cdot} \pm t_{n-k}^{1-\alpha/2}\left(\frac{\hat{\sigma}}{\sqrt{n_i}}\right)$$

8.3 Anova in R

First we must define a data frame with the appropriate columns. We start with two vectors, one of which has the leaf area data and the other vector denotes the species. Our response variable must be a continuous random variable and the explanatory is a discrete variable. In R discrete variables are called `factors` and can you can change a numerical variable to be a factor using the function `factor()`.

The analysis of variance method is an example of a linear model which can be fit in a variety of ways. We can use either `lm()` or `aov()` to fit this model, and the following we will concentrate on using `aov()`. The first argument to this function is a formula that describes the relationship between the explanatory variables and the response variable. In this case it is extremely simple, that `LAI` is a function of the categorical variable `Species`.

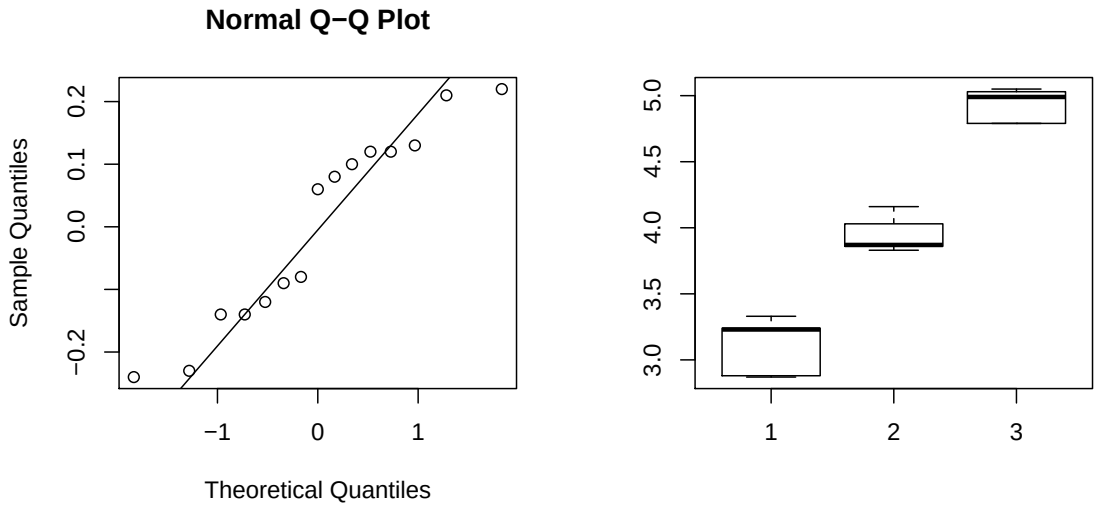
```
data <- data.frame(LAI = c(2.88, 2.87, 3.23, 3.24, 3.33,
                           3.83, 3.86, 4.03, 3.87, 4.16,
                           4.79, 5.03, 4.99, 4.79, 5.05),
                   Species = factor( rep(1:3, each=5) ) )
str(data)

## 'data.frame': 15 obs. of  2 variables:
##  $ LAI    : num  2.88 2.87 3.23 3.24 3.33 3.83 3.86 4.03 3.87 4.16 ...
##  $ Species: Factor w/ 3 levels "1","2","3": 1 1 1 1 1 2 2 2 2 2 ...

model <- aov(LAI ~ Species, data=data)
```

The `aov()` command is the command that does all the calculations necessary to fit an ANOVA model. This command returns a list object that is useful for subsequent analysis and it is up to the use to know what subsequent functions to call that answer questions of interest.

In the call to `aov()` we created a formula. Formulas in R always are of the form `Y ~ X` where `Y` is the dependent variable and the `X` variables are the independent variables. In the formula we passed to `aov()`, we used a `LAI ~ Species`.

Before we examine the anova table and make any conclusion, we should double check that the anova assumptions have been satisfied. To check the normality assumption, we will look at the qqplot of the residuals $e_{ij} = y_{ij} - \bar{y}_{i\cdot}$. These residuals are easily accessed in R using the `resid` function on the object `model`. To check the variance assumption, we will examine the boxplot of the data

```
par(mfrow=c(1,2)) # side-by-side plots...
qqnorm( resid(model) )
qqline( resid(model) )
boxplot(LAI~Species, data=data)
```



The qqplot doesn't look too bad, with only two observations far from the normality line. The equal variance assumption seems acceptable as well. To get the Analysis of Variance table, we'll extract it from the `model` object using the function `anova()`.

```
model <- aov(LAI ~ Species, data=data)
anova(model)

## Analysis of Variance Table
##
## Response: LAI
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Species    2 8.2973  4.1487  147.81 3.523e-09 ***
## Residuals 12 0.3368  0.0281
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Notice that R does not give you the third line in the ANOVA table. This was a deliberate choice by the Core Development Team of R, but one that is somewhat annoying[3]. Because the third line is just the total of the first two, it isn't hard to calculate, if necessary.

The row labeled `Species` corresponds to the difference between the simple and complex models, while the `Residuals` row corresponds to the complex model. Notice that SSE_{diff} is quite large, but to decide if it is large enough to justify the use of the complex model, we must go through the calculations to get the p-value, which is quite small. Because the p-value is smaller than any reasonable α-level, we can reject the null hypothesis and conclude that at least one of the means is different than the others.

But which mean is different? The first thing to do is to look at the point estimates and confidence intervals for μ_i. These are

$$\hat{\mu}_i = \bar{y}_{i\cdot}$$

[3]The package `NCStats` modifies the print command for an `aov` object to create the missing third row.

$$\hat{y}_{i\cdot} \pm t_{n_t-k}^{1-\alpha/2}\left(\frac{\hat{\sigma}}{\sqrt{n_i}}\right)$$

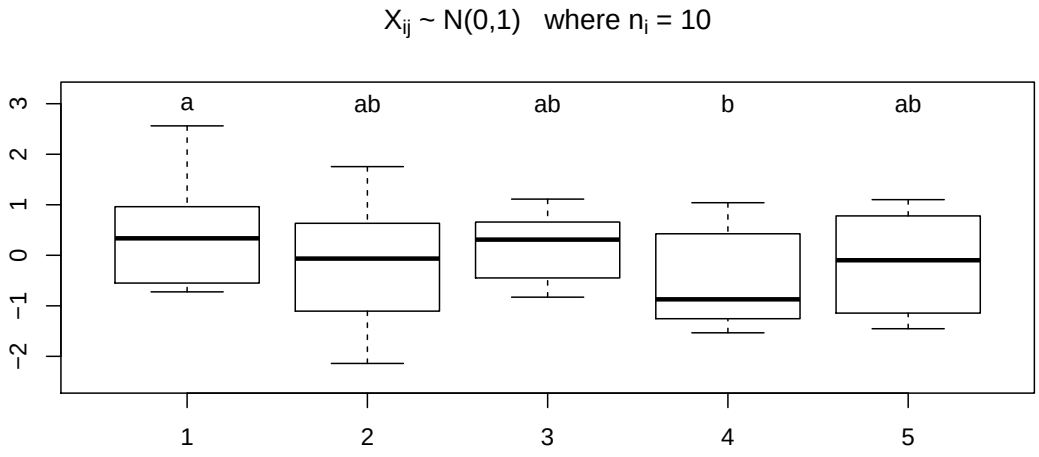
and can be found using the `coef()` and `confint()` functions.

```
# To get coefficients in the way we have represented the
# complex model (which we call the cell means model), we
# must add a -1 to the formula passed to aov()
# We'll explore this more in section 5 of this chapter.
model.2 <- aov(LAI ~ Species - 1, data=data)
coef(model.2)

## Species1 Species2 Species3
##     3.11     3.95     4.93

confint(model.2)

##              2.5 %   97.5 %
## Species1 2.946759 3.273241
## Species2 3.786759 4.113241
## Species3 4.766759 5.093241
```

Are the all the species different from each other? In practice I will want to examine each group and compare it to all others and figure out if they are different. How can we efficiently do all possible t-tests and keep the correct α level correct?

8.4 Multiple comparisons

Recall that for every statistical test there is some probability of making a type I error and we controlled that probability by setting a desired α-level. If I were to do 20 t-tests of samples with identical means, I would expect, on average, that one of them would turn up to be significantly different just by chance. If I am making a large number of tests, each with a type I error rate of α, I am practically guaranteed to make at least one type I error.



With 5 groups, there are 10 different comparisons to be made, and just by random chance, one of those comparisons might come up significant. In this sampled data, performing 10 different two

sample t-tests without making any adjustments to our α-level, we find one statistically significant difference even though all of the data came from a standard normal distribution.

I want to be able to control the family-wise error rate so that the probability that I make one or more type I errors in the set of m of tests I'm considering is α. One general way to do this is called the Bonferroni method. In this method each test is performed using a significance level of α/m. (In practice I will multiple each p-value by m and compare each p-value to my desired family-wise α-level). Unfortunately for large m, this results in unacceptably high levels of type II errors. Fortunately there are other methods for addressing the multiple comparisons issue and they are built into R.

John Tukey's test of "Honestly Significant Differences" is commonly used to address the multiple comparisons issue when examining all possible pairwise contrasts. This method is available in R by the function `TukeyHSD`. This test is near optimal when each group has the same number of samples (which is often termed "a balanced design"), but becomes more conservative (fails to detect differences) as the design becomes more unbalanced. In extremely unbalanced cases, it is preferable to use a Bonferroni adjustment.

Using TukeyHSD, the adjusted p-value for the difference between groups 1 and 4 is no longer significant.

```
# TukeyHSD is very picky and will not accept Y ~ Group - 1
model <- aov(Y~Group, mydata)
TukeyHSD(model)

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Y ~ Group, data = mydata)
##
## $Group
##            diff        lwr       upr     p adj
## 2-1 -0.55735682 -1.8277879 0.7130742 0.7244152
## 3-1 -0.23996214 -1.5103932 1.0304689 0.9830031
## 4-1 -0.98855350 -2.2589845 0.2818775 0.1943377
## 5-1 -0.62440394 -1.8948350 0.6460271 0.6330050
## 3-2  0.31739468 -0.9530364 1.5878257 0.9531756
## 4-2 -0.43119668 -1.7016277 0.8392344 0.8695429
## 5-2 -0.06704712 -1.3374782 1.2033839 0.9998817
## 4-3 -0.74859136 -2.0190224 0.5218397 0.4596641
## 5-3 -0.38444180 -1.6548728 0.8859892 0.9099064
## 5-4  0.36414956 -0.9062815 1.6345806 0.9248234
```

Likewise if we are testing the ANOVA assumption of equal variance, we cannot rely on doing all pairwise F-tests and we must use a method that controls the overall error rate. The multiple comparisons version of `var.test()` is Bartlett's test which is called similarly to `aov()`.

```
bartlett.test(Y~Group, mydata)

##
##  Bartlett test of homogeneity of variances
##
## data:  Y by Group
## Bartlett's K-squared = 3.1397, df = 4, p-value = 0.5347
```

Example 13. (Example 8.2 from the Ott and Longnecker) A clinical psychologist wished to compare three methods for reducing hostility levels in university students, and used a certain test (HLT) to measure the degree of hostility. A high score on the test indicated great hostility. The psychologist

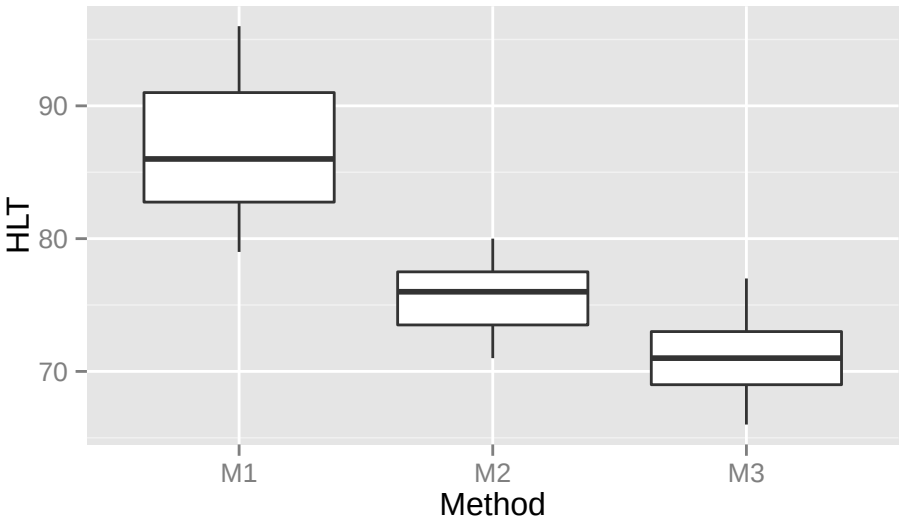
used 24 students who obtained high and nearly equal scores in the experiment. Eight subjects were selected at random from among the 24 problem cases and were treated with method 1, seven of the remaining 16 students were selected at random and treated with method 2 while the remaining nine students were treated with method 3. All treatments were continued for a one-semester period. Each student was given the HLT test at the end of the semester, with the results show in the following table. Use these dat to perform an analysis of variance to determine whether there are differences among the mean scores for the three methods using a significance level of $\alpha = 0.05$.

Method	Test Scores								
1	96	79	91	85	83	91	82	87	
2	77	76	74	73	78	71	80		
3	66	73	69	66	77	73	71	70	74

```r
# define the data
Hostility <- data.frame(
  HLT = c(96,79,91,85,83,91,82,87,
          77,76,74,73,78,71,80,
          66,73,69,66,77,73,71,70,74),
  Method = c( rep('M1',8), rep('M2',7), rep('M3',9) ) )
```

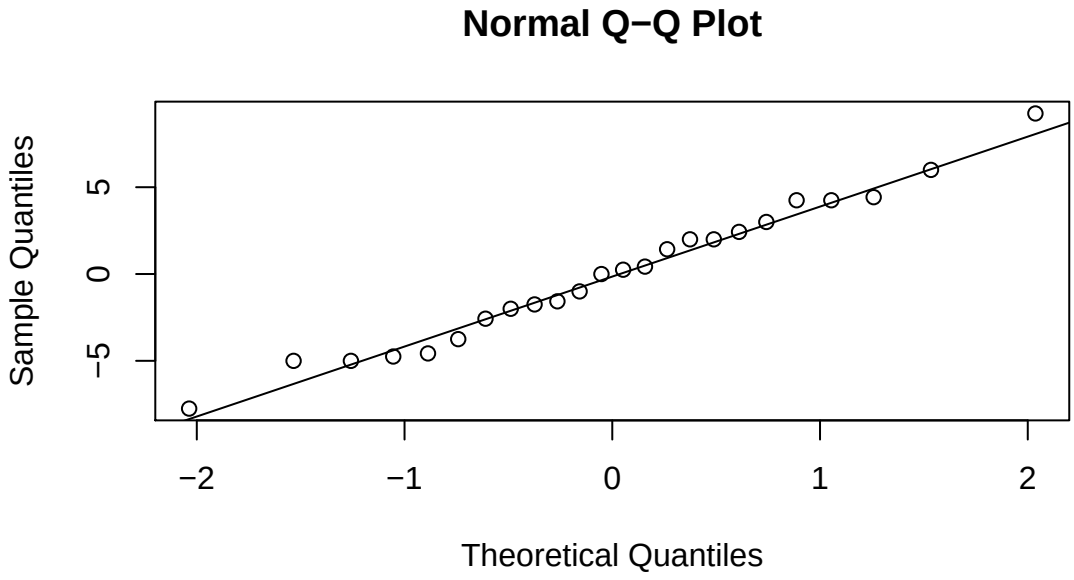
The first thing we will do (as we should do in all data analyses) is to graph our data.

```r
library(ggplot2)
ggplot(Hostility, aes(x=Method, y=HLT)) +
  geom_boxplot()
```



These box plots make it clear that there is a difference between the three groups (at least group `M1` is different from `M2` or `M3`). An ANOVA model assumes equal variance between groups and that the residuals are normally distributed. Based on the box plot, the equal variance assumption might be suspect (although with only ≈ 8 observations per group, it might not be bad). We'll examine a QQ-plot of the residuals to consider the normality.

```
# Do the model assumptions hold?
model <- aov( HLT ~ Method, data=Hostility )
qqnorm( resid(model) )
qqline( resid(model) )
```



To examine the Normality of the residuals, we'll use a Shapiro-Wilk's test and we'll also use Bartlett's test for homogeneity of variances.

```
# Test for Normality
shapiro.test(resid(model))

##
##  Shapiro-Wilk normality test
##
## data:  resid(model)
## W = 0.9836, p-value = 0.9516

# Test for equal variances between groups
bartlett.test(HLT~Method, data=Hostility)

##
##  Bartlett test of homogeneity of variances
##
## data:  HLT by Method
## Bartlett's K-squared = 2.4594, df = 2, p-value = 0.2924
```

The results of the Shapiro-Wilks test agree with the QQ-plot, and Bartlett's test fails to detect differences in the variances between the two groups. This is not to say that there might not be a difference, only that we do not detect one.

```
model <- aov( HLT ~ Method, data=Hostility )
anova(model)

## Analysis of Variance Table
##
## Response: HLT
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## Method     2 1090.62  545.31  29.574 7.806e-07 ***
## Residuals 21  387.21   18.44
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Because the p-value in the ANOVA table is smaller than $\alpha = 0.05$, we can reject the null hypothesis of equal means and conclude that at least one of the means is different from the others. Our estimate of σ^2 is 18.44 so the estimate of $\sigma = \sqrt{18.44} = 4.294$.

To find out which means are different we first look at the group means and confidence intervals.

```
# To get the group means from aov, we must
# use the -1 in the formula command
model.2 <- aov( HLT ~ Method - 1, data=Hostility )
coef(model.2)

## MethodM1 MethodM2 MethodM3
## 86.75000 75.57143 71.00000

confint(model.2)

##             2.5 %   97.5 %
## MethodM1 83.59279 89.90721
## MethodM2 72.19623 78.94663
## MethodM3 68.02335 73.97665
```

To control for the multiple comparisons issue we again look at all possible group comparisons using the `TukeyHSD` function.

```
# Remember TukeyHSD is picky and doesn't like the -1...
TukeyHSD(model)

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = HLT ~ Method, data = Hostility)
##
## $Method
##             diff       lwr         upr     p adj
## M2-M1 -11.178571 -16.78023  -5.5769151 0.0001590
## M3-M1 -15.750000 -21.00924 -10.4907592 0.0000006
## M3-M2  -4.571429 -10.02592   0.8830666 0.1113951
```

If we feel uncomfortable with the equal variance assumption, we can do each pairwise t-test using non-pooled variance and then correct for the multiple comparisons using Bonferroni's p-value correction. If we have $k = 3$ groups, the we have $k(k-1)/2 = 3$ different comparisons, so I will calculate each p-value and multiply by 3.

```
pairwise.t.test(Hostility$HLT, Hostility$Method,
                pool.sd=FALSE, p.adjust.method='none')

##
##  Pairwise comparisons using t tests with non-pooled SD
##
## data:  Hostility$HLT and Hostility$Method
##
##    M1      M2
## M2 0.0005  -
## M3 2.2e-05 0.0175
##
## P value adjustment method: none

pairwise.t.test(Hostility$HLT, Hostility$Method,
                pool.sd=FALSE, p.adjust.method='bonferroni')

##
##  Pairwise comparisons using t tests with non-pooled SD
##
## data:  Hostility$HLT and Hostility$Method
##
##    M1      M2
## M2 0.0015  -
## M3 6.7e-05 0.0525
##
## P value adjustment method: bonferroni
```

Using the Bonferroni adjusted p-values, we continue to detect a statistically significant difference between Method 1 and both Methods 2 & 3, but do not detect a difference between Method 2 and Method 3.

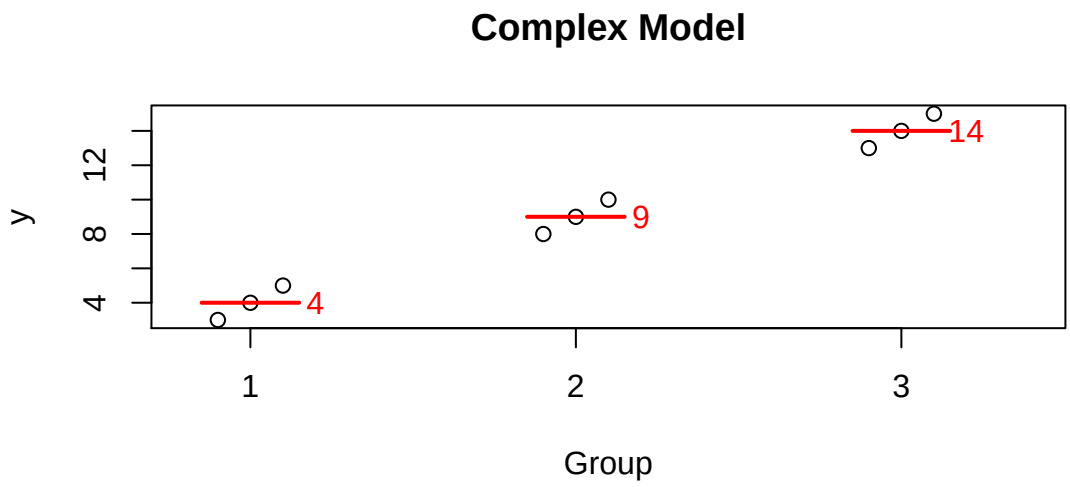
8.5 Different Model Representations

8.5.1 Theory

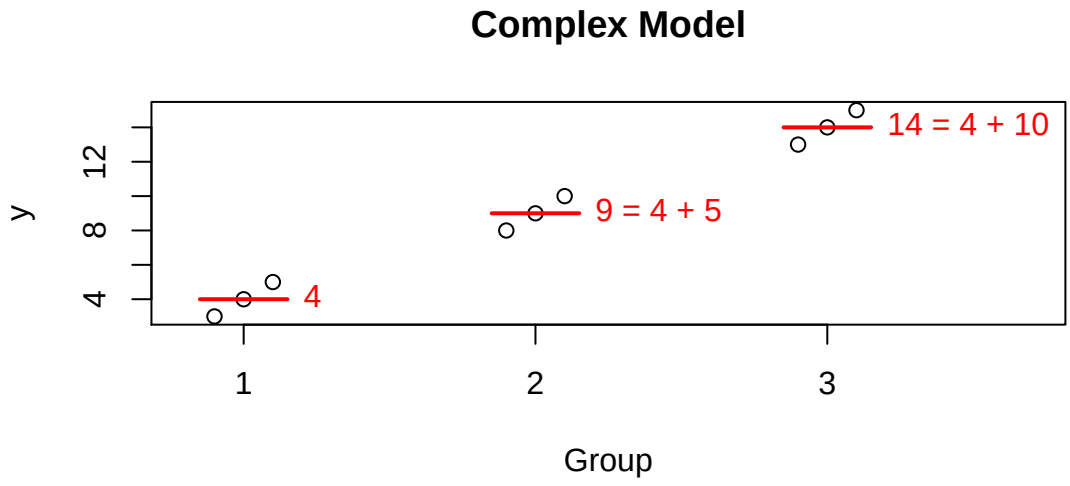
We started with what I will call the "cell means model"

$$Y_{ij} = \mu_i + \epsilon_{ij} \quad \text{where} \quad \epsilon_{ij} \overset{iid}{\sim} N\left(0, \sigma^2\right)$$

so that the $E\left(Y_{ij}\right) = \mu_i$ where I interpret μ_i as the mean of each population. Given some data, we the following graph where the red lines and numbers denote the observed mean of the data in each group :



But I am often interested in the difference between one group and another. For example, suppose this data comes from an experiment and group 1 is the control group. Then perhaps what I'm really interested is not that group 2 has a mean of 9, but rather that it is 5 units larger than the control. In this case perhaps what we care about is the differences. I could re-write the group means in terms of these differences from group 1. So looking at the model this way, the values that define the group means are the mean of group 1 (here it is 4), and the offsets from group 1 to group 2 (which is 5), and the offset from group 1 to group 3 (which is 10).



I could write this interpretation of the model as the "offset" model which is

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij}$$

where μ is the mean of group 1 and τ_i is each population's offset from group 1. Since group 1 can't be offset from itself, this forces $\tau_1 = 0$.

Notice that this representation of the complex model has 4 parameters (aside from σ), but it has an additional constraint so we still only have 3 parameters that can vary (just as the cell means model has 3 means).

The cell means model and the offset model really are the same model, just looked at slightly differently. They have the same number of parameters, and produce the same predicted values for

$\hat{y}_{ij}$ and therefore have the same sum of squares, etc. The only difference is that one is might be more convenient depending on the question the investigator is asking.

Another way to write the cell means model is as

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij}$$

but with the constraint that $\mu = 0$. It doesn't matter which constraint you use so long as you know which is being used because the interpretation of the values changes (group mean versus an offset from the reference group).

8.5.2 Model Representations in R

To obtain the different representations within R, you must use the `-1` term that we have already seen

```
fake.data <- data.frame(   y =        c( 3,4,5,  8,9,10, 13,14,15),
                         grp = factor(c( 1,1,1,  2,2,2,   3,3,3 )) )
# Offset representation
#   Unless you have a -1, R implicitly
#   adds a "+1" to the formula, so
#   so the following statements are equivalent
#c.model.1 <- aov(y ~ grp  , data=fake.data)
 c.model.1 <- aov(y ~ grp+1, data=fake.data)
coef(c.model.1)

## (Intercept)        grp2        grp3
##           4           5          10
```

In the above case, we see R giving the mean of group 1 and then the two offsets.

To force R to use the cell means model, we force R to use the constraint that $\mu = 0$ by including a `-1` in the model formula.

```
c.model.1 <- aov(y ~ grp -1, data=fake.data)
coef(c.model.1)

## grp1 grp2 grp3
##    4    9   14
```

Returning the hostility example, recall we used the cell means model and we can extract parameter coefficient estimates using the `coef` function and ask for the appropriate confidence intervals using `confint()`.

```
model <- aov(HLT ~ Method - 1, data=Hostility)
coef(model)

## MethodM1 MethodM2 MethodM3
## 86.75000 75.57143 71.00000

confint(model)

##             2.5 %   97.5 %
## MethodM1 83.59279 89.90721
## MethodM2 72.19623 78.94663
## MethodM3 68.02335 73.97665
```

We can use the intercept model by removing `-1` term from the formula.

```
model <- aov(HLT ~ Method, data=Hostility)
coef(model)

## (Intercept)    MethodM2    MethodM3
##    86.75000   -11.17857   -15.75000

confint(model)

##                 2.5 %     97.5 %
## (Intercept)  83.59279  89.907212
## MethodM2    -15.80026  -6.556886
## MethodM3    -20.08917 -11.410827
```

The intercept term in the offset representation corresponds to `Method1` and the coefficients and confidence intervals are the same as in the cell means model. However in the offset model, `Method2` is the *difference* between `Method1` and `Method2`. Notice the coefficient is negative, thus telling us that `Method2` has a smaller mean value than the reference group `Method1`. Likewise `Method3` has a negative coefficient indicating that the `Method3` group is lower than the reference group.

Similarly the confidence intervals for `Method2` and `Method3` are now confidence intervals for the *difference* between these methods and the reference group `Method1`.

Why would we ever want the offset model vs the cell means model? Often we are interested in testing multiple treatments against a control group and we only care about the change from the control. In that case, setting the control group to be the reference makes sense.

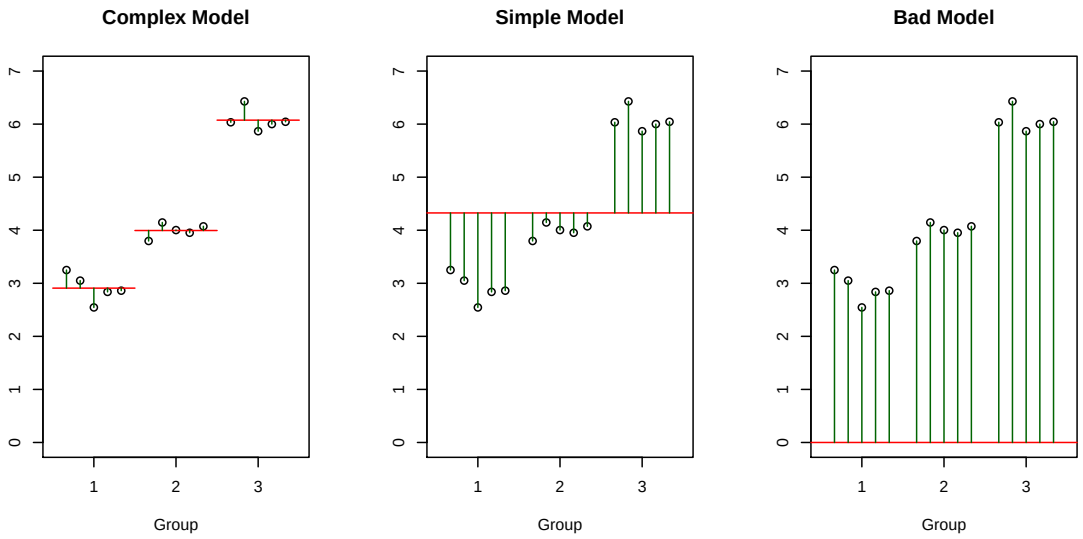
Neither representation is more powerful because on a very deep mathematical level, they are exactly the same model. Superficially though, one representation might be more convenient than the other in a given situation.

8.5.3 Implications on the ANOVA table

We have been talking about the complex and simple models for our data but there is one more possible model, albeit not a very good one. I will refer to this as the **bad model** because it is almost always a poor fitting model.

$$Y_{ij} = \epsilon_{ij}$$

where $\epsilon_{ij} \overset{iid}{\sim} N\left(0, \sigma^2\right)$.



Notice that the complex model has three parameters that define "signal" part of the model (i.e. the three group means). The simple has one parameter that defines the "signal" (the overall mean). The bad model has *no* parameters that define the model (i.e. the red line is always at zero).

These three models can be denoted in R by:

- Complex:
  - offset representation: `Y ~ group` which R will recognize as `Y ~ group + 1`
  - cell means representation: `Y ~ group - 1`
- Simple: `Y ~ 1`
- Bad: `Y ~ -1`

In the analysis of variance table calculated by `anova()`, R has to decide which simple model to compare the complex model to. If you used the offset representation, then when `group` is removed from the model, we are left with the model `Y ~ 1`, which is the simple model. If we wrote the complex model using the cell means representation, then when `group` is removed, we are left with the model `Y ~ -1` which is the bad model.

When we produce the ANOVA table compare the complex to the bad model, the difference in number of parameters between the models will be 3 (because I have to add three parameters to go from a signal line of 0, to three estimated group means). The ANOVA table comparing simple model to the complex will have a difference in number of parameters of 2 (because the simple mean has 1 estimated value compared to 3 estimated values).

Example. Hostility Scores

We return to the hostility scores example and we will create the two different model representations in R and see how the ANOVA table produced by R differs between the two.

```
offset.representation <- aov(HLT ~ Method, data=Hostility)
cell.representation   <- aov(HLT ~ Method -1, data= Hostility)
#
#
# This is the ANOVA table we want, comparing Complex to Simple
# Notice the df of the difference between the models is 3-1 = 2
anova(offset.representation)

## Analysis of Variance Table
##
## Response: HLT
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## Method     2 1090.62  545.31  29.574 7.806e-07 ***
## Residuals 21  387.21   18.44
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#
#
# This is the ANOVA table comparing the Complex to the BAD model
# Noice the df of the difference between the models is 3-0 = 3
anova(cell.representation)

## Analysis of Variance Table
##
## Response: HLT
```

```
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Method     3 145551   48517  2631.2 < 2.2e-16 ***
## Residuals 21    387      18
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Because the bad model is *extremely* bad in this case, the F-statistic for comparing the complex to the bad model is extremely large (F=2631). The complex model is also superior to the simple model, but not by as emphatically (F=29).

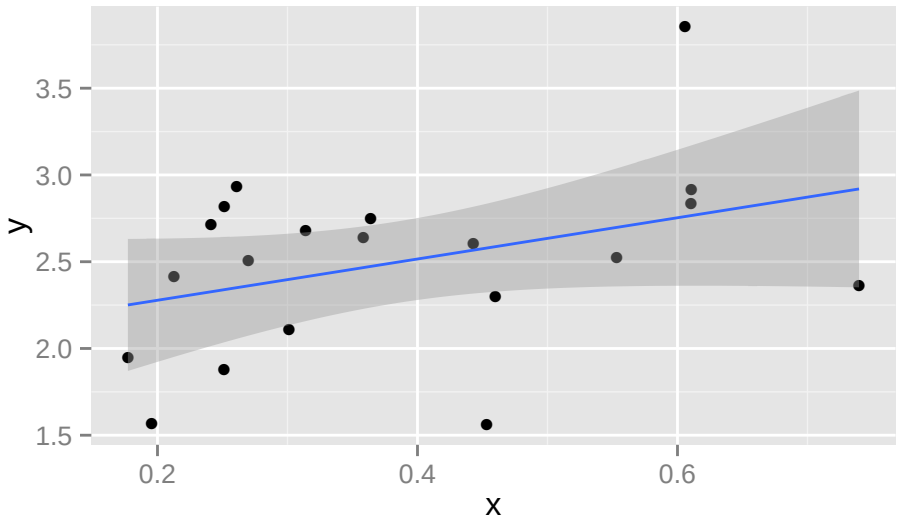
One way to be certain which models you are comparing is to explicitly choose the two models.

```
simple <- aov(HLT ~ 1, data=Hostility)

# create the ANOVA table comparing the complex model (using the
# cell means representation) to the simple model.
# The output shown in the following contains all the
# necessary information, but is arranged slightly differently.
anova(simple, cell.representation)

## Analysis of Variance Table
##
## Model 1: HLT ~ 1
## Model 2: HLT ~ Method - 1
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     23 1477.83
## 2     21  387.21  2    1090.6 29.574 7.806e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Chapter 9

Regression

We continue to want to examine the relationship between a predictor variable and a response but now we consider the case that the predictor is continuous and the response is also continuous. In general we are going to be interested in finding the line that best fits the observed data and determining if we should include the predictor variable in the model.

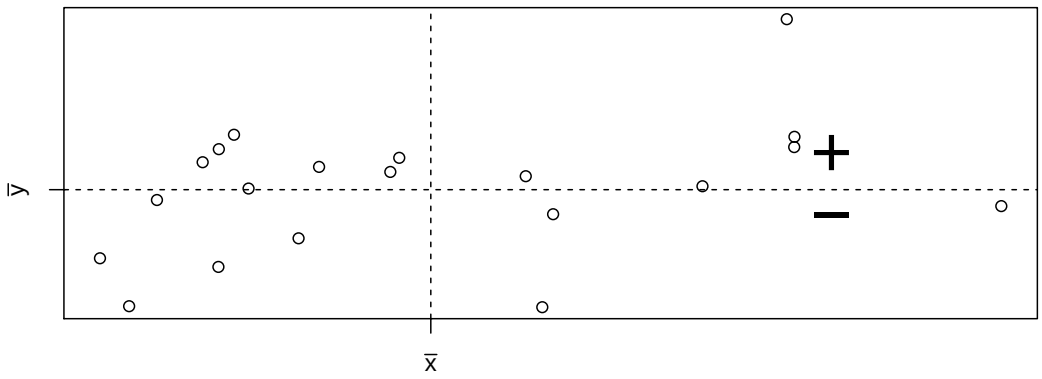


9.1 Pearson's Correlation Coefficient

We first consider Pearson's correlation coefficient, which is a statistics that measures the strength of the linear relationship between the predictor and response. Consider the following Pearson's correlation statistic

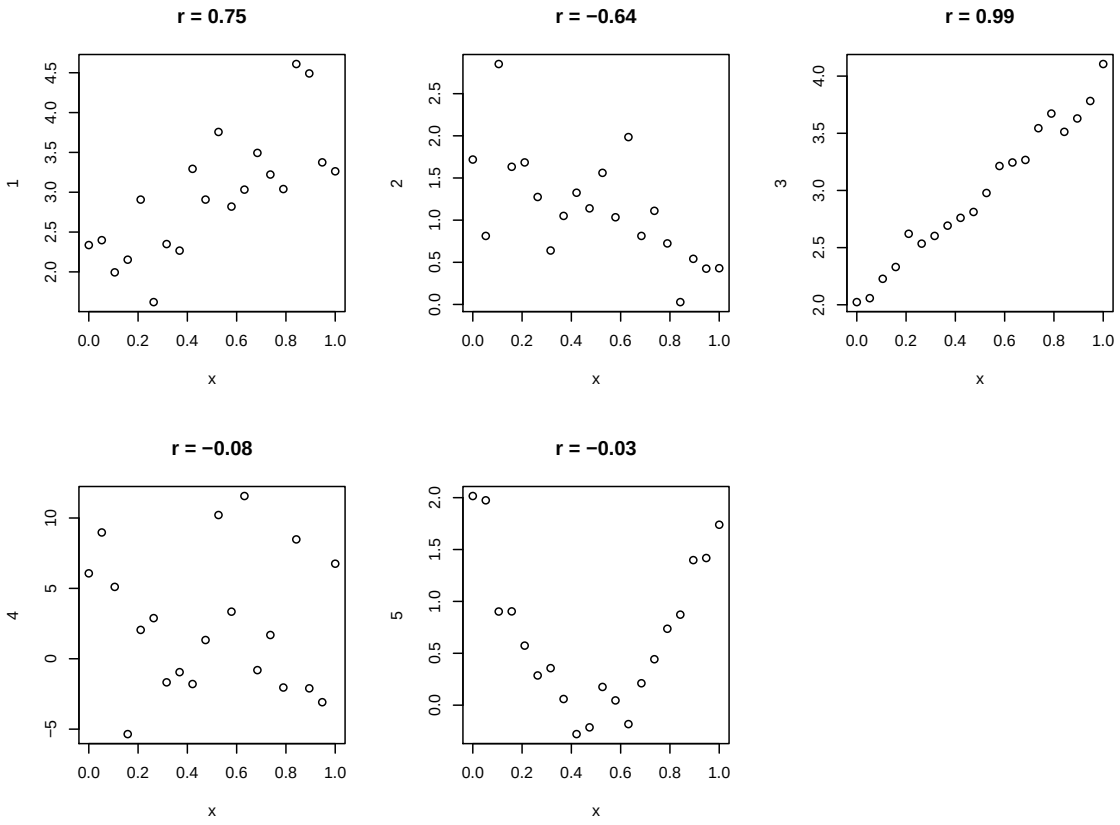
$$r = \frac{\sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)}{n-1}$$

where x_i and y_i are the x and y coordinate of the ith observation. Notice that each parenthesis value is the standardized value of each observation. If the x-value is big (greater than $\bar{x}$) and the y-value is large (greater than $\bar{y}$), then after multiplication, the result is positive. Likewise if the x-value is small and the y-value is small, both standardized values are negative and therefore after multiplication the result is positive. If a large x-value is paired with a small y-value, then the first value is positive, but the second is negative and so the multiplication result is negative.



The following are true about Pearson's correlation coefficient:

1. r is unit-less because we have standardized the x and y values.
2. $-1 \leq r \leq 1$ because of the scaling by $n - 1$
3. A negative r denotes a negative relationship between x and y, while a positive value of r represents a positive relationship.
4. r measures the strength of the *linear* relationship between the predictor and response.



9.2 Model Theory

To scatterplot data that looks linear we often want to fit the model

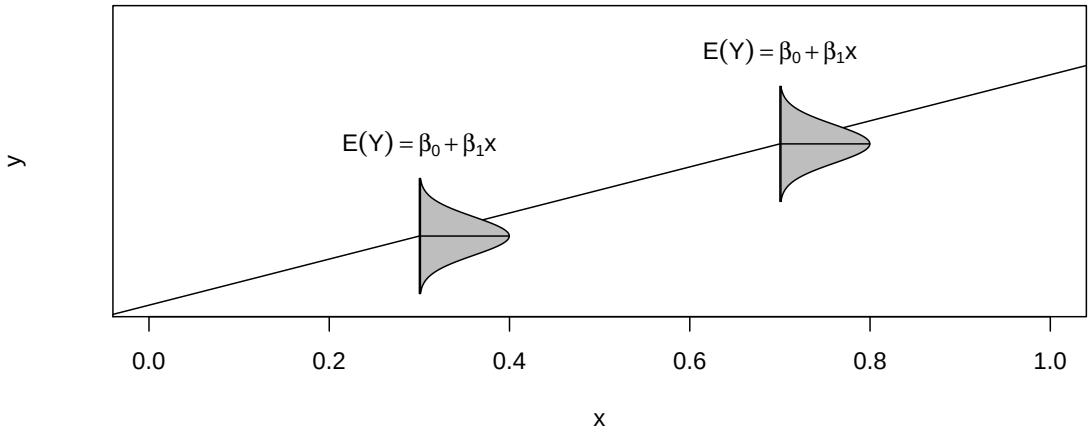
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad \text{where } \epsilon_i \overset{iid}{\sim} N\left(0, \sigma^2\right)$$

where β_0 is the y-intercept term[1] and β_1 is the slope term[2]. The assumptions of this model are:

1. *The relationship between the predictor and response is actually linear*
2. *The error terms come from a normal distribution*
3. *The variance of the errors is the same for every value of x (homoscedasticity)*
4. *The error terms are independent*

Under this model, the expected value of an observation with covariate $X = x$ is $E\left(Y \mid X = x\right) = \beta_0 + \beta_1 x$ and has a standard deviation of σ.

[1] The y-intercept is the height of the line when $x = 0$.
[2] The slope is the change in y for every one-unit change in x



Given this model, how do we find estimates of β_0 and β_1? In the past we have always relied on using some sort of sample mean, but it is not obvious what we can use here. Instead of a mean, we will use the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the sum of squared error (SSE) where

$$\begin{aligned} \hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 x_i \\ e_i &= y_i - \hat{y}_i \\ SSE &= \sum_{i=1}^{n} e_i^2 \end{aligned}$$

Fortunately there are simple closed form solutions for $\hat{\beta}_0$ and $\hat{\beta}_1$

$$\begin{aligned} \hat{\beta}_1 &= r\left(\frac{s_y}{s_x}\right) \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \end{aligned}$$

and using these estimates several properties can be shown

1. $\hat{\beta}_0$ and $\hat{\beta}_1$ are the slope and intercept values that minimize SSE.
2. The regression line goes through the center of mass of the data $(\bar{x}, \bar{y})$.
3. The sum of the residuals is 0. That is: $\sum e_i = 0$
4. $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators of β_0 and β_1

We are also interested in an estimate of σ^2 and we will use our usual estimation scheme of

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ &= \frac{\sum_{i=1}^{n} e_i^2}{n-2} \\ &= \frac{SSE}{n-2} \\ &= MSE \end{aligned}$$

where the -2 comes from having to estimate β_0 and β_1 before we can estimate σ^2. As in the ANOVA case, we can interpret σ as the typical distance an observation is from it's predicted value.

As always we are also interested in knowing the estimated standard deviation (which we will now call *Standard Error)* of the model parameters β_0 and β_1 and it can be shown that

$$StdErr\left(\hat{\beta}_0\right) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$$

and

$$StdErr\left(\hat{\beta}_1\right) = \hat{\sigma}\sqrt{\frac{1}{S_{xx}}}$$

where $S_{xx} = \sum (x_i - \bar{x})^2$. These intervals can be used to calculate confidence intervals for β_0 and β_1 using the formulas:

$$\hat{\beta}_i \pm t_{n-2}^{1-\alpha/2} StdErr\left(\hat{\beta}_i\right)$$

Again we consider the `iris` dataset that is available in R. I wish to examine the relationship between sepal length and sepal width in the species *setosa*.

```
library(ggplot2)
library(dplyr)
setosa <- filter( iris, Species == 'setosa' )
ggplot(setosa, aes(x=Sepal.Length, y=Sepal.Width)) +
  geom_point() +
  labs(x="Sepal Length", y="Sepal Width", title='Setosa Irises') +
  theme_bw()
```

```
x <- setosa$Sepal.Length
y <- setosa$Sepal.Width
n <- length(x)
r <- sum( (x-mean(x))/sd(x) * (y-mean(y))/sd(y) ) / (n-1)
b1 <- r*sd(y)/sd(x)
b0 <- mean(y) - b1*mean(x)
cbind(r, b0, b1)

##              r         b0        b1
## [1,] 0.7425467 -0.5694327 0.7985283

yhat <- b0 + b1*x
resid <- y - yhat
SSE <- sum( resid^2 )
s2 <- SSE/(n-2)
s2

## [1] 0.06580573

Sxx <- sum( (x-mean(x))^2 )
stderr.b0 <- sqrt(s2) * sqrt( 1/n + mean(x)^2 / Sxx)
stderr.b1 <- sqrt(s2) * sqrt(1 / Sxx )
cbind(stderr.b0, stderr.b1)

##      stderr.b0 stderr.b1
## [1,] 0.5217119 0.1039651

t.star <- qt(.975, df=n-2)
c(b0-t.star*stderr.b0, b0+t.star*stderr.b0)

## [1] -1.6184048  0.4795395

c(b1-t.star*stderr.b1, b1+t.star*stderr.b1)

## [1] 0.5894925 1.0075641
```

Of course, we don't want to have to do these calculations by hand. Fortunately statistics packages will do all of the above calculations. In R, we will use `lm()` to fit a linear regression model and then call various accessor functions to give me the regression output I want.

```
cor( setosa$Sepal.Width,  setosa$Sepal.Length )

## [1] 0.7425467

model <- lm(Sepal.Width ~ Sepal.Length, data=setosa)
coef(model)

##  (Intercept) Sepal.Length
##   -0.5694327    0.7985283

confint(model)

##                   2.5 %    97.5 %
## (Intercept)  -1.6184048 0.4795395
## Sepal.Length  0.5894925 1.0075641
```

In general, most statistics programs will give a table of output summarizing a regression and the table is usually set up as follows:

Coefficient	Estimate	Standard Error	t-stat	p-value
Intercept	$\hat{\beta}_0$	$StdErr\left(\hat{\beta}_0\right)$	$t_0 = \frac{\hat{\beta}_0}{StdErr\left(\hat{\beta}_0\right)}$	$2 * P\left(T_{n-2} > \lvert t_0 \rvert\right)$
Slope	$\hat{\beta}_1$	$StdErr\left(\hat{\beta}_1\right)$	$t_1 = \frac{\hat{\beta}_1}{StdErr\left(\hat{\beta}_1\right)}$	$2 * P\left(T_{n-2} > \lvert t_1 \rvert\right)$

This table is printed by R by using the `summary()` function:

```
model <- lm(Sepal.Width ~ Sepal.Length, data=setosa)
summary(model)

##
## Call:
## lm(formula = Sepal.Width ~ Sepal.Length, data = setosa)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.72394 -0.18273 -0.00306  0.15738  0.51709
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.5694     0.5217  -1.091    0.281
## Sepal.Length   0.7985     0.1040   7.681 6.71e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2565 on 48 degrees of freedom
## Multiple R-squared:  0.5514,Adjusted R-squared:  0.542
## F-statistic: 58.99 on 1 and 48 DF,  p-value: 6.71e-10
```

The first row is giving information about the y-intercept. In this case the estimate is -0.5694 and the standard error of the estimate is 0.5217. The t-statistic and associated p-value is testing the hypotheses: $H_0 : \beta_0 = 0$ vs $H_a : \beta_0 \neq 0$. This test is not usually of much interest. However since the equivalent test in the slope row testing $\beta_1 = 0$ vs $\beta_1 \neq 0$, the p-value of the slope row is *very* interesting because it tells me if I should include the slope variable in the model. If β_1 could be zero, then we should drop the predictor from our model and use the simple model $y_i = \beta_0 + \epsilon_i$ instead.

There are a bunch of other statistics that are returned by `summary()`. The `Residual standard error` is just $\hat{\sigma} = \sqrt{MSE}$ and the degrees of freedom for that error is also given. The rest are involved with the ANOVA interpretation of a linear model.

9.2.1 Anova Interpretation

Just as in the ANOVA analysis, we really have a competition between two models. The full model

$$y_i = \beta_0 + \beta_1 x + \epsilon_i$$

vs the simple model where x does not help predict y

$$y_i = \mu + \epsilon_i$$

where I've rewritten $\beta_0 = \mu$ to try to keep our notation straight. If I were to look at the simple model I would use $\bar{y} = \hat{\mu}$ as an estimate of μ and my Sum of Squared Error in the simple model will

be

$$SSE_{simple} = \sum_{i=1}^{n} (y_i - \hat{\mu})^2$$

and the appropriate Mean Squared Error is

$$MSE_{simple} = \frac{1}{n-1} \sum (y_i - \hat{\mu})^2$$

We can go through the same sort of calculations for the full complex model and get

$$SSE_{complex} = \sum_{i=1}^{n} \left(y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i \right) \right)^2$$

and

$$MSE_{complex} = \frac{1}{n-2} \sum_{i=1}^{n} \left(y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i \right) \right)^2$$

Just as in the AVOVA analysis, if we often like to look at the difference between $SSE_{simple} - SSE_{comples} = SSE_{diff}$ and think of this quantity as the amount of variability that is explained by adding the slope parameter to the model. Just as in the AVOVA case we'll calculate

$$MSE_{diff} = SSE_{diff}/df_{diff}$$

where df_{diff} is the number of parameters that we added to the simple model to create the complex one. In the simple linear regression case, $df_{diff} = 1$.

Just as in the ANOVA case, we will calculate an f-statistic to test the null hypothesis that the simple model suffices vs the alternative that the complex model is necessary. The calculation is

$$f = \frac{MSE_{diff}}{MSE_{complex}}$$

and the associated p-value is $P(F_{1,n-2} > f)$. Notice that this test is *exactly* testing if $\beta_1 = 0$ and therefore the p-value for the F-test and the t-test for β_1 are the same. It can easily be shown that $t_1^2 = f$.

The Analysis of Variance table looks the same as what we have seen, but now we recognize that the rows actually represent the complex and simple models and the difference between them.

Source	df	Sum of Squares	Mean Squared	F-value	P-value
Difference	1	SSE_{diff}	$MSE_{diff} = \frac{SSE_{diff}}{1}$	$f = \frac{MSE_{diff}}{MSE_{complex}}$	$P(F_{1,n-2} > f)$
Complex	$n-2$	$SSE_{complex}$	$MSE_{complex} = \frac{SSE_{complex}}{n-2}$		
Simple	$n-1$	SSE_{simple}			

As usual, the ANOVA table for the regression is available in R using the `anova()` command.

```
model <- lm(Sepal.Width ~ Sepal.Length, data=setosa)
anova(model)

## Analysis of Variance Table
##
## Response: Sepal.Width
##              Df Sum Sq Mean Sq F value   Pr(>F)
## Sepal.Length  1 3.8821  3.8821  58.994 6.71e-10 ***
## Residuals    48 3.1587  0.0658
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

But we notice that R chooses not to display the row corresponding to the simple model.

I could consider SSE_{simple} as a baseline measure of the amount of variability in the data. It is interesting to look at how much of that baseline variability has been explained by adding the additional parameter to the model. Therefore we'll define the ratio R^2 as:

$$R^2 = \frac{SSE_{diff}}{SSE_{simple}} = \frac{SSE_{simple} - SSE_{complex}}{SSE_{simple}} = r^2$$

where r is Pearson's Correlation Coefficient. R^2 has the wonderful interpretation of the percent of variability in the response variable that can be explained by the predictor variable x.

9.2.2 Confidence Intervals vs Prediction Intervals

There are two different types of questions that we might ask about predicting the value for some x-value x_{new}.

We might be interested in a confidence interval for regression line. For this question we want to know how much would we expect the sample regression line move if we were to collect a new set of data. In particular, for some value of x, say x_{new}, how variable would the regression line be? To answer that we have to ask what is the estimated variance of $\hat{\beta}_0 + \hat{\beta}_1 x_{new}$? The variance of the regression line will be a function of the variances of $\hat{\beta}_0$ and $\hat{\beta}_1$ and thus the standard error looks somewhat reminiscent of the standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$. Recalling that $S_{xx} = \sum (x_i - \bar{x})^2$, we have:

$$\hat{Var}\left(\hat{\beta}_0 + \hat{\beta}_1 x_{new}\right) = \hat{\sigma}^2 \left(\frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{S_{xx}}\right)$$

and therefore its $StdErr(\hat{\beta}_0 + \hat{\beta}_1 x_{new})$ is

$$StdErr\left(\hat{\beta}_0 + \hat{\beta}_1 x_{new}\right) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{S_{xx}}}$$

We can use this value to produce a confidence interval for the regression line for any value of x_{new}.

$$\begin{aligned} Estimate &\pm t\ StdErr\,(Estimate) \\ \left(\hat{\beta}_0 + \hat{\beta}_1 x_{new}\right) &\pm t_{n-2}^{1-\alpha/2}\ \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{S_{xx}}} \end{aligned}$$

the expected value of new observation $\hat{E}\,(Y \mid X = x_{new})$. This expectation is regression line but since the estimated regression line is a function of the data, then the line isn't the exactly the same as the true regression line. To reflect that, I want to calculate a confidence interval for where the true regression line should be.

I might instead be interested calculating a confidence interval for y_{new}, which I will call a *prediction interval* in an attempt to keep from being confused with the confidence interval of the regression line. Because we have

$$y_{new} = \beta_0 + \beta_1 x_{new} + \epsilon_{new}$$

Then my prediction interval will still be centered at $\hat{\beta}_0 + \hat{\beta}_1 x_{new}$ but the the uncertainty should be the sum of the uncertainty associated with the estimates of β_0 and β_1 and the additional variability associated with ϵ_{new}. In short,

$$\begin{aligned} \hat{Var}\left(\hat{\beta}_0 + \hat{\beta}_1 x_{new} + \epsilon\right) &= \hat{Var}\left(\hat{\beta}_0 + \hat{\beta}_1 x_{new}\right) + \hat{Var}\,(\epsilon) \\ &= \hat{\sigma}^2 \left(\frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{S_{xx}}\right) + \hat{\sigma}^2 \end{aligned}$$

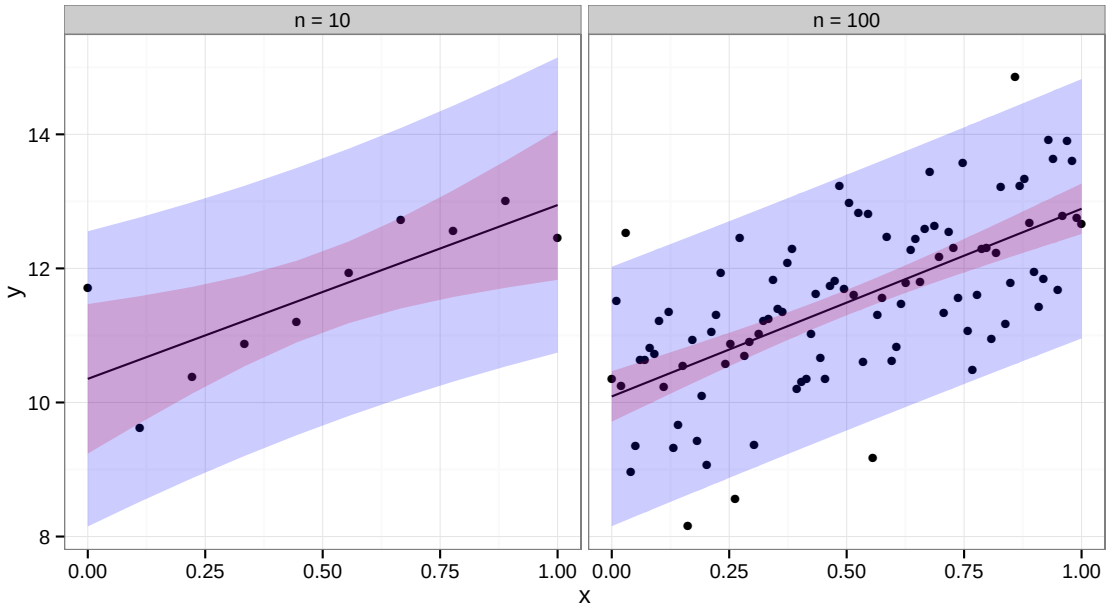
and the $StdErr\,()$ of a new observation will be

$$StdErr\,(\hat{y}_{new}) = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{S_{xx}}}$$

So the prediction interval for a new observation will be:

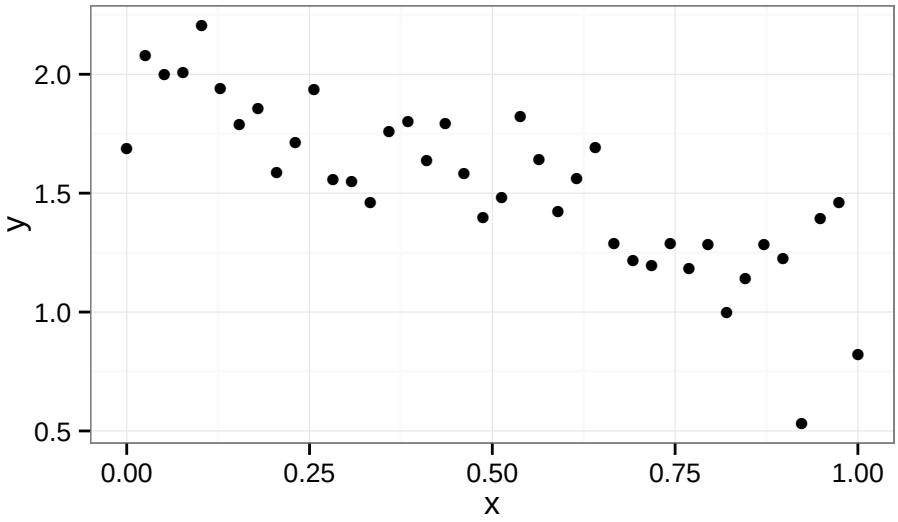
$$\left(\hat{\beta}_0 + \hat{\beta}_1 x_{new}\right) \pm t_{n-2}^{1-\alpha/2}\ \ \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{S_{xx}}}$$

To emphasize the difference between confidence regions (capturing where we believe the regression line to lay) versus prediction regions (where new data observations will lay) we note that as the sample size increases, the uncertainty as to where the regression line lays decreases, but the prediction intervals will always contain a minimum width due to the error associated with an individual observation. Below are confidence (red) and prediction (blue) regions for two different sample sizes.



In general, you will not want to calculate the confidence intervals and prediction intervals by hand. Fortunately R makes it easy to calculate the intervals. The function `predict()` will calculate the point estimates along with confidence and prediction intervals. The function requires the `lm()` output along with an optional data frame (if you want to predict values not in the original data).

```r
# make up some data and graph it
n <- 40
sim.data <- data.frame( x = seq(0,1, length=n) ) %>%
  mutate( y = 2 - 1*x + rnorm(n, sd=.2) )

ggplot(sim.data, aes(x=x, y=y)) + geom_point() + theme_bw()
```

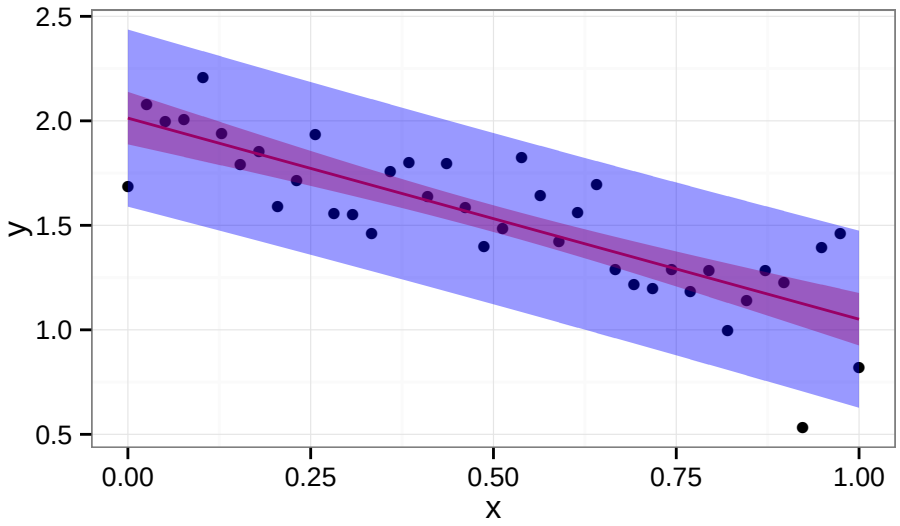



```r
# fit the regression
model <- lm(y~x, data=sim.data)

# display the first few predictions
head( predict(model, interval="confidence") )

##        fit      lwr      upr
## 1 2.013046 1.887414 2.138677
## 2 1.988372 1.867479 2.109265
## 3 1.963699 1.847472 2.079925
## 4 1.939025 1.827384 2.050666
## 5 1.914352 1.807206 2.021497
## 6 1.889678 1.786926 1.992430

# predict at x = 0.75
predict(model, interval="prediction", newdata=data.frame(x=0.75))

##        fit       lwr      upr
## 1 1.291345 0.8780134 1.704677
```

We can create a nice graph of the regression line and associated confidence and prediction regions using the following code in R:

```
library(ggplot2)
library(dplyr)
# ask for the confidence and prediction intervals
conf.region <- predict(model, interval='confidence')
pred.region <- predict(model, interval='prediction')

# add them to my original data frame
sim.data <- sim.data %>%
  mutate( fit = fitted(model),
          conf.lwr = conf.region[,2],
          conf.upr = conf.region[,3],
          pred.lwr = pred.region[,2],
          pred.upr = pred.region[,3])
```

```
# make a nice plot
ggplot(sim.data) +
  geom_point( aes(x=x, y=y) ) +
  geom_line(  aes(x=x, y=fit), col='red' ) +
  geom_ribbon( aes(x=x, ymin=conf.lwr, ymax=conf.upr), fill='red',  alpha=.4) +
  geom_ribbon( aes(x=x, ymin=pred.lwr, ymax=pred.upr), fill='blue', alpha=.4) +
  theme_bw()
```

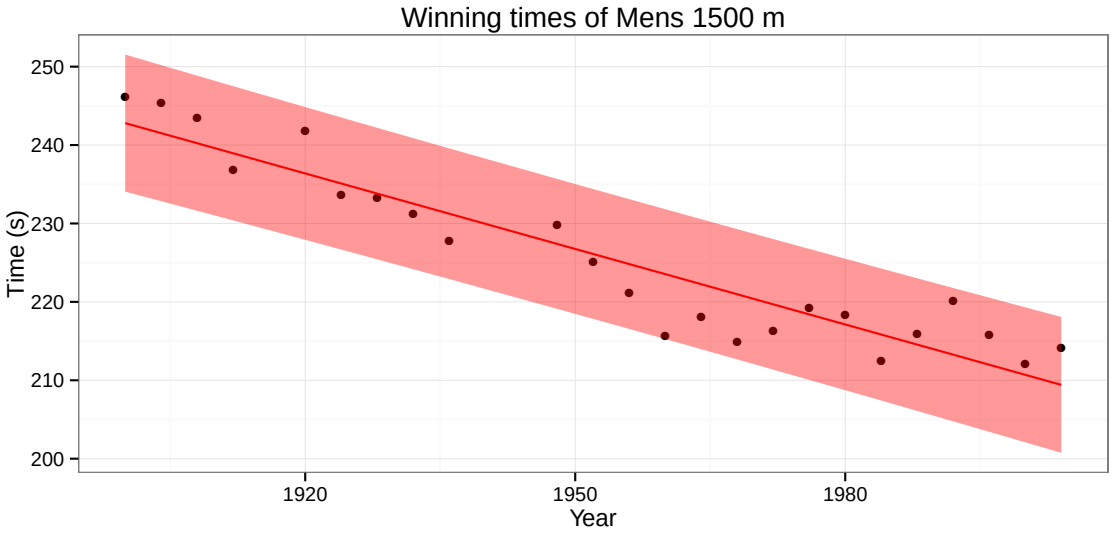


It is worth noting that these confidence intervals are all *point-wise* confidence intervals. If I want to calculate confidence or prediction intervals for a large number of x_{new} values, then I have to deal with the multiple comparisons issue. Fortunately this is easy to do in the simple linear regression case. Instead of using the $t_{n-2}^{1-\alpha/2}$ quantile in the interval formulas, we should use $W = \sqrt{2 * F_{1-\alpha,\,2,\,n-2}}$. Your book ignores this issue as does the `predict()` function in R.

9.3 Extrapolation

The data observed will inform a researcher about the relationship between the x and y variables, but *only in the range for which you have data!* Below are the winning times of the men's 1500 meter Olympic race.

```
library(HSAUR2)
data(men1500m)
small <- filter( men1500m, year != 1896 )  # Remove the 1896 Olympics

# fit the model and get the prediction interval
model <- lm( time ~ year, data=small )
small <- cbind( small, predict(model, interval='prediction') )

ggplot(small, aes(x=year, y=time, ymin=lwr, ymax=upr)) +
  geom_point() +
  geom_line( aes(y=fit), col='red' ) +
  geom_ribbon( fill='red',  alpha=.4) +
  labs( x='Year', y='Time (s)', title='Winning times of Mens 1500 m' ) + theme_bw()
```



If we are interested in predicting the results of the 2008 and 2012 Olympic race, what would we predict?

```
predict(model,
        newdata=data.frame(year=c(2008, 2012)),
        interval="prediction")

##        fit      lwr      upr
## 1 208.1293 199.3971 216.8614
## 2 206.8451 198.0450 215.6453
```

We can compare the predicted intervals with the time actually recorded by the winner of the men's 1500m. In Beijing 2008, Rashid Ramzi from Brunei won the event in 212.94 seconds and in London 2012 Taoufik Makhloufi from Algeria won in 214.08 seconds. Both times are within the corresponding prediction intervals, but clearly the linear relationship must eventually change and therefore our regression could not possibly predict the winning time of the 3112 race.

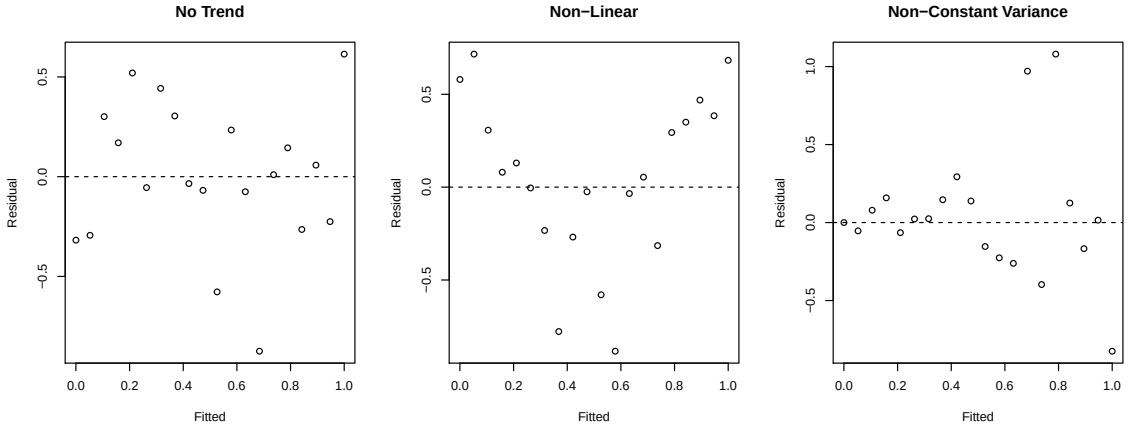
```
predict(model, newdata=data.frame(year=c(3112)), interval="prediction")

##         fit       lwr       upr
## 1 -146.2973 -206.7705 -85.82402
```

9.4 Checking Model Assumptions

As in the anova analysis, we want to be able to check the model assumptions. To do this, we will examine the residuals

$$e_i = y_i - \hat{y}_i$$

for normality using a QQ-plot as we did in Anova. To address the constant variance and linearity assumptions we will look at scatterplots of the residuals vs the fitted values $\hat{y}_i$. For the regression to be valid, we want the scatterplot to show no discernible trend. There are two patterns that commonly show up that indicate a violation of the regression assumptions.

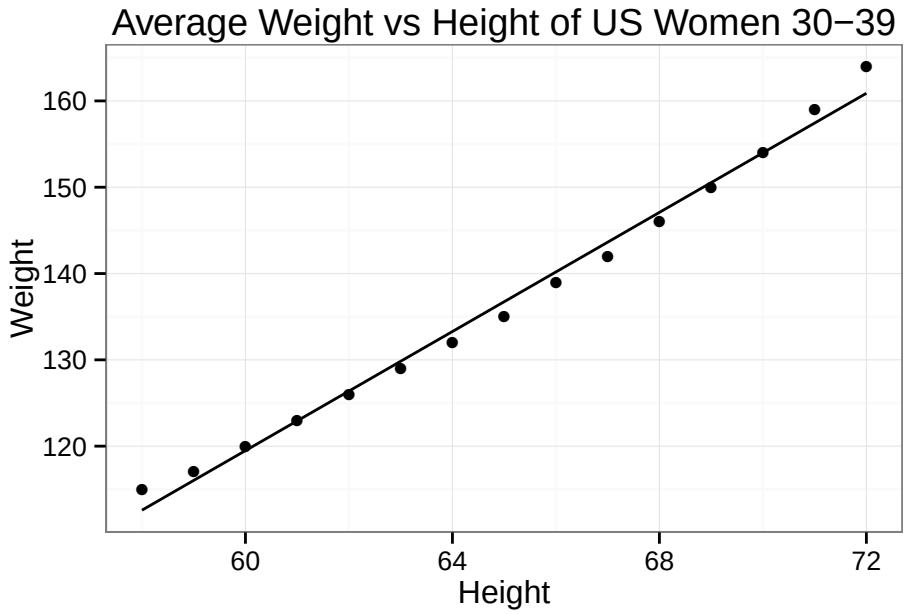


Non-Linearity

To illustrate this, we'll look at data about the height and weight values for women between 30 and 39. (The data presented is actually the average weight for women of given heights, but is a useful example).

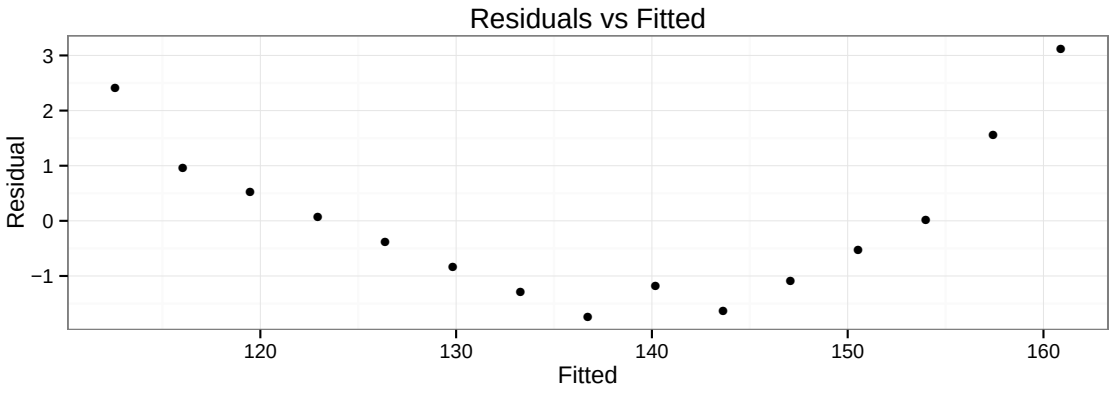
```
data('women')
str(women)

## 'data.frame': 15 obs. of  2 variables:
##  $ height: num  58 59 60 61 62 63 64 65 66 67 ...
##  $ weight: num  115 117 120 123 126 129 132 135 139 142 ...

# fit the regression line and add that info to my data frame
model <- lm(weight ~ height, data=women)
women <- women %>%
  mutate( fit   = fitted(model),
          resid = resid(model)  )
```

```
ggplot(women) +
  geom_point(aes(x=height, y = weight )) +
  geom_line( aes(x=height, y = fit    )) +
  labs( x='Height', y='Weight',
        title='Average Weight vs Height of US Women 30-39' ) +
  theme_bw()
```



If we squint at this graph, we see that the data are not perfectly linear and there appears to be some curvature. It isn't particularly clear from this graph however. However, when we look at the residuals vs fitted values and immediately conclude that the linearity assumption is violated.

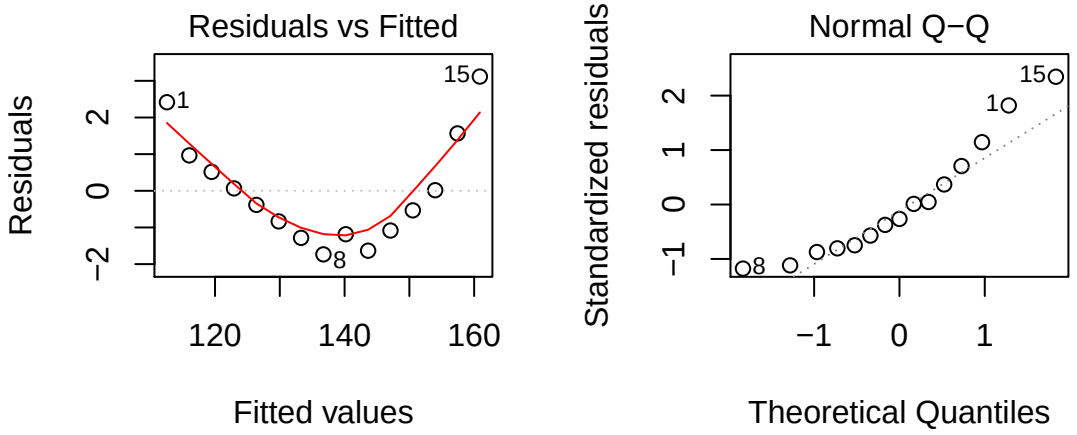
```
ggplot(women, aes( x=fit, y=resid )) +
  geom_point() +
  labs(title='Residuals vs Fitted', x='Fitted', y='Residual') + theme_bw()
```



This sort of graph is useful enough that R provides a quick way to create it. The function `plot( lm.object )` will take the input linear model and produce 6 diagnostic plots. This function uses the base R graphics system, so all the `ggplot2` options don't work.[3]

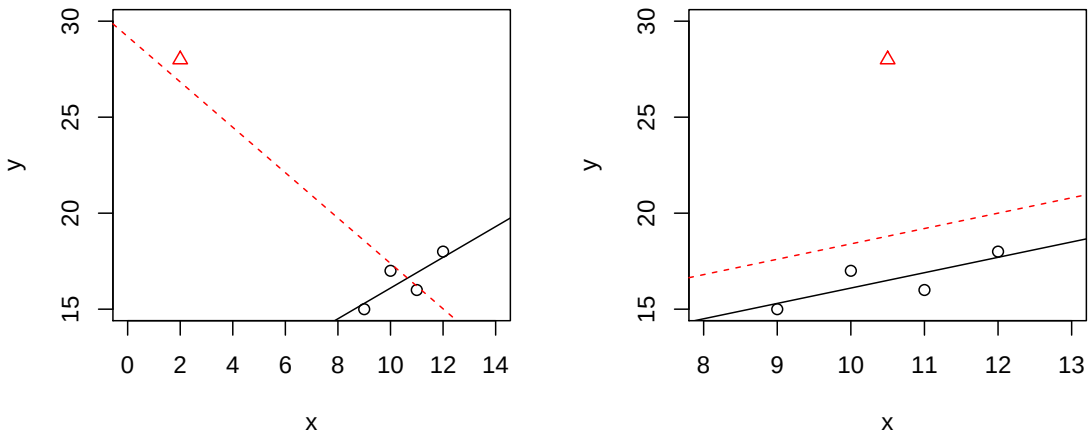
[3]There are a couple of packages that re-create these diagnostic plots using `ggplot2` but we won't use them here.

```
par(mfrow=c(1,2)) # base graphics, two plots side-by-side
plot(model, which=c(1,2))
```



9.5 Influential Points

Sometimes a dataset will contain one observation that has a large effect on the outcome of the model. Consider the following datasets where the red denotes a highly influential point and the red line is the regression line including the point.

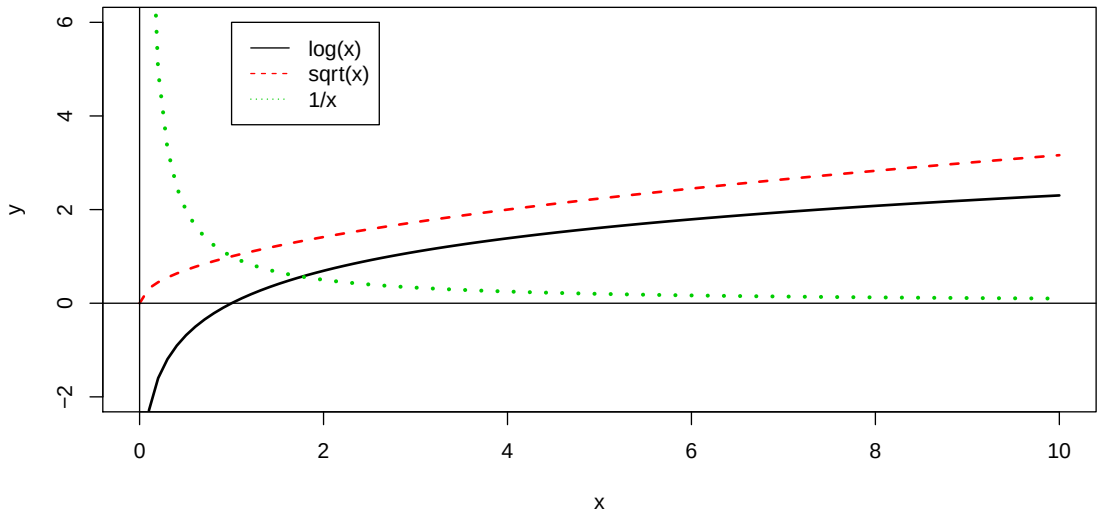


The question of what to do with influential points is not easy to answer. Sometimes these are data points that are a result of lab technician error and should be removed. Sometimes they are the result of an important process that is not well understood by the researcher. It is up to the scientist to figure out which is the case and take appropriate action.

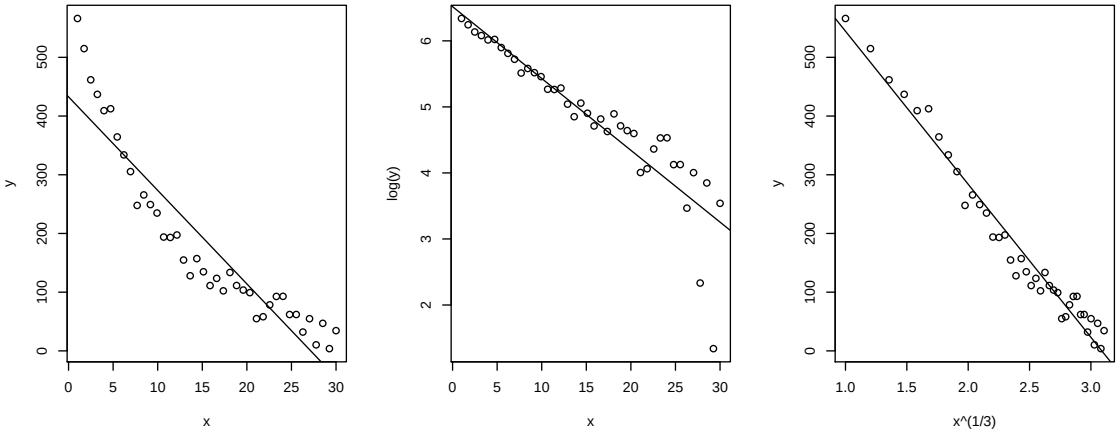
One solution is to run the analysis both with and without the influential point and see how much it affects your inferences.

9.6 Transformations

When the normality or constant variance assumption is violated, sometimes it is possible to *transform* the data to make it satisfy the assumption. Often times count data is analyzed as log(count) and weights are analyzed after taking a square root or cube root transform.



We have the option of either transforming the x-variable or transforming the y-variable or possibly both. One thing to keep in mind, however, is that transforming the x-variable only effects the linearity of the relationship. Transforming the y-variable effects both the linearity and the variance.



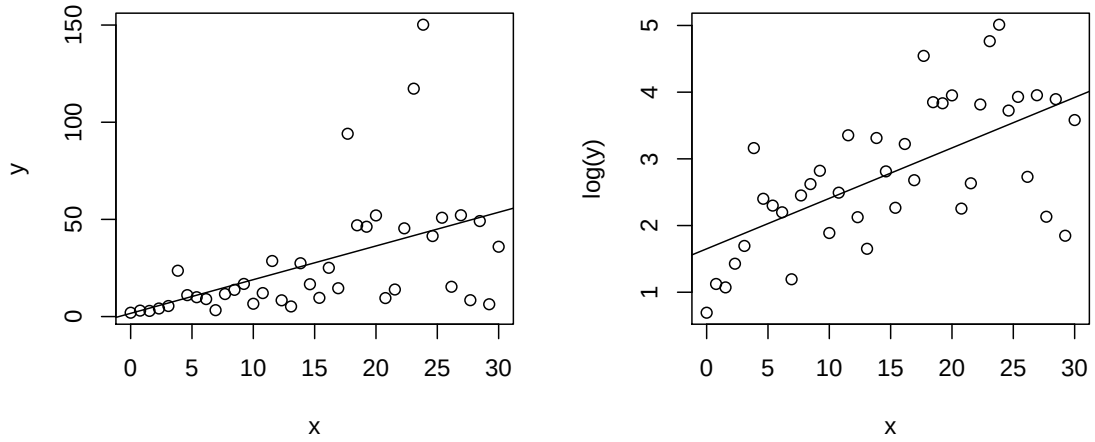
Unfortunately it is not always obvious what transformation is most appropriate. The Box-Cox family of transformations for the y-variable is

$$f(y \mid \lambda) \quad = \quad \begin{cases} y^{\lambda} & \text{if } \lambda \neq 0 \\ \log y & \text{if } \lambda = 0 \end{cases}$$

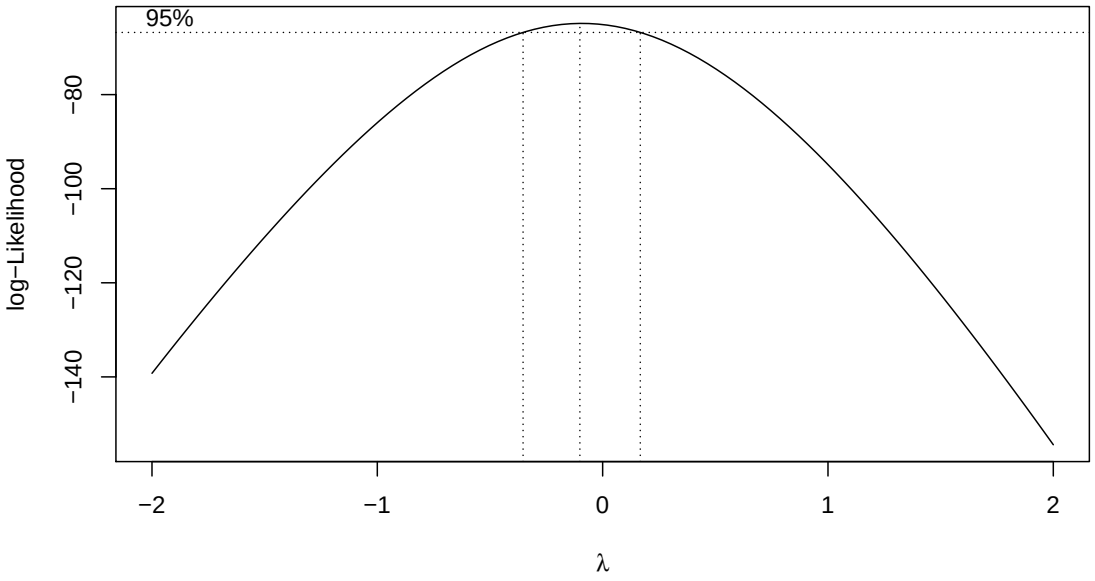
which includes squaring ($\lambda = 2$), square root ($\lambda = 1/2$) and as $\lambda \to 0$ the transformation converges to $\log y$. (To do this correctly we should define the transformation in a more complicated fashion, but that level of detail is unnecessary here.) The transformation is selected by looking at the

profile log-likelihood value of different values of λ and we want to use the λ that maximizes the log-likelihood.

Of course, we also want to use a transformation that isn't completely obscure and is commonly used in the scientific field, so square roots, reciprocals, and logs are preferred.




```
library(MASS)
str(mydata)

## 'data.frame': 40 obs. of  2 variables:
##  $ x: num  0 0.769 1.538 2.308 3.077 ...
##  $ y: num  2 3.08 2.92 4.17 5.44 ...

boxcox(y~x, data=mydata, plotit=TRUE)
```



Here we see the resulting confidence interval for λ contains 0, so a log transformation would be most appropriate.

In general, deciding on a transformation to use is often a trade-off between statistical pragmatism and interpretability. In cases that a transformation is not possible, or the interpretation is difficult, it is necessary to build more complicated models that are interpretable.

Chapter 10

Bootstrapping Linear Models

The last several chapters have introduced a number of parametric models where we assume that the error terms are normally distributed.

$$\begin{aligned}
&\text{One-sample t-test:} && Y_i = \mu + \epsilon_i \quad \text{where } \epsilon_i \overset{iid}{\sim} N(0, \sigma) \\
&\text{Two-sample t-test} && Y_{ij} = \mu_i + \epsilon_{ij} \quad \text{where } \epsilon_{ij} \overset{iid}{\sim} N(0, \sigma) \quad i \in \{1, 2\} \\
&\text{ANOVA} && Y_{ij} = \mu_i + \epsilon_{ij} \quad \text{where } \epsilon_{ij} \overset{iid}{\sim} N(0, \sigma) \quad i \in \{1, 2, \ldots, k\} \\
&\text{Regression} && Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad \text{where } \epsilon_i \overset{iid}{\sim} N(0, \sigma)
\end{aligned}$$

We developed hypothesis tests and confidence intervals for the model parameters assuming that the error terms were normally distributed and, in the event that they are normally distributed, those tests and confidence intervals are the best we can do. However, if the errors are not normally distributed, what should we do?

Previously we used bootstrapping to estimate the sampling distribution of the sampling statistic when we didn't know the distribution. We will use the same bootstrapping method, but we'll simplify all of the above cases to the the same simple linear model

$$Y_i = E(Y_i) + \epsilon_i \quad \text{where } \epsilon_i \overset{iid}{\sim} N(0, \sigma)$$

and $E(Y_i)$ takes on some form of the parameters depending on the model specified. It turns out that R can do all of these analyses using the same `lm()` function we used in for regression.

10.1 Using `lm()` for many analyses

10.1.1 One-sample t-tests

In this model we are concerned with testing

$$\begin{aligned}
H_0 &: \quad \mu = \mu_0 \\
H_a &: \quad \mu \neq \mu_0
\end{aligned}$$

for some μ_0. For example, suppose we have the following data and we want to test $H_0 : \mu = 5$ vs $H_a : \mu \neq 5$. The R code we used previously was

```
# How we previously did a t.test
library(mosaic)
test.data <- data.frame( y=c(3,5,4,5,7,13) )
t.test( test.data$y, mu=5 )

##
##  One Sample t-test
##
## data:  x
## t = 0.7936, df = 5, p-value = 0.4634
## alternative hypothesis: true mean is not equal to 5
## 95 percent confidence interval:
##  2.387727 9.945607
## sample estimates:
## mean of x
##  6.166667
```

but we can just as easily consider this a linear model with only an intercept term.

```
m1 <- lm(y ~ 1, data=test.data)
summary(m1)

##
## Call:
## lm(formula = y ~ 1, data = test.data)
##
## Residuals:
##       1       2       3       4       5       6
## -3.1667 -1.1667 -2.1667 -1.1667  0.8333  6.8333
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.167      1.470   4.195  0.00853 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.601 on 5 degrees of freedom

confint(m1)

##                2.5 %   97.5 %
## (Intercept) 2.387727 9.945607
```

Notice that we get the same point estimate and confidence interval for μ, but the p-value is different because the `t.test()` p-value is testing $H_0: \mu = 5$ vs $H_a: \mu \neq 5$ while the `lm()` function is testing $H_0: \mu = 0$ vs $H_a: \mu \neq 0$.

10.1.2 Two-sample t-tests

This model is concerned with testing

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

```
# How we previously did a t.test
test.data <- data.frame( y=c(3, 5, 4,  5,  7, 13,
                             8, 9, 4, 16, 12, 13 ),
                         group=rep(c('A','B'), each=6) )

t.test( y ~ group, data=test.data, var.equal=TRUE )

##
##  Two Sample t-test
##
## data:  y by group
## t = -1.838, df = 10, p-value = 0.09591
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -9.2176608  0.8843275
## sample estimates:
## mean in group A mean in group B
##        6.166667       10.333333
```

This analysis gave use the mean of each group and the confidence interval for the difference $\mu_2 - \mu_1$. We could get the same analysis using either the `aov()`, because a two-sample t-test with pooled variance is equivalent to an ANOVA with $k = 2$ groups. Furthermore, both of these are just cases of the linear model and we could use the `lm()` command.

```
m2 <- aov(y ~ group, data=test.data)
summary(m2)

##             Df Sum Sq Mean Sq F value Pr(>F)
## group        1  52.08   52.08   3.378 0.0959 .
## Residuals   10 154.17   15.42
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

coef(m2)

## (Intercept)      groupB
##    6.166667    4.166667

confint(m2)

##                  2.5 %   97.5 %
## (Intercept)  2.5950745 9.738259
## groupB      -0.8843275 9.217661
```

```
m2 <- lm(y ~ group, data=test.data)
summary(m2)

##
## Call:
## lm(formula = y ~ group, data = test.data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.333 -2.208 -1.167  1.917  6.833
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.167      1.603   3.847  0.00323 **
## groupB         4.167      2.267   1.838  0.09591 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.926 on 10 degrees of freedom
## Multiple R-squared:  0.2525,Adjusted R-squared:  0.1778
## F-statistic: 3.378 on 1 and 10 DF,  p-value: 0.09591

coef(m2)

## (Intercept)      groupB
##    6.166667    4.166667

confint(m2)

##                  2.5 %   97.5 %
## (Intercept)  2.5950745 9.738259
## groupB      -0.8843275 9.217661
```

Aside from `t.test()` reporting $\mu_2 - \mu_1$ while the other two calculate $\mu_1 - \mu_2$, the estimates are identical.

10.2 Creating Simulated Data

The basic goal of statistics is that we are interested in some population (which is described by some parameter μ, δ, τ, β, or generally, θ) and we take a random sample of size n from the population of interest and we *truly believe* that the sample is representative of the population of interest. Then we use some statistic of the data $\hat{\theta}$ as an estimate θ. However we know that this estimates, $\hat{\theta}$, vary from sample to sample. Previously we've used that the Central Limit Theorem gives

$$\hat{\theta} \dot{\sim} N\left(\theta, \sigma_{\hat{\theta}}\right)$$

to construct confidence intervals and perform hypothesis tests, but we don't necessarily like this approximation. If we could somehow take repeated samples (call these repeated samples $\mathbb{Y}_j$ for $j \in 1, 2, \ldots, M$) from the population we would understand the distribution of $\hat{\theta}$ by just examining the distribution of many observed values of $\hat{\theta}_j$ where $\hat{\theta}_j$ is the statistic calculated from the ith sample data $\mathbb{Y}_j$.

However, for practical reasons, we can't just take 1000s of samples of size n from the population. However, because we *truly believe* that $\mathbb{Y}$ is representative of the entire population, then our best guess of what the population is just many repeated copies of our data.

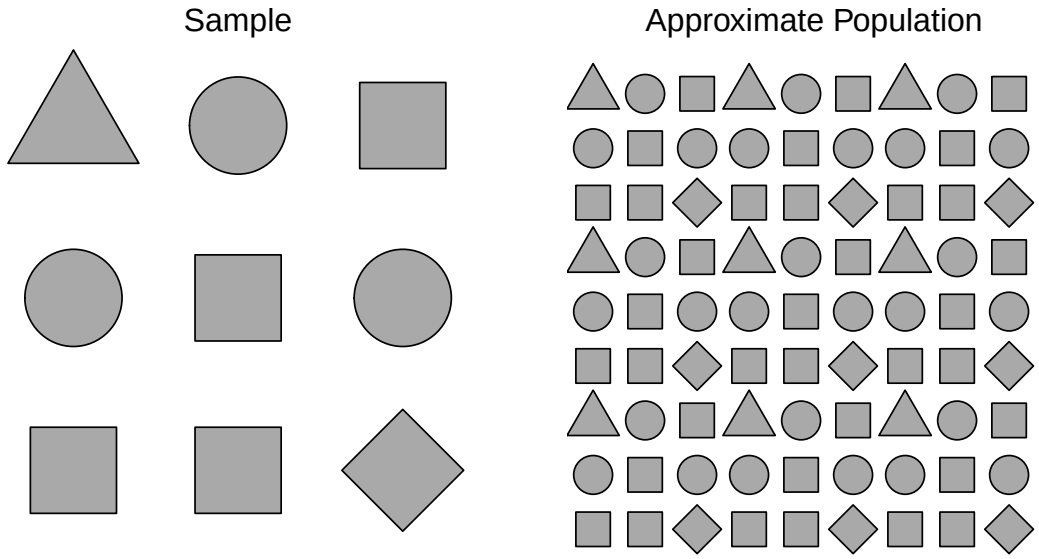


Figure 10.2.1: A possible sample from a population of shapes. Because 4/9 of our sample were squares, our best estimate is that the population is also approximately 4/9 squares. We can think of the approximated population as just many many copies of the observed sample data.

Suppose we were to sample from a population of shapes, and we observed 4/9 of the sample were squares, 3/9 were circles, and a triangle and a diamond. Then our best guess of what the population that we sampled from was a population with 4/9 squares, 3/9 circles, and 1/9 of triangles and diamonds.

Using this approximated population (which is just many many copies of our sample data), we can take many samples of size n. We denote these bootstrap samples as $\mathbb{Y}_j^*$, where the star denotes that the sample was taken from the approximate population, not the actual population. From each bootstrap sample $\mathbb{Y}_j^*$ a statistic of interest can be taken $\hat{\theta}_j^*$.

Because our approximate population is just an infinite number of copies of our sample data, then sampling from the approximate population is equivalent to sampling *with replacement* from our sample data. If I take n samples from n distinct objects with replacement, then the process can be thought of as mixing the n objects in a bowl and taking an object at random, noting which it is, replace it into the bowl, and then draw the next sample. Practically, this means some objects will be selected more than once and some will not be chosen at all. To sample our observed data with replacement, we'll use the `resample()` function in the `mosaic` package. We see that some rows will be selected multiple times, and some will not be selected at all.

We can resample the data in two different ways.

1. Resample the rows of the data frame which is called *case* resampling.
2. Resample only the residuals which is called *residual* sampling.

```
library(dplyr)
library(mosaic)
Testing.Data <- data.frame(
  x = c(3,5,7,9),
  y = c(3,7,7,11))
Testing.Data

##   x  y
## 1 3  3
## 2 5  7
## 3 7  7
## 4 9 11
```

```
# Case resampling
Boot.Data <- resample(Testing.Data)
Boot.Data

##     x  y orig.ids
## 1   3  3        1
## 4   9 11        4
## 2   5  7        2
## 2.1 5  7        2
```

Notice that we've sampled $\{x = 5, y = 7\}$ twice and did not get the $\{7, 7\}$ data point.

Residual sampling is done by resampling the residuals and calling them $\hat{\epsilon}^*$ and then the new y-values will be

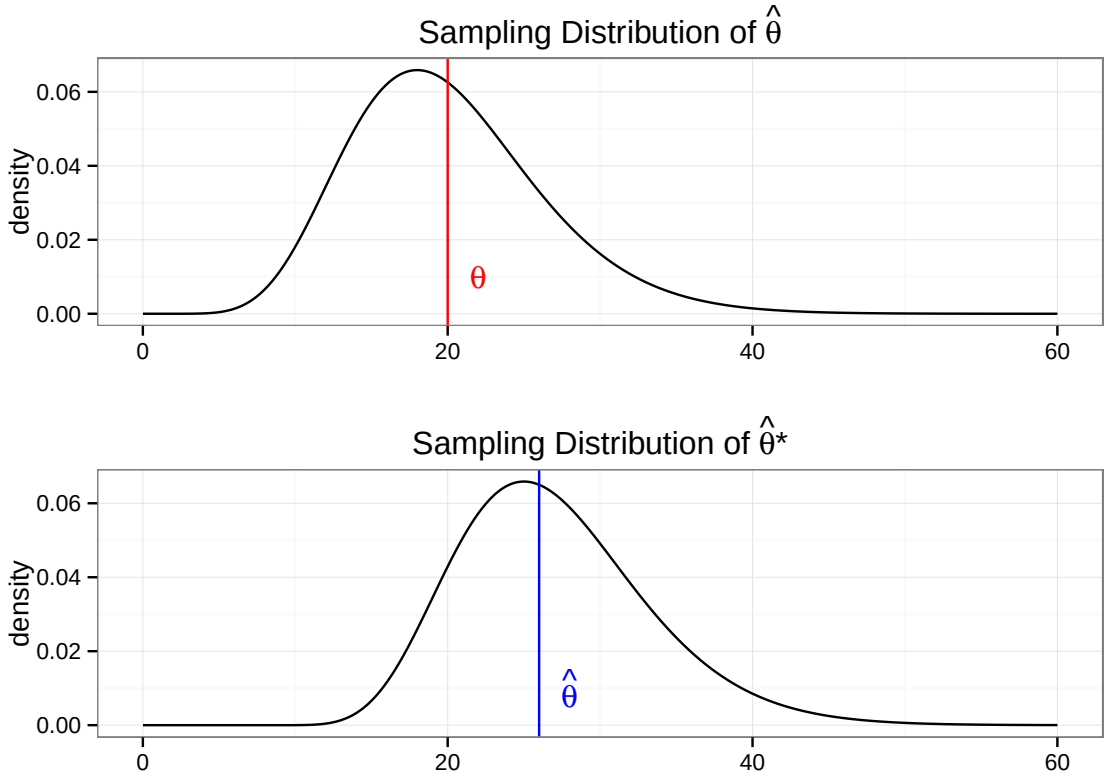
$$y_i^* = \hat{y}_i + \hat{\epsilon}_i^*$$

```
# Residual resampling
model <- lm( y ~ x, data=Testing.Data)
Boot.Data <- Testing.Data %>%
  mutate( fit = fitted(model),
          resid = resid(model),
          resid.star = resample(resid),
          y.star = fit + resid.star )
Boot.Data

##   x  y  fit resid resid.star y.star
## 1 3  3  3.4  -0.4       -1.2    2.2
## 2 5  7  5.8   1.2       -1.2    4.6
## 3 7  7  8.2  -1.2       -0.4    7.8
## 4 9 11 10.6   0.4        1.2   11.8
```

Notice that the residuals resampling results in a data set where each of the x-values is retained, but a new y-value (possibly not seen in the original data) is created from the predicted value $\hat{y}$ and a randomly selected residual.

10.3 Confidence Interval Types

We want to understand the relationship between the sample statistic $\hat{\theta}$ to the population parameter θ. We create an estimated population using many repeated copies of our data. By examining how the simulated $\hat{\theta}^*$ vary relative to $\hat{\theta}$, we will understand how possible $\hat{\theta}$ values vary relative to θ.



We will outline several methods for producing confidence intervals (in the order of most assumptions to fewest).

10.3.1 Normal intervals

This confidence interval assumes the sampling distribution of $\hat{\theta}$ is approximately normal (which is often true due to the central limit theorem). We can use the bootstrap replicate samples to get an estimate of the standard error of the statistic of interest by just calculating the sample standard deviation of the replicated statistics.

Let $\theta()$ be the statistic of interest and $\hat{\theta}$ be the value of that statistic calculated from the observed data. Define $\hat{SE}^*$ as the sample standard deviation of the $\hat{\theta}^*$ values.

Our first guess as to a confidence interval is

$$\hat{\theta} \pm z_{1-\alpha/2}\hat{SE}^*$$

which we could write as

$$\left[\hat{\theta} - z_{1-\alpha/2}\hat{SE}^*, \quad \hat{\theta} + z_{1-\alpha/2}\hat{SE}^*\right] \tag{10.3.1}$$

10.3.2 Percentile intervals

The percentile interval doesn't assume normality but it does assume that the bootstrap distribution is symmetric and unbiased for the population value. This is the method we used to calculate confidences intervals in the first several chapters. It is perhaps the easiest to calculate and understand. This method only uses $\hat{\theta}^*$, and is

$$\left[\hat{\theta}^*_{\alpha/2}\ , \quad \hat{\theta}^*_{1-\alpha/2}\right]$$

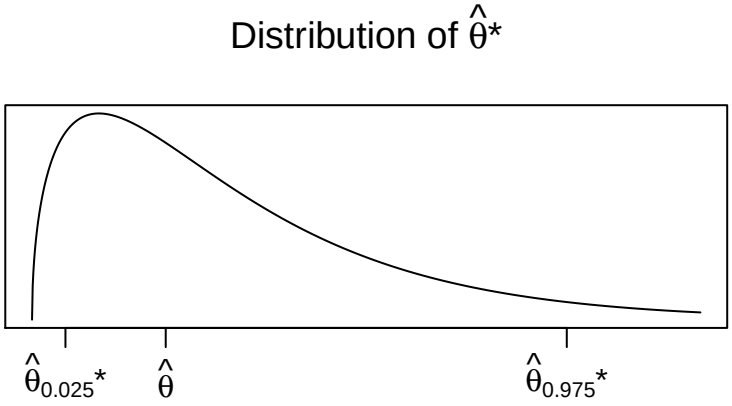
10.3.3 Basic intervals

Unlike the percentile bootstrap interval, the basic interval does not assume the bootstrap distribution is symmetric but does assume that $\hat{\theta}$ is an unbiased estimate for θ.

To address this, we will using the observed distribution of our replicates $\hat{\theta}^*$. Let $\hat{\theta}^*_{\alpha/2}$ and $\hat{\theta}^*_{1-\alpha/2}$ be the $\alpha/2$ and $1-\alpha/2$ quantiles of the replicates $\hat{\theta}^*$. Then another way to form a confidence interval would be

$$\left[\hat{\theta} - \left(\hat{\theta}^*_{1-\alpha/2} - \hat{\theta}\right), \quad \hat{\theta} - \left(\hat{\theta}^*_{\alpha/2} - \hat{\theta}\right)\right]$$

where the minus sign on the upper limit is because $\left(\hat{\theta}^*_{\alpha/2} - \hat{\theta}\right)$ is already negative. The idea behind this interval is that the sampling variability of $\hat{\theta}$ from θ is the same as the sampling variability of the replicates $\hat{\theta}^*$ from $\hat{\theta}$, and that the distribution of $\hat{\theta}$ is possibly skewed, so we can't add/subtract the same amounts. Suppose we observe the distribution of $\hat{\theta}^*$ as



Then any particular value of $\hat{\theta}^*$ could be *much* larger than $\hat{\theta}$. Therefore $\hat{\theta}$ could be *much* larger than θ. Therefore our confidence interval should be $\left[\hat{\theta} - \text{big},\ \hat{\theta} + \text{small}\right]$.

This formula can be simplified to

$$\left[\hat{\theta} - \left(\hat{\theta}^*_{1-\alpha/2} - \hat{\theta}\right)\ , \hat{\theta} + \left(\hat{\theta} - \hat{\theta}^*_{\alpha/2}\right)\right]$$
$$\left[2\hat{\theta} - \hat{\theta}^*_{1-\alpha/2}\ ,\ \ 2\hat{\theta} - \hat{\theta}^*_{\alpha/2}\right]$$

10.3.4 Towards bias-corrected and accelerated intervals (BCa)

Different schemes for creating confidence intervals can get quite complicated. There is a thriving research community investigating different ways of creating intervals and which are better in what instances. The BCa interval is the most general of the bootstrap intervals and makes the fewest assumptions. Unfortunately is can sometimes fail to converge. The details of this method are too complicated to be presented here but can be found in texts such as chapter 12 in Efron and Tibshirani's book *An Introduction to the Bootstrap* (1998).

10.4 Using `car::Boot()` function

For every model we've examined we can create simulated data sets using either case or residual resampling and produce confidence intervals for any of the parameters of interest. We won't bother

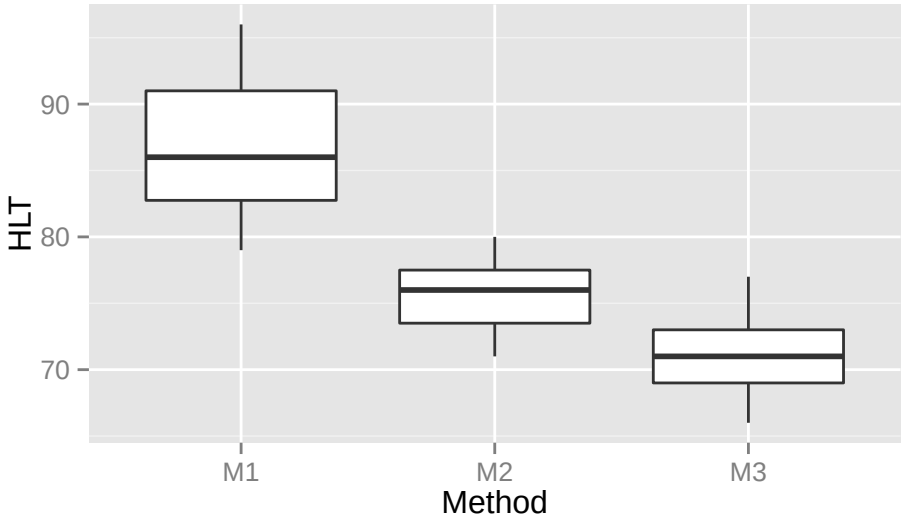
to do this by hand, but rather let R do the work for us. The package that contains most of the primary programs for bootstrapping is the package `boot`. The functions within this package are quite flexible but they are a little complex. While we will use this package directly later, for now we will use the package `car` which has a very convenient function `Boot()`.

We return to our ANOVA example of hostility scores after three different treatment methods.

```
# define the data
Hostility <- data.frame(
  HLT = c(96,79,91,85,83,91,82,87,
          77,76,74,73,78,71,80,
          66,73,69,66,77,73,71,70,74),
  Method = c( rep('M1',8), rep('M2',7), rep('M3',9) ) )
```

The first thing we will do (as we should do in all data analyses) is to graph our data.

```
library(ggplot2)
ggplot(Hostility, aes(x=Method, y=HLT)) +
  geom_boxplot()
```



We can fit the cell-means model and examine the summary statistics using the following code.

```
model <- lm( HLT ~ -1 + Method, data=Hostility )
summary(model)

##
## Call:
## lm(formula = HLT ~ -1 + Method, data = Hostility)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.750 -2.866  0.125  2.571  9.250
##
## Coefficients:
##          Estimate Std. Error t value Pr(>|t|)
## MethodM1   86.750      1.518   57.14   <2e-16 ***
## MethodM2   75.571      1.623   46.56   <2e-16 ***
## MethodM3   71.000      1.431   49.60   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.294 on 21 degrees of freedom
## Multiple R-squared:  0.9973,Adjusted R-squared:  0.997
## F-statistic:  2631 on 3 and 21 DF,  p-value: < 2.2e-16
```

Confidence intervals using the $\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma)$ assumption are given by

```
confint(model)

##             2.5 %   97.5 %
## MethodM1 83.59279 89.90721
## MethodM2 72.19623 78.94663
## MethodM3 68.02335 73.97665
```

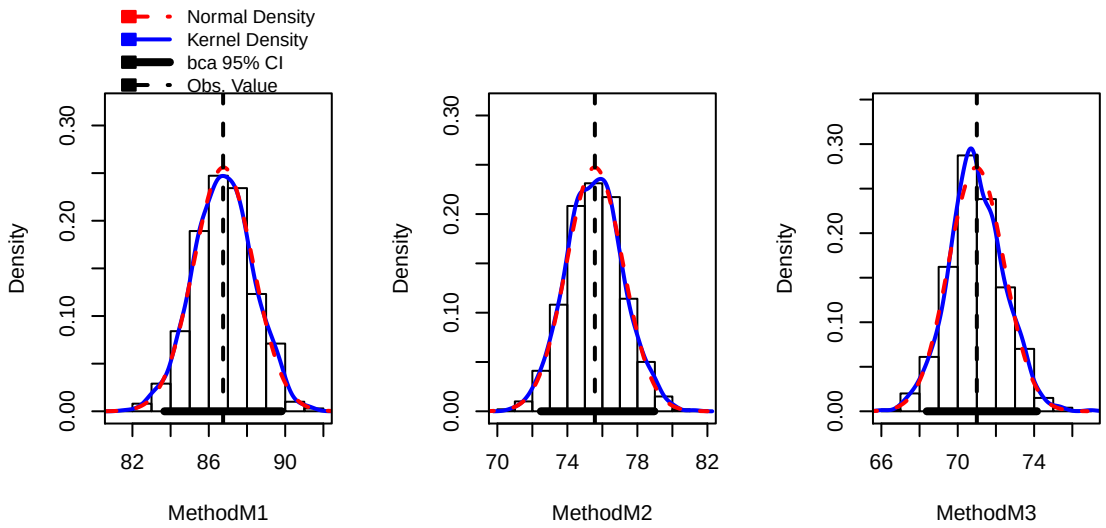
To utilize the bootstrap confidence intervals, we will load the package `car` and use the function `Boot`. It defaults to using case resampling, but `method='residual'` will cause it to use residual resampling. We can control the number of bootstrap replicates it using with the `R` parameter.

```
library(car)
boot.model <- Boot(model, method='case', R=999) # default values
boot.model <- Boot(model, method='residual', R=999) # residual resampling
```

The `Boot()` function has done all work of doing the resampling and storing values of $\hat{\mu}_1, \hat{\mu}_2$, and $\hat{\mu}_3$ for each bootstrap replicate data set created using case resampling. To look at the bootstrap estimate of the sampling distribution of these statistics, we use the `hist()` function.[1]

[1] The `hist()` function is actually overloaded and will act differently depending on the type of object. We will send it an object of class `boot` and the `hist()` function looks for a function name `hist.boot()` and when it finds it, just calls it with the function arguments we passed.

```
hist(boot.model, layout=c(1,3)) # 1 row, 3 columns of plots
```



While this plot is aesthetically displeasing (we could do so much better using `ggplot2`!) this shows the observed bootstrap histogram of $\hat{\mu}_i^*$, along with the normal distribution centered at $\hat{\mu}_i$ with spread equal to the $StdDev\,(\hat{\mu}_i^*)$. In this case, the sampling distribution looks very normal and the bootstrap confidence intervals should line up well with the asymptotic intervals. The function `confint()` will report the BCa intervals by default, but you can ask for "bca", "norm", "basic", "perc".

```
confint(boot.model)

## Bootstrap quantiles, type =  bca
##
##             2.5 %   97.5 %
## MethodM1 83.67302 89.81350
## MethodM2 72.48592 78.97034
## MethodM3 68.37861 74.14954

confint(boot.model, type='perc')

## Bootstrap quantiles, type =  percent
##
##             2.5 %   97.5 %
## MethodM1 83.66222 89.80938
## MethodM2 72.43179 78.82929
## MethodM3 68.04848 73.89276

confint(model)

##             2.5 %   97.5 %
## MethodM1 83.59279 89.90721
## MethodM2 72.19623 78.94663
## MethodM3 68.02335 73.97665
```

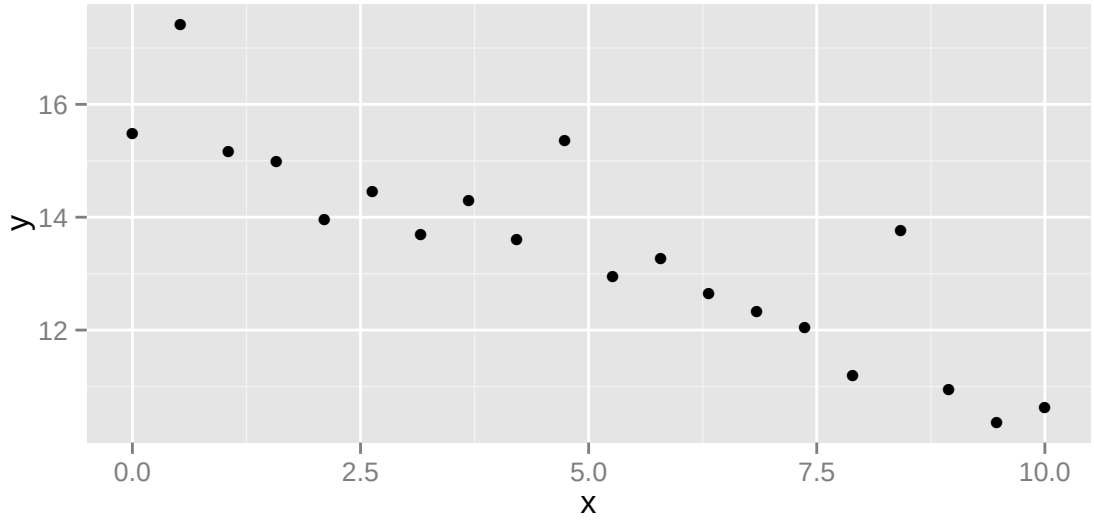
In this case we see that the confidence intervals match up very well with asymptotic intervals.

The `Boot()` function will work for a regression model as well. In the following example, the data was generated from

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

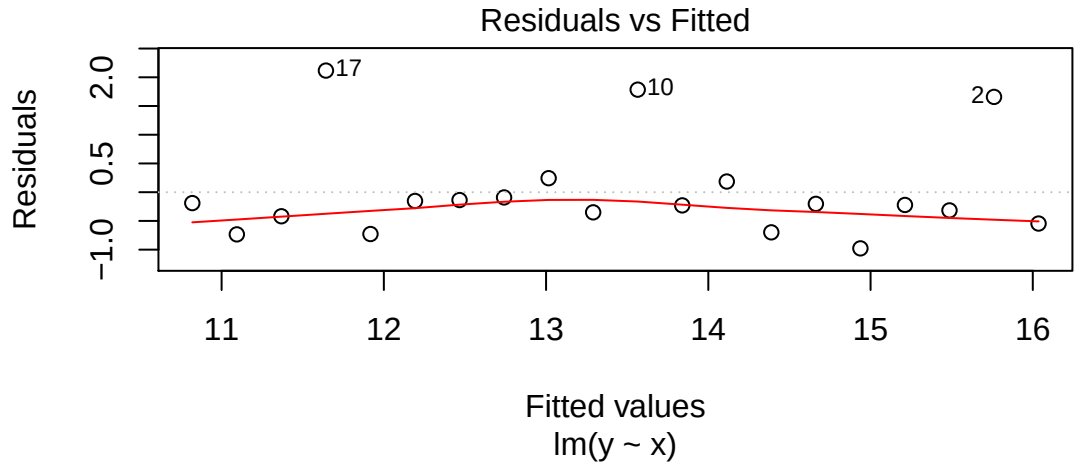
but the ϵ_i terms have a strong positive skew and are not normally distributed.

```
my.data <- data.frame(
  x = seq(0,10, length=20),
  y = c( 15.49, 17.42, 15.17, 14.99, 13.96,
         14.46, 13.69, 14.30, 13.61, 15.35,
         12.94, 13.26, 12.65, 12.33, 12.04,
         11.19, 13.76, 10.95, 10.36, 10.63))
ggplot(my.data, aes(x=x, y=y)) + geom_point()
```



Fitting a linear model, we see a problem that the residuals don't appear to be balanced. The large residuals are all positive. The Shapiro-Wilks test firmly rejects normality of the residuals.

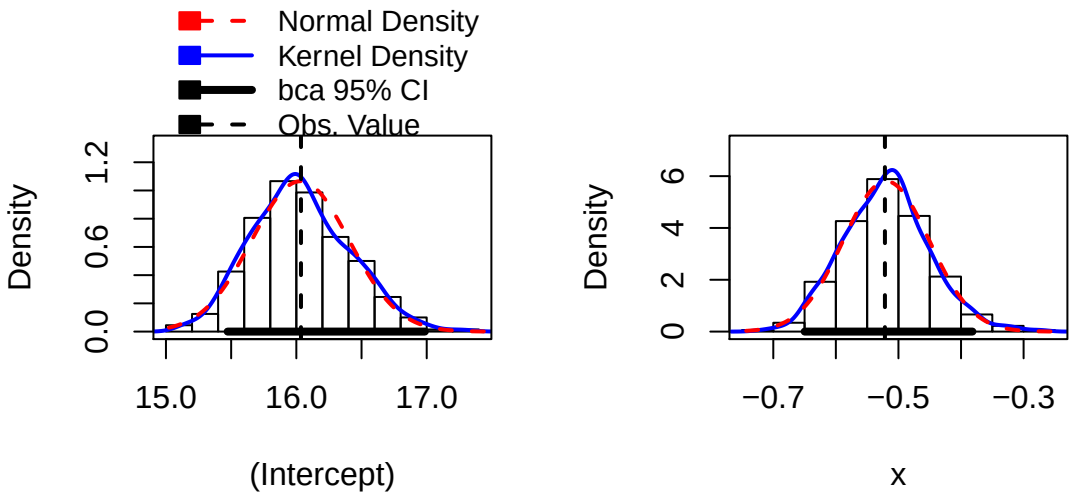
```
model <- lm( y ~ x, data=my.data)
plot(model, which=1)
```




```
shapiro.test( resid(model) )

##
##  Shapiro-Wilk normality test
##
## data:  resid(model)
## W = 0.7732, p-value = 0.0003534
```

As a result, we don't might not feel comfortable using the asymptotic distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$ for the creation of our confidence intervals. The bootstrap procedure can give reasonable good intervals, however.

```
boot.model <- Boot( model )
hist( boot.model )
```




```
confint( boot.model )

## Bootstrap quantiles, type =  bca
##
##                   2.5 %     97.5 %
## (Intercept) 15.4729799 16.9857123
## x           -0.6498092 -0.3817759
```

Notice that both of the bootstrap distribution for both $\hat{\beta}_0^*$ and $\hat{\beta}_1^*$ are skewed, and the BCa intervals are likely to be the most appropriate intervals to use.

10.5 Using the `boot` package

The `car::Boot()` function is very handy, but it lacks flexibility; it assumes that you just want to create bootstrap confidence intervals for the model coefficients[2]. The `car::Boot()` function is actually a nice simple user interface to the `boot` package which is more flexible, but requires the user to be more precise about what statistic should be stored and how the bootstrap samples should be created. We will next examine how to use this package.

Case resampling

Suppose that we have `n` observations in our sample data. Given some vector of numbers resampled from `1:n`, we need to either resample those cases or those residuals and then using the new dataset calculate some statistic. The function `boot()` will require the user to write a function that does this.

[2] Actually it is a little more flexible, but we might as well use see the how to use the `boot` package sooner rather than later.

```
library(boot)

my.data <- data.frame(
  x = seq(0,10, length=20),
  y = c( 15.49, 17.42, 15.17, 14.99, 13.96,
         14.46, 13.69, 14.30, 13.61, 15.35,
         12.94, 13.26, 12.65, 12.33, 12.04,
         11.19, 13.76, 10.95, 10.36, 10.63))
model <- lm( y ~ x, data=my.data )

# Do case resampling with the regression example
my.stat <- function(sample.data, indices){
  data.star <- sample.data[indices, ]
  model.star <- lm(y ~ x, data=data.star)
  output <- coef(model.star)
  return(output)
}

# original group means
my.stat(my.data, 1:20)

## (Intercept)           x
##  16.0355714  -0.5216143

# one bootstrap replicate
my.stat(my.data, resample(1:20))

## (Intercept)           x
##  16.3069314  -0.5786995
```

Notice that the function we write doesn't need to determine the random sample of the indices to use. Our function will be told what indices to use (possibly to calculate the statistic of interest $\hat{\theta}$, or perhaps a bootstrap replicate $\hat{\theta}^*$. For example, the BCa method needs to know the original sample estimates $\hat{\theta}$ to calculate how far the mean $\hat{\theta}^*$ value is from $\hat{\theta}$. The studentized method calculates a bootstrap estimate of variance for each $\hat{\theta}^*$ and so bootstraps the bootstrap. To avoid the user having to see all of that, we just need to take the set of indices given and calculate the statistic of interest.

```
boot.model <- boot(my.data, my.stat, R=1000)
boot.ci(boot.model, type='bca', index=1) # CI for Intercept

## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 1000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot.model, type = "bca", index = 1)
##
## Intervals :
## Level       BCa
## 95%   (15.45, 17.10 )
## Calculations and Intervals on Original Scale
## Some BCa intervals may be unstable
```

```
boot.ci(boot.model, type='bca', index=2) # CI for the Slope

## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 1000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot.model, type = "bca", index = 2)
##
## Intervals :
## Level       BCa
## 95%   (-0.6544, -0.3563 )
## Calculations and Intervals on Original Scale
```

Residual Resampling

We will now consider the ANOVA problem and in this case we will resample the residuals.

```
library(boot)

# Fit the ANOVA model to the Hostility Data
model <- lm( HLT ~ Method, data=Hostility )

# now include the predicted values and residuals to the data frame
Hostility <- Hostility %>% mutate(
  fitted = fitted(model),
  resid  = resid(model))

# Do residual resampling with the regression example
my.stat <- function(sample.data, indices){
  data.star <- sample.data %>% mutate(HLT = fitted + resid[indices])
  model.star <- lm(HLT ~ Method, data=data.star)
  output <- coef(model.star)
  return(output)
}

boot.model <- boot(Hostility, my.stat, R=1000)
```

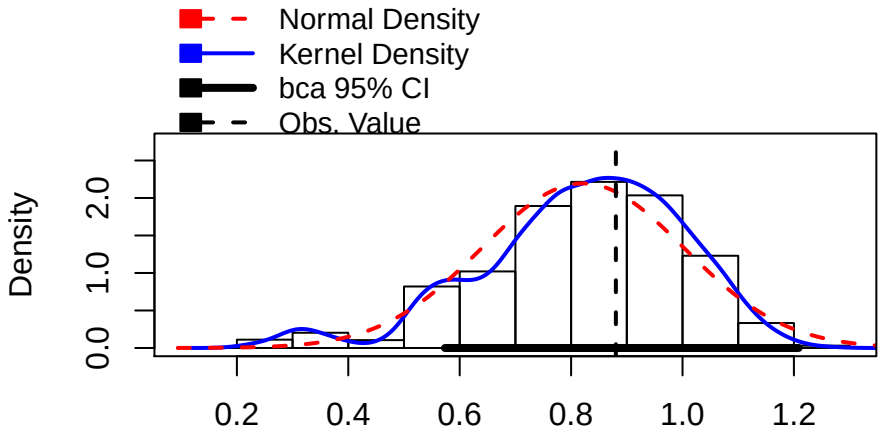
```
boot.ci(boot.model, type='bca', index=1) # Mean of Method 1

## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 1000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot.model, type = "bca", index = 1)
##
## Intervals :
## Level       BCa
## 95%   (84.12, 89.42 )
## Calculations and Intervals on Original Scale
```

```
boot.ci(boot.model, type='bca', index=2) # Mean of Method 2

## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 1000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot.model, type = "bca", index = 2)
##
## Intervals :
## Level       BCa
## 95%   (-14.76,  -6.89 )
## Calculations and Intervals on Original Scale
```

```
boot.ci(boot.model, type='bca', index=3) # Mean of Method 3

## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 1000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot.model, type = "bca", index = 3)
##
## Intervals :
## Level       BCa
## 95%   (-19.80, -12.36 )
## Calculations and Intervals on Original Scale
```

Notice that we don't need to have the model coefficients $\hat{\mu}_i$ be our statistic of interest, we could just as easily produce a confidence interval for the residual standard error $\hat{\sigma}$.

```
# Do residual resampling with the regression example
model <- lm( y ~ x, data=my.data )
my.data <- my.data %>% mutate(
  fitted = fitted(model),
  resid  = resid(model))

# Define the statisitc I care about
my.stat <- function(sample.data, indices){
  data.star <- sample.data %>% mutate(y = fitted + resid[indices])
  model.star <- lm(y ~ x, data=data.star)
  output <- summary(model.star)$sigma
  return(output)
}

boot.model <- boot(my.data, my.stat, R=10000)
hist(boot.model)
```




```
boot.ci(boot.model, type='bca')

## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 10000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot.model, type = "bca")
##
## Intervals :
## Level       BCa
## 95%   ( 0.5736,  1.2085 )
## Calculations and Intervals on Original Scale
```

The decision to resample the cases or the residuals should be made to reflect the original design of the study.

- If the study made no attempt to guarentee a particular range of the x-values or to force a particular number of observations in each of the predictor groups and instead just collected

pairs (x_i, y_i) then the resampling method should reflect that and we should use *case resampling.*

- If the study chose the levels of the predictor variable (x) or decided how many observations in each predictor group, then we should use *residual* resampling.

Chapter 11

Nonparametric Rank-Based Tests

In the most common statistical methods that are introduced in introductory classes rely on the assumption that the distribution of the sample mean is either normal or approximately normal[1]. We then use this distribution to create confidence intervals and do t-tests. If the normality assumption does not hold, then we must turn to some alternative analysis that has fewer assumptions. There is a large body of work to address this situation and we will only present only a few key ideas and methods.

Rank based methods rely on looking *at the order* of the data and not the magnitude. For example, if we have 10 independent observations from a population, all of which are greater than zero, I feel comfortable concluding that the population median is greater than zero. If nine out of ten are larger, I'd still feel comfortable about concluding the population median is greater than zero, but if I had only six out of ten observations larger than zero then I would not reject a null hypothesis that the population median was equal to ten.

Bootstrapping is another method often used in these sorts of problems. Instead of ignoring the observations magnitude, we use observed distribution of data as an estimate of the model distribution. We will discuss this method in the following chapter.

These methods are typically referred to as *nonparametric* methods and care should be taken to recognize that these tests are not assumption-less as we will require the observations to be independent and identically distributed.

Finally, there is a price to be paid for using a more general method. If the normality assumption is true, these nonparametric tests will have less power to reject the null hypothesis than the corresponding method that uses normality. Therefore, the standard methods should be used when appropriate and the nonparametric alternative only used when the normality assumption is substantially violated.

11.1 Alternatives to one sample and paired t-tests

We often want to take a sample of observed values and make statistical inference about the mean or median of the population that the observations came from. Suppose we have a sample of data z_i coming from a non-normal distribution and want to test if the mean μ or median M is equal to some specified value μ_0 or M_0.

The literature commonly introduces these tests as alternatives to the paired t-test. However, recall that the paired t-test was just a single sample t-test performed on the differences between paired observations. In that case our observed data is just

$$z_i = x_i - y_i$$

[1] If the population that the data is drawn from is normal, then the sample mean is normal. If the sample size is large ($n > 30$ is usually sufficient) then the Central Limit Theorem states that the sample mean is approximately normally distributed.

for $i = 1 \ldots n$. Keeping with standard practice and the most likely use of these tests, we present these tests in the paired t-test context, and note that the modification to a one-sampled t-test is usually trivial.

11.1.1 Sign Test

This is the most easily understood of the rank based tests, but suffers from a lack of power. Typically the Wilcoxon Sign Rank test is preferred, but we present the Sign Test as it is an extremely flexible test and is a good introduction to thinking about rank based tests.

Hypothesis

We are interested in testing if the medians of two populations are equal versus an alternative of not equal.

$$\begin{aligned} H_0 : M_1 - M_2 &= 0 \\ H_a : M_1 - M_2 &\neq 0 \end{aligned}$$

One sided tests are also possible,

$$H_a : M_1 - M_2 > 0$$

Assumptions

One very nice aspect of the Sign Test is that it has very few assumptions, only that the paired observations (x_i, y_i) are independent and identically distributed. In particular we note that there is no symmetric assumption on the distribution of z_i.

Calculation

Calculate $z_i = x_i - y_i$ and observe the sign[2]. We define our test statistic T to be the number of positive values of z_i. If an observation $z_i = 0$, we'll remove it from the analysis.

Sampling Distribution

Under the null hypothesis, the two samples have the same median and so the sign of the difference z_i should be negative approximately half the time and positive half the time. In fact, under the null hypothesis, if we define a positive z_i value to be a success, then our test statistic T has a binomial distribution with success probability $\pi = 1/2$.

$$T \sim Binomial\left(m, \pi = \frac{1}{2}\right)$$

where m is the number of non-zero z_i values.

Example

Suppose we have data for 7 students from two exams of a class and we want to evaluate if the first exam was harder than the second.

[2] In the case one sample case, we observe the sign of $z_i = x_i - M_0$.

Student	Exam 1	Exam 2	$z_i = Exam_1 - Exam_2$
1	66	71	-5
2	74	76	-2
3	85	84	1
4	81	85	-4
5	93	93	0
6	88	90	-2
7	79	78	1

Here we have $t = 2$ positive values out of $m = 6$ nonzero observations.

Recall that a p-value is the probability of seeing your data or something more extreme given the null hypothesis is true. In this case our p-value is the probability that $T \leq 2$. Using the binomial distribution, the p-value for this test is

$$p - value = P(T \leq 2) = \sum_{i=0}^{2} P(T = i) = 0.34375$$

which can be found using R

```
dbinom(0, size=6, prob=1/2) + dbinom(1,6,1/2) +          dbinom(2,6,1/2)

## [1] 0.34375
```

As usual, if we had been interested in a two-sided alternative, we would multiply the p-value by two.

11.1.2 Wilcoxon Sign Rank Test

While the sign test is quite flexible, ignoring the magnitude of the differences is undesirable. The Wilcoxon Sign Rank test will utilize that information and is typically a more powerful test.

Hypothesis

As with the Sign Test, we are interested in testing if the medians of two populations are equal versus an alternative of not equal.

$$\begin{aligned} H_0 : M_1 - M_2 &= 0 \\ H_a : M_1 - M_2 &\neq 0 \end{aligned}$$

One sided tests are also possible,

$$\begin{aligned} H_a : M_1 - M_2 &> 0 \\ H_a : M_1 - M_2 &< 0 \end{aligned}$$

Assumptions

As with the Sign Test, we require that the paired observations (x_i, y_i) are independent and identically distributed. We further impose an additional assumption the the differences are symmetric around some value.

Calculation

As with the Sign Test, we calculate[3] $z_i = x_i - y_i$. Next order the absolute values $|z_i|$, and as in the Sign Test, observations with $z_i = 0$ are removed from the data set. Using the sorted values calculate

[3] In the case one sample case, calculate and order the values $z_i = x_i - M_0$.

the rank R_i of each observation where the rank of 1 is the observation with the smallest magnitude, and m is corresponds to the largest observation. In the case of ties, use the average rank.

Next define

$$\phi_i = \begin{cases} 0 & \text{if } z_i < 0 \\ 1 & \text{if } z_i > 0 \end{cases}$$

to be an indicator function denoting if $z_i > 0$. Finally we define

$$W_+ = \sum_{i=1}^{m} \phi_i R_i$$

and

$$W_- = \sum_{i=1}^{m} (1 - \phi_i) R_i$$

so that W_+ is the sum of the ranks of the positive z_i values and W_- is the sum of the ranks of the negative z_i values. If there are no positive ranks, then define $W_+ = 0$. Likewise if there are no negative ranks, define $W_- = 0$. Let[4]

$$S = \min [W_+, W_-]$$

Sampling Distribution

Under the null hypothesis, we would expect W_+ and W_- to be approximately the same. Unfortunately the distribution of S under the null hypothesis is not a distribution that we recognize, but it can be calculated. The quantiles of the distribution can be found in tables in statistics books or using R.

Example

We again use the student test data and we wish to test if median of Exam 1 is less than the median of Exam 2.

Student	Exam 1	Exam 2	$z_i = Exam_1 - Exam_2$
1	66	71	-5
2	74	76	-2
3	85	84	1
4	81	85	-4
5	93	93	0
6	88	90	-2
7	79	78	1

We now sort the absolute values and remove the zero observations

z_i	$\lvert z_i \rvert$	R_i	R_i after accounting for ties
-5	5	6	6
-4	4	5	5
-2	2	4	3.5
-2	2	3	3.5
1	1	2	1.5
1	1	1	1.5

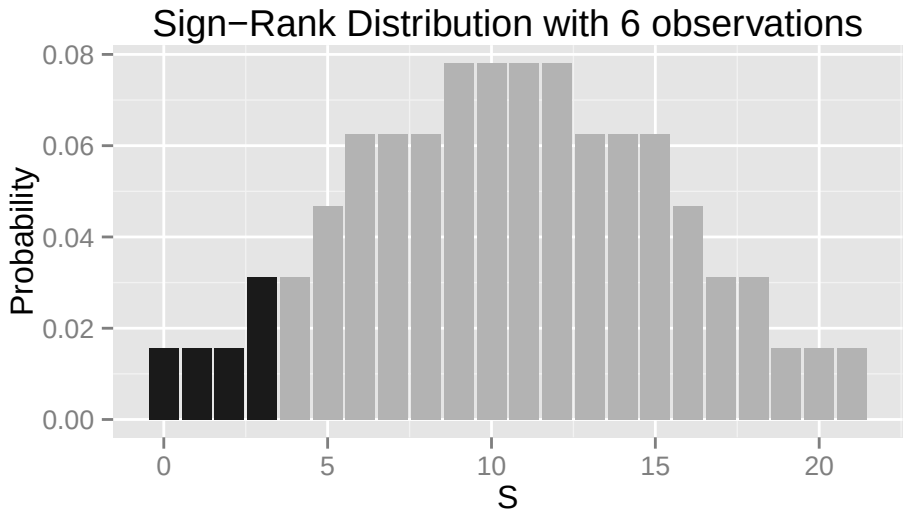
and then calculate

$$\begin{aligned} W_- &= 6 + 5 + 3.5 + 3.5 = 18 \\ W_+ &= 1.5 + 1.5 = 3 \end{aligned}$$

[4]For the alternative $M_1 - M_2 > 0$ then $S = W_-$. For the alternative $M_1 - M_2 < 0$ use $S = W_+$

and thus we will use $S = 3$.

To calculate a p-value we want to find $P(S \leq 3)$.



which we do using a table look up in R. Notice I could look up either the probability of observing a 3 or less or the probability of observing 18 or more.

```
# less than or equal to 3
psignrank(3, 6)

## [1] 0.078125

# greater than or equal to 18
1 - psignrank( 17, 6 )

## [1] 0.078125
```

Example in R

The function that we will use for both Wilcoxon's Sign Rank and Rank Sum tests is `wilcox.text()`. You can pass the function either one vector of data or two and can indicate if the test should be a paired test.

```
exam.1 <- c(66, 74, 85, 81, 93, 88, 79)
exam.2 <- c(71, 76, 84, 85, 93, 90, 78)
wilcox.test(exam.1, exam.2, paired=TRUE, alternative='less')

## Warning in wilcox.test.default(exam.1, exam.2, paired = TRUE, alternative = "less"):
cannot compute exact p-value with ties
## Warning in wilcox.test.default(exam.1, exam.2, paired = TRUE, alternative = "less"):
cannot compute exact p-value with zeroes

##
##  Wilcoxon signed rank test with continuity correction
##
## data:  exam.1 and exam.2
## V = 3, p-value = 0.07001
## alternative hypothesis: true location shift is less than 0
```

Notice that the p-value is slightly different than when we used the `signrank()` distribution. This is due to an approximation to the actual sign rank distribution being used whenever ties occur in the data. Because the only tie occurred in the same group, we could have used the actual distribution, but the function `wilcox.test` immediately jumped to the approximation.

Also notice how we are interested in testing if exam 1 was harder than exam 2 and so we want the alternative to be

$$H_a: \ exam_1 < exam_2$$

so because I input `exam.1` first and `exam.2` second, then the appropriate alternative is `'less'` because I want to test is `first.argument < second.argument`. If we had changed the order of the exam vectors, I would have to also switch the alternative to `'greater'`.

11.2 Alternatives to the two sample t-test

11.2.1 Wilcoxon Rank Sum Test

The Wilcoxon Rank Sum Test is the nonparametric alternative to the two sample t-test. We are interested in testing if the medians of two populations are equal versus an alternative of not equal, but we have independent samples from each population and there is no way to pair an observations from the two populations.

Let n_1 be the number of observations from the first group, and n_2 be the number from the second group.

Assumptions The assumptions for the Rank Sum Test are that all the observations are independent (both between and within samples).

Hypothesis Again our hypotheses

$$\begin{aligned} H_0: M_1 - M_2 &= 0 \\ H_a: M_1 - M_2 &\neq 0 \end{aligned}$$

One sided tests are also possible,

$$\begin{aligned} H_a: M_1 - M_2 &> 0 \\ H_a: M_1 - M_2 &< 0 \end{aligned}$$

Calculation Combine observations from both samples and order them. In the case of ties, assign the average rank. Next define T_1 as the sum of the ranks for observations in sample 1 and likewise define T_2.

Sampling Distribution Under the null hypothesis, T_1 and T_2 should be approximately equivalent and if they have an extremely large difference. We compare the smaller of T_1 and T_2 against the null distribution and the null distribution quantiles can be found in tables in various statistics books or using R.

Example Ten tents using plain camouflage (group 1) and ten using patterned camouflage (group 2) are set up in a wooded area, and a team of observers is sent out to find them. The team reports the distance at which they first sight each tent until all 20 tents are found. The distances at which each tent is detected are reported:

Distance	10	12	14	16	16	18	20	20	21	21	22	25	26	28	29	32	34	36	38	43
Group	2	2	2	1	2	2	2	2	1	2	2	1	2	1	1	1	1	1	1	1
Rank	1	2	3	4.5	4.5	6	7.5	7.5	9.5	9.5	11	12	13	14	15	16	17	18	19	20

We calculated

$$\begin{aligned} T_1 &= 4.5 + 9.5 + 12 + 14 + 15 + 16 + 17 + 18 + 19 + 20 = 145 \\ T_2 &= 1 + 2 + 3 + 4.5 + 6 + 7.5 + 7.5 + 9.5 + 11 + 13 = 65 \end{aligned}$$

and compare T_2 to the sampling distribution of under the null hypothesis.

Unfortunately the literature is somewhat inconsistent as to the definition of T_1 and T_2. It seems that Wilcoxon's original paper used the unadjusted ranks while subsequent tables subtracted the minimum rank. Further complicating the matter is that there are corrections that should be made if there are too many ties. The end result is that calculating the test statistic by hand and comparing it to the "right" Wilcoxon distribution is troublesome.

The Wilcoxon Rank Sum test is completely equivalent to the Mann-Whitney test and the Mann-Whitney test became more widely used because it dealt with unequal sample sizes more easily. Since the tests are equivalent, disturbingly, some software programs will return the test statistic for one when the user asked for the other. While the p-values will be identical, the test statistic will not.[5]

11.2.2 Mann-Whitney

We have the same assumptions and hypotheses as the Wilcoxon Rank Sum Test. For notational convenience, let x_i be an observation from sample 1 and y_j be an observation from sample 2.

Calculation For all $n_1 n_2$ combinations of pairs of observations (x_i, y_j), let U be the number of times $x_i > y_j$. If they are equal, count the combination as $1/2$.

Sampling Distribution Under the null hypothesis, we would expect $U \approx n_1 n_2/2$. If U is too big or too small, we should reject the null hypothesis.

Example - Camouflage Tents

The Mann-Whitney U statistic is 10, with 9 instances where a group 1 observation is less than a group 2 observation and two instances of ties.

```
patt  <- c(10,12,14,16,18,20,20,21,22,26)
plain <- c(16,21,25,28,29,32,34,36,38,43)
wilcox.test(patt, plain, paired=FALSE, alternative='less')

## Warning in wilcox.test.default(patt, plain, paired = FALSE, alternative = "less"):
cannot compute exact p-value with ties

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  patt and plain
## W = 10, p-value = 0.001398
## alternative hypothesis: true location shift is less than 0
```

[5] R returns the results of the Mann-Whitney test from the function `wilcox.test()` in the two sample case.