Introduction to Statistics for Researchers

Derek Sonderegger

April 18, 2014

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.

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. I attempt to introduce the software with as little pain as possible, but some frustration is inevitable.

Since 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, but these notes are mature enough that I don't anticipate large changes.

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

Derek Sonderegger
Department of Mathematics and Statistics
Northern Arizona University

Contents

1 Preliminaries **5**

2 Summary Statistics and Graphing **7**
2.1 Measures of Centrality . . . 8
2.2 Measures of Variation . . . 9
2.3 Graphical summaries of Data . . . 12
2.3.1 Univariate - Categorical . . . 12
2.3.2 Univariate - Continuous . . . 13
2.3.3 Bivariate - Categorical vs Continuous . . . 15
2.3.4 Bivariate - Continuous vs Continuous . . . 17
2.3.5 Notable Graphs . . . 18

3 Probability **20**
3.1 Introduction to Set Theory . . . 20
3.1.1 Venn Diagrams . . . 20
3.1.2 Composition of events . . . 21
3.2 Probability Rules . . . 22
3.2.1 Simple Rules . . . 22
3.2.2 Conditional Probability . . . 24
3.2.3 Bayes' Formula . . . 26
3.2.4 Summary of Probability Rules . . . 28
3.3 Discrete Random Variables . . . 28
3.3.1 Introduction to Discrete Random Variables . . . 28
3.3.2 Binomial Distribution . . . 30
3.3.3 Poisson Distribution . . . 33
3.4 Continuous Random Variables . . . 35
3.4.1 Uniform Distribution . . . 35
3.4.2 Exponential Distribution . . . 35
3.4.3 Normal Distribution . . . 37

4 Maximum Likelihood **41**
4.1 Likelihood Function . . . 41
4.1.1 Binomial Distribution . . . 41
4.1.2 Poisson Distribution . . . 43
4.2 Maximization . . . 45
4.2.1 Calculus Solution . . . 46
4.2.2 Numerical Solution . . . 46
4.3 Normal Distribution . . . 47
4.3.1 Theory . . . 47
4.4 Asymptotic Results for MLEs . . . 49
4.5 Connection to Bayesian Analysis . . . 49

5 Sampling Distributions **50**
5.1 Mean and Variance of the Sample Mean . . . 52
5.2 Distribution of $\bar{X}$ if the samples were drawn from a normal distribution . . . 54

6 Confidence Intervals and T-tests **58**
6.1 Confidence Intervals assuming σ is known . . . 58
6.2 Confidence interval for μ assuming σ is unknown . . . 61
6.2.1 t-distributions . . . 61
6.2.2 Sample Size Selection . . . 63
6.3 Hypothesis Testing . . . 63
6.3.1 Writing Hypotheses . . . 64
6.3.2 Calculating p-values . . . 68
6.3.3 Calculating p-values vs cutoff values . . . 69
6.3.4 t-tests in R . . . 69
6.4 Relationship between Confidence Intervals and Hypothesis Tests . . . 70
6.4.1 One-Sided Hypothesis Tests and their associated Confidence Intervals . . . 71
6.5 Type I and Type II Errors . . . 73
6.5.1 Power and Sample Size Selection . . . 74
6.6 Variations of the t-test: Comparing two population means . . . 77
6.6.1 Paired t-Tests . . . 78
6.6.2 Two Sample t-test . . . 79
6.6.3 Two sample t-test using a pooled variance estimator . . . 81

7 Testing Model Assumptions **84**
7.1 Testing Normality . . . 84
7.1.1 Visual Inspection - QQplots . . . 84
7.1.2 Tests for Normality . . . 88
7.2 Testing Equal Variance . . . 88
7.2.1 Visual Inspection . . . 88
7.2.2 Tests for Equal Variance . . . 89

8 Analysis of Variance **94**
8.1 Model . . . 94
8.2 Theory . . . 96
8.2.1 Anova Table . . . 96
8.2.2 ANOVA using Simple vs Complex models. . . . 97
8.2.3 Parameter Estimates and Confidence Intervals . . . 98
8.3 Anova in R . . . 99
8.4 Multiple comparisons . . . 101
8.5 Different Model Representations . . . 105
8.5.1 Theory . . . 105
8.5.2 Model Representations in R . . . 107
8.5.3 Implications on the ANOVA table . . . 109

9 Regression **112**
9.1 Pearson's Correlation Coefficient . . . 112
9.2 Model Theory . . . 114
9.2.1 Anova Interpretation . . . 118
9.2.2 Confidence Intervals vs Prediction Intervals . . . 119
9.3 Extrapolation . . . 123
9.4 Checking Model Assumptions . . . 124
9.5 Influential Points . . . 126
9.6 Transformations . . . 127

10 Nonparametric Rank-Based Tests **130**
10.1 Alternatives to one sample and paired t-tests 130
10.1.1 Sign Test 131
10.1.2 Wilcoxon Sign Rank Test 132
10.2 Alternatives to the two sample t-test 135
10.2.1 Wilcoxon Rank Sum Test 135
10.2.2 Mann-Whitney 136

Chapter 1

Preliminaries

The goal of statistics is to obtain data that relates to some process of interest and use those data to make inferences about the unobserved process. For example, we could be interested in the relationship between a person's income level and their level of education. The observed data will be summarized by some quantity called a statistic. If we had observed a different sample, then that statistic would have been different, but hopefully not too different and both sample statistics should be "close" to the true value of the process. The quantification of "close" is the heart of statistics.

To make this concrete, suppose that I have a jar of 1000 M&Ms that is composed of 28% brown and 72% green candies. Suppose that I didn't know the true proportion of green M&Ms, which I will denote as $\pi = 0.72$. The quantity is called a parameter and not a statistic because it is calculated from the full population of interest (statistics are calculated from sampled data) I will take a sample of size $n = 20$ and use this sample to calculate the sample proportion statistic. This sample proportion should be close to the true population proportion, but won't be exactly correct (with 20 observations, the sample proportion could be 0.7 or 0.75, but not in between!)

The mathematics of probability will tell us how close our sample proportion can be expected to be from the population parameter and in turn, yield a region in which π is likely to be contained.

It is traditional for introductory statistics books to start with a chapter about gathering data, but this often seems out of place. Instead we will begin by emphasizing that gathering high quality data is *hard* but a well designed collection scheme is absolutely critical.

Example - A rather improbable example should show the differences between an observational study and an experiment. Suppose that in the human genome there is a gene that causes lung cancer. If a person has this gene, they will always get lung cancer. But genes often do more than one thing, and suppose that this gene also compels a person to smoke cigarettes. This situation is unlikely but possible. An observational study would gather data from people and observe whether they smoke and if they have cancer. An association between smoking and cancer will show up. If we preformed an experiment were done we would get a large group of individuals and then randomly assign each individual to a treatment level, in this case, smoking or non-smoking. Since the people with the gene for cancer will be even mixed among the two groups, the cancer rates between the two treatments will be equivalent. This random assignment is how statisticians control for these unknown, "lurking" variables. As a result the experiment shows that there is no cause/effect relationship

Example - For years it was thought that prescribing hormone therapy (estrogen and progestin) reduced the rate of rate of breast cancer in post-menopausal women. This belief was the result of many observational studies. The thing that those observational studies were missing 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.

To 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 (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 are evenly spread across treatments.

Two variables are said to be or **confounding** if the design of the study is such that you cannot distinguish between the effects. In both observational studies and experiments, the researcher should try to design the study to avoid confounding between known explanatory variables. The random assignment of treatments in an experiments allows for the researcher to avoid confounding between known explanatory variables and unidentified lurking variables.

When designing an experiment or sample, there are three things to keep in mind:

1. How can independent data be gathered? If that is not possible, how can it be gathered in a fashion such that the dependence structure is known?

2. The sampled population should be representative of the population of interest.

3. The design should be free of confounding effects.

Chapter 2

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 as much as possible.

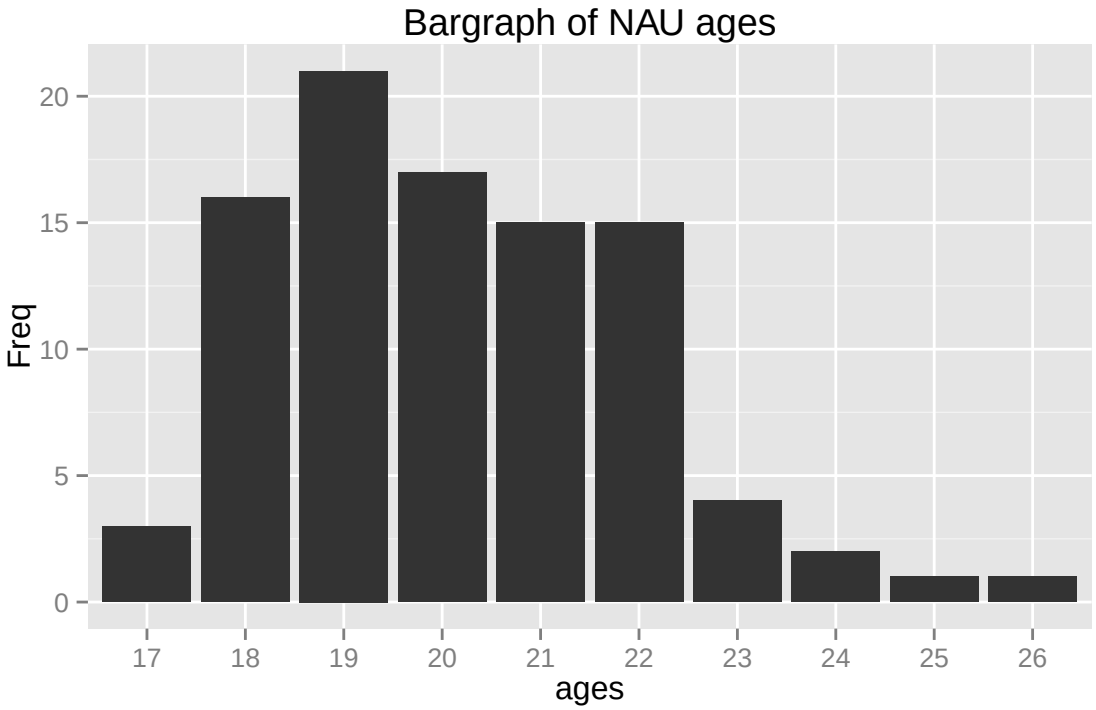
For this section, suppose we wish to summarize a dataset of ages of NAU undergraduates. We have $n = 95$ observations.

```
ages <- c(20, 19, 22, 18, 19, 19, 21, 20, 20, 22, 19, 23,
          19, 20, 19, 18, 20, 18, 19, 22, 17, 19, 18, 21,
          22, 21, 24, 19, 20, 19, 22, 22, 24, 18, 21, 22,
          18, 21, 21, 18, 17, 23, 19, 18, 19, 22, 25, 19,
          18, 21, 19, 20, 22, 19, 18, 26, 22, 21, 18, 17,
          18, 20, 20, 21, 21, 18, 18, 23, 19, 23, 22, 21,
          19, 18, 20, 22, 19, 20, 19, 22, 20, 19, 20, 19,
          22, 21, 21, 20, 20, 20, 22, 21, 21, 18, 20)
n <- length(ages)
```

Perhaps the best summary of the data is in a graphical form[1].

```
library(ggplot2)
temp <- data.frame(table(ages))
ggplot(temp, aes(x=ages, y=Freq)) +
  geom_bar() +
  labs(title='Bargraph of NAU ages')
```

[1] At first glance the code used to produce this graph is quite complicated. Don't worry too much about how the graph is created for now, as the code is only here for reference.



2.1 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

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

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

```
sum(ages) / n
```

```
## [1] 20.15
```

```
mean(ages)
```

```
## [1] 20.15
```

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()`.

```
median(ages)

## [1] 20
```

Mode

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

Examples

- If my father were to become bored with retirement and enroll as an undergraduate and I included him in my dataset, which measure of centrality would change the most and why?
  - The answer is that the mean would move much more than the median. Since the 47th and 48th largest observations are both 20, the median would remain 20. 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.

2.2 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(ages) - min(ages)

## [1] 9
```

Inter-Quartile Range

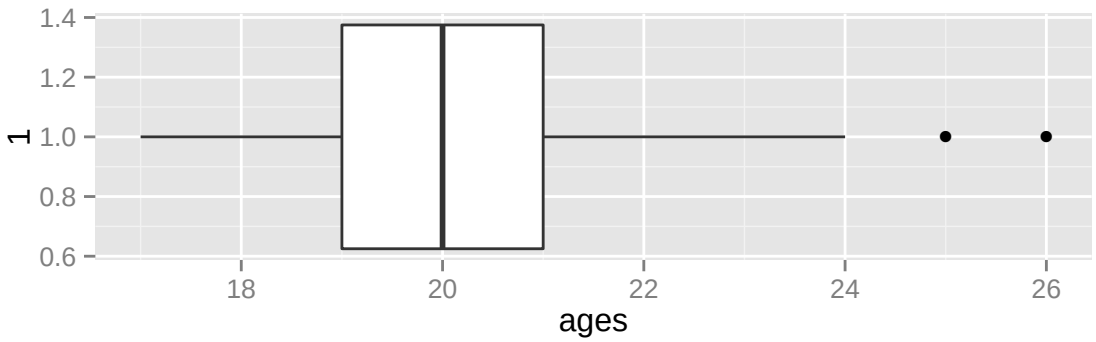
The **p-th** percentile is the observation (or observations) that has at most p percent of the observations below it and $(1 - p)$ above it, where p is between 0 and 100. The median is the 50th percentile. Often we are interested in splitting the data into four equal sections using the 25th, 50th, and 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.

```
quantile(ages)

##   0%  25%  50%  75% 100%
##   17   19   20   21   26

quantile(ages, 0.33)

## 33%
##  19
```

The inter-quartile range is defined as the distance from the 3rd quartile to the 1st. Often this is denoted as IQR.

```
ggplot( data.frame(ages), aes(y=ages, x=1)) +
  geom_boxplot() +
  coord_flip()
```



The above graph is a box-and-whisker plot of undergraduate student ages[2]. The solid bar in the middle of the box is the median, the edges of the box are the 25th and 75th percentile (sometimes called the 1st and 3rd quartiles). The whiskers extend out the to the maximum and minimum non-outlier data points. Conventionally outliers are defined as data points that lay farther than $1.5 * IQR$ from the nearest quartile.

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 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.

[2] It is surprisingly hard to get `ggplot2` to make such a simple plot.

Absolute values are a gigantic pain to deal with. So we square it 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(ages)

## [1] 3.34
```

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(ages)

## [1] 1.827
```

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:

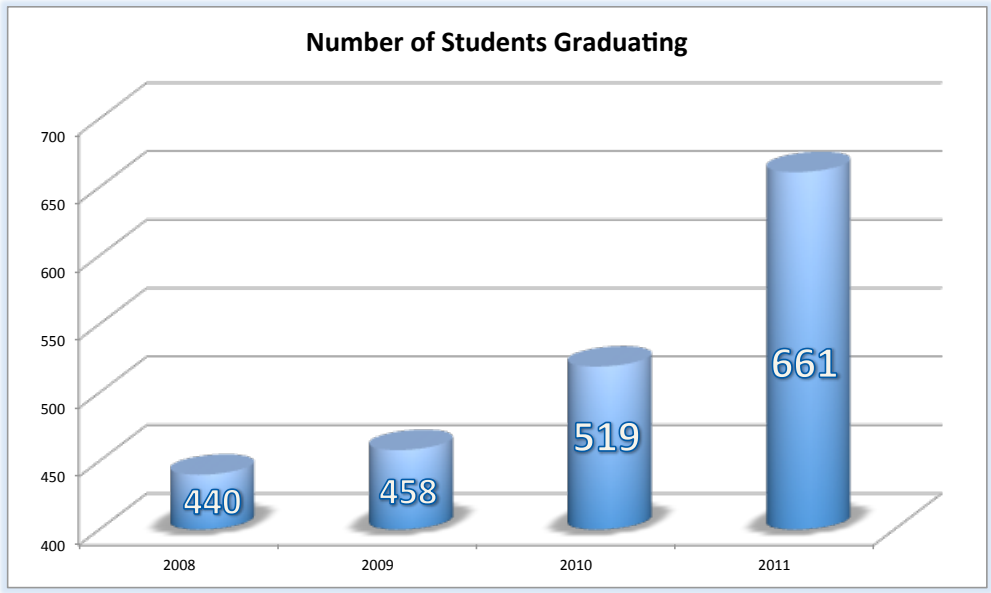
Interval	Approximate percent of measurements
$\bar{x} \pm s$	68%
$\bar{x} \pm 2s$	95%
$\bar{x} \pm 3s$	99.7%

2.3 Graphical summaries of Data

2.3.1 Univariate - Categorical

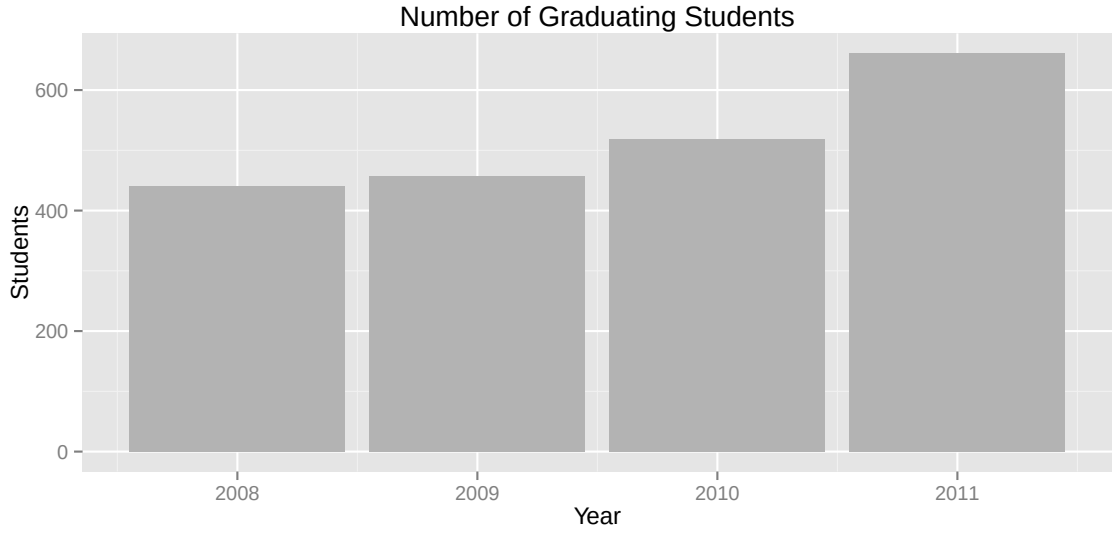
If we have univariate data about a number of groups, often the best way to display it is using barplots. They have the advantage over pie-charts that groups are easily compared. The only tricky thing about barcharts is how you scale the y-axis. In general, I prefer that the y-axis is scaled from zero because an offset scale tends to over-emphasize differences.

Consider the following graph. At first glance, it appears that the number of graduating students in 2011 is 5 times that of 2008, when in fact it is approximately 1.5 times as large. Furthermore, the 3-D effect distorts the graph making the 2011 year appear to have only 650 graduating students.



This graph should have been made as follows:

```
Year <- c(2008,2009,2010,2011)
Students <- c(440,458,519,661)
data <- data.frame( Year=Year, Students=Students)
ggplot(data, aes(x=Year, y=Students)) +
   geom_bar(stat='identity', fill='grey70') +
   labs(title='Number of Graduating Students')
```



2.3.2 Univariate - Continuous

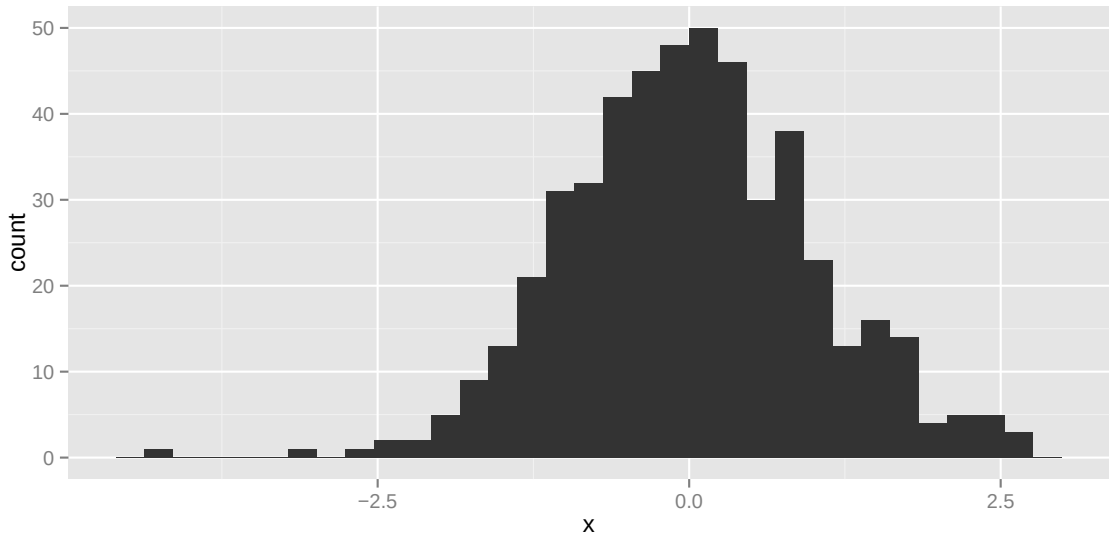
Suppose we are looking at test scores from a class of students and we are interested in summarizing the distribution. If there is a small number of observations, then a stem-and-leaf plot is quite nice because you still see all of the data, but the data is arranged nicely.

```
grades <- c(89,79,83,81,92,95,85,96,91,86,84,85,76,92,71,88,91,80)
stem(grades)

##
##   The decimal point is 1 digit(s) to the right of the |
##
##   7 | 169
##   8 | 013455689
##   9 | 112256
```

In this plot, the left hand digit represents the tens place, and the right hand side represents the ones. So 1 in the first row represents the 71, and the 6 represents the 76.

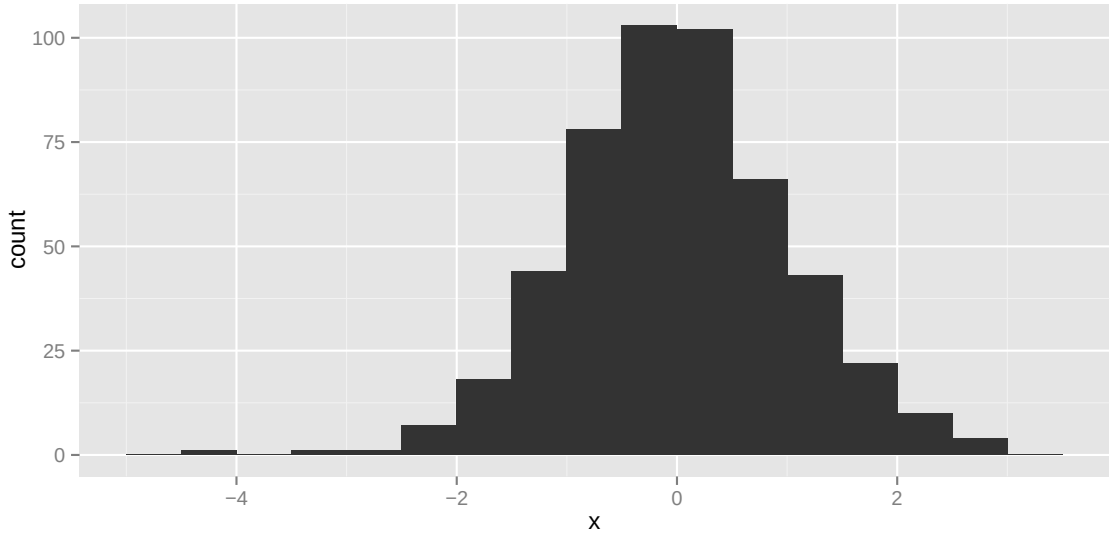
When we have large numbers of observations, the stem-and-leaf plot becomes overwhelming and we must turn to histogram. A histogram looks very similar to a bar plot, but is used to represent continuous data instead of categorical and therefore the bars will actually be touching. In this example, we draw a random sample of 500 observations from a normal distribution and plot their histogram.

```
x <- rnorm(500, mean=0, sd=1)
qplot(x)
```



Unfortunately histograms can be somewhat misleading due to the selection of break points between bins. It is advisable to look at several different resolutions of the data.

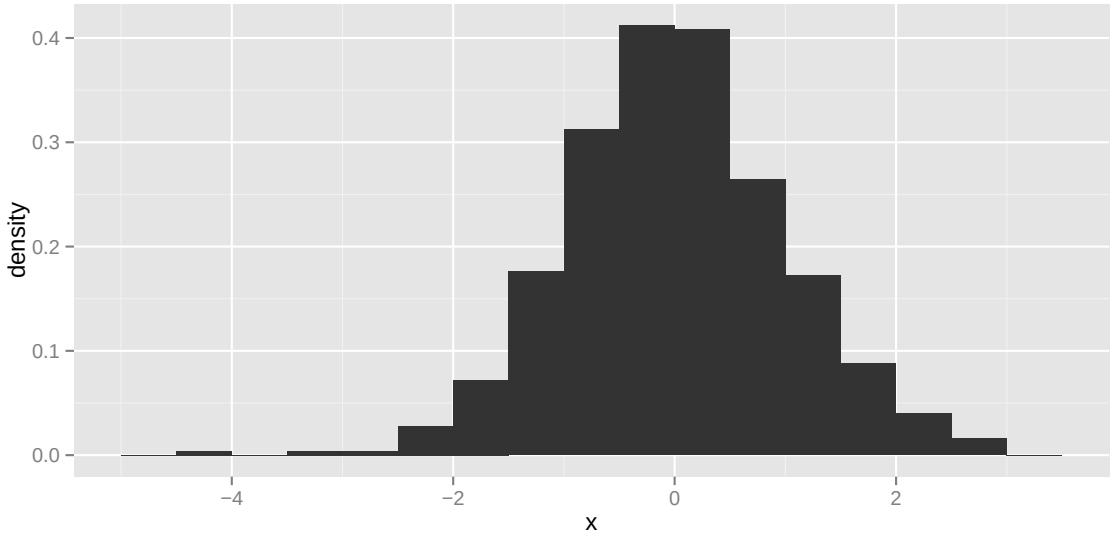
```
qplot(x, binwidth=.5)
```



The final change to the histogram that you might see is a change in the y-axis from frequency (number of observations in a bin) to a number associated with the percent of observations in a particular bin. But it is also 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

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

```
data <- data.frame(x=x)
ggplot(data, aes(x=x, y=..density..)) +
  geom_histogram(binwidth=.5)
```

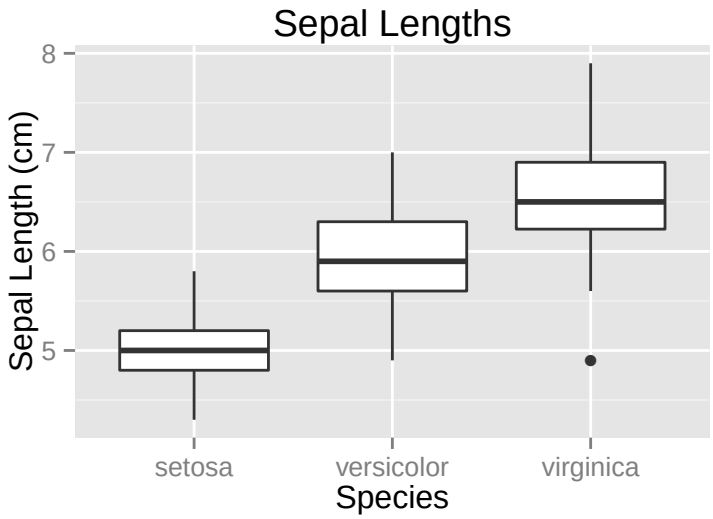


2.3.3 Bivariate - Categorical vs Continuous

We often wish to compare response levels from two or more groups of interest. To do this, we often use side-by-side boxplots. Boxplots were first introduced in the 1960's by John Tukey. Notice that each observation is associated with a continuous response value and a categorical value.

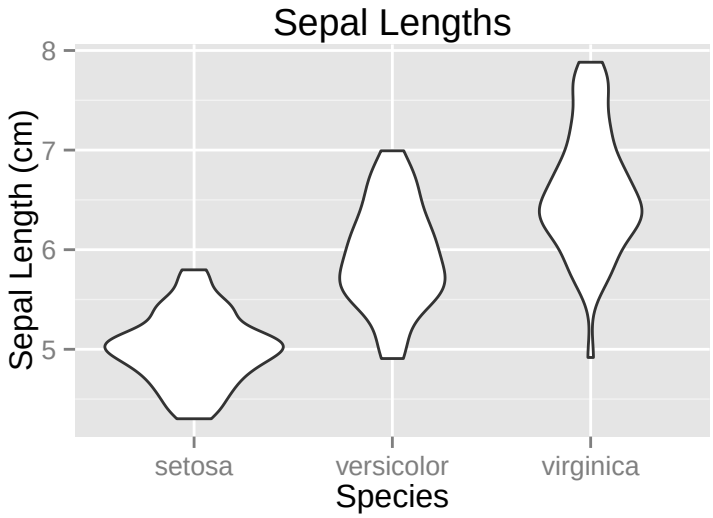
In this example, we will examine the pedal length of three different species of iris. The data we will use is a built in dataset in R called `iris`.

```
str(iris)

## 'data.frame': 150 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...

qplot(data=iris, x=Species, y=Sepal.Length, geom='boxplot',
      main='Sepal Lengths', ylab='Sepal Length (cm)')
```



There have been several attempts to create a better boxplot and the best is one that allows for easy comparison of (mirrored) histograms. This plot is called a *violin plot*.

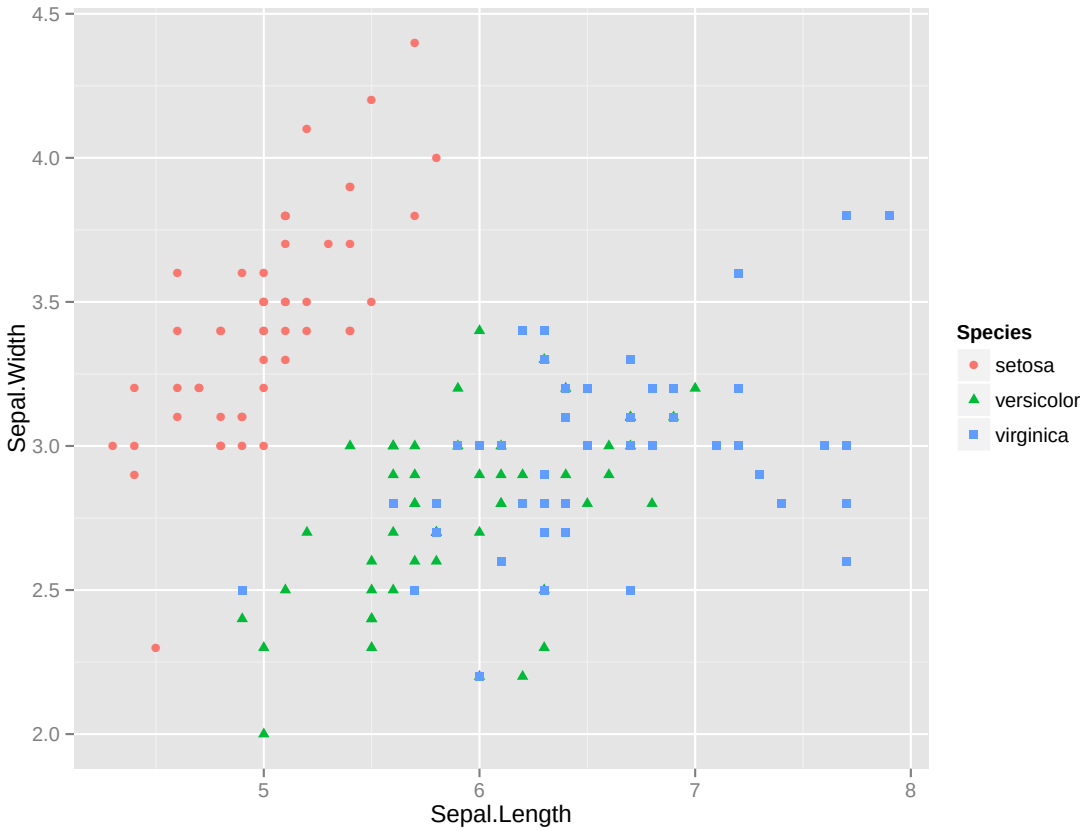
```
qplot(data=iris, x=Species, y=Sepal.Length, geom='violin',
   main='Sepal Lengths', ylab='Sepal Length (cm)')
```



2.3.4 Bivariate - Continuous vs Continuous

Comparing a continuous variable vs another continuous variable is done using a scatter plot.

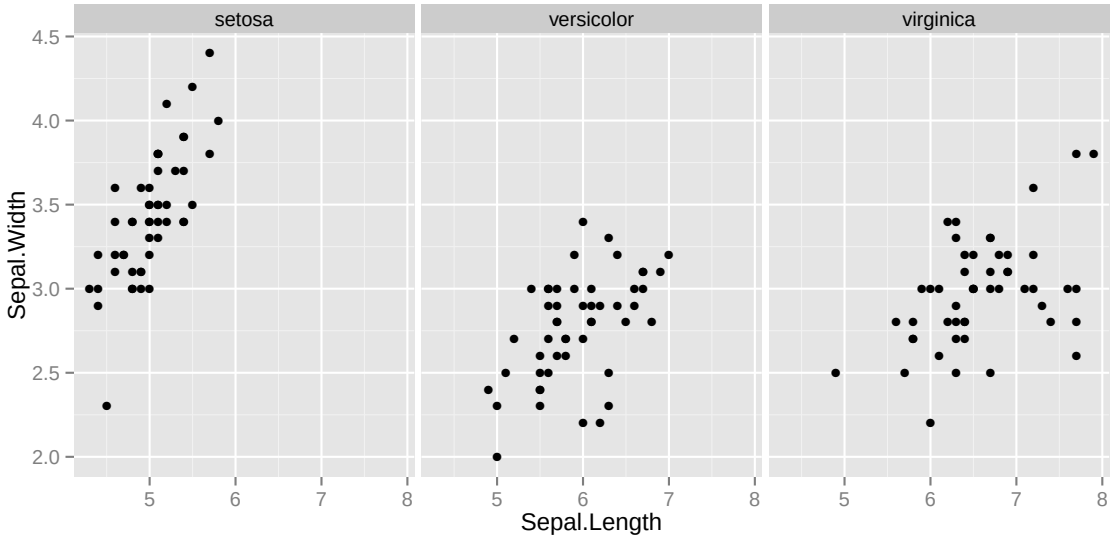
```
qplot(Sepal.Length, Sepal.Width, pch=Species, col=Species, data=iris)
```



Alternatively, I could have made the same group for each species and compared them. Here I want my graphs to vary by species. The `facets` parameter takes a formula y ~ x and the plots will

vary according to the given variable. In this case, I only want 1 row of graphs, but a column for each species.

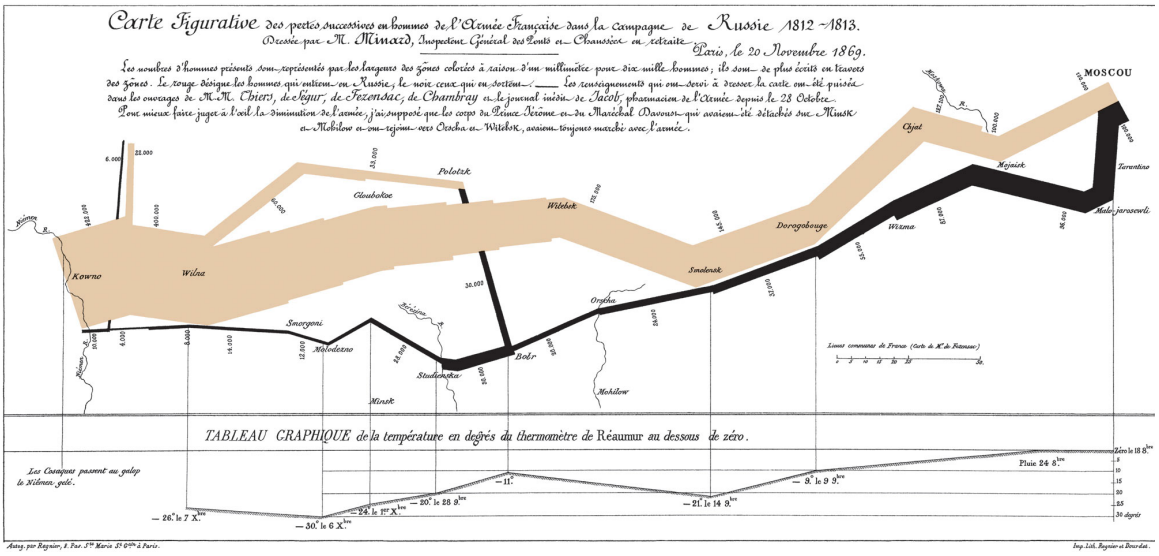
```
qplot(Sepal.Length, Sepal.Width, facets= . ~ Species, data=iris)
```



2.3.5 Notable Graphs

There are several statistical graphs that are extremely well known and provide amazing insight to the issue that the graphs address.

The first is Charles Minard's graph of the size of Napoleon's army as it marches to Moscow and retreats.



Tufte (1983): "Beginning at the left on the Polish-Russian border near the Niemen River, the thick band shows the size of the army (422,000 men) as it invaded Russian in June 1812. The width of the band indicates the size of the army at each place on the map. In September, the army reached Moscow, which was by then sacked and deserted, with 100,000 men. The path of Napoleon's retreat from Moscow is depicted by the darker, lower band, which is linked to a temperature scale and dates at the bottom of the chart. It was a bitterly cold winter, and many froze on the march out

of Russia. As the graphic shows, the crossing of the Berezina River was a disaster, and the army finally struggled back into Poland with only 10,000 men remaining. Also shown are the movements of auxiliary troops, as they sought to protect the rear and the flank of the advancing army. Minard's graphic tells a rich, coherent story with its multivariate data, far more enlightening than just a single number bouncing along over time. Six variables are plotted: the size of the army, its location on a two-dimensional surface, direction of the army's movement, and temperature on various dates during the retreat from Moscow."

Another famous set of graphics are a "Ted Talk" given by Hans Rosling which can be viewed at: Hans Rosling: Stats that reshape your world view

Chapter 3

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.

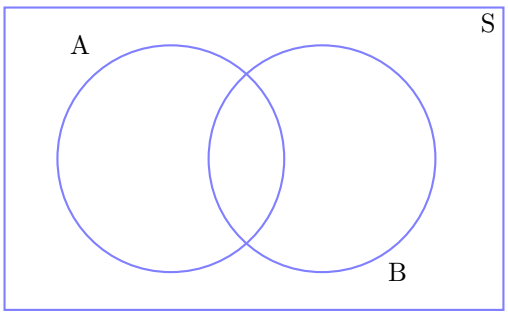
3.1 Introduction to Set Theory

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

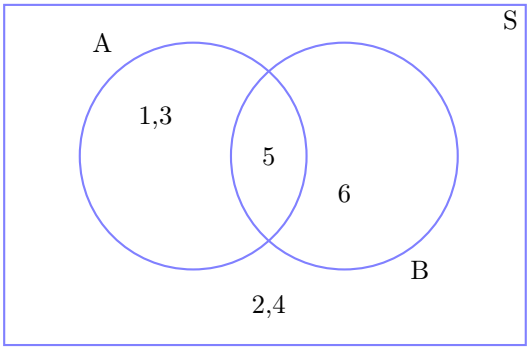
Events are properties of a particular outcome. For a coin flip, the event "Heads" would be the event that a heads was flipped. For the single roll of a six-sided die, a possible event might be that the result is even. For the NAU student, we might be interested in the event that the student is a biology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

3.1.1 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 $A = \textit{roll is odd}$ and $B = \textit{roll is 5 or greater}$. In this case our picture is:



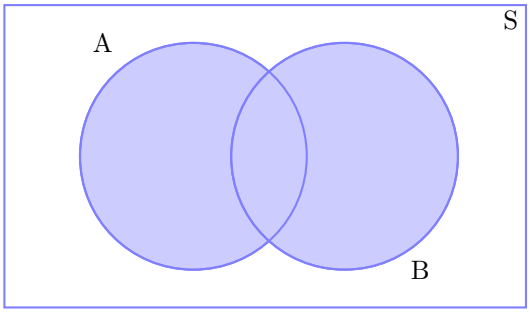
All of our possible events are present, and distributed amongst our possible events.

3.1.2 Composition of events

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

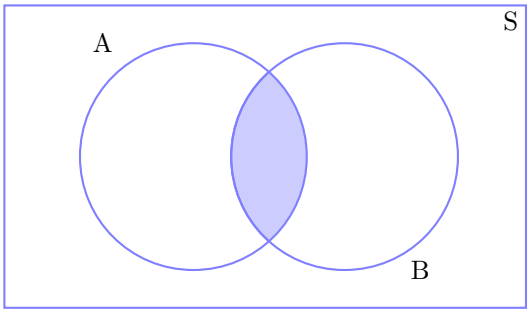
- Union: Denote the event that either A or B occurs as $A \cup B$.

$A \cup B$



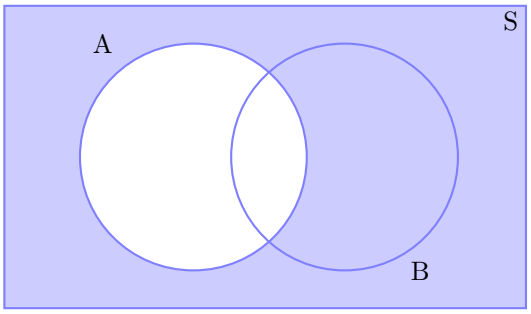
- Denote the event that both A and B occur as $A \cap B$

$A \cap B$



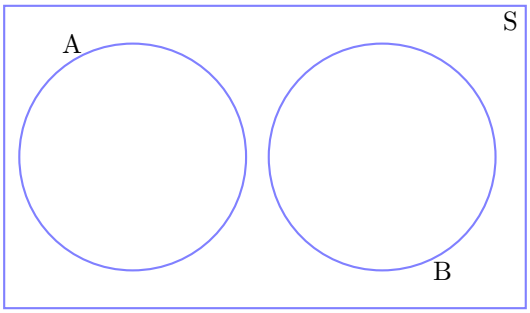
- Denote the event that A does not occur as $\bar{A}$ or A^C (different people use different notations)

$\bar{A}$ 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



3.2 Probability Rules

3.2.1 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.

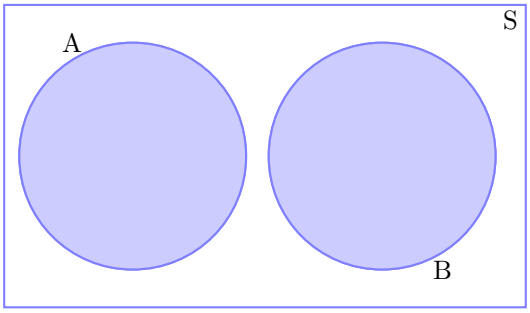
Definition 2. Probability is the proportion of times an event occurs in many repeated trials of a random phenomenon. In other words, probability is the long-term relative frequency.

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

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



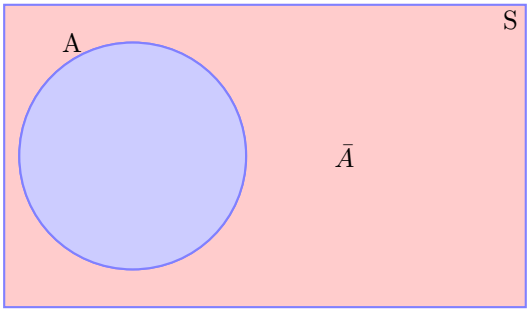
Example. Let S be the sum of two different colored dice. Suppose we are interested in $P(S \leq 4)$. Notice that the pair of dice can fall 36 different ways (6 ways for the first die and six for the second results in 6x6 possible outcomes, and each way has equal probability 1/36. Since the dice cannot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

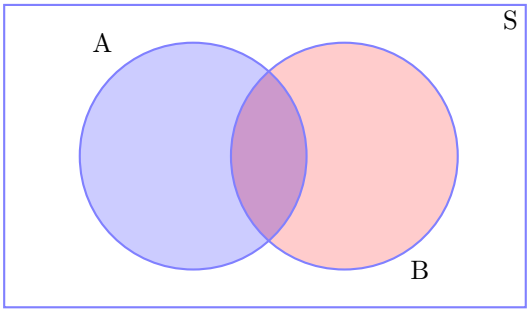
$$P(A) + P(\bar{A}) = 1$$



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.

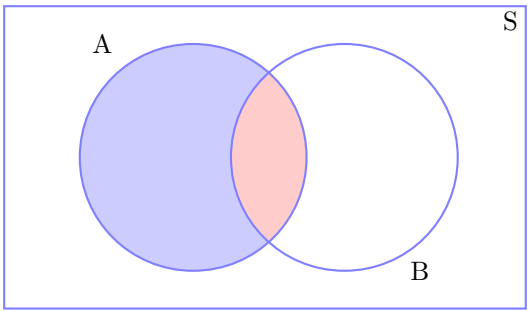
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$



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.

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



3.2.2 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.

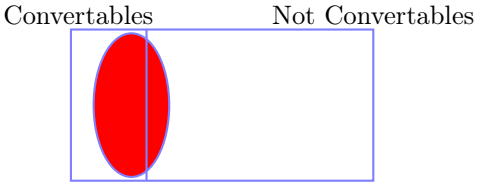
Fact 5. *If A and B are independent events, then*

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

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.

So now we might ask what is the probability that a car is a convertible given that it is red. That is, what is $P($ *Convertible* $|$ *Red* $)$? Since my known values are in the form $P($ *color* $|$ *car type* $)$, I'll need to do some algebra to manipulate the desired value into terms that I have.

$$\begin{aligned} P(\, Convertable \,|\, Red\,) &= \frac{P(\, Convertable \,\cap\, Red\,)}{P\left(\, Red\,\right)} \\ &= \frac{0.06}{0.195} \\ &= 0.308 \end{aligned}$$

It turns out that this re-ordering of the conditioning is very common and we want to develop a systematic way of doing it. The result is Bayes' Formula.

3.2.3 Bayes' Formula

Often we want to re-order a conditional probability

$$\begin{aligned} P(A|B) &= \frac{P\left(B \cap A\right)}{P(B)} \\ &= \frac{P\left(B \cap A\right)}{P(B \cap A) + P\left(B \cap \bar{A}\right)} \\ &= \frac{P\left(B|A\right) P(A)}{P\left(B|A\right) P(A) + P\left(B|\bar{A}\right) P\left(\bar{A}\right)} \end{aligned}$$

Notice that A and $\bar{A}$ are disjoint and span the sample space (i.e., one of the two events has to happen). We can write a generalized version of Bayes' formula.

Fact. *Let $A_1 \dots A_k$ be mutual exclusive events that span the sample space (i.e. one of the A_i must happen), and B some event such that $P(B) > 0$. Then*

$$P(A_i|B) = \frac{P(B|A_i)\,P(A_i)}{\sum_{j=1}^{k} P(B|A_j)\,P(A_j)}$$

Changing the ordering of the conditioning is extremely important for allowing us to manipulate probabilities.[1]

Example 7. Inexpensive enzyme immunoassay screening tests for HIV have the following rates of success/failure (Notice the rows sum to 100%) and therefore we are given data in the form $P(\,TestResult \mid TrueCondition\,)$:

		Test Returns	
		Negative	Positive
True Condition	Negative	98.5%	1.5%
	Positive	0.03%	99.97%

Given that the rate of HIV infection in Sub-Saharan Africa is 5%, what is the probability that a randomly selected person has HIV given that they have tested positive?

$$\begin{aligned} P(Pos|TestPos) &= \frac{P(TestPos|Pos)P(Pos)}{P(TestPos|Pos)P(Pos) + P(TestPos|Neg)P(Neg)} \\ &= \frac{0.9997(0.05)}{0.9997(0.05) + 0.015(0.95)} \\ &= 0.778 \end{aligned}$$

Example 8. Inferring populations from genetic markers. Suppose we have done genotype sampling at three nesting sites for a migratory bird and we know the following demographic information:

	Percent of Nesting Pairs	Allele 1	Allele 2	Allele3	Total
Site 1	50%	60%	25%	15%	100%
Site 2	40%	43%	40%	17%	100%
Site 3	10%	12%	8%	80%	100%

Notice that the data is given in the form $P(A|S)$!

Suppose we then go to the wintering grounds and sample a bird and find that it has allele 3. What is the probability it came from nesting site 3? Find

$$\begin{aligned} P(S_3|A_3) &= \frac{P(A_3|S_3)P(S_3)}{\sum_{j=1}^{3} P(A_3|S_j)P(S_j)} \\ &= \frac{0.8(0.1)}{0.15(0.5) + 0.17(0.4) + 0.8(0.1)} \\ &= \frac{0.08}{0.223} \\ &= 0.359 \end{aligned}$$

[1] This formula is the foundation of Bayesian statistical inference; the distribution of $Data|Parameters$ becomes $Parameters|Data$.

Similarly you can calculate $P(S_1|A_3) = 0.336$ and $P(S_2|A_3) = 0.305$.

A similar flipping of conditions often can be found in stable isotope experiments where you know the starting ratios of heavy isotopes and infer something about a biological process based off of the ending ratio.

3.2.4 Summary of Probability Rules

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

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

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

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

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B \mid A)\, P(A)}{P(B \mid A)\, P(A) + P\left(B \mid \bar{A}\right) P\left(\bar{A}\right)}$$

$$P(A_i \mid B) = \frac{P(B|A_i)\, P(A_i)}{\sum_{j=1}^{k} P(B|A_j)\, P(A_j)} \qquad \text{Assuming } A_j \text{ are all mutually exclusive and span}$$

3.3 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.

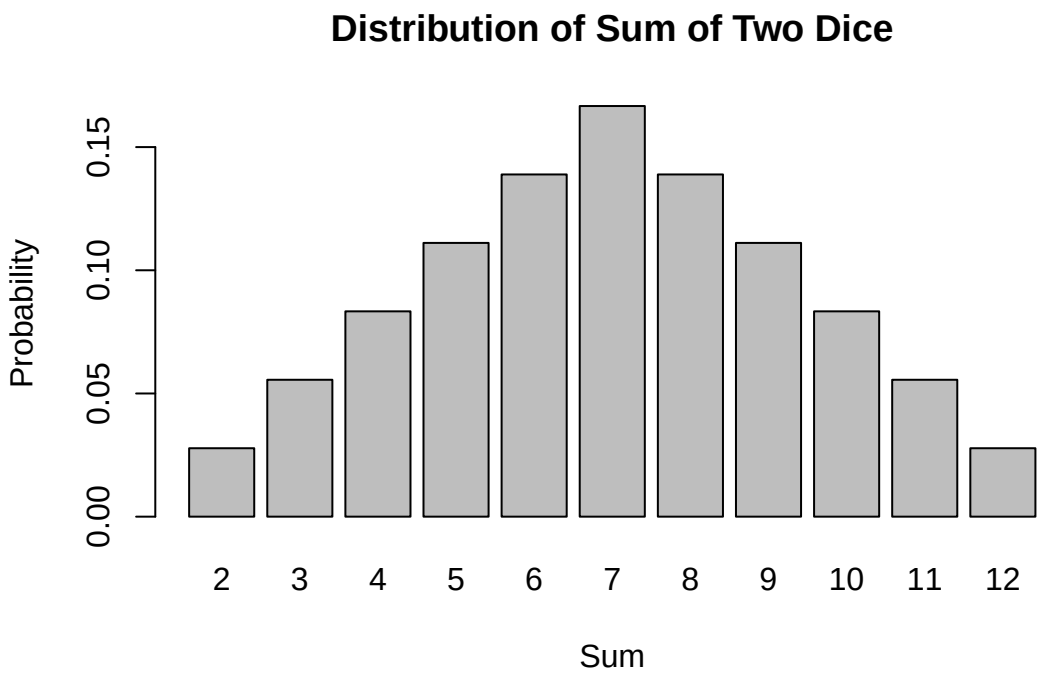
Examples:

1. Presence or Absence of wolves in a State?
2. Number of Speeding Tickets received?
3. Tree girth (in cm)?
4. Photosynthesis rate?

3.3.1 Introduction to Discrete Random Variables

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

```
barplot( c(1:6,5:1)/36, names.arg=2:12,
         main='Distribution of Sum of Two Dice',
         ylab='Probability', xlab='Sum')
```



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., $P(3 \text{ or } 4) = P(3) + P(4)$

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.*

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

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.*

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

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 9. 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?

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

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.

3.3.2 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 $\binom{n}{y} = \frac{n!}{y!(n-y)!}$. For our small mammal example:

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

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

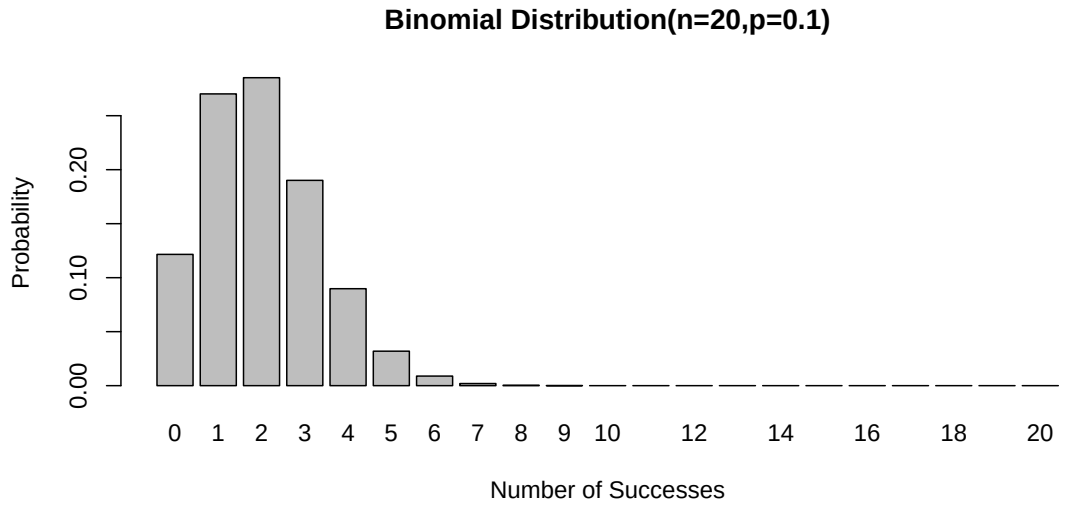
$$\begin{aligned} E[Y] &= \sum_{y=0}^{n} y\, P(Y=y) \\ &= \sum_{y=0}^{n} y\, \frac{n!}{y!\,(n-y)!}\pi^y\,(1-\pi)^{n-y} \\ &= \vdots \\ &= n\pi \end{aligned}$$

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 10. 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

```
barplot( dbinom(0:20,size=20,prob=0.1), names.arg=0:20,
         main='Binomial Distribution(n=20,p=0.1)',
         ylab='Probability', xlab='Number of Successes')
```



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)
dbinom(0, size = 20, prob = 0.1) + dbinom(1, size = 20, prob = 0.1) + dbinom(2,
    size = 20, prob = 0.1)

## [1] 0.6769

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

## [1] 0.6769
```

3.3.3 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 11. 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`).

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

## [1] 0.1169

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

## [1] 0.9162

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

## [1] 0.08375
```

3.4 Continuous Random Variables

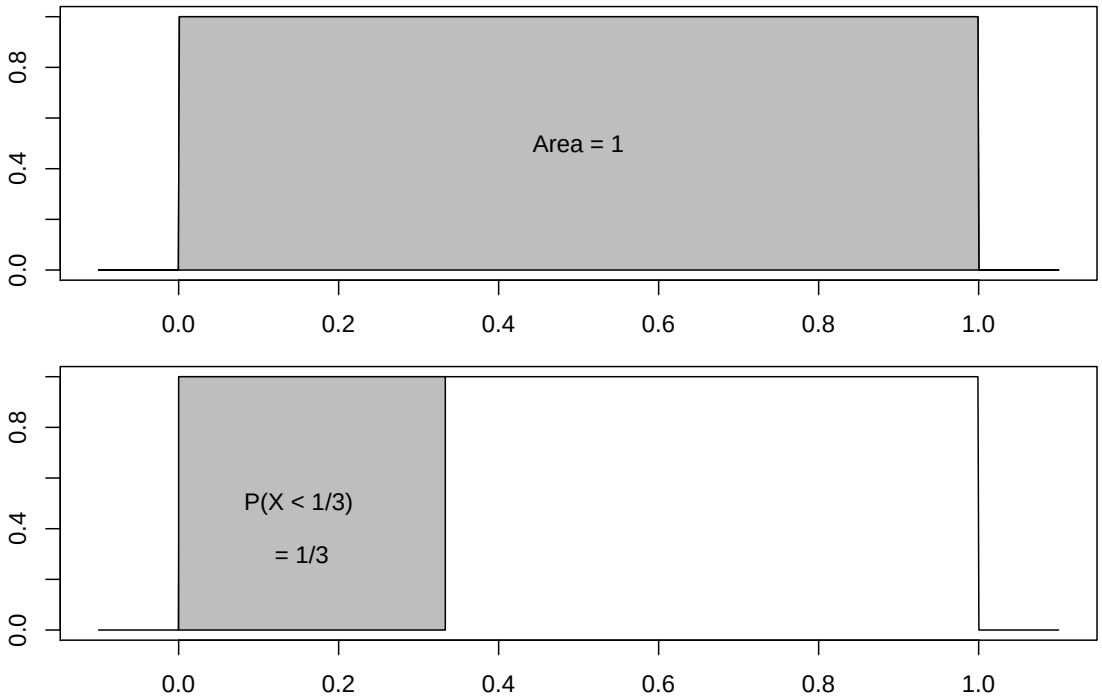
Finding the area under the curve of a particular density function $f(x)$ requires the use of calculus, but since this isn't a calculus course, we will resort to using R or tables of calculated values.

3.4.1 Uniform 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.

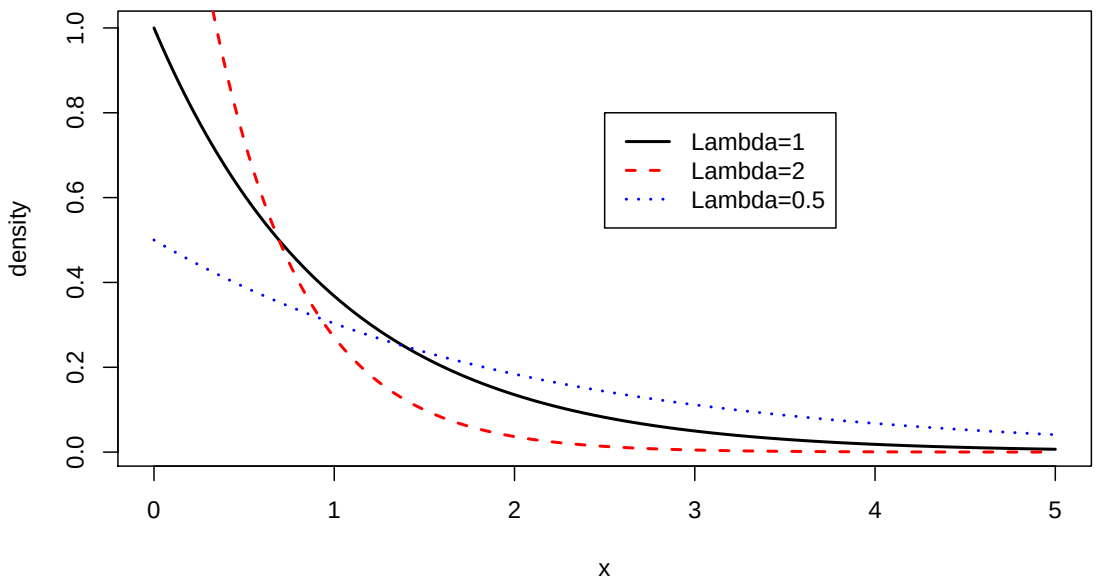


Finding the area under the curve of a particular density function $f(x)$ usually requires the use of calculus, but since this isn't a calculus course, we will resort to using R or tables of calculated values.

3.4.2 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

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



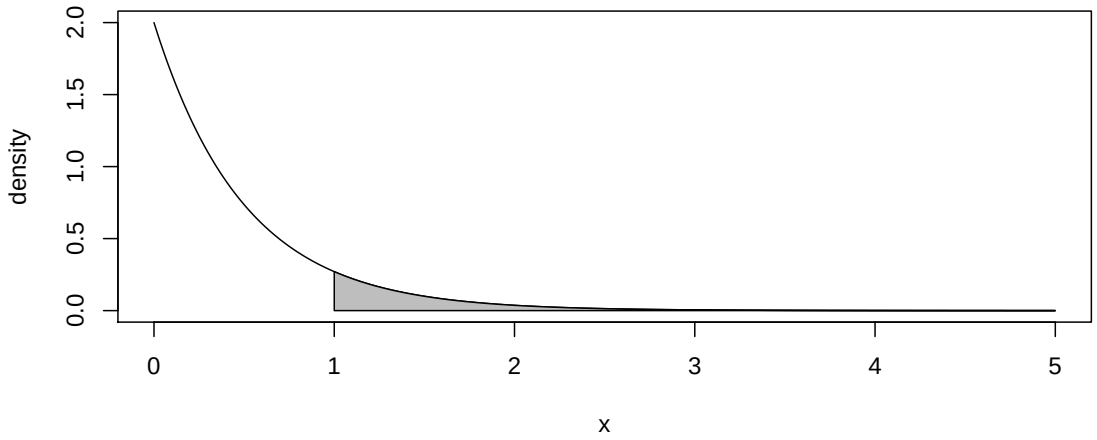
Analogous to the discrete distributions, we can define the Expectation and Variance of these distributions by replacing the summation with an integral

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

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 12. 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] 0.1353
```

3.4.3 Normal Distribution

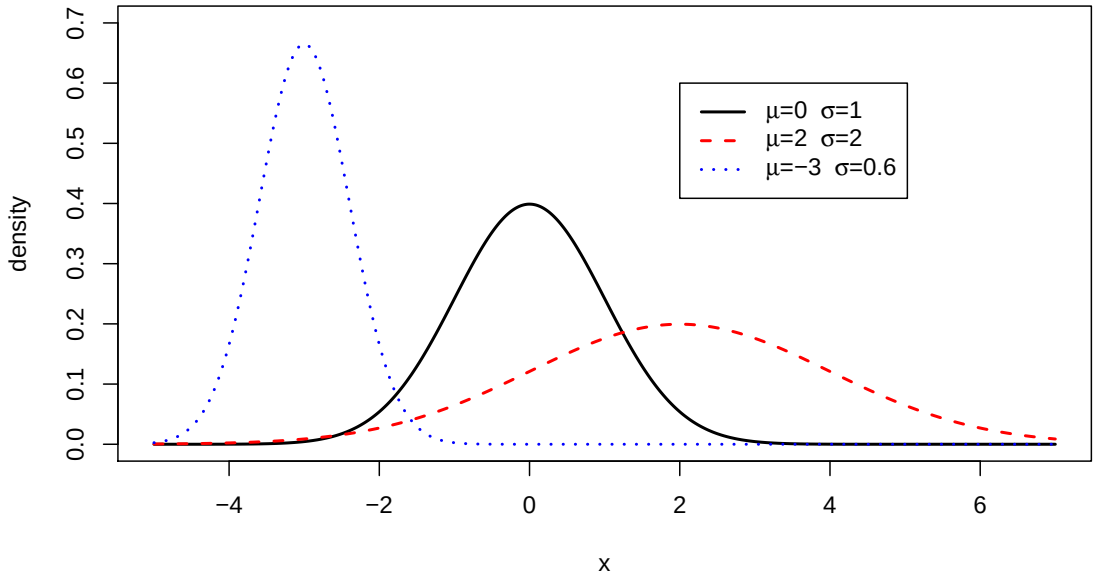
Undoubtably the most important distribution in statistics is the normal distribution. If my RV X is normally distributed with mean μ and standard deviation σ, its probability distribution function is given by

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

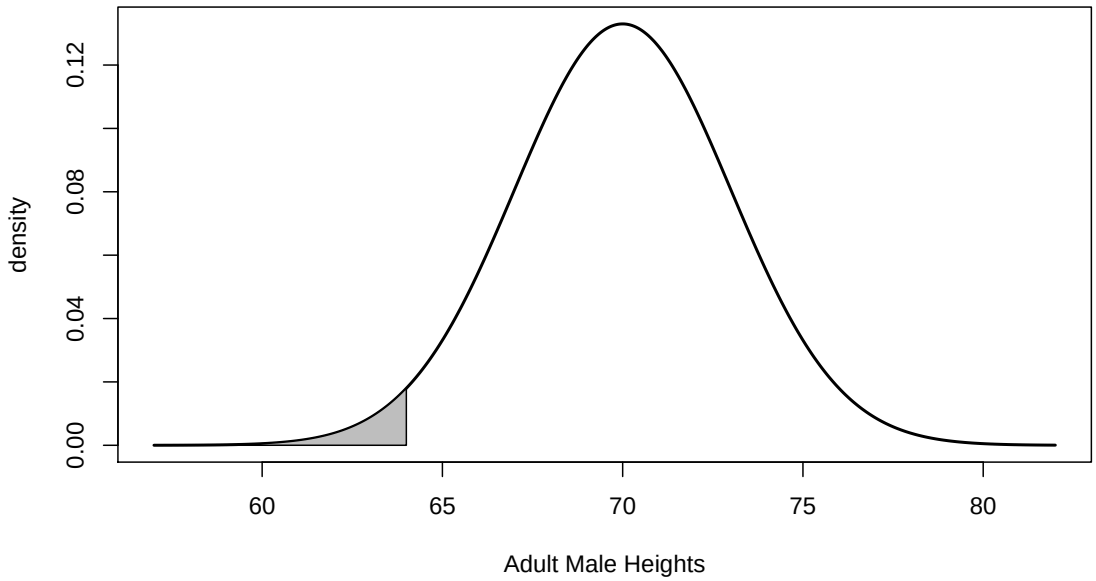
where $\exp[y]$ is the exponential function e^y. We could slightly rearrange the function to

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

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 13. 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] 0.02275
```

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.

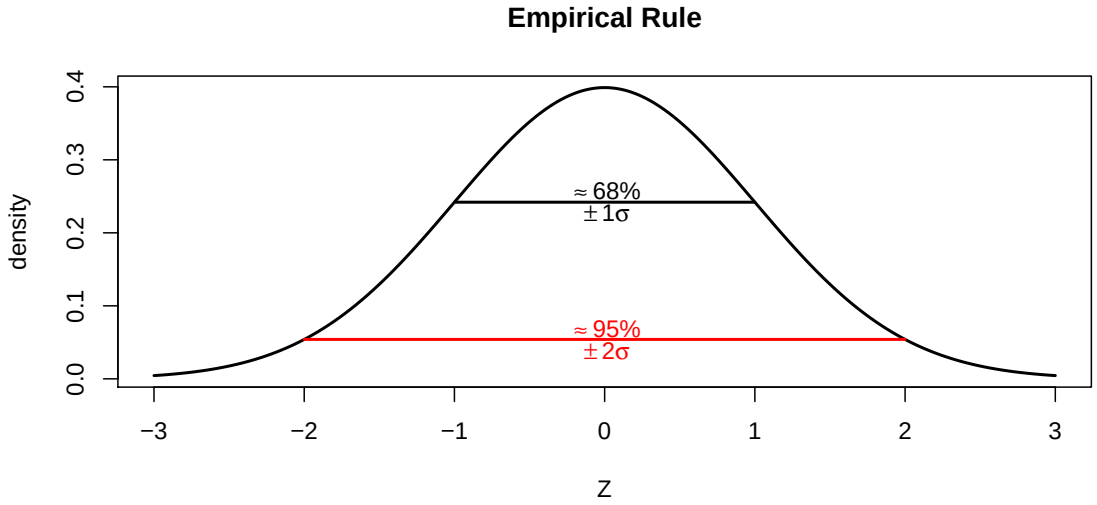
```
qnorm(0.8, mean = 0, sd = 1)
```

```
## [1] 0.8416
```

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

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

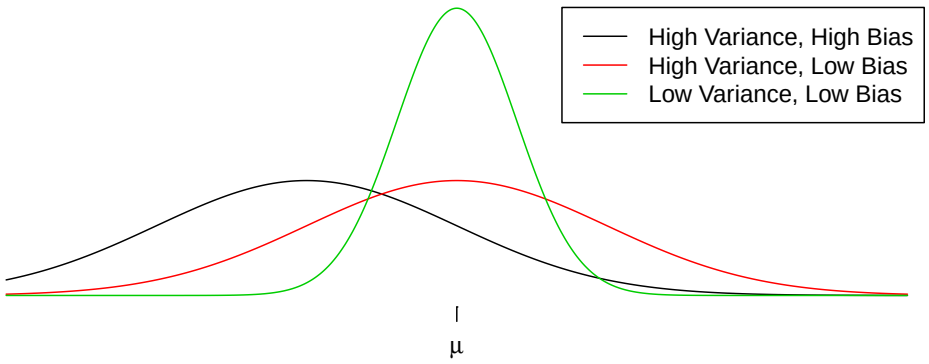
$$\begin{aligned} P(\mu - \sigma \leq X \leq \mu + \sigma) &= P(-1 \leq Z \leq 1) \\ &= P(Z \leq 1) - P(Z \leq -1) \\ &\approx 0.8413 - 0.1587 \\ &= 0.6826 \end{aligned}$$


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 4

Maximum Likelihood

Previously we used the sample mean $\bar{x}$ to estimate μ but why did we chose $\bar{x}$ instead of the median? Or perhaps $(max + min)/2$? We want to find estimators that are unbiased (not systematically over or under estimating the true value) and have small variance.



But how do we go about finding estimators that satisfy these? In STA 673 students will learn *much* more about finding estimators, but here we will only introduce **Maximum Likelihood Estimators**.

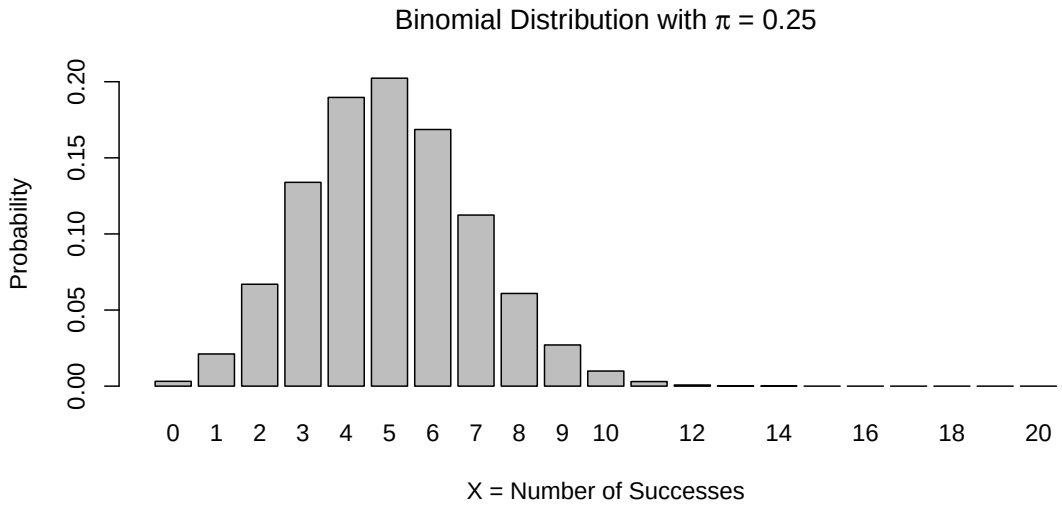
4.1 Likelihood Function

4.1.1 Binomial Distribution

Recall the Binomial distribution gave the number of success given the number of trials and the probability of success. For example, suppose the probability of success is $\pi = 0.25$ and we have $n = 20$ experiments. Recall that the probability mass function was

$$\begin{aligned} P\left(X \mid \pi = 0.25, n = 20\right) &= \binom{n}{x} \pi^x \left(1-\pi\right)^{n-x} \\ &= \binom{20}{x} 0.25^x (0.75)^{20-x} \end{aligned}$$

which has the following barplot:



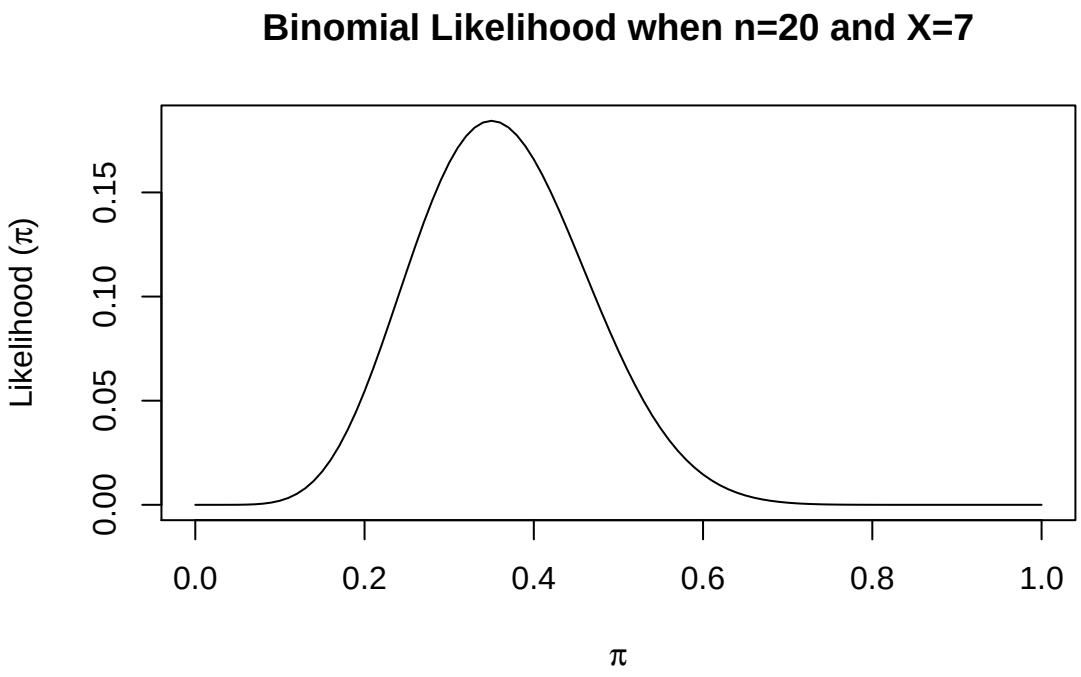
It is clear that the most probable value to observe is $X = 5$, but I wouldn't be surprised to see $X = 4$ or $X = 6$.

The relationship between the observed data (X) and the parameter of interest[1] (π) is entirely given by this probability mass function. Since our ultimate goal is to use data to give estimates for parameter values, we should use the same probability mass function, but interpret it as a function of parameters given the data. To denote this change in reference, we'll take the the probability mass function and call it the *likelihood function* where

$$L(\pi \mid X) = \binom{n}{x} \pi^x (1-\pi)^{n-x}$$

Suppose that we observed $X = 7$ successes out of $n = 20$. We plot the likelihood function:

[1] Here we assume that n is known and uninteresting.



Definition 14. The maximum likelihood estimator (MLE) is parameter value that maximizes the likelihood function.

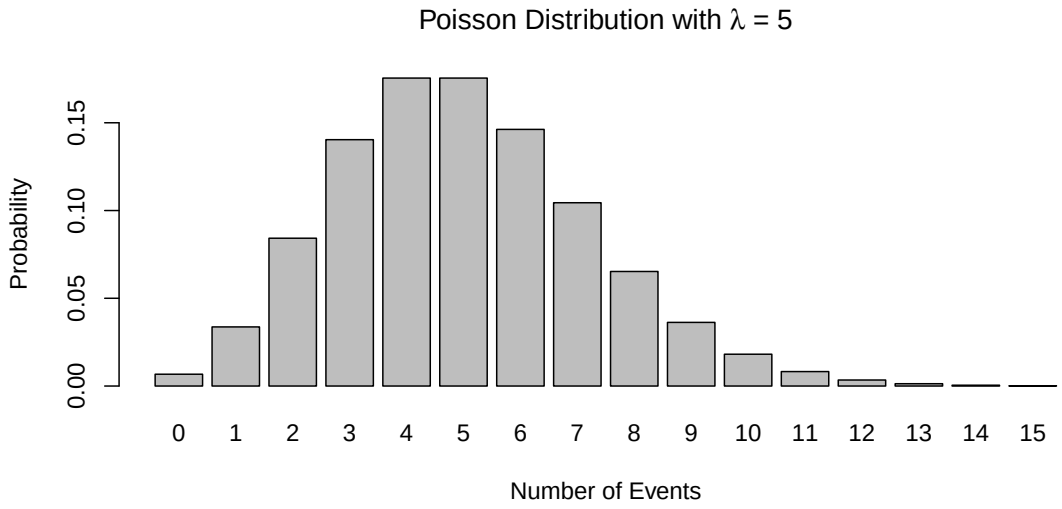
The value that maximizes this likelihood function is $\hat{\pi}_{MLE} = 7/20 = 0.35$ and I will refer to this as the *Maximum Likelihood Estimator* (MLE).

4.1.2 Poisson Distribution

Recall the Poisson distribution is often used to model count data over a some duration time. It is defined by the rate parameter λ and we want a good estimator of λ using a random sample of n independent observations taken from this distribution. Recall that the probability mass function is

$$f(x \mid \lambda) = \frac{e^{-\lambda}\lambda^x}{x!}$$

and notice that the function is highest is where the observation has the highest probability of occurring.



Next notice since the observations are independent then the probability function for our sample $x_1, x_2, \ldots, x_n$ is

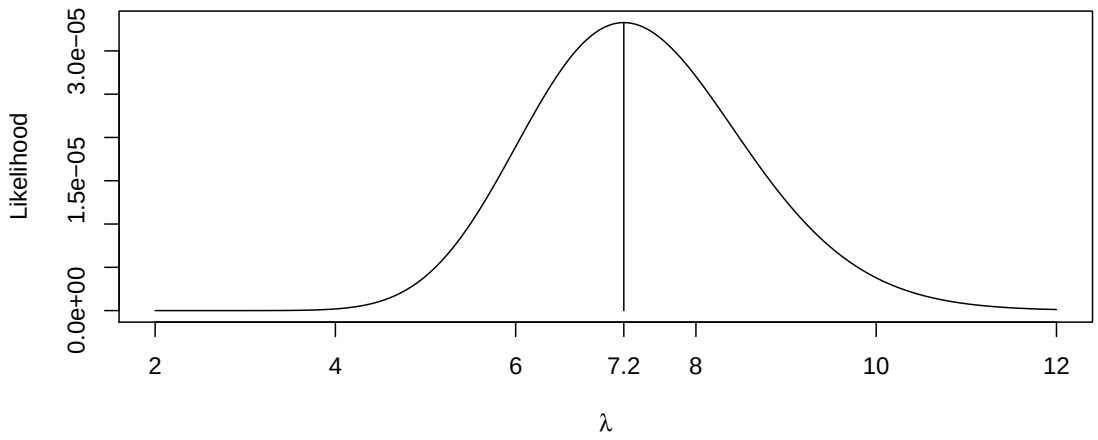
$$\begin{aligned} f(x_1, \ldots x_n \,|\, \lambda) &= f(x_1|\lambda)\, f(x_2|\lambda) \ldots f(x_n|\lambda) \\ &= \frac{e^{-\lambda}\lambda^{x_1}}{x_1!}\frac{e^{-\lambda}\lambda^{x_2}}{x_2!} \cdots \frac{e^{-\lambda}\lambda^{x_n}}{x_n!} \\ &= \frac{e^{-n\lambda}\lambda^{\sum x_i}}{x_1! x_2! \cdots x_n!} \end{aligned}$$

If we regard this as a function of $\lambda \,|\, data$ then we can use this as a function where the height of the function is representative of how likely a parameter value is. We formally recognize this inversion by calling this new function the *likelihood function.*

$$L(\lambda \,|\, x_1, \ldots, x_n) = \frac{e^{-n\lambda}\lambda^{\sum x_i}}{x_1! x_2! \cdots x_n!}$$

Suppose that we have observed the data $\{5, 9, 8, 6, 8\}$ then the graph of the likelihood function looks like

```
x <- c(5, 9, 8, 6, 8)
n <- length(x)
L <- function(lambda){
        temp <- 1
        for(i in 1:n){
                temp <- temp * dpois(x[i], lambda)
        }
        return(temp)
}
lambda <- seq(2,12, length=1000)
plot(lambda, L(lambda), xlab=expression(lambda), ylab='Likelihood', type='l')
lines(rep(mean(x),2), c(0, L(mean(x))))
axis(1, at=mean(x))
```

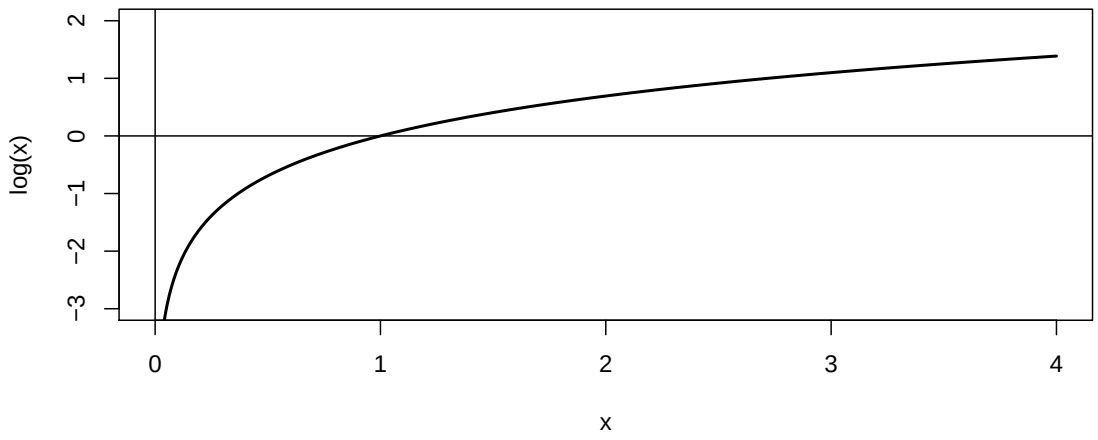


and the maximum value occurs at $\hat{\lambda}_{mle} = \bar{x} = 7.2$. But how did I know that it occurred at the sample mean?

4.2 Maximization

There are several approaches to finding the maximum of a function. The best case scenario is if we can algebraically solve for the maximum. Failing that we can turn to numerical methods for finding a maximum.

Maximizing a function $g(x)$ is equivalent to maximizing $\log[g(x)]$ because if $a < b$ then $\log a < \log b$.



Algebraically it is easier to find a maximum of $\log[L(\lambda \mid data)]$ and numerically it is more stable, so we will use the log-likelihood for our calculations.

Remark. The simplification provided by using the log-likelihood is only applicable if a student remembers the rules for logs. The basic rule is that every operation becomes one operation simpler, i.e. exponentiation becomes multiplication, and multiplication becomes addition. Assuming that the following exist (i.e. $a, b > 0$) and recognizing that in this course (and mathematics in general)

we will always use the natural log (base e) the rules are:

$$\begin{aligned}
\log\left(e^a\right) &= a \\
e^{\log a} &= a \\
\log\left(a^b\right) &= b\,\log a \\
\log\left(ab\right) &= \log a + \log b \\
\log\left(\frac{a}{b}\right) &= \log a - \log b
\end{aligned}$$

4.2.1 Calculus Solution

In calculus we learned how to maximize this function by finding the derivative and setting it equal to zero and solving for the parameter value.

First we note that the data are constant and therefore I only have to maximize

$$L\left(\lambda|x_1\ldots x_n\right) \propto e^{-n\lambda}\lambda^{\sum x_i}$$

because the division by $x_1!\ldots x_n!$ is the same for all values of λ. The log-likelihood can now be simplified to

$$\begin{aligned}
\log L\left(\lambda\,|\,x_1\ldots x_n\right) &\propto \log\left[e^{-n\lambda}\lambda^{\sum x_i}\right] \\
&= \log\left(e^{-n\lambda}\right) + \log\left(\lambda^{\sum x_i}\right) \\
&= -n\lambda + \left(\sum x_i\right)\log\lambda
\end{aligned}$$

and the derivative is

$$\frac{d}{d\lambda}\log L\left(\lambda|x_1\ldots x_n\right) = -n + \frac{1}{\lambda}\sum x_i \stackrel{set}{=} 0$$

thus

$$\hat{\lambda}_{mle} = \frac{\sum x_i}{n} = \bar{x}$$

The downsides of this method is that we have to be able to manipulate the log-likelihood function and take derivatives and that is not always possible.

4.2.2 Numerical Solution

In the cases where calculus has failed us, we can turn to numerical solutions to finding the maximum. The most basic of numerical methods is called a *hill-climbing* method. Given an initial guess, we'll look in a nearby neighborhood and look for higher values and use the highest nearby value and by next guess. In this manner I will continue to find larger and larger values until I am in a neighborhood with no higher values. This is analogous to climbing a mountain or hill by always moving in the direction that is steepest. There are two primary ways that a hill climbing method can be lead astray:

- We can get stuck in local maximums. This can be partly remedied by having many spread out starting locations.
- If we find a plateau, it can be very hard to find what direction to move "uphill".

4.3 Normal Distribution

4.3.1 Theory

Recall that the probability density function of a normal random variable was

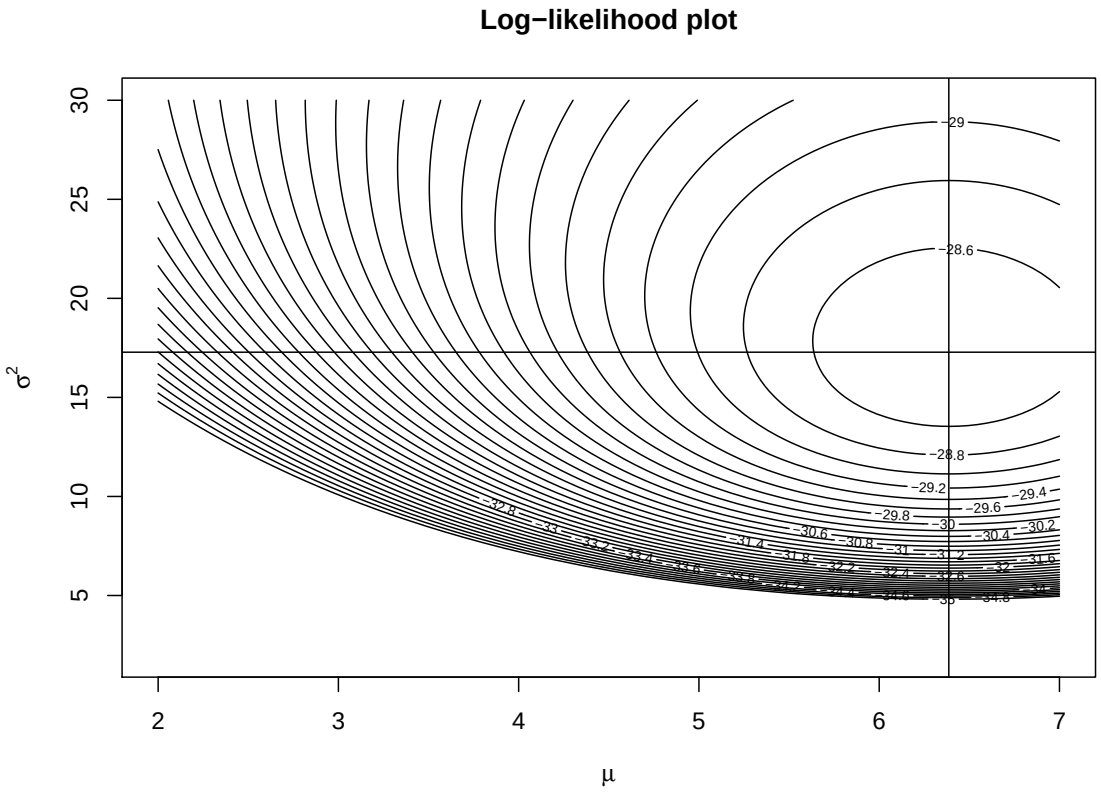
$$f\left(y \,|\, \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}\left(y-\mu\right)^2\right]$$

and we notice that there is a squared-error term. As a result, maximizing the likelihood with respect to μ is exactly minimizing the squared-error. This is why in ANOVA and regression models (where we had normal error terms) we chose model parameter estimates that minimized the sum of squared error. So for the signal part of the model (i.e. $\mu, \tau_i, \beta_0, \beta_1$), the maximum likelihood estimators are the same ones that minimized the sum of squared error.

However, the maximum likelihood estimator for the variance term σ^2 is different than the one we have been using. The MLE is $\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum\left(y_i - \hat{y}\right)^2$. In the following we give a numerical example of this where the signal part of the model is just a constant. That is we are looking at the model

$$Y_i = \mu + \epsilon_i \quad \text{where } \epsilon_i \overset{iid}{\sim} N\left(0, \sigma^2\right)$$

```
x <- rnorm(10, mean=5, sd=4)
n <- length(x)
mu.hat <- mean(x)
sigma2.hat <- sum( (x-mean(x))^2 ) / n
logL <- function(params){
        return( sum( dnorm(x, mean=params[1], sd=sqrt( params[2] ), log=TRUE) ) )
}
n.grid <- 100
mu.grid <- seq(2,7, length=n.grid)
sigma2.grid <- seq(2, 30, length=n.grid)
logL.grid <- matrix(0, nrow=n.grid, ncol=n.grid)
for(i in 1:n.grid){
        for(j in 1:n.grid){
                logL.grid[i,j] <- logL( c(mu.grid[i], sigma2.grid[j] ) )
        }
}
contour(mu.grid, sigma2.grid, logL.grid, levels=seq(-35,-10, by=.2), xlab=expression(mu), ylab=expre
abline(h=sigma2.hat)
abline(v=mu.hat)
```



A variety of hill climbing functions are available in software packages such as R. For a function with multiple dimensions, we will use the function `optim()`. Most of these sorts of functions assume you want to minimize a function rather than maximize it, so we will define a negative log-likelihood function `neglogL` and minimize that. Finally we will need a starting point, and we will use $\mu = 3$, $\sigma^2 = 20$.

```
neglogL <- function(params){
        return( -logL(params) )
}
optim(c(3,20), neglogL)

## $par
## [1]  6.387 17.283
##
## $value
## [1] 28.44
##
## $counts
## function gradient
##       53       NA
##
## $convergence
## [1] 0
##
## $message
## NULL
```

4.4 Asymptotic Results for MLEs

There are many asymptotic results for maximum likelihood estimators and we summarize them here. Let θ be some parameter of interest and $\hat{\theta}_{mle}$ be the maximum likelihood estimator.

- $\hat{\theta}_{mle}$ is a consistent estimator of θ
  - as the sample size increases, the $Var\left(\hat{\theta}_{mle}\right) \to 0$
  - $\hat{\theta}_{mle}$ is asymptotically unbiased
- $\hat{\theta}_{mle}$ has an asymptotically normal distribution.
  - Similar result as the Central Limit Theorem
- Functional Invariance: For any transformation of $g(\theta)$ then $g\left(\hat{\theta}_{mle}\right)$ is the MLE of $g(\theta)$
  - This is important for scale transformations such as taking logs of your data.

4.5 Connection to Bayesian Analysis

I would like to interpret the likelihood function as a probability density function for my parameter θ, but recall that a probability density function must integrate to one. Unfortunately there is no guarantee that the likelihood function integrates to something less than infinity. Mathematically we'll say that the function is *integrable* if the integral can be calculated and is less the infinity. Even if the likelihood is integrable, it still must be rescaled to integrate to one.

One way to address the integrability is to multiply the likelihood by another function that is integrable. In this case I'll multiply the likelihood by a probability density function of the parameter $P(\theta)$, which Bayesians call the prior distribution. Therefore

$$L(\theta \mid data)\, P(\theta)$$

is integrable. This can be verified by recognizing that

$$\begin{aligned}\int_{\Theta} L(\theta|data)\, P(\theta)\, d\theta &\leq \int_{\Theta} \max\left[L(\theta|data)\right] P(\theta)\, d\theta \\ &= \max\left[L(\theta|data)\right] \int_{\Theta} P(\theta)\, d\theta \\ &= \max\left[L(\theta|data)\right]\end{aligned}$$

We can now rescale our function to integrate to one by dividing by whatever it integrates to. Therefore we can define the posterior distribution of our parameter value θ by the probability density function

$$P(\theta \mid data) = \frac{L(\theta \mid data)\, P(\theta)}{\int_{\Theta} L(\theta \mid data)\, P(\theta)\; d\theta}$$

This posterior distribution is a probability distribution and can be used to draw inferences about the parameter of interest θ.

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

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

Now doing the same calculation for my cost random variable,

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

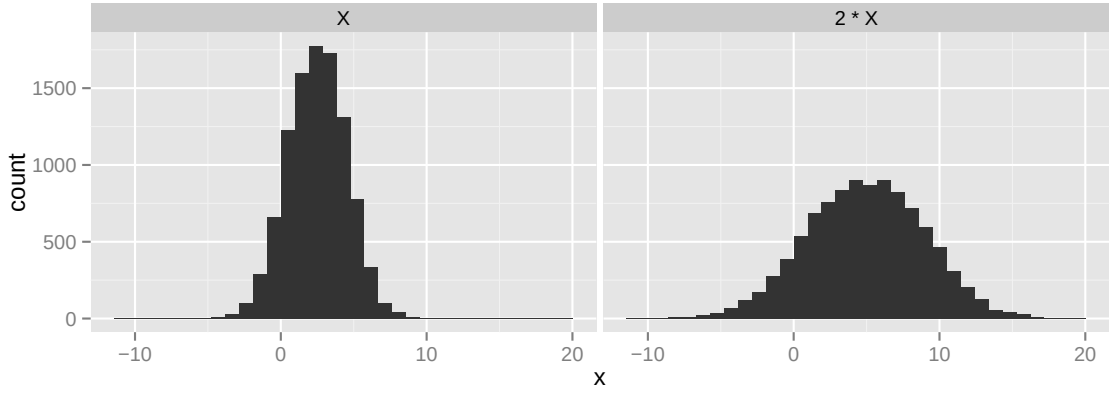
A similar calculation for variance can be done.

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

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.

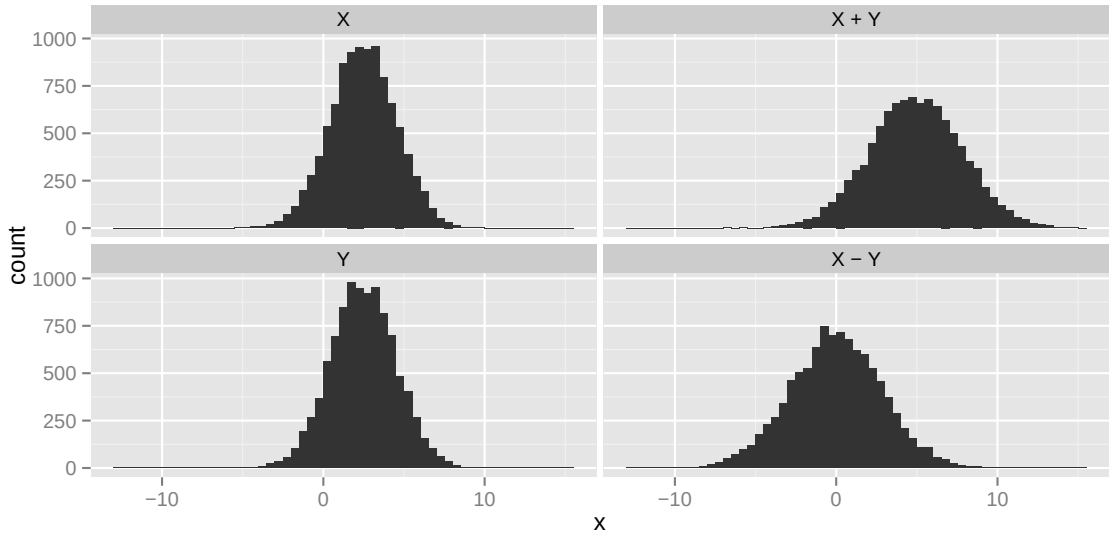
```
n <- 10000
x <- rnorm(n, mean=2.5, sd=2)
mean(x)

## [1] 2.465

mean(2*x)

## [1] 4.929

var(x)

## [1] 4.032

var(2*x)

## [1] 16.13
```

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] 4.934

mean( x-y )

## [1] -0.004531

var( x+y )

## [1] 8.088

var( x-y )

## [1] 8.065
```

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

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

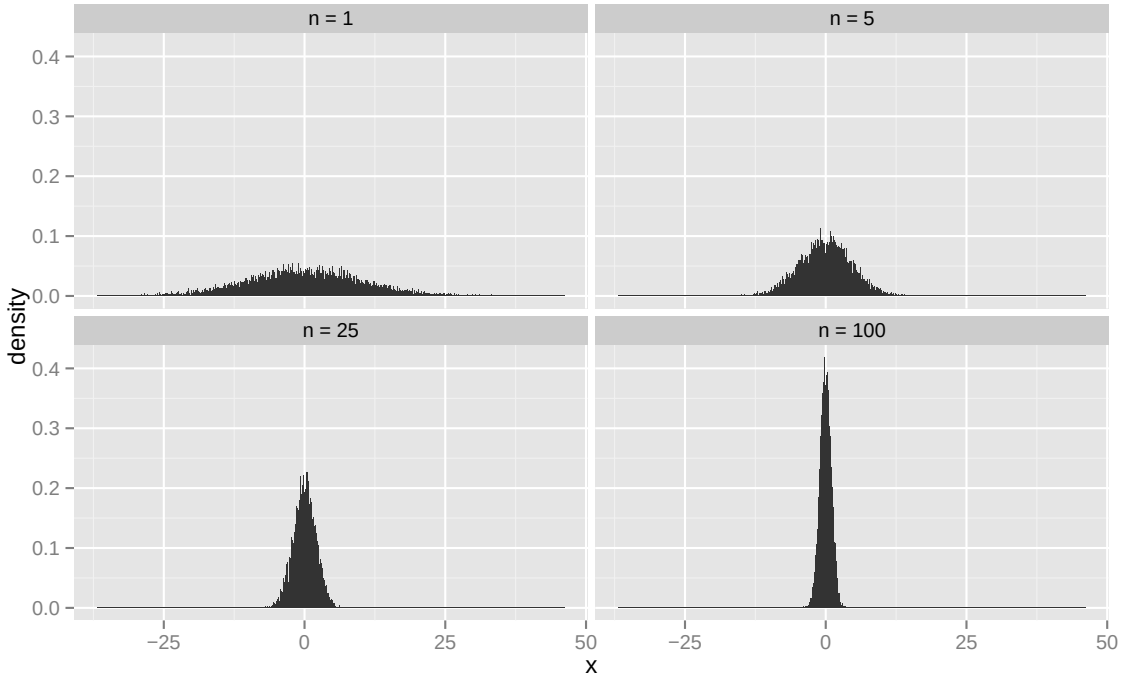
and

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

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] -0.0001187

var(samples)

## [1] 20.46
```

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?

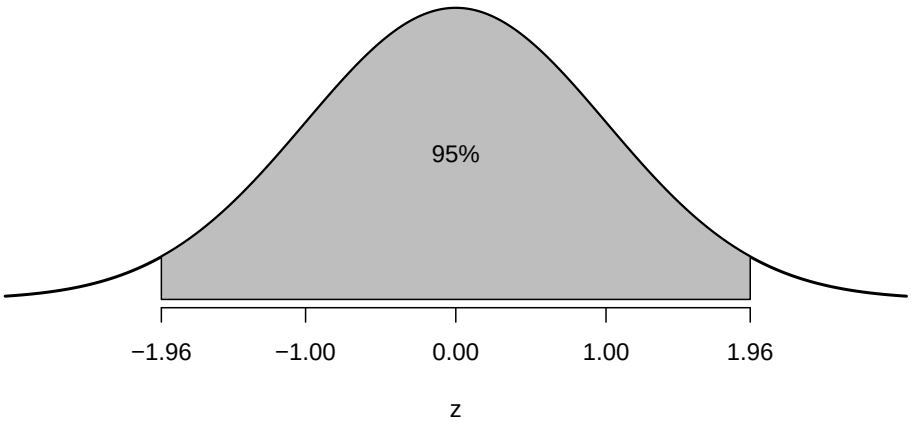
$$\begin{aligned} P\left(\bar{X} \geq 7\right) &= P\left(\frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} \geq \frac{7-6}{\sqrt{\frac{1}{3}}}\right) \\ &= P\left(Z \geq \sqrt{3}\right) \\ &= P\left(Z \geq 1.73\right) \\ &= 0.0418 \end{aligned}$$

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.

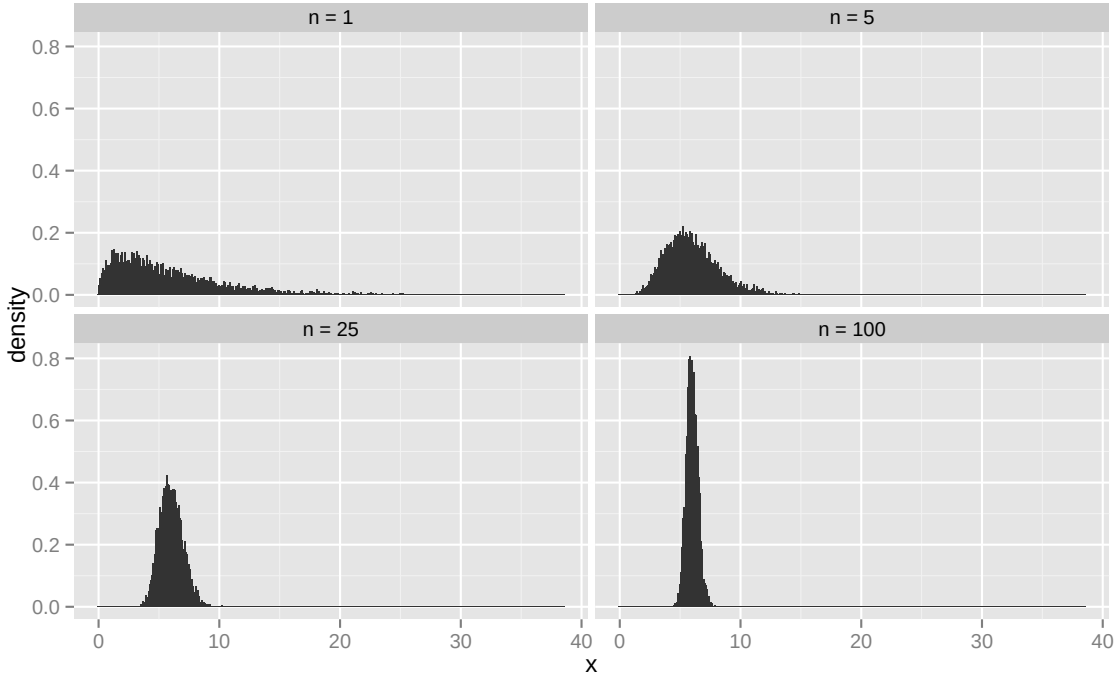
Central Limit Theorem. *Let $X_1, \ldots X_n$ be independent observations collected from a distribution with expectation μ and variance σ^2. Then the distribution of $\bar{X}$ converges to a normal distribution with expectation μ and variance σ^2/n as $n \to \infty$.*

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

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

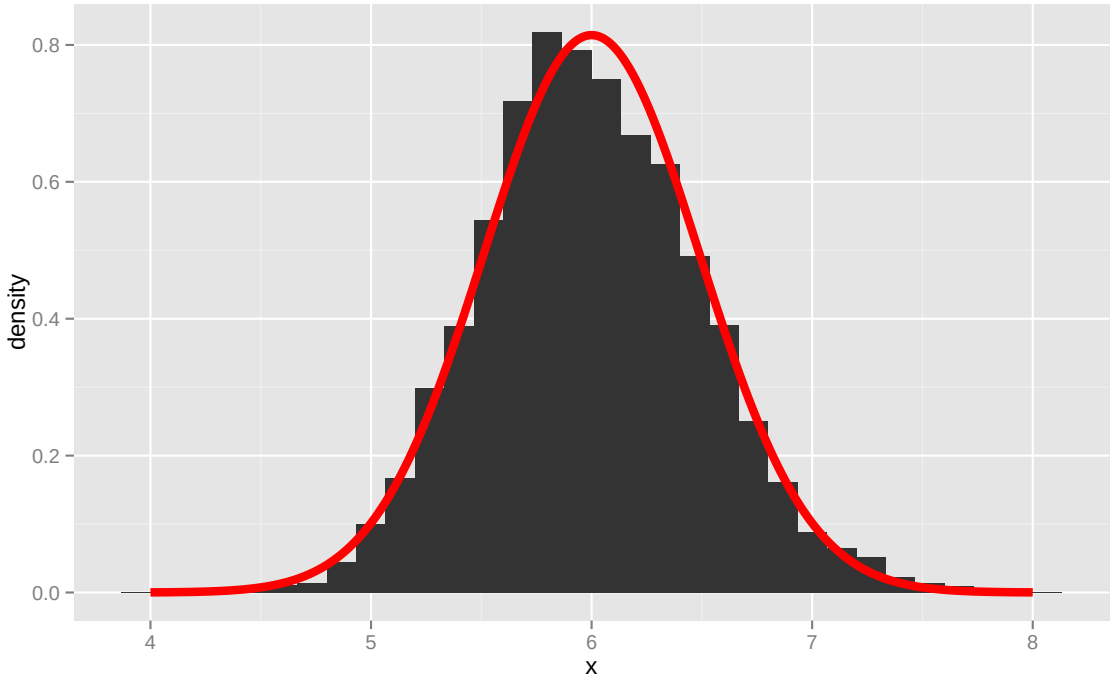
Evidence:

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?

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

Chapter 6

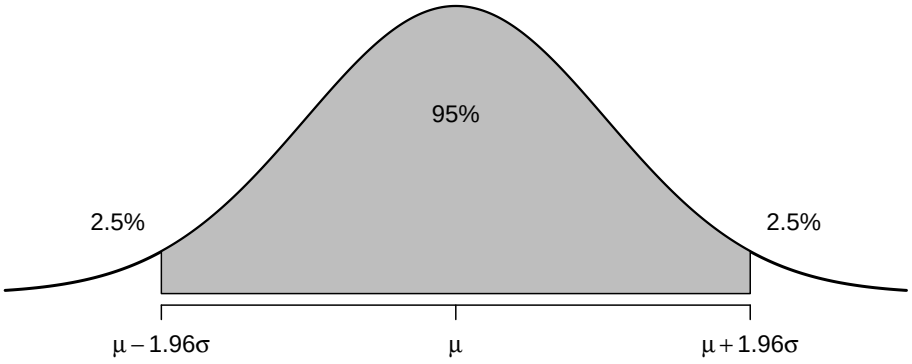
Confidence Intervals and T-tests

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.

6.1 Confidence Intervals assuming σ is known

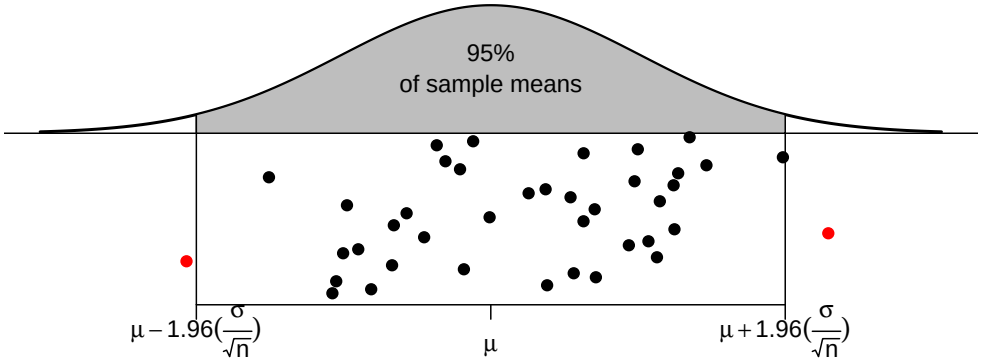
Recall that if $Z \sim N(\mu = 0,\ \sigma = 1)$ then we can find the middle $1-\alpha$ percent of the distribution by finding the $1-\alpha/2$ quantile. For example, I might be interested in the middle 95% of the distribution and thus $\alpha = 0.05$ the 97.5^{th} quantile of the standard normal distribution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

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



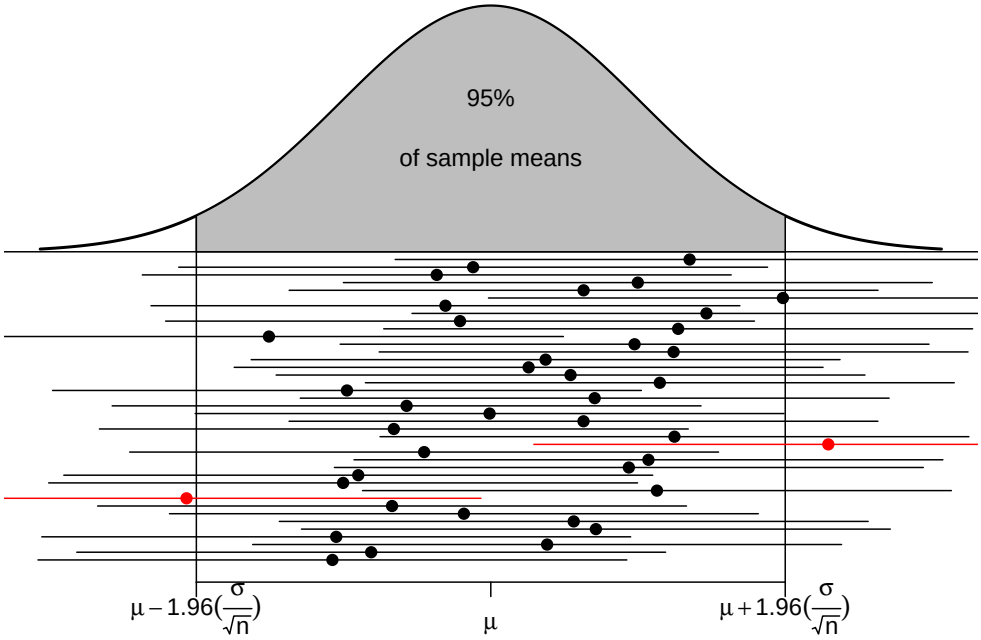
Generalizing this statement to its logical conclusion, the middle $1-\alpha$ of any normal distribution is found by $\mu \pm z_{1-\alpha/2}\,\sigma$.

If a sample was taken from a distribution with expectation μ and standard deviation σ, then $\bar{X}$ has expectation μ and standard deviation $\sigma/\sqrt{n}$. If the data were sampled from a normal distribution (or if the sample size is large) then the sample mean $\bar{X}$ has a normal distribution (or approximately normal) distribution. 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.

6.2 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.

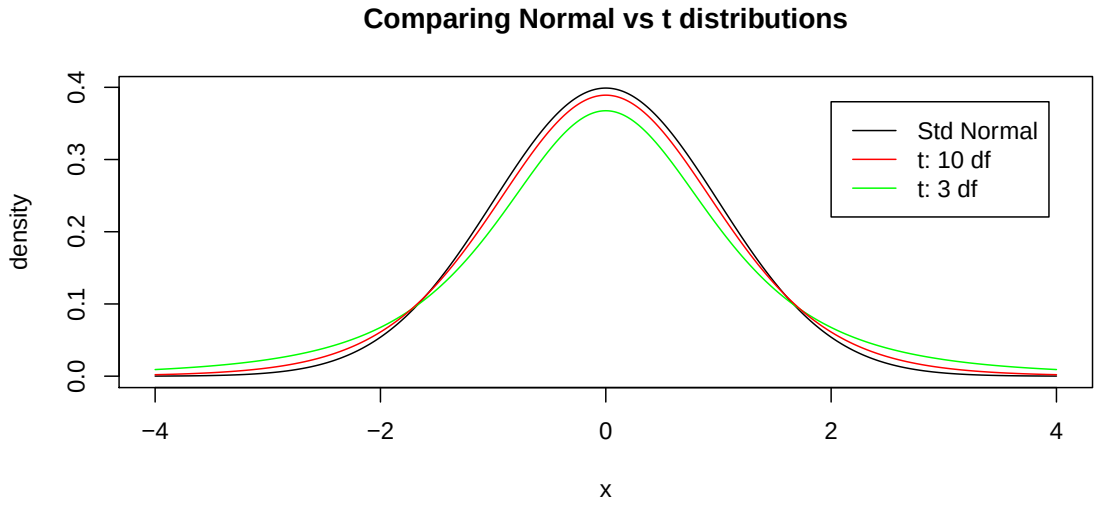
Assume that $X_1, \ldots, X_n \overset{iid}{\sim} N\left(\mu, \sigma^2\right)$ and therefore $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ and

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

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. Since Guineas 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 normally distributed.

Substituting the sample standard deviation into the confidence interval formula, we must 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 sample mean $\bar{x} = 350$ pounds and a sample standard deviation $s = 60$ pounds. Since 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] 2.023
```

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

$$\begin{aligned} \bar{x} &\pm t_{n-1}^{1-\alpha/2}\left(\frac{s}{\sqrt{n}}\right) \\ 350 &\pm 2.02\left(\frac{60}{\sqrt{40}}\right) \\ 350 &\pm 19.16 \end{aligned}$$

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."

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 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 E be the half-width desired (so the confidence interval would be $\bar{x} \pm E$). To do this calculation, we must also have some estimate of the population standard deviation. Denote this estimate as s. Then we must solve

$$E = t_{n-1}^{1-\alpha/2} \left(\frac{s}{\sqrt{n}} \right)$$

for n. Perhaps the easiest way to solve this (because $t_{n-1}^{1-\alpha/2}$ keeps changing) is to resort to guess-and-check. A good starting point would be to use $z_{1-\alpha/2}$ and solve for n.

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 $s = 60$ lbs is a reasonable standard deviation number to use.

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

So we'll look at half-widths for $n \in [97, 102]$ where the half-width is $t_{n-1}^{1-\alpha/2} \left(\frac{s}{\sqrt{n}} \right)$

n	97	98	99	100	101	102
$t_{n-1}^{0.95} \left(\frac{60}{\sqrt{n}} \right)$	10.11	10.06	10.01	9.96	9.91	9.86

and conclude that we should use $n = 100$ to achieve the desired confidence interval width. Notice that for large samples, the difference between the estimate using the standard normal quantile was quite close to the actual value. Finally we should emphasize that this calculation is only as accurate as the estimation of σ by the estimate s.

Your book uses the estimate s as the actual parameter value σ and just uses the standard normal quantile. This avoids the computationally tedious final step and could be justified by our desire is only to get to an approximate sample size and the final number (n of 97 vs 100) will be determined by sampling logistics. However software packages can handle the tedious calculations and will use the t-distribution.

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.

Newton proposed his 3 Laws of Motion which matched observed data quite well. However, as the accuracy of measuring instruments improved, eventually we were able to record data that did not agree with Newtonian physics, and a new theory was proposed by Einstein.

The point is that it is critical to have a method by which we can state if the data does not support a 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 $s = 50$. There are two possibilities. Either Bryan was just unlucky in his 10 cows, or the true average weight is less than 500.

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

Assuming[1] the true mean is 500, how likely is it to get a sample mean of 475? Since we know the distribution of $\bar{x}$ we can (and will) calculate how far into the tails this sample is.

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{475 - 500}{50/\sqrt{10}} = -1.58$$

This value is in the tail of the distribution, so in this case the data do not tend to support H_0, however if my hypotheses where different, this t might support it so I want to describe this t in terms of the null and alternative hypothesis.

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] 0.07428
```

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.

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.

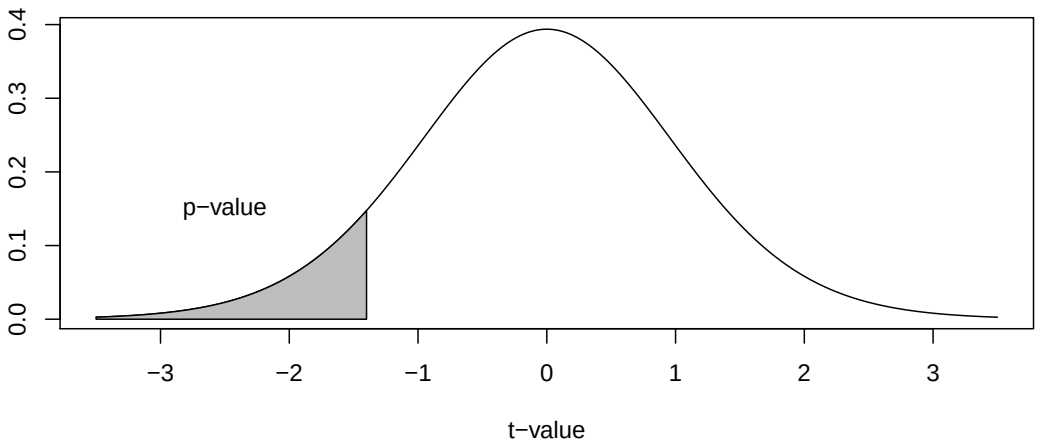
H_0 is often a statement of no effect, or no difference between the claimed and observed.

[1] For this calculation we'll assume the weight of a steer is normally distributed, and therefore $\bar{X}$ is also normally distributed.

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



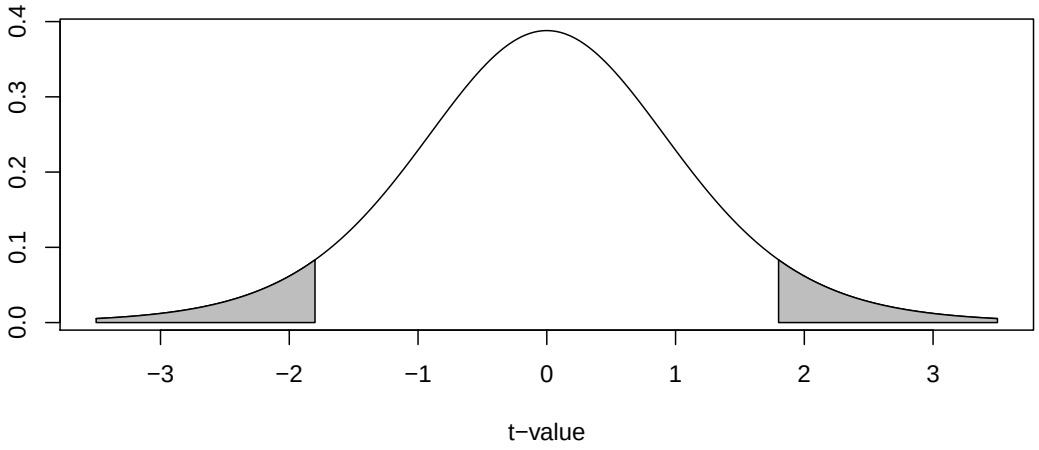
and we calculate

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

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

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

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* $\bar{x}$. Since we calculate 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. Since 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.
There is not 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.[2] 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.

[2]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.

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. Since 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.

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.

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

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

amd calculate

```
x <- c(31.8, 32.1, 32.5, 30.9, 31.3)
mean(x)

## [1] 31.72

sd(x)

## [1] 0.634
```

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

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

```
t.test(x, mu=31, alternative="two.sided")

##
##  One Sample t-test
##
## data:  x
## t = 2.539, df = 4, p-value = 0.06403
## alternative hypothesis: true mean is not equal to 31
## 95 percent confidence interval:
##  30.93 32.51
## sample estimates:
## mean of x
##     31.72
```

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 Relationship between Confidence Intervals and Hypothesis Tests

In many respects the calculations done to create a confidence interval are the same calculations done for a hypothesis test so it should not be surprising that there is a relationship between them. Notice that in our CO_2 example that the function `t.test()` also returns a confidence interval. Any value in that $100(1-\alpha)\%$ confidence interval would not be rejected with an two-sided α-level hypothesis test.

Returning to our fuel economy example, a value of $\mu = 31$ mpg should not be rejected at the $\alpha = 0.05$ level because it is in the 95% confidence interval. Likewise a value of $\mu = 32$ would not be rejected, but a value of $\mu = 32.75$ would.

```
t.test(x, mu=32.75, alternative="two.sided")

##
##  One Sample t-test
##
## data:  x
## t = -3.632, df = 4, p-value = 0.02211
## alternative hypothesis: true mean is not equal to 32.75
## 95 percent confidence interval:
##  30.93 32.51
## sample estimates:
## mean of x
##     31.72
```

6.4.1 One-Sided Hypothesis Tests and their associated Confidence Intervals

Since we can think about a confidence interval as the set of parameter values that would not be rejected in a t-test, it is natural to ask what this means in a one-sided t-test.

A one-sided confidence interval takes the form:

$H_a: \mu < \mu_0$	$H_a: \mu > \mu_0$
$\left(-\infty,\ \bar{x} + t_{n-1}^{1-\alpha}\left(\frac{s}{\sqrt{n}}\right)\right]$	$\left[\bar{x} - t_{n-1}^{1-\alpha}\left(\frac{s}{\sqrt{n}}\right),\ \infty\right)$

To understand why the intervals are this way we will again use the fuel economy example. Recall, we were testing

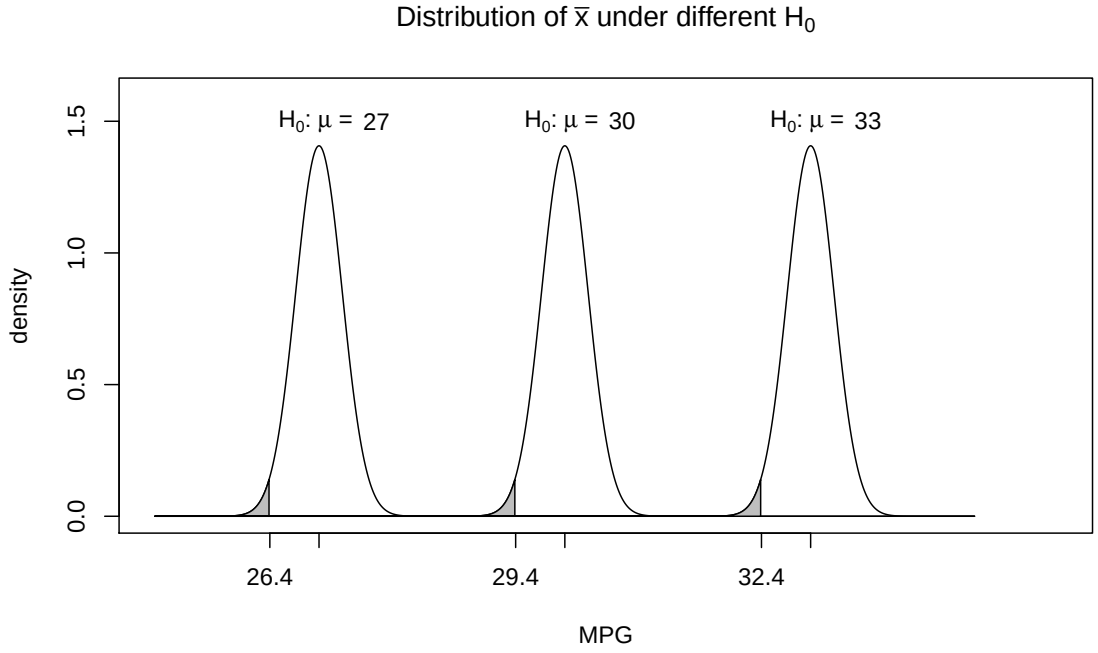
$$\begin{aligned} H_0: \mu &= \mu_0 \\ H_a: \mu &\leq \mu_0 \end{aligned}$$

where $\mu_0 = 31$ mpg, using a t-test with $n = 5$ at the $\alpha = 0.05$ rate. Given a sample mean of $\bar{x} = 31.72$ and sample standard deviation $s = 0.63$, we can calculate what values of μ_0 would not be rejected.

Since the critical t-value for this test is $t_4^{0.05} = -2.13$ then to reject H_0, $\bar{x}$ must be less than some critical x-value. We can find that via the following calculation.

$$\begin{aligned} -2.13 &> \frac{\bar{x}_{crit} - \mu_0}{s/\sqrt{n}} \\ -2.13\left(\frac{s}{\sqrt{n}}\right) + \mu_0 &> \bar{x}_{crit} \\ \bar{x}_{crit} &< \mu_0 - 2.13\left(\frac{s}{\sqrt{n}}\right) \\ \bar{x}_{crit} &< \mu_0 - 2.13\left(\frac{0.634}{\sqrt{5}}\right) \\ \bar{x}_{crit} &< \mu_0 - 0.60 \end{aligned}$$

In the case of $\mu_0 = 27$, then $\bar{x} < 26.4$ would lead us to reject the null hypothesis. If $\mu_0 = 30$, then $\bar{x} < 29.4$ would lead us to reject the null hypothesis. If $\mu_0 = 33$, then $\bar{x} < 32.4$ would lead us to reject the null hypothesis. If would lead us to reject the null hypothesis.



Since we actually observed $\bar{x} = 31.72$, then $\mu_0 = 27$ and $\mu_0 = 30$ would not be rejected but $\mu = 33$ would be. Finally we can generalize this idea and find that for μ_0 to not be rejected, the following must by true

$$\begin{aligned} t_{n-1}^{\alpha} &\leq \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \\ -2.13 &\leq \frac{31.72 - \mu_0}{0.63/\sqrt{5}} \\ -2.13\left(\frac{0.63}{\sqrt{5}}\right) &\leq 31.72 - \mu_0 \\ \mu_0 &\leq 31.72 + 2.13\left(\frac{0.63}{\sqrt{5}}\right) \\ \mu_0 &\leq 32.32 \end{aligned}$$

So any hypothesized value $\mu_0 \leq 32.32$ would not be rejected given $\bar{x} = 31.72$ and $s = 0.63$. So my one-sided confidence interval for μ is

$$(-\infty, 32.32]$$

which seems a little odd, but is the set of values for μ that would not be rejected in a one-sided hypothesis test.

```
x <- c(31.8, 32.1, 32.5, 30.9, 31.3)
t.test(x, mu=31, alternative="less")

##
##  One Sample t-test
##
## data:  x
## t = 2.539, df = 4, p-value = 0.968
## alternative hypothesis: true mean is less than 31
## 95 percent confidence interval:
```

```
##   -Inf 32.32
## sample estimates:
## mean of x
##     31.72
```

6.5 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

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

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.

6.5.1 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 $\hat{\sigma}$.

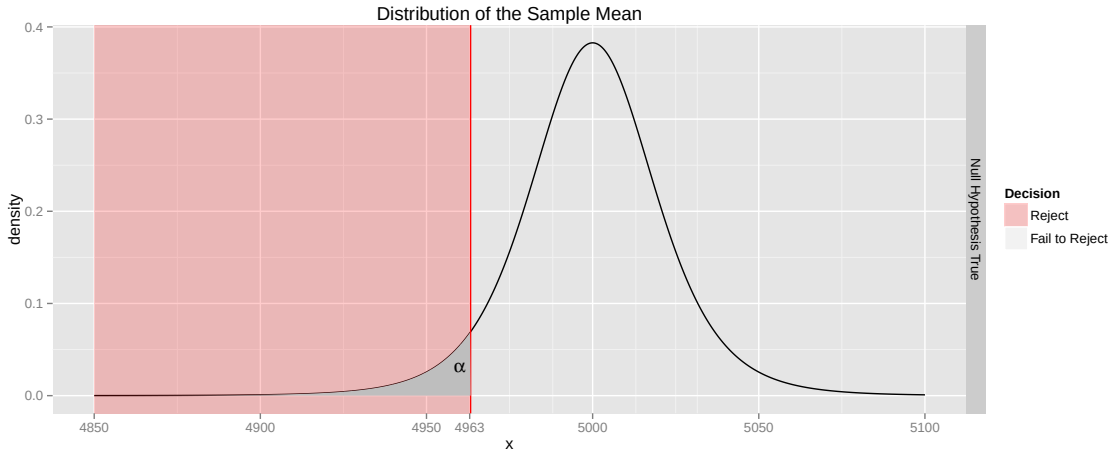
Example. Suppose that I work in Quality Control for a company that manufactures a type of rope. This rope is supposed to have a mean breaking strength of 5000 pounds and long experience with the process suggests that the standard deviation is approximately $s = 50$. As with many manufacturing processes, 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 $\bar{x}_{crit}$

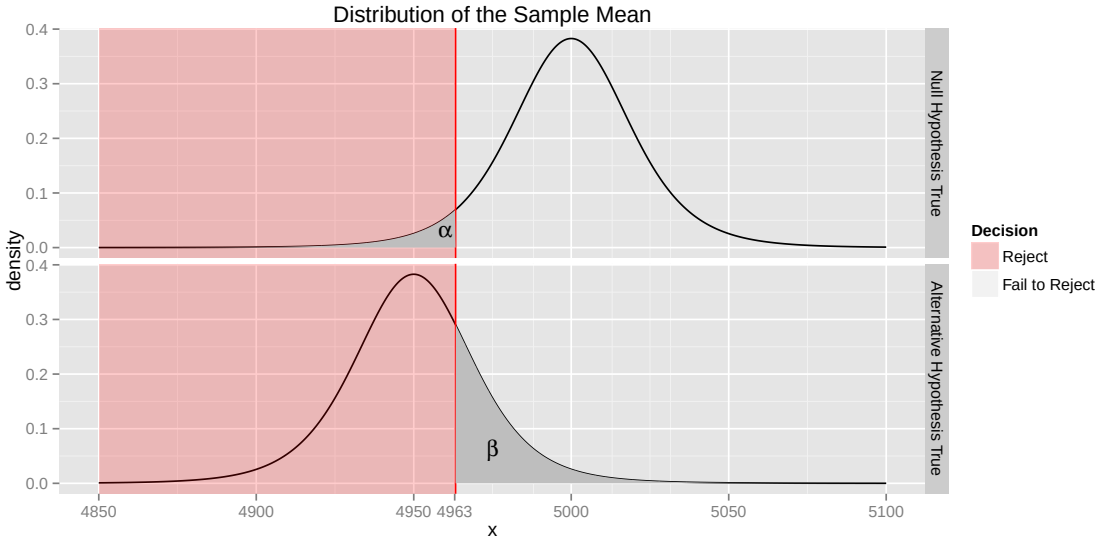
$$\begin{aligned} t_{crit} &= \frac{\bar{x}_{crit} - \mu_0}{\frac{s}{\sqrt{n}}} \\ t_{crit}\left(\frac{s}{\sqrt{n}}\right) + \mu_0 &= \bar{x}_{crit} \\ -1.943\left(\frac{50}{\sqrt{7}}\right) + 5000 &= \bar{x}_{crit} \\ 4963 &= \bar{x}_{crit} \end{aligned}$$

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 $\mu = 4950$.

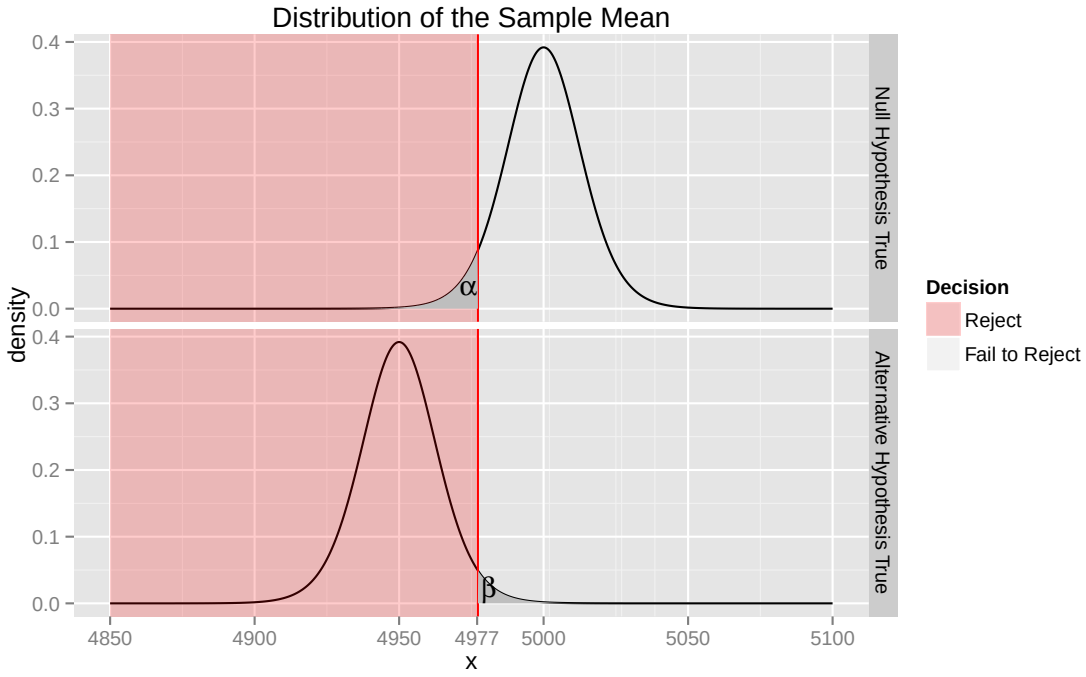


The the probability of a type II error is

$$\begin{aligned} \beta &= P\left(\bar{X} > 4963.3 \,|\, \mu = 4950\right) \\ &= P\left(\frac{\bar{X} - 4950}{50/\sqrt{7}} > \frac{4963.3 - 4950}{50/\sqrt{7}}\right) \\ &= P\left(T_6 > 0.703\right) \\ &= 0.254 \end{aligned}$$

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.[3] 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 $n = 15$.

[3]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()` available in the package `pwr`. It might be necessary for you to install the package from the Comprehensive R Archive Network (CRAN) before the following code will work.

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.

```
library(pwr)
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 = 0.7544
##     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 = 12.32
##           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%.

6.6 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.

6.6.1 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.

6.6.2 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

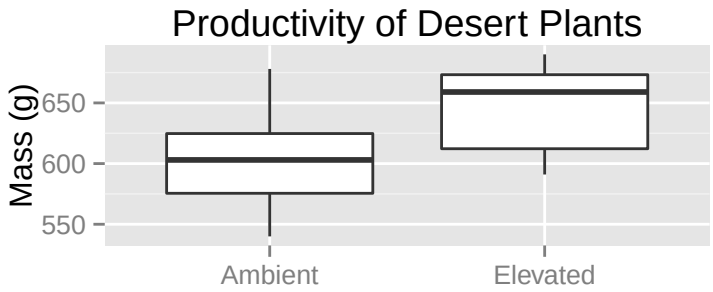
$$\begin{aligned} p - value &= 2 \cdot P\left(T_{37} < -4.15\right) \\ &= 0.000187 \end{aligned}$$

The equivalent analysis in R is as follows:

```
ambient <- c(540, 634, 620, 606, 598, 627, 593, 541, 577, 638, 571,
             649, 678, 604, 559, 624, 553, 614, 602, 594)
elevated <- c(685, 677, 610, 601, 682, 659, 638, 687, 609, 607, 690,
              591, 613, 647, 672, 664, 659, 618, 669, 659)
t.test(ambient, elevated, var.equal=FALSE)

##
##  Welch Two Sample t-test
##
## data:  ambient and elevated
## t = -4.157, df = 37.58, p-value = 0.0001797
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -68.04 -23.46
## sample estimates:
## mean of x mean of y
##     601.1     646.9
```

6.6.3 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_{pooled}^2 instead of s_1^2 and s_2^2. So our test statistic becomes

$$\begin{aligned} t_{df=n_1+n_2-2} &= \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_{pool}^2}{n_1} + \frac{s_{pool}^2}{n_2}}} \\ &= \frac{(\bar{x}_1 - \bar{x}_2) - 0}{s_{pool}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \end{aligned}$$

Example. If we had decided to pool the variance in the elevated CO_2 example we would have

$$\begin{aligned} s_{pooled}^2 &= \frac{1}{n_1 + n_2 - 2}\left[(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2\right] \\ &= \frac{1}{38}\left[(19)\, 36.6^2 + (19)\, 32.92^2\right] \\ &= 1211.643 \\ s_{pooled} &= 34.81 \end{aligned}$$

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(ambient, elevated, var.equal=TRUE)

##
##  Two Sample t-test
##
## data:  ambient and elevated
## t = -4.157, df = 38, p-value = 0.000177
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
```

```
##  -68.03 -23.47
## sample estimates:
## mean of x mean of y
##     601.1     646.9
```

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?

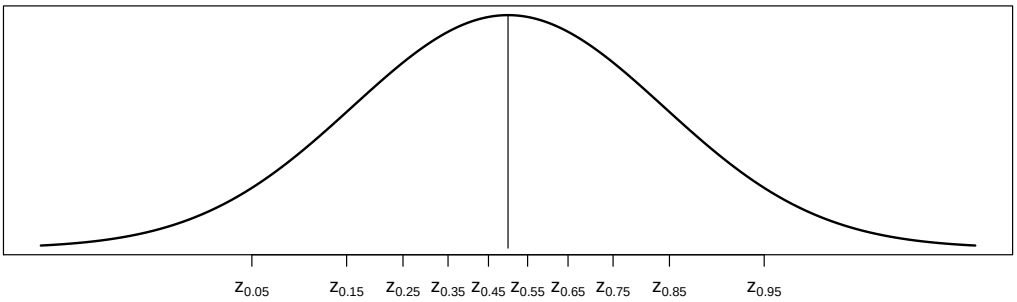
If we are using a model that assumes equal variance between groups, how should we test if that assumption is true?

7.1 Testing Normality

7.1.1 Visual Inspection - QQplots

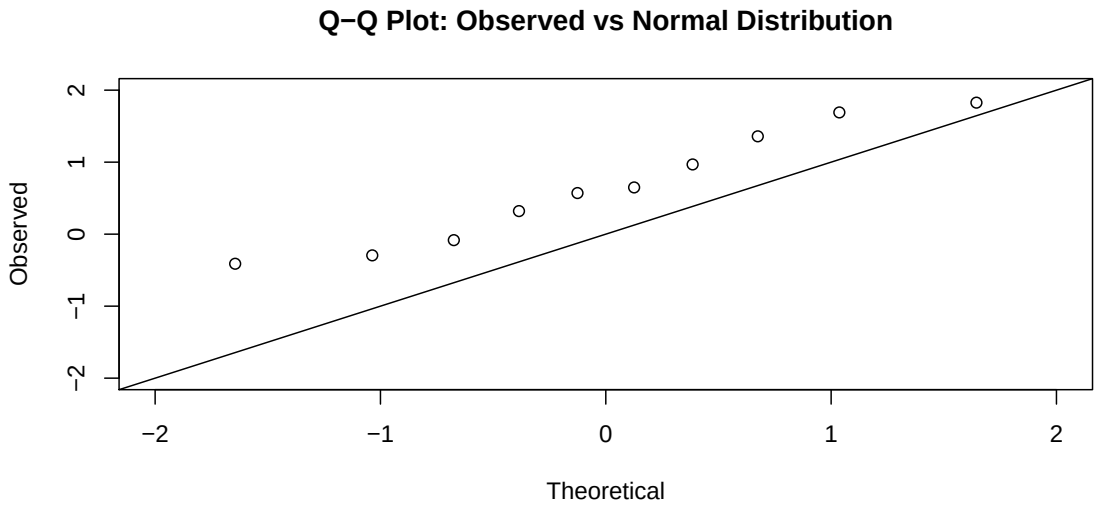
If we are taking a sample of size $n = 10$ from a standard normal distribution, then I should expect that the smallest observation will be negative. Intuitively, you would expect the smallest observation to be near the 10th percentile of the standard normal, and likewise the second smallest should be 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*.



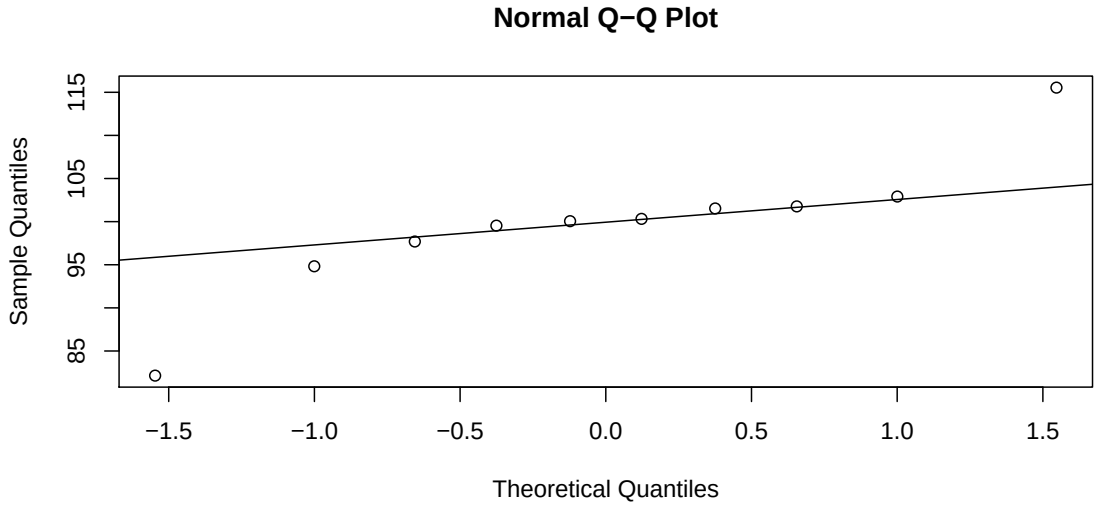
I can then graph the theoretical quantiles vs my observed values and if they lie on the 1-to-1 line, then my data comes from a standard normal distribution.

```
n <- 10
x <- rnorm(n, mean=0, sd=1)
theoretical <- qnorm( (1:n - .5)/(n), mean=0, sd=1)
plot(theoretical, sort(x),
     xlab='Theoretical', ylab='Observed',
     xlim=c(-2,2), ylim=c(-2,2),
     main="Q-Q Plot: Observed vs Normal Distribution")
abline(0, 1)
```

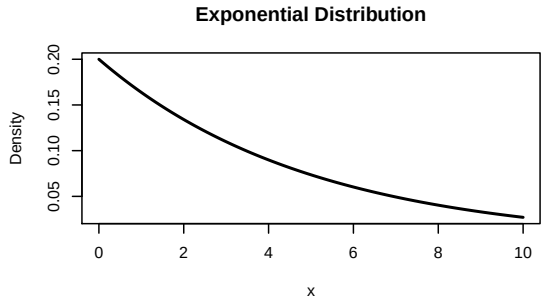


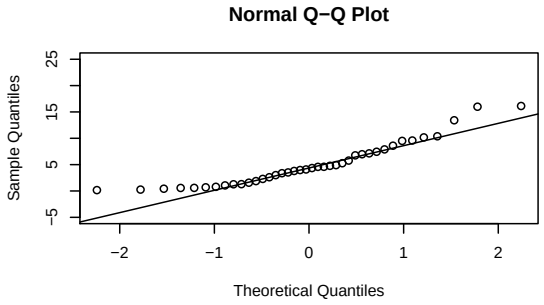
If I think my data are normal, but with some mean μ and standard deviation σ, we still make the same graph, but the 1-to-1 line will be moved to pass through the 1st and 3rd quartiles. Again, the data points should be near the line. This is common enough that R has built in functions to make this graph:

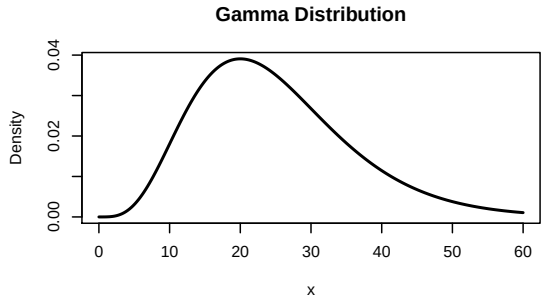
```
n <- 10
x <- rnorm(n, mean=100, sd=10)
qqnorm(x)
qqline(x)
```

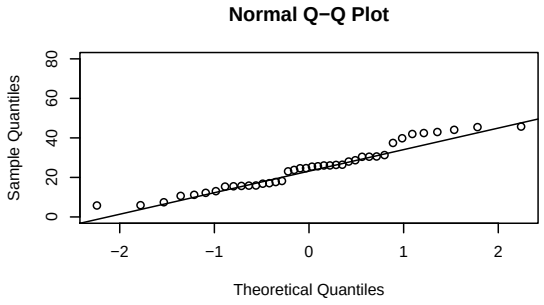


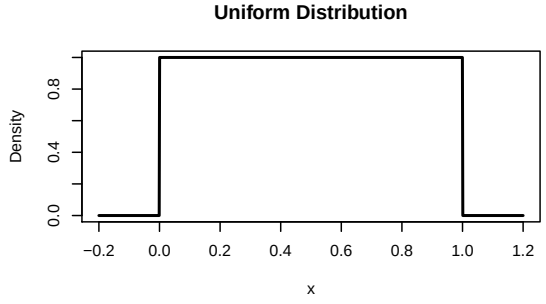
We now will examine a sample of $n = 40$ from a bunch of different distributions that are not normal and see what the normal QQ plot looks like. In the following graphs, pay particular attention to the tails. Notice the the T-distribution has significantly heavier tails than the normal distribution and 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.

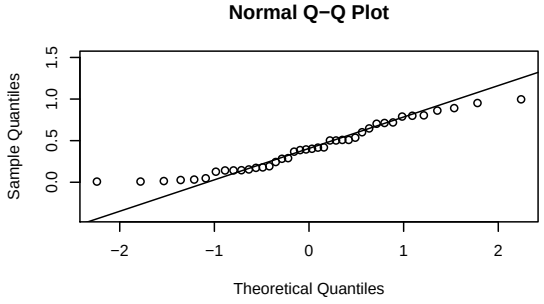

Exponential Distribution
Density
x

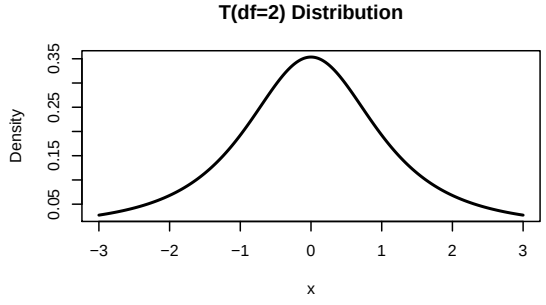

Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles

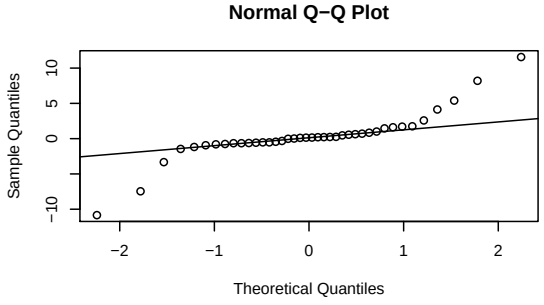

Gamma Distribution
Density
x

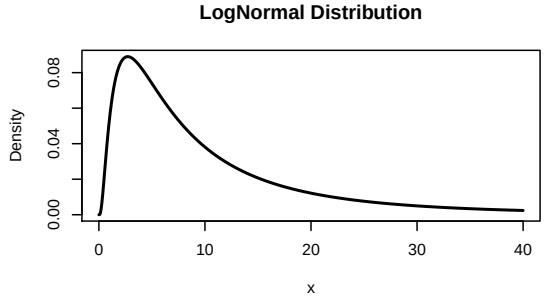

Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles

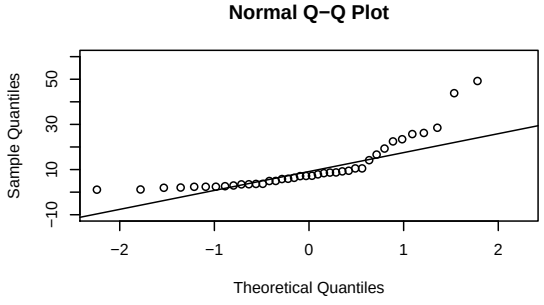

Uniform Distribution
Density
x


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 = 0.3934, p-value = 2.094e-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 = 0.9491, p-value = 0.658
```

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 that do not make distributional assumptions.

One general technique that is applicable in these circumstances is bootstrap methods, which we will cover in STA 571.

7.2 Testing Equal Variance

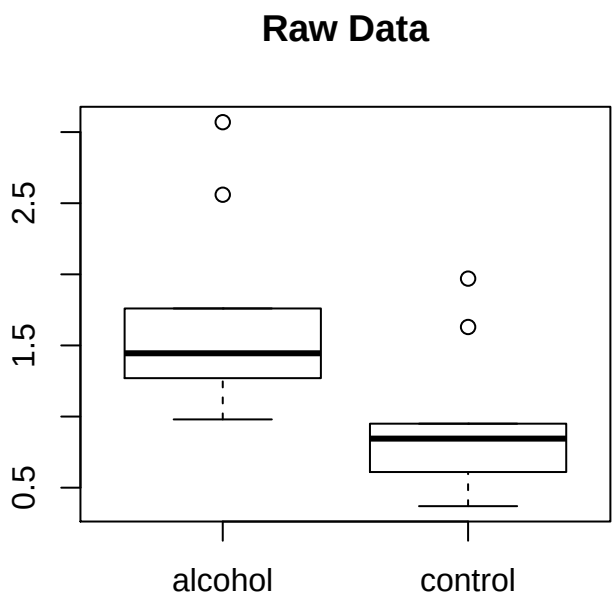
7.2.1 Visual Inspection

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]

```
control <- c(0.90, 0.37, 1.63, 0.83, 0.95, 0.78, 0.86, 0.61, 0.38, 1.97)
alcohol <- c(1.46, 1.45, 1.76, 1.44, 1.11, 3.07, 0.98, 1.27, 2.56, 1.32)
data <- data.frame(time=c(control, alcohol),
trt=rep(c("control","alcohol"),each=10))
str(data)

## 'data.frame': 20 obs. of  2 variables:
##  $ time: num  0.9 0.37 1.63 0.83 0.95 0.78 0.86 0.61 0.38 1.97 ...
##  $ trt : Factor w/ 2 levels "alcohol","control": 2 2 2 2 2 2 2 2 2 2 ...

boxplot(time ~ trt, data=data, main='Raw Data')
```



7.2.2 Tests for Equal Variance

Consider having samples drawn from normal distributions

$$X_{ij} \sim N\left(\mu_i, \sigma_i^2\right)$$

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 $\sigma_i^2 = \sigma_j^2$.

[1] This study was long enough ago that review boards let this sort of thing be done.

Let's consider the simplest case of two populations and consider the null and alternative hypotheses:

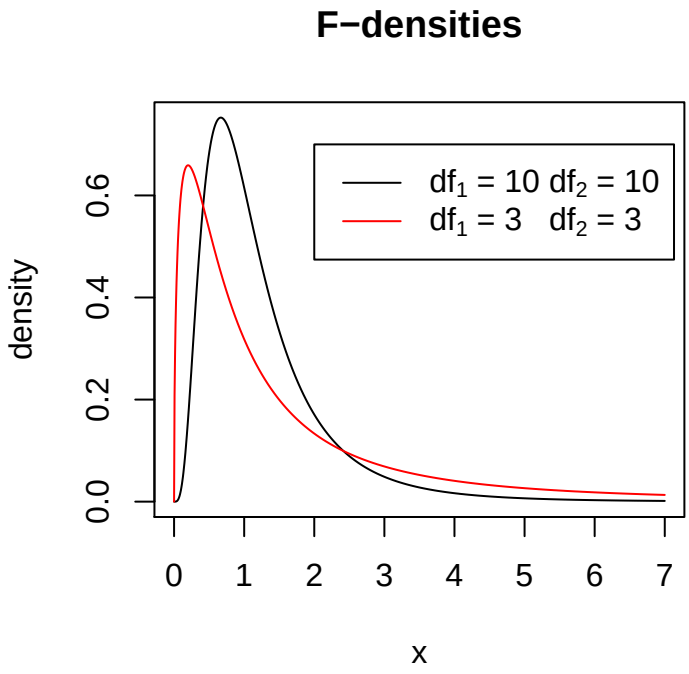
$$\begin{aligned} H_0 : \sigma_1^2 &= \sigma_2^2 \\ H_a : \sigma_1^2 &\neq \sigma_2^2 \end{aligned}$$

If the null hypothesis is true, then the ratio s_1^2/s_2^2 should be approximately one. It can be shown 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:



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] 0.181
```

```
qf(0.975, 6, 9)
```

```
## [1] 4.32
```

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] 0.2655
```

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

$$P\left(F_{df1,df2} > \frac{s_1^2}{s_2^2}\right) = P\left(F_{df2,df1} < \frac{s_2^2}{s_1^2}\right)$$

For example, suppose that $n_1 = 5$ and $n_2 = 20$ and $s_1^2 = 6$ and $s_2^2 = 3$ then

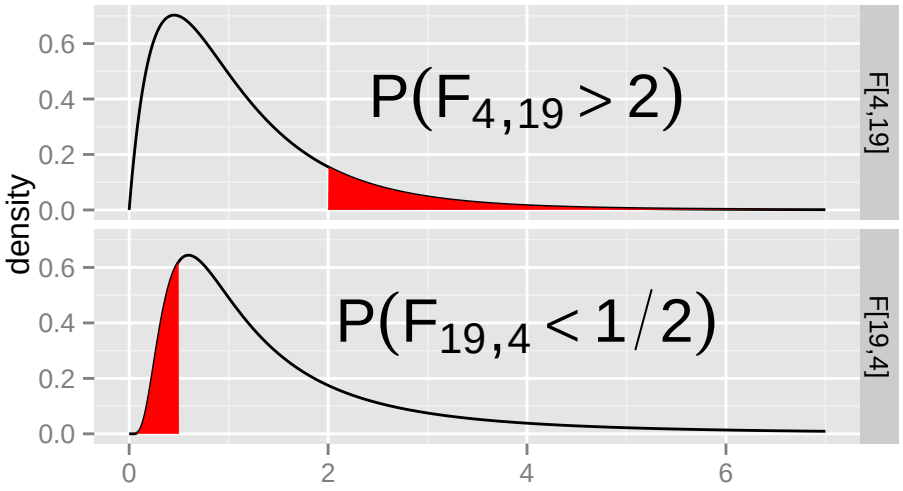
$$P\left(F_{4,19} > \frac{6}{3}\right) = P\left(F_{19,4} < \frac{3}{6}\right)$$

```
1 - pf(6/3, 4, 19)
```

```
## [1] 0.1354
```

```
pf(3/6, 19, 4)
```

```
## [1] 0.1354
```



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] 0.1143
```

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] 4.276e-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.

Returning to the research example with the alcohol and control group, an F-test for different variances results in a p-value of

```
v1 <- var(alcohol)
v2 <- var(control)
f <- v1/v2
f

## [1] 1.705

2*(1-pf(f,9,9))

## [1] 0.439
```

F-tests in R

Calculating test statistics by hand is often error prone so one naturally wonders how to do the F-test in R. Here we will use the function `var.test()` which takes two vectors as arguments.

```
var.test(alcohol, control)

##
##  F test to compare two variances
##
## data:  alcohol and control
## F = 1.705, 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:
##  0.4234 6.8633
## sample estimates:
## ratio of variances
##              1.705
```

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 what predictors should be included in my model of the response variable. Given twenty different predictors, I want to pare it 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.

First we will start small and come up with a test that helps me evaluate if a model that has a categorgical 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.

It is useful to notice that this is a categorical variable (which population) being used to predict a continuous response.

Suppose that my data can be written as

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

where μ_i is the mean of group i and $\epsilon_{ij} \overset{iid}{\sim} N\left(0, \sigma^2\right)$ 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.

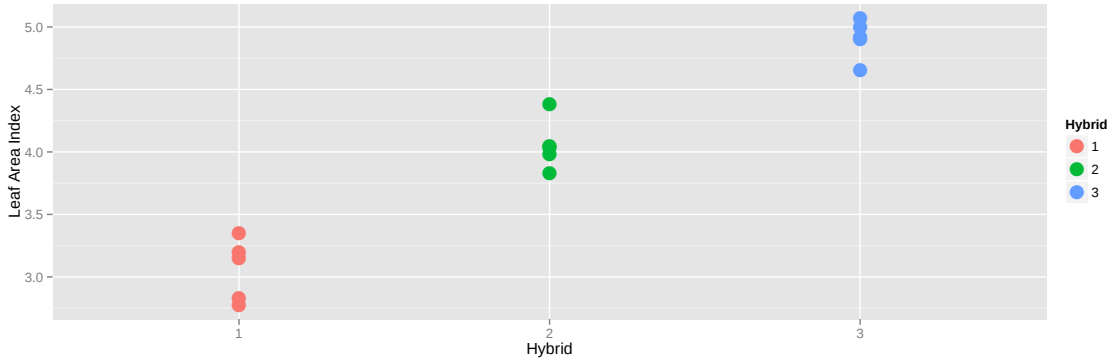
Assumptions:

1. *The error terms come from a normal distribution*
2. *The variance of each group is the same*
3. *The observations are independent*

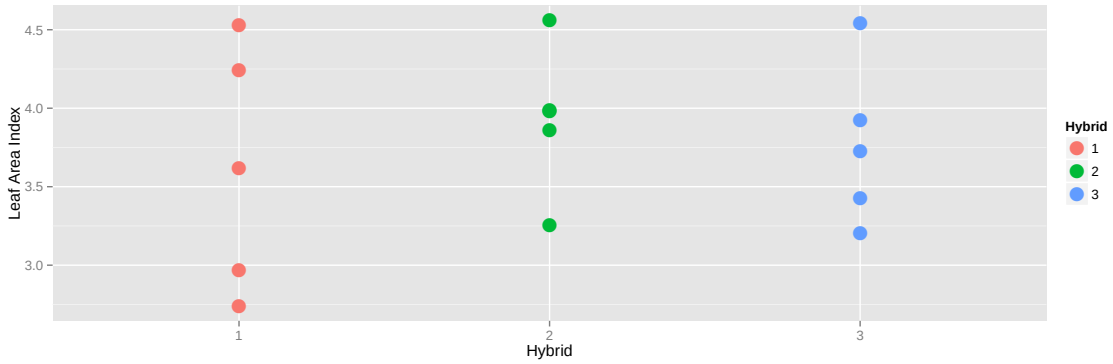
In general I want to test the hypotheses

$$\begin{aligned} H_0: &\quad \mu_1 = \mu_2 = \cdots = \mu_k \\ H_a: &\quad \text{at least on mean is different than the others} \end{aligned}$$

Example 15. Suppose that we have three hybrids of a particular plant and we measure the leaf area for each hybrid. In this case, it looks like there is a difference in the means of each hybrid:



However the following graph does not appear to have a difference between the hybrid means:



What is the difference between these two?

1. If there is small variance *within* a hybrid compared to the variance *between* hybrid means, 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.
2. If the *within* hybrid variance is huge compared to the differences between the hybrid means, then I would fail to reject the null hypothesis of equal means (this would be the second 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.

8.2 Theory

Notation: denote $n_t = n_1 + n_2 + \cdots + n_k$ as the total number of observations and $\bar{y}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$ as the sample mean from the ith group and $\bar{y}_{\cdot\cdot}$ be the mean of all the observations.

Under the null hypothesis there are two different ways to estimate σ^2. First 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.

$$\begin{aligned} s_W^2 &= \frac{\sum_{j=1}^{n_1}(y_{1j}-\bar{y}_{1\cdot})^2 + \sum_{j=1}^{n_2}(y_{2j}-\bar{y}_{2\cdot})^2 + \cdots + \sum_{j=1}^{n_k}(y_{kj}-\bar{y}_{k\cdot})^2}{(n_1-1)+(n_2-1)+\cdots+(n_k-1)} \\ &= \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2 + \cdots + (n_k-1)s_k^2}{n_t-k} \end{aligned}$$

The 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 $N\left(\mu, \frac{\sigma^2}{n_i}\right)$ and therefore the k sample means could be used to estimate σ^2.

$$s_B^2 = \frac{1}{k-1}\sum_{i=1}^{k} n_i(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})^2$$

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 $n_1 + n_2 + \cdots + n_k - k$ 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 variablity that we are dealing with.

SSW: Sum of Squares Within - This is the variability within sample groups. It has an associated $df_W = n_t - k$

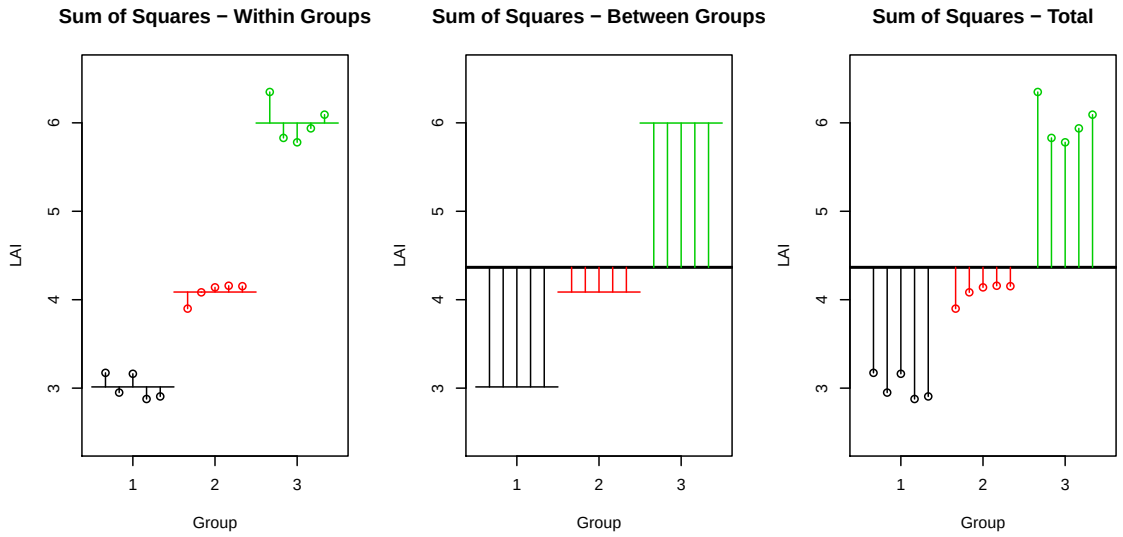
$$SSW = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot})^2$$

SSB: Sum of Squares Between - This is the variability between sample groups. It has an associated $df_B = k-1$

$$SSB = \sum_{i=1}^{k} n_i(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})^2$$

SST: Sum of Squares Total - This is the total variability in the data set. It has an associated $df = n_t - 1$ because under the null hypothesis there is only one mean μ.

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_j}(y_{ij}-\bar{y}_{\cdot\cdot})^2$$



An anova table is usually set up the in the following way (although the total row is sometimes removed):

Source	Sum of Squares	df	Mean Squares	F-stat	P-value
Between Samples	SSB	$k-1$	$s_B^2 = SSB/(k-1)$	$f = s_B^2/s_W^2$	$P(F_{k-1,n_t-k} > f)$
Within Samples	SSW	n_t-k	$s_W^2 = SSW/(n_t-k)$		
Total	SST	n_t-1			

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.[1]

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 $\bar{y}_{..}$ as the estimate of μ and therefore our error terms are

$$e_{ij} = y_{ij} - \bar{y}_{..}$$

The sum of squared error associated with the simple model is thus

$$\begin{aligned} SSE_{simple} &= \sum_{i=1}^{k}\sum_{j=1}^{n_i} e_{ij}^2 \\ &= \sum_{i=1}^{k}\sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_{..}\right)^2 \\ &= SST \end{aligned}$$

[1] 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.

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 $\bar{y}_{i\cdot}$ as estimates for μ_i and thus the error terms are

$$e_{ij} = y_{ij} - \bar{y}_{i\cdot}$$

and the sum of squared error associated with the complex model is thus

$$\begin{aligned} SSE_{complex} &= \sum_{i=1}^{k}\sum_{j=1}^{n_i} e_{ij}^2 \\ &= \sum_{i=1}^{k}\sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_{i\cdot}\right)^2 \\ &= SSW \end{aligned}$$

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	Sum of Squares	df	Mean Squares	F-stat	P-value
Difference	SSE_{diff}	$k-1$	$MSE_{diff} = \frac{SSE_{diff}}{k-1}$	$f = \frac{MSE_{diff}}{MSE_{complex}}$	$P\left(F_{k-1,n_t-k} > f\right)$
Complex	$SSE_{complex}$	n_t-k	$MSE_{complex} = \frac{SSE_{complex}}{n_t-k}$		
Simple	SSE_{simple}	n_t-1			

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

$$\hat{\sigma}^2 = s_W^2 = MSE_{complex} = \frac{1}{n_t - k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{i\cdot}\right)^2$$

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

$$\bar{y}_{i\cdot} \pm t_{n_t-k}^{1-\alpha/2}\left(\frac{\hat{\sigma}}{\sqrt{n_i}}\right)$$

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 is fit using the function `lm()`. 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`.

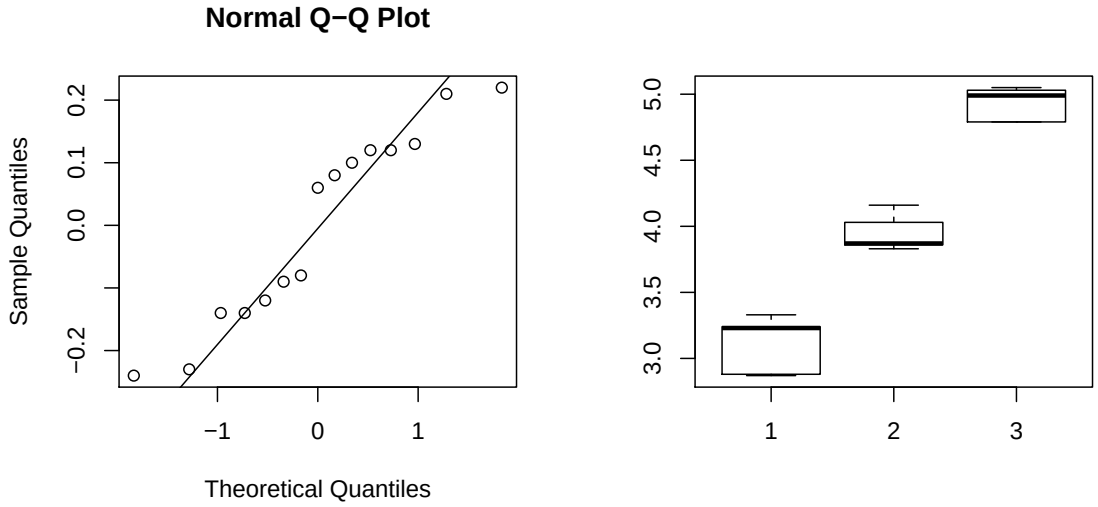
```
leaf.area.index <- c(2.88, 2.87, 3.23, 3.24, 3.33, 3.83,
   3.86, 4.03, 3.87, 4.16, 4.79, 5.03, 4.99, 4.79, 5.05)
population <- rep(1:3, each=5)
population <- factor(population)
data <- data.frame(LAI=leaf.area.index, Species=population)
str(data)

## '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 -1, 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 - 1` and that `-1` term is important and we will address why it is there later in the chapter.

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))
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()`.

```
anova(model)

## Analysis of Variance Table
##
## Response: LAI
##           Df Sum Sq Mean Sq F value Pr(>F)
## Species    3  247.9    82.6    2944 <2e-16 ***
## Residuals 12    0.3     0.0
## ---
## 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. Since 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. Since 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

$$\hat{\mu}_i = \bar{y}_{i\cdot}$$

$$\hat{y}_{i\cdot} \pm t_{n_t-k}^{1-\alpha/2}\left(\frac{\hat{\sigma}}{\sqrt{n_i}}\right)$$

and can be found using the `coef()` and `confint()` functions.

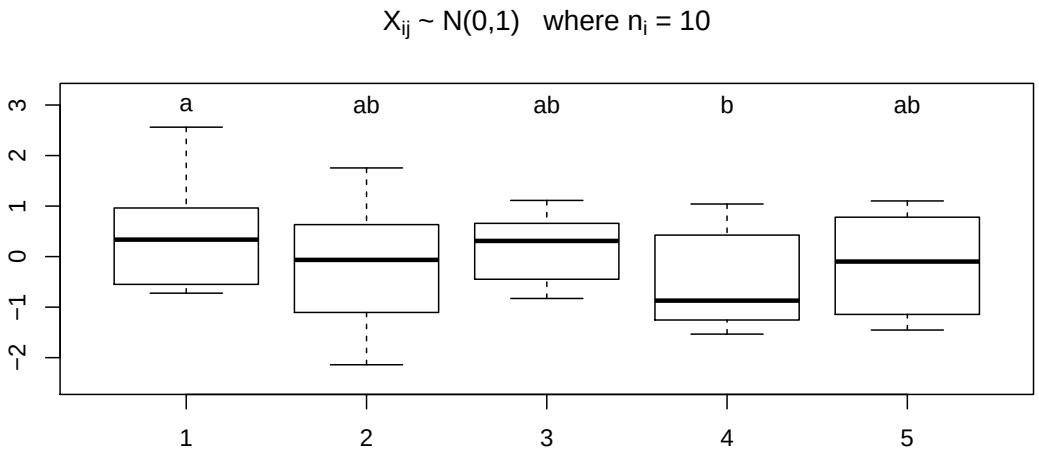
```
# To get coefficients in the way we have represented the
# complex model (which we call the cell means model), we
# must add a -1 to the formula passed to aov()
```

```
# 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 2.947  3.273
## Species2 3.787  4.113
## Species3 4.767  5.093
```

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 -0.55736 -1.8278 0.7131 0.7244
## 3-1 -0.23996 -1.5104 1.0305 0.9830
## 4-1 -0.98855 -2.2590 0.2819 0.1943
## 5-1 -0.62440 -1.8948 0.6460 0.6330
## 3-2  0.31739 -0.9530 1.5878 0.9532
## 4-2 -0.43120 -1.7016 0.8392 0.8695
## 5-2 -0.06705 -1.3375 1.2034 0.9999
## 4-3 -0.74859 -2.0190 0.5218 0.4597
## 5-3 -0.38444 -1.6549 0.8860 0.9099
## 5-4  0.36415 -0.9063 1.6346 0.9248
```

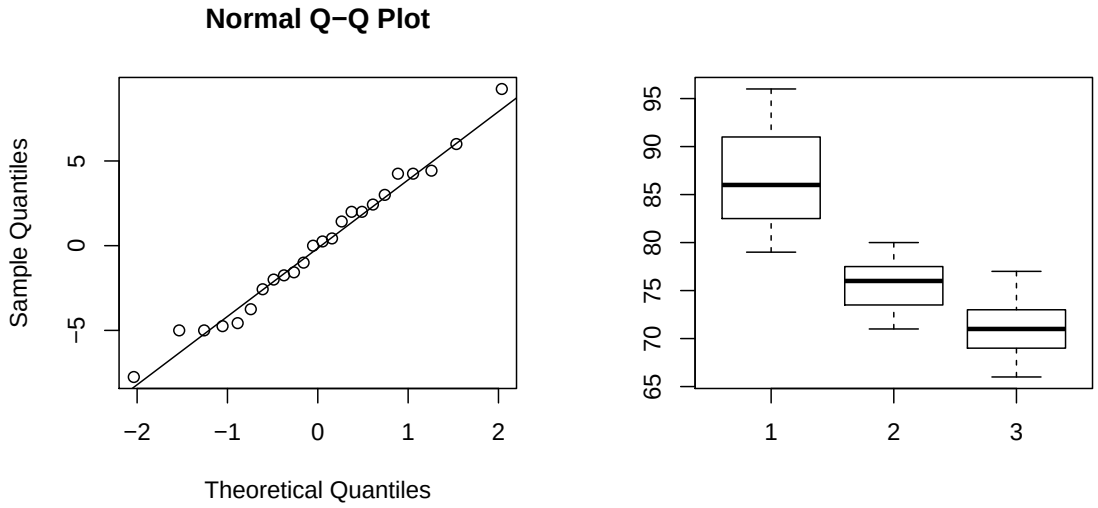
Likewise if we are testing the ANOVA assumption of equal variance, we cannot rely on doing all pairwise F-tests and we must use a method that controls the overall error rate. The multiple comparisons version of `var.test()` is Bartlett's test which is called similarly to `aov()`.

```
bartlett.test(Y~Group, mydata)

##
##  Bartlett test of homogeneity of variances
##
## data:  Y by Group
## Bartlett's K-squared = 3.14, df = 4, p-value = 0.5347
```

Example 16. (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 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. 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

```
x1 <- data.frame(HLT=c(96,79,91,85,83,91,82,87),    Method=1)
x2 <- data.frame(HLT=c(77,76,74,73,78,71,80),       Method=2)
x3 <- data.frame(HLT=c(66,73,69,66,77,73,71,70,74), Method=3)
hostility.data <- rbind(x1,x2,x3)
hostility.data$Method <- factor(hostility.data$Method)
model <- aov( HLT ~ Method, data=hostility.data )
par(mfrow=c(1,2))
qqnorm( resid(model) )
qqline( resid(model) )
boxplot(HLT~Method, data=hostility.data)
```



Looking at QQ-plot I'm content with the normality assumption, but the equal variance assumption might be suspect. The box length for the first group is quite a bit larger than the others, but that might be due to one observation, since our sample size is so small.

```
shapiro.test(resid(model))

##
##  Shapiro-Wilk normality test
##
## data:  resid(model)
## W = 0.9836, p-value = 0.9516

bartlett.test(HLT~Method, data=hostility.data)

##
##  Bartlett test of homogeneity of variances
##
## data:  HLT by Method
## Bartlett's K-squared = 2.459, 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.

```
anova(model)

## Analysis of Variance Table
##
## Response: HLT
##           Df Sum Sq Mean Sq F value  Pr(>F)
## Method     2   1091     545    29.6 7.8e-07 ***
## Residuals 21    387      18
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since 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.data )
coef(model.2)

## Method1 Method2 Method3
##   86.75   75.57   71.00

confint(model.2)

##         2.5 % 97.5 %
## Method1 83.59  89.91
## Method2 72.20  78.95
## Method3 68.02  73.98
```

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.data)
##
## $Method
##        diff    lwr      upr  p adj
## 2-1 -11.179 -16.78  -5.5769 0.0002
## 3-1 -15.750 -21.01 -10.4908 0.0000
## 3-2  -4.571 -10.03   0.8831 0.1114
```

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.data$HLT, hostility.data$Method,
                pool.sd=FALSE, p.adjust.method='none')

##
##  Pairwise comparisons using t tests with non-pooled SD
##
## data:  hostility.data$HLT and hostility.data$Method
##
##   1       2
## 2 0.0005  -
## 3 2.2e-05 0.0175
##
## P value adjustment method: none

pairwise.t.test(hostility.data$HLT, hostility.data$Method,
                pool.sd=FALSE, p.adjust.method='bonferroni')

##
##  Pairwise comparisons using t tests with non-pooled SD
##
## data:  hostility.data$HLT and hostility.data$Method
##
##   1       2
## 2 0.0015  -
## 3 6.7e-05 0.0525
##
## P value adjustment method: bonferroni
```

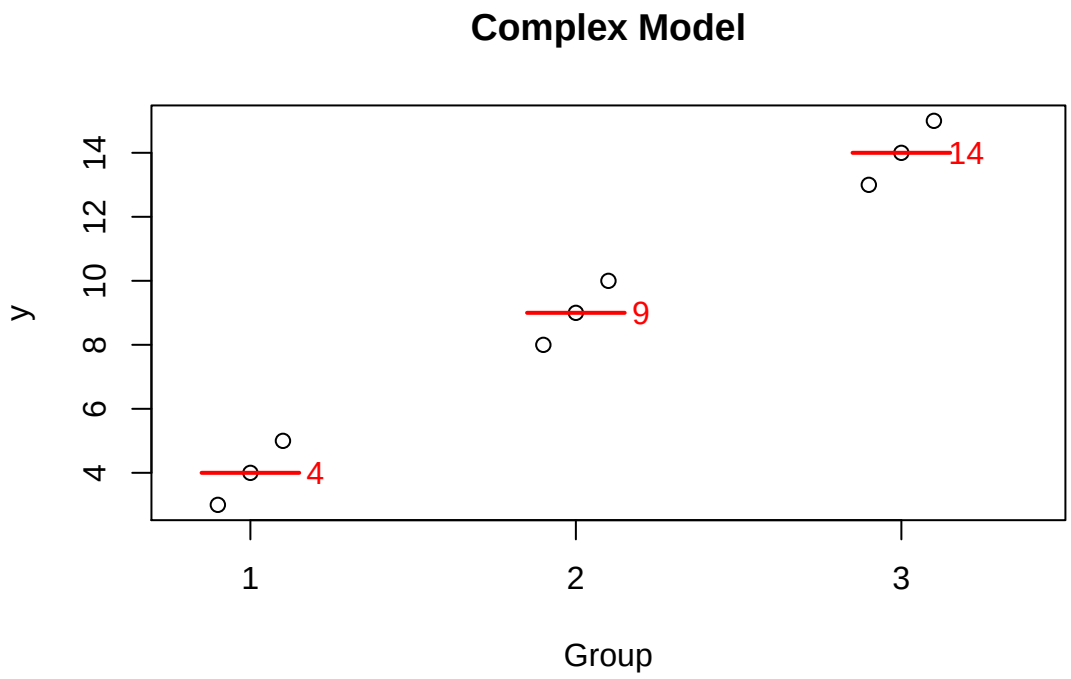
8.5 Different Model Representations

8.5.1 Theory

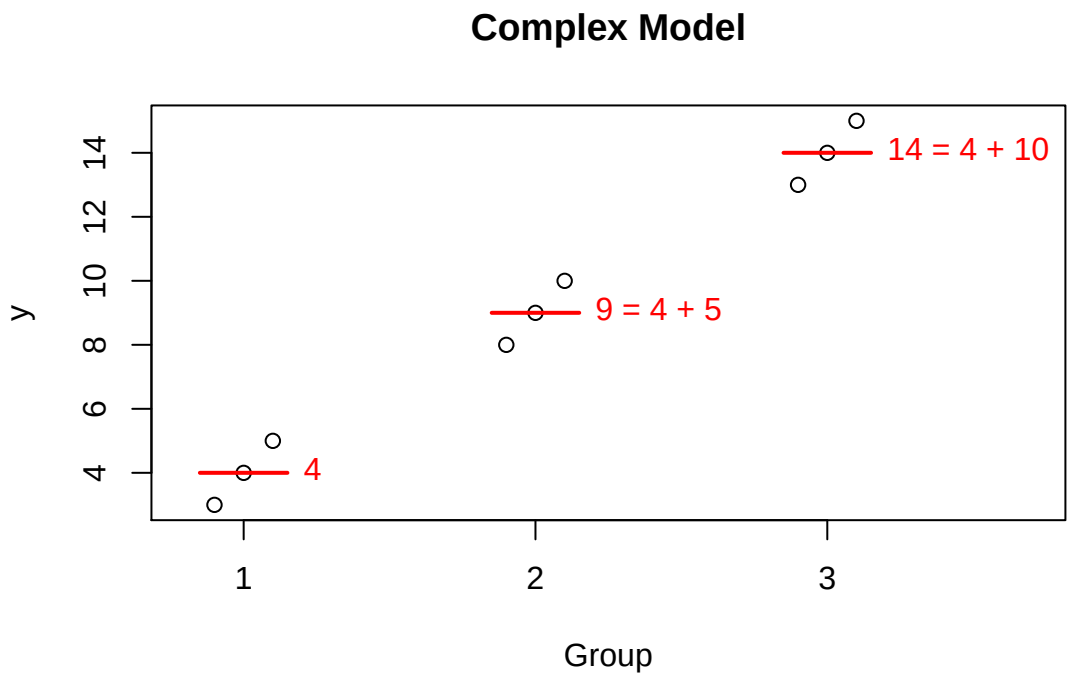
We started with what I will call the "cell means model"

$$Y_{ij} = \mu_i + \epsilon_{ij} \quad \text{where} \quad \epsilon_{ij} \overset{iid}{\sim} N\left(0, \sigma^2\right)$$

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 useful 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 (from the group mean to the 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

```
y <- c(3,4,5, 8,9,10, 13,14,15)
grp <- factor(c(1,1,1,2,2,2,3,3,3))
fake.data <- data.frame(y=y, grp=grp)
# 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.data)
coef(model)

## Method1 Method2 Method3
##   86.75   75.57   71.00

confint(model)

##         2.5 % 97.5 %
## Method1 83.59  89.91
## Method2 72.20  78.95
## Method3 68.02  73.98
```

We can use the intercept model by removing `-1` term from the formula.

```
model <- aov(HLT ~ Method, data=hostility.data)
coef(model)

## (Intercept)     Method2     Method3
##       86.75      -11.18      -15.75

confint(model)

##              2.5 %  97.5 %
## (Intercept)  83.59  89.907
## Method2     -15.80  -6.557
## Method3     -20.09 -11.411
```

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.

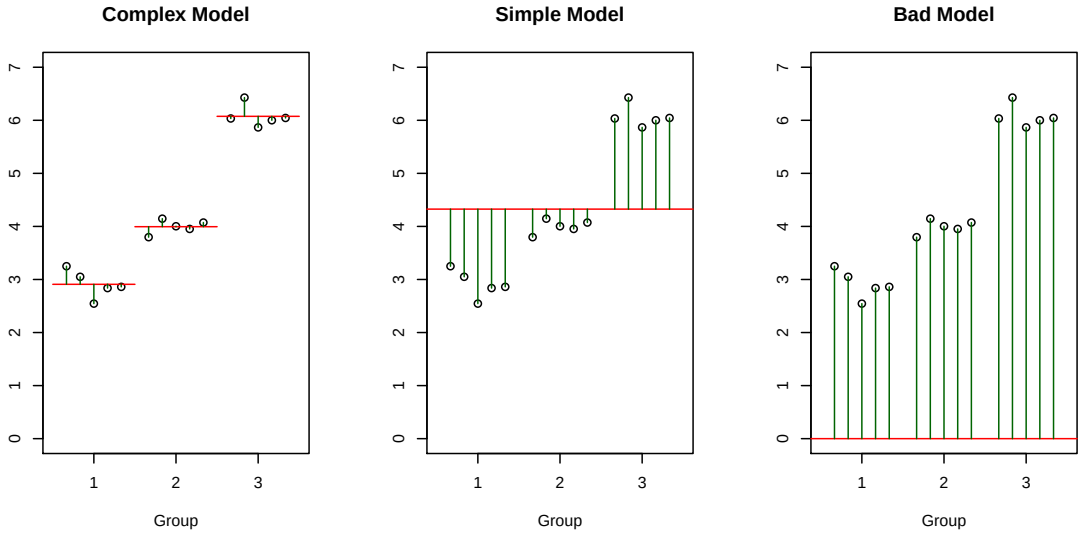
Neither representation is more powerful because on a very deep mathematical level, they are exactly the same model. Superficially though, one representation might be more convenient than the other in a given situation.

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.data)
cell.representation   <- aov(HLT ~ Method -1, data= hostility.data)
#
#
# 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   1091     545    29.6 7.8e-07 ***
## Residuals 21    387      18
## ---
## 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 <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.

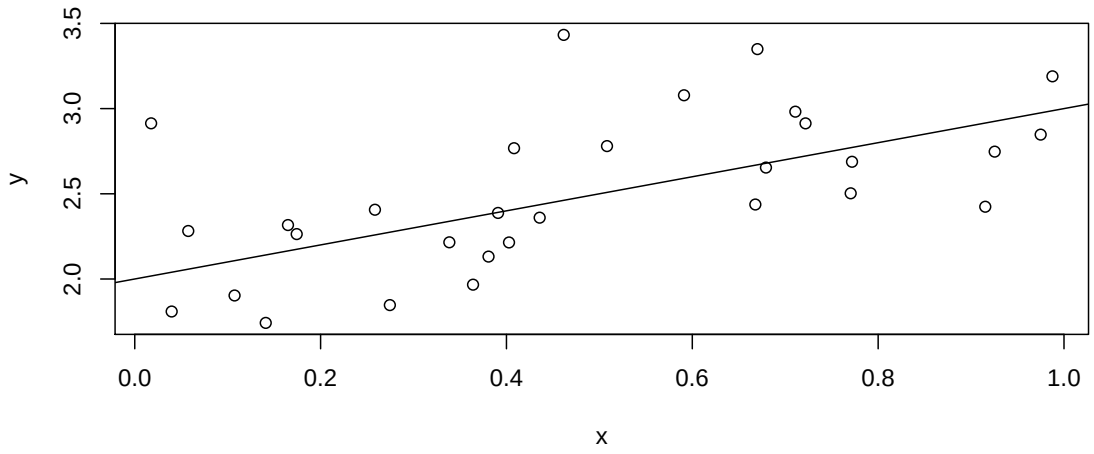
```
simple                <- aov(HLT ~ 1, data=hostility.data)

# 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 1478
## 2     21  387  2      1091 29.6 7.8e-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.

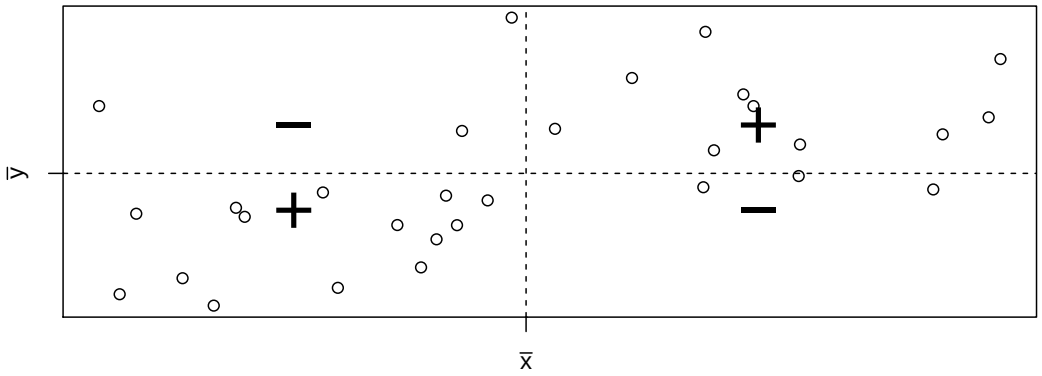


9.1 Pearson's Correlation Coefficient

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

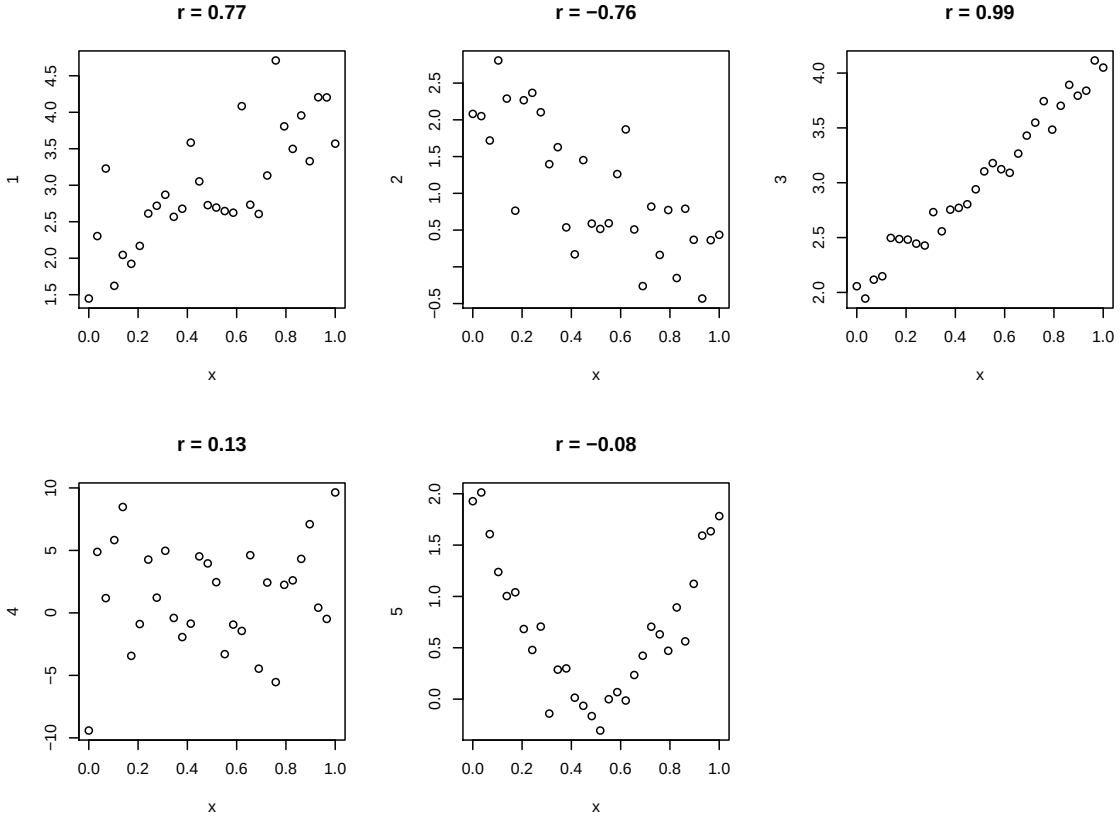
$$r = \frac{\sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)}{n-1}$$

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

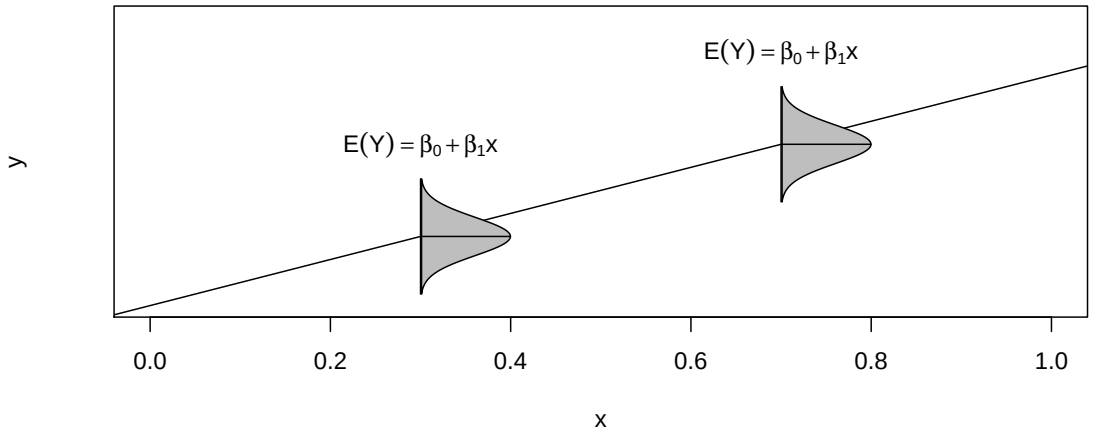
To scatterplot data that looks linear we often want to fit the model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad \text{where } \epsilon_i \overset{iid}{\sim} N\left(0, \sigma^2\right)$$

where β_0 is the y-intercept term and β_1 is the slope term. 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.



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 we will use the maximum likelihood estimates (MLE) of β_0 and β_1. Since we are assuming that the error terms are normally distributed, finding the MLEs is equivalent to finding 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

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

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:

$$\hat{\beta}_i \pm t_{n-2}^{1-\alpha/2} StdErr\left(\hat{\beta}_i\right)$$

Example. 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.*

```
setosa <- iris[which(iris$Species == 'setosa'),];
plot(setosa$Sepal.Length, setosa$Sepal.Width,
        xlab='Sepal Length', ylab='Sepal Width', main='Setosa');
```

```
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)
r

## [1] 0.7425

b0

## [1] -0.5694

b1

## [1] 0.7985

yhat <- b0 + b1*x
resid <- y - yhat
SSE <- sum( resid^2 )
s2 <- SSE/(n-2)
s2

## [1] 0.06581

Sxx <- sum( (x-mean(x))^2 )
stderr.b0 <- sqrt(s2) * sqrt( 1/n + mean(x)^2 / Sxx)
stderr.b1 <- sqrt(s2) * sqrt(1 / Sxx )
stderr.b0

## [1] 0.5217

stderr.b1

## [1] 0.104

t.star <- qt(.975, df=n-2)
c(b0-t.star*stderr.b0, b0+t.star*stderr.b0)

## [1] -1.6184  0.4795

c(b1-t.star*stderr.b1, b1+t.star*stderr.b1)

## [1] 0.5895 1.0076
```

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] 0.7425

model <- lm(Sepal.Width ~ Sepal.Length, data=setosa)
coef(model)

##  (Intercept) Sepal.Length
##      -0.5694       0.7985

confint(model)

##                2.5 % 97.5 %
## (Intercept)  -1.6184 0.4795
## Sepal.Length  0.5895 1.0076
```

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
## -0.7239 -0.1827 -0.0031  0.1574  0.5171
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -0.569      0.522   -1.09     0.28
## Sepal.Length    0.799      0.104    7.68  6.7e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.257 on 48 degrees of freedom
## Multiple R-squared:  0.551,Adjusted R-squared:  0.542
## F-statistic:   59 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   3.88    3.88      59 6.7e-10 ***
## Residuals    48   3.16    0.07
## ---
## 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,

$$\begin{aligned} \hat{Var}\left(\hat{\beta}_0 + \hat{\beta}_1 x_{new} + \epsilon\right) &= \hat{Var}\left(\hat{\beta}_0 + \hat{\beta}_1 x_{new}\right) + \hat{Var}\left(\epsilon\right) \\ &= \hat{\sigma}^2 \left(\frac{1}{n} + \frac{\left(x_{new} - \bar{x}\right)^2}{S_{xx}}\right) + \hat{\sigma}^2 \end{aligned}$$

and the $StdErr\left(\right)$ of a new observation will be

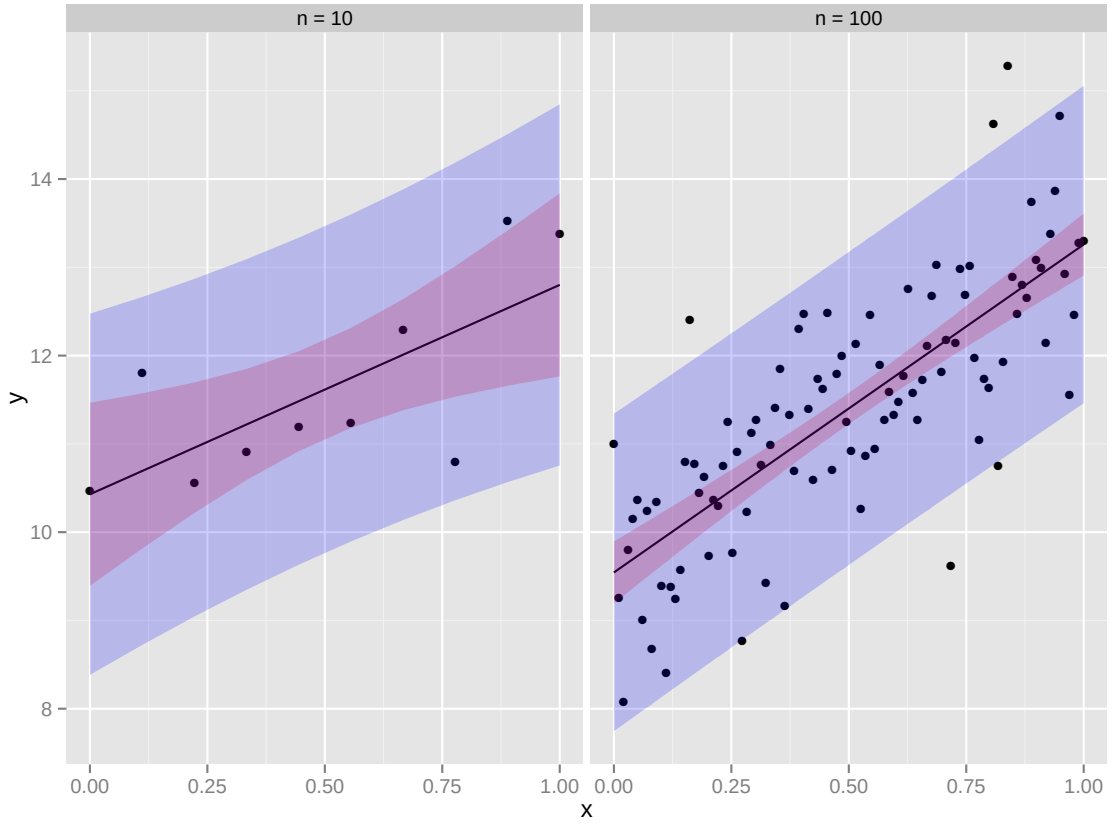
$$StdErr\left(\hat{y}_{new}\right) = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{\left(x_{new} - \bar{x}\right)^2}{S_{xx}}}$$

So the prediction interval for a new observation will be:

$$\left(\hat{\beta}_0 + \hat{\beta}_1 x_{new}\right) \pm t_{n-2}^{1-\alpha/2} \;\; \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{\left(x_{new} - \bar{x}\right)^2}{S_{xx}}}$$

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.

```
## Warning:  predictions on current data refer to _future_ responses
```

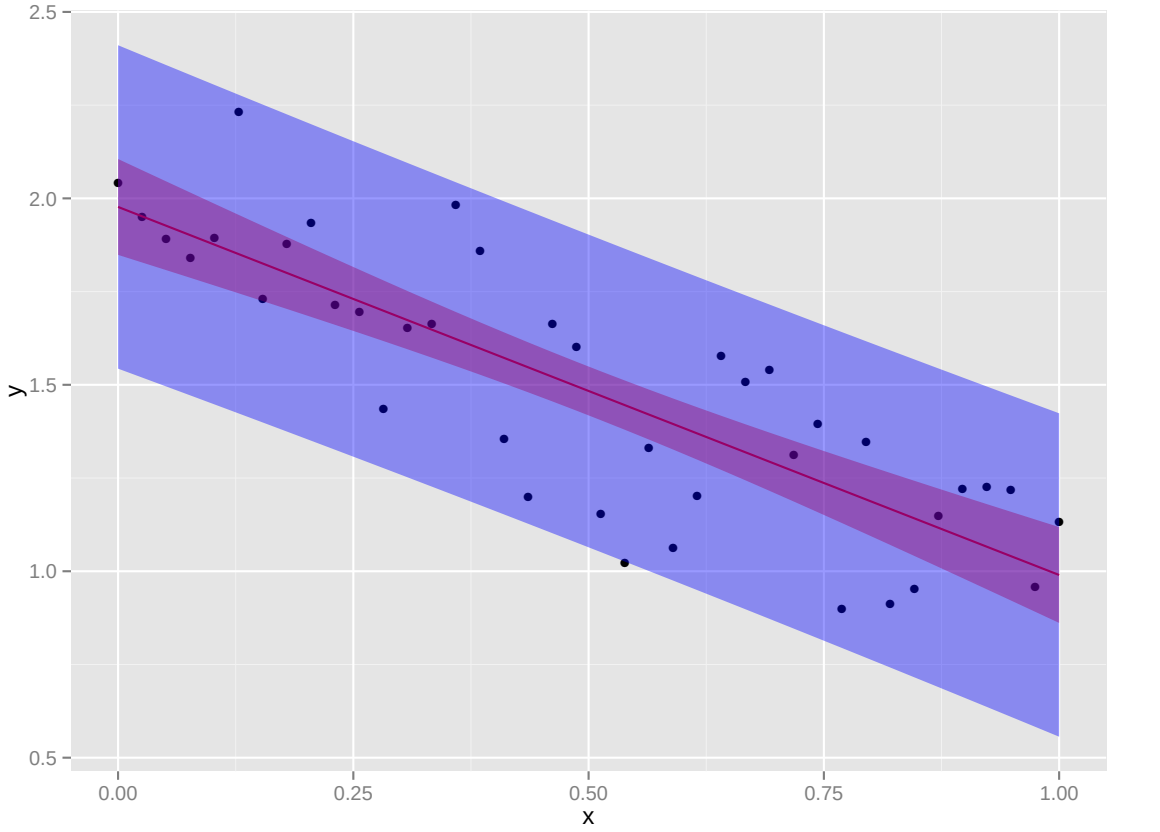


In general, you will not want to calculate the confidence intervals and prediction intervals by hand. Fortunately R makes it easy to calculate the intervals. The function `predict()` will calculate the point estimates along with confidence and prediction intervals. The function requires the `lm()` output along with an optional data frame (if you want to predict values not in the original data).

```
n <- 40
x <- seq(0,1,length=n)
y <- 2 - 1*x + rnorm(n, sd=.2)
model <- lm(y~x)
predict(model, interval="confidence")[1:5, ]  # Only the first 5 rows...

##     fit   lwr   upr
## 1 1.977 1.848 2.106
## 2 1.952 1.828 2.075
## 3 1.926 1.807 2.045
## 4 1.901 1.787 2.015
## 5 1.876 1.766 1.985

predict(model, interval="prediction", newdata=data.frame(x=0.75))

##     fit    lwr  upr
## 1 1.237 0.8138 1.66
```

We can create a nice graph of the regression line and associated confidence and prediction regions using the following code in R:

```
library(ggplot2)
y.hat <- fitted(model)
conf.region <- predict(model, interval='confidence')
pred.region <- predict(model, interval='prediction')

## Warning:  predictions on current data refer to _future_ responses

data <- data.frame(x=x, y=y,
      y.fitted = y.hat,
      y.conf.lwr = conf.region[,2],
      y.conf.upr = conf.region[,3],
      y.pred.lwr = pred.region[,2],
      y.pred.upr = pred.region[,3])
ggplot(data) +
   geom_point( aes(x=x, y=y) ) +
   geom_line(  aes(x=x, y=y.fitted), col='red' ) +
   geom_ribbon( aes(x=x, ymin=y.conf.lwr, ymax=y.conf.upr), fill='red',  alpha=.4) +
   geom_ribbon( aes(x=x, ymin=y.pred.lwr, ymax=y.pred.upr), fill='blue', alpha=.4)
```

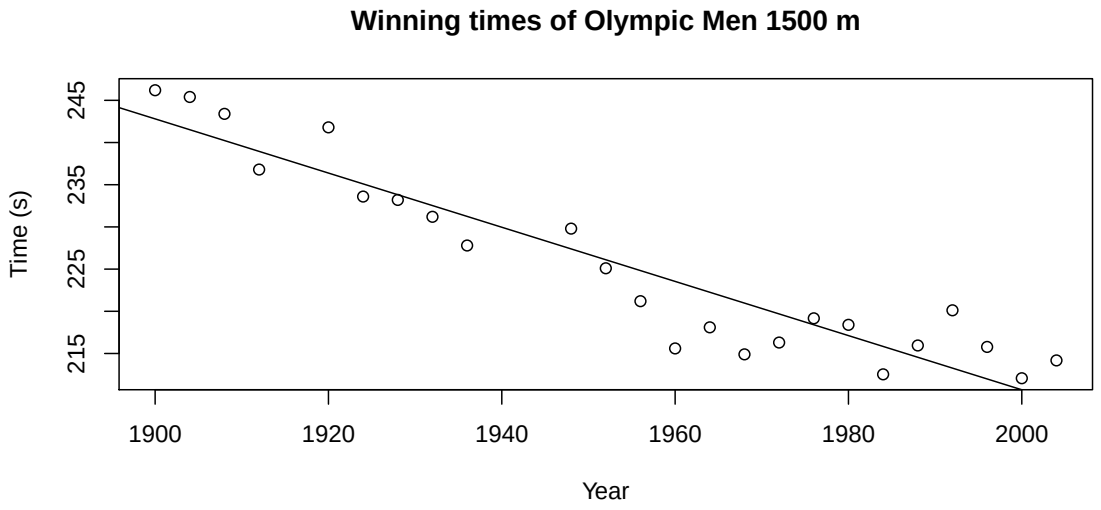


It is worth noting that these confidence intervals are all *point-wise* confidence intervals. If I want to calculate confidence or prediction intervals for a large number of x_{new} values, then I have to deal with the multiple comparisons issue. Fortunately this is easy to do in the simple linear regression 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 is the winning time of the men's 1500 meter Olympic race.

```
library(HSAUR2);
small <- men1500m[-1,]  # Remove the 1896 Olympics
plot(small$year, small$time, xlab='Year', ylab='Time (s)',
        main='Winning times of Olympic Men 1500 m')
model <- lm( time ~ year, data=small )
abline(coef(model))
```



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
## 1 208.1 199.4 216.9
## 2 206.8 198.0 215.6
```

We can compare the first interval with the time actually recorded by the winner of the men's 1500m in Beijing 2008, Rashid Ramzi from Brunei, who won the event in 212.94 seconds. Clearly the linear relationship must eventually change and therefore our regression could not possibly predict the winning time of the 3112 race.

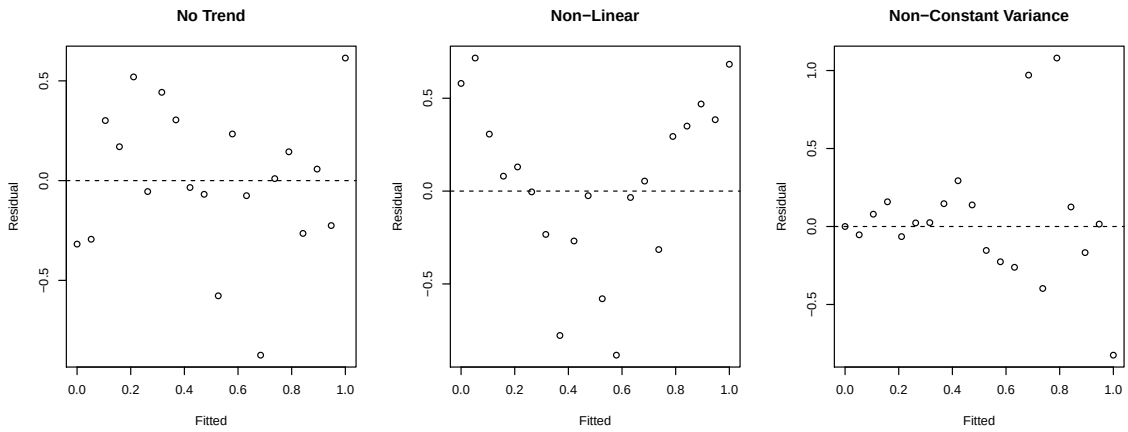
```
predict(model, newdata=data.frame(year=c(3112)), interval="prediction")

##      fit    lwr    upr
## 1 -146.3 -206.8 -85.82
```

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

$$e_i = y_i - \hat{y}_i$$

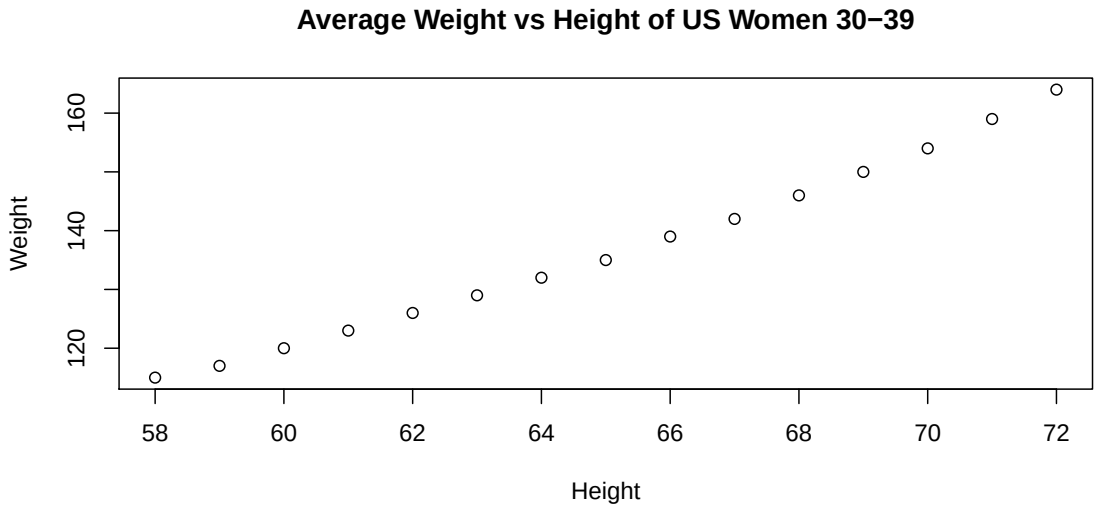
for normality using a QQ-plot as we did in Anova. To address the constant variance and linearity assumptions we will look at scatterplots of the residuals vs the fitted values $\hat{y}_i$. For the regression to be valid, we want the scatterplot to show no discernible trend. There are two patterns that commonly show up that indicate a violation of the regression assumptions.



Example. 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).

```
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 ...

plot(women$height, women$weight, xlab='Height', ylab='Weight', main='Average Weight vs Height of US
```

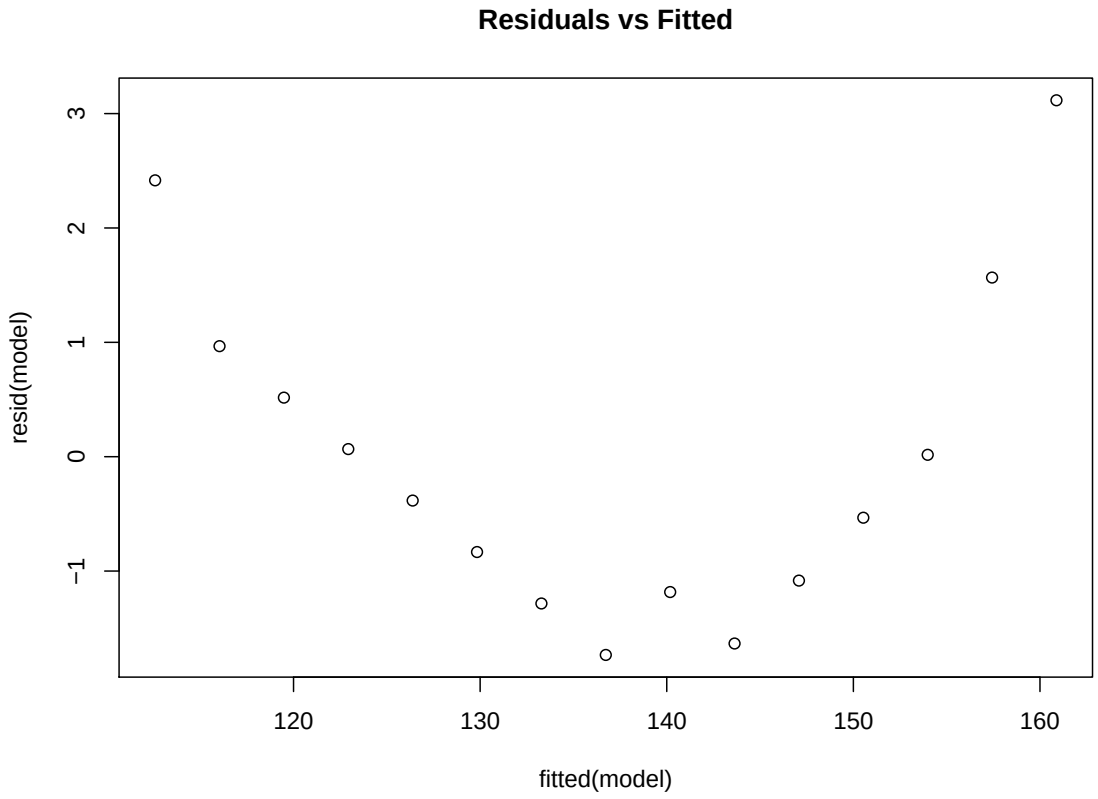


Next we will fit the linear model and look at the regression summary table.

```
model <- lm(weight ~ height, data=women)
summary(model)

##
## Call:
## lm(formula = weight ~ height, data = women)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.733 -1.133 -0.383  0.742  3.117
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -87.5167     5.9369   -14.7  1.7e-09 ***
## height        3.4500     0.0911    37.9  1.1e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.53 on 13 degrees of freedom
## Multiple R-squared:  0.991,Adjusted R-squared:  0.99
## F-statistic: 1.43e+03 on 1 and 13 DF,  p-value: 1.09e-14
```

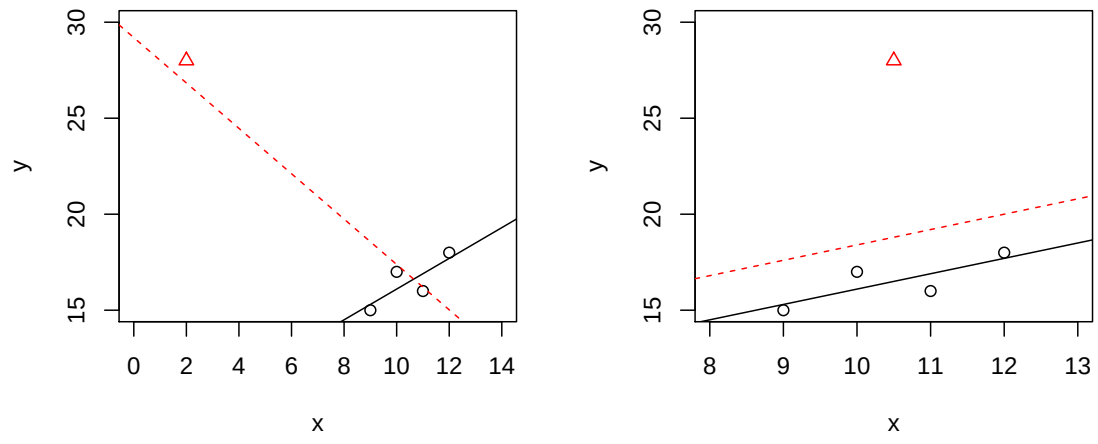
And then look at the residuals vs fitted values and immediately conclude that the linearity assumption is violated.

```
plot(x=fitted(model), y=resid(model), main='Residuals vs Fitted')
```



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.



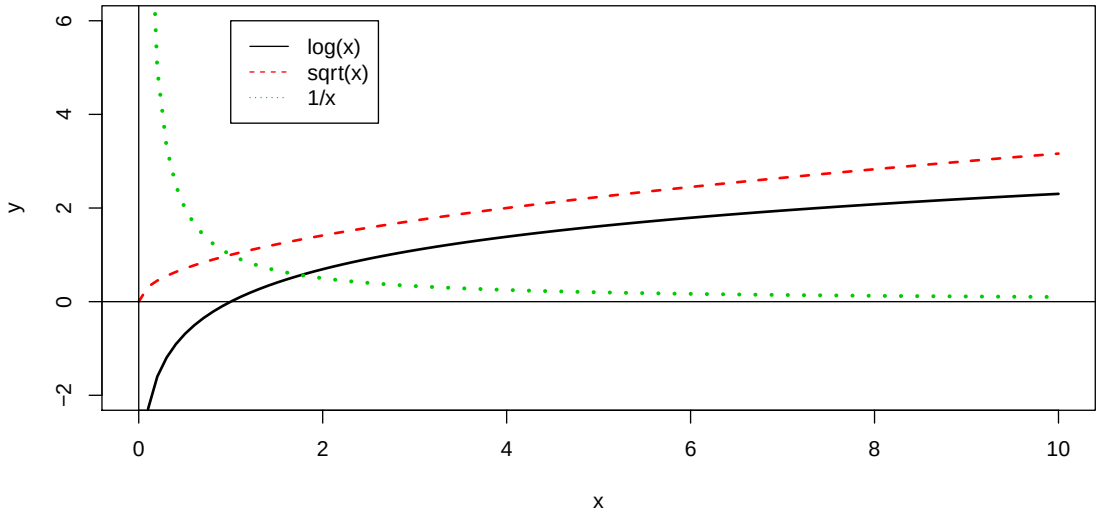
The question of what to do with influential points is not easy to answer. Sometimes these are data points that are a result of lab technician error and should be removed. Sometimes they are the

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.

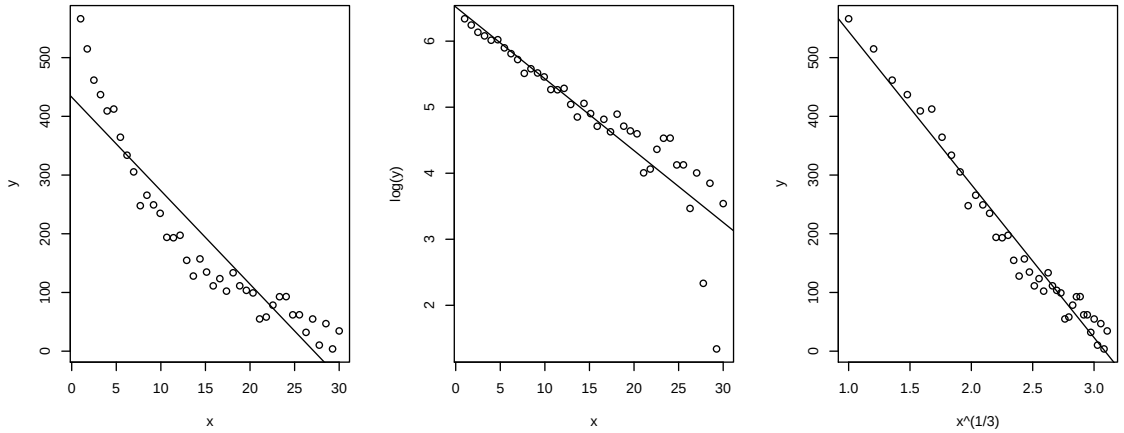
One solution is to run the analysis both with and without the influential point and see how much it affects your inferences.

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.



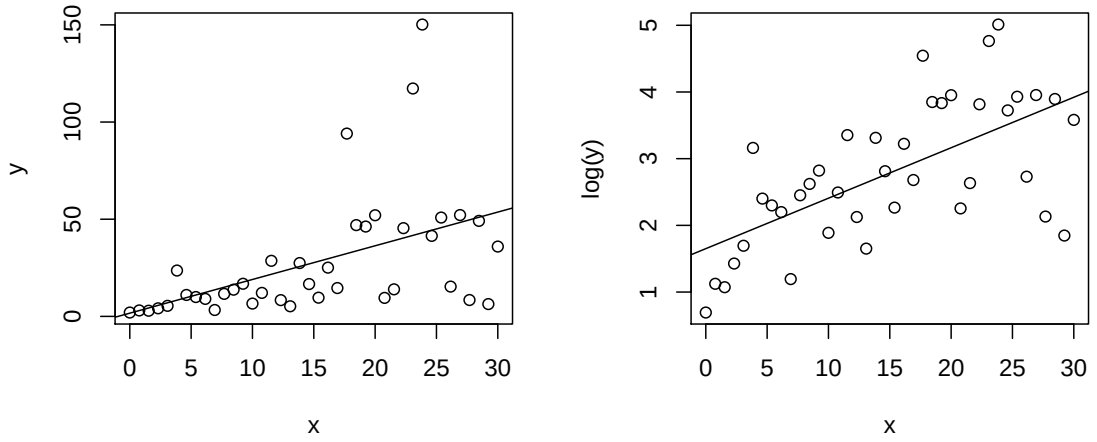
Unfortunately it is not always obvious what transformation is most appropriate. The Box-Cox

family of transformations for the y-variable is

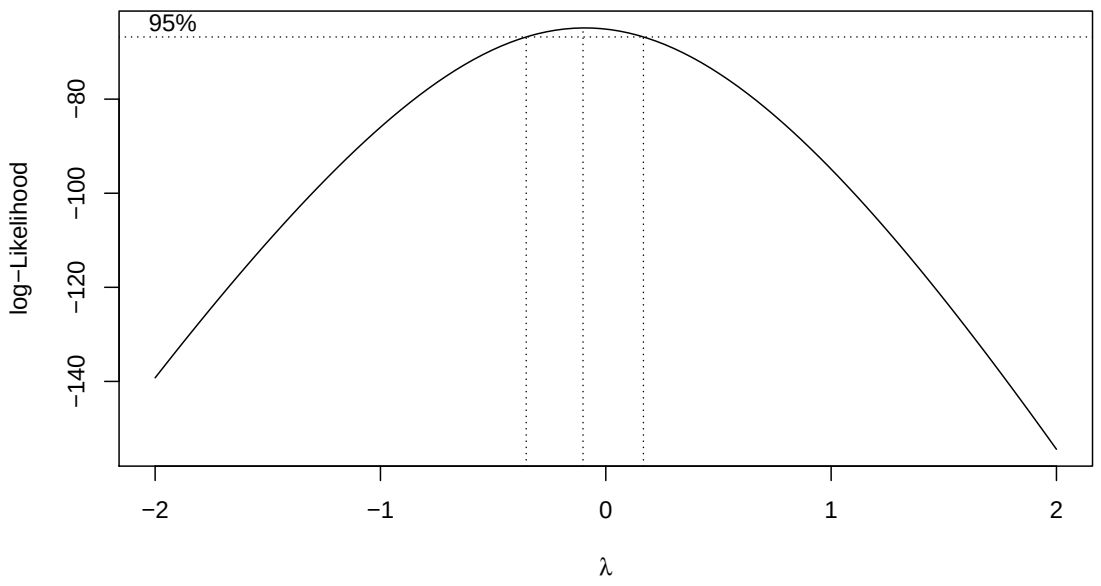
$$f(y\,|\,\lambda) \quad = \quad \begin{cases} y^\lambda & \text{if } \lambda \neq 0 \\ \log y & \text{if } \lambda = 0 \end{cases}$$

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

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.

10.1 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.

10.1.1 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] 0.3438
```

As usual, if we had been interested in a two-sided alternative, we would multiply the p-value by two.

10.1.2 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 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

[3] In the case one sample case, calculate and order the values $z_i = x_i - M_0$.

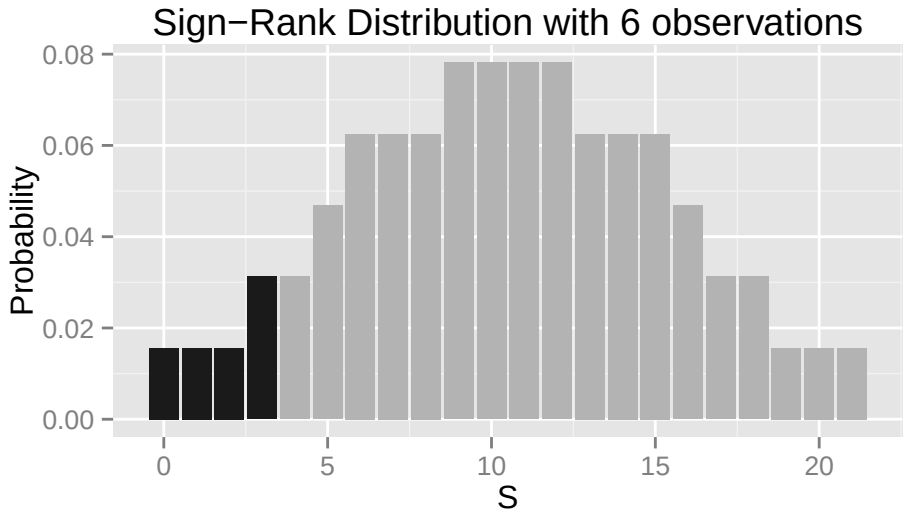
[4] For the alternative $M_1 - M_2 > 0$ then $S = W_-$. For the alternative $M_1 - M_2 < 0$ use $S = W_+$

and then calculate

$$\begin{aligned} W_- &= 6 + 5 + 3.5 + 3.5 = 18 \\ W_+ &= 1.5 + 1.5 = 3 \end{aligned}$$

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] 0.07813

# greater than or equal to 18
1 - psignrank( 17, 6 )

## [1] 0.07812
```

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:  cannot compute exact p-value with ties
## Warning:  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'`.

10.2 Alternatives to the two sample t-test

10.2.1 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]

10.2.2 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:  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.