Introduction to Statistics for Researchers

Derek L. Sonderegger

February 26, 2016

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

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

Statistical software has progressed by leaps and bounds over the last decades. Scientists need access to reliable software that is flexible enough to handle new problems, with minimal headaches. R has become a widely used, and extremely robust Open Source platform for statistical computing and most new methodologies will appear in R before being incorporated into commercial software. Second, data exploration is the first step of any analysis and a user friendly yet powerful mechanism for graphing is a critical component in a researchers toolbox. R succeeds in this area as R has the most flexible graphing library of any statistical software I know of and the basic plots can created quickly and easily. The only downside is that there is a substantial learning curve to learning a scripting language, particularly for students without any programming background.

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

As these notes are in a continual state of being re-written, I endeavor to keep the latest version available on the GitHub repository for this book at https://github.com/dereksonderegger/STA_570_Book/raw/master/Stat_570.pdf. In general, I recommend printing the chapter we are currently covering in class. If you wish to submit a bug report, or submit a patch, feel free to log an issue on the GitHub site or to fix it and submit a pull request. If you wish to use these notes for your own class, feel free to do so, but please acknowledge the source.

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

Contents

The scientific method requires us to make observations about the world and then use those observations to make reasonable predictions. However the way we make those observations can have a profound effect on how well understand the phenomena or how well we predict some event.

During the 1936 US presidential election between Franklin Roosevelt (D) and Alfred Landon (R), the magazine *The Literary Digest* included a card that asked its readers who they were to vote for and to send it back to the magazine, which would tabulate and report the predicted winner. This scheme correctly predicted the 1920, 1924, 1928, and 1932 races and so many people respected the forecast. The supposed strength of the prediction was how many people responded (2.4 million!). However, the readers of *The Literary Digest* were typically more affluent and could afford a magazine subscription during the Great Depression. As a result, the magazine performed a highly biased survey and predicted that Landon would win. The actual outcome was that Roosevelt won 61% of the vote.

Another person also interested in predicting the 1936 election was the statistician George Gallup. He had a much smaller sample, (with 50,000 respondents) and correctly predicted the outcome. This was a clear demonstration of the power of a well selected random sample compared to a massive set of data collected in a biased fashion.

The goals of statistics can be

1. Collection of data

   (a) Sampling Design: Observational studies rely on the assumption that your sample is representative of the population of interest. Unfortunately, it is often difficult to randomly select individuals in a non-biased way and this branch of statistics focuses on how to leverage known population level data with sampled data to account for bias.

   (b) Experimental Design: The scientific gold standard is to perform an experiment where the factors of interest are manipulated, but how the manipulations are done to maximize the amount of information obtained from the experiment is a classic issue in statistics.

2. Process Modeling - Often we have some mathematical model that is a function of some population level parameters that we think represents some process of interest. We can use sample statistics calculated from our data to estimate those parameters of interest in the model.

   (a) Because our sample statistics will vary from sample to sample, we want to understand how much faith should we have in our estimate.

   (b) We often want confidence intervals for those parameters. For example, we might have a sample where 54% of the respondents support Bernie Sanders for president, but I don't expect that to be *exactly* the population level percent. Instead we report that 52-56% of the population of likely voters supports Bernie Sanders.

   (c) We might ask if some particular value for a parameter is plausible. For example, could the association between person's income level and their probability of voting republican be zero?

3. Predictions - Once we have a process model (even if it is extremely complicated and not easily interpretable), we will often want to make predictions about future or unobserved observations.

   (a) We will want to consider how well we predict new observations, but, by definition, those are unavailable.

   (b) Quality of prediction can be used to update the process modeling steps.

The course covered by these notes will discuss some of the data collection step, but will focus primarily on the process modeling step.

Chapter 1

Summary Statistics and Graphing

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

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

```
library(mosaicData) # library of datasets we'll use
library(ggplot2)    # graphing functions
library(dplyr)      # data summary tools
head(TenMileRace)   # examine the first few rows of the data
```

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

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

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

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

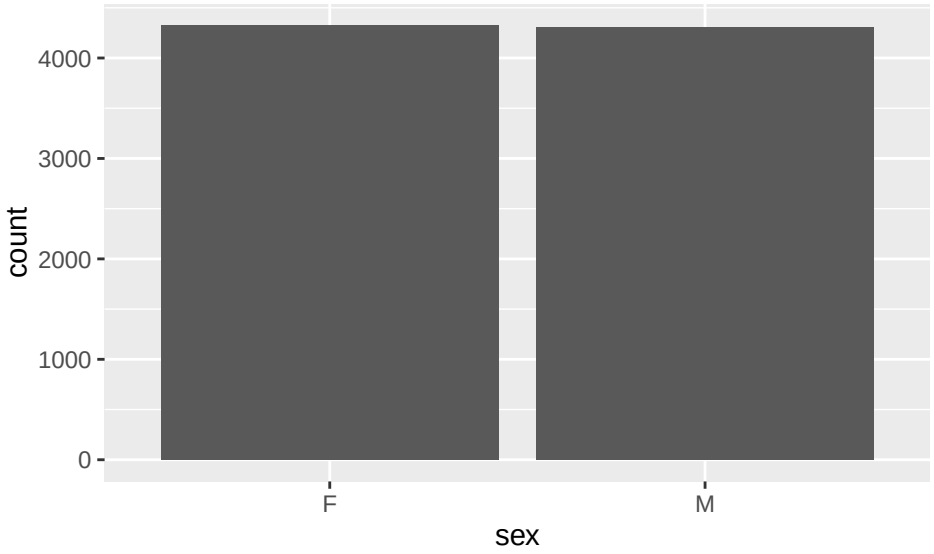
[1] Actually this isn't true as both gender and sex are far more complex. However from a statistical point of view it is often useful to simplify our model of the world. George Box famously said, "All models are wrong, but some are useful."

1.1 Graphical summaries of data

1.1.1 Univariate - Categorical

If we have univariate data about a number of groups, often the best way to display it is using barplots. They have the advantage over pie-charts that groups are easily compared. [2]

```
ggplot(TenMileRace, aes(x=sex)) + geom_bar()
```

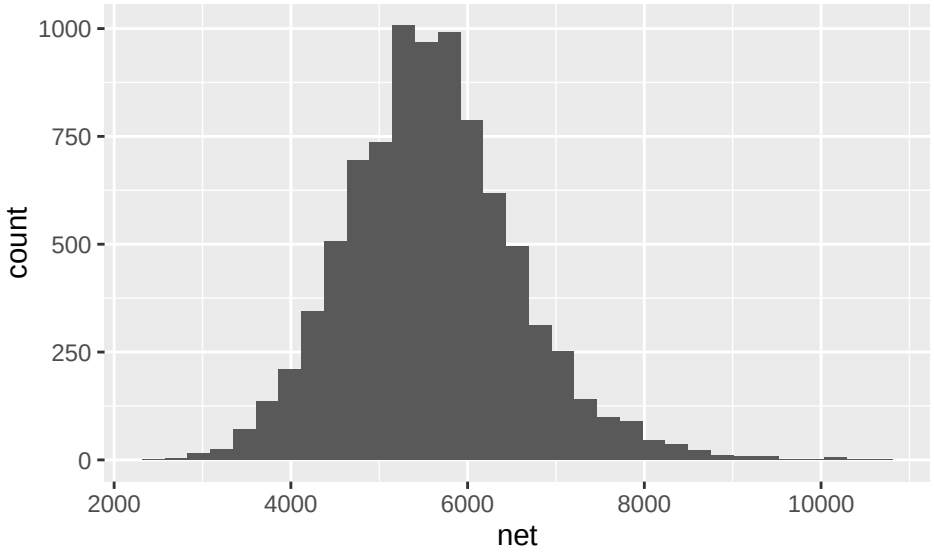


1.1.2 Univariate - Continuous

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

[2] This is an example of a poorly labeled covariate, this really ought to be gender.

```
ggplot(TenMileRace, aes(x=net)) + geom_histogram()

## `stat_bin()` using `bins = 30`.  Pick better value with `binwidth`.
```



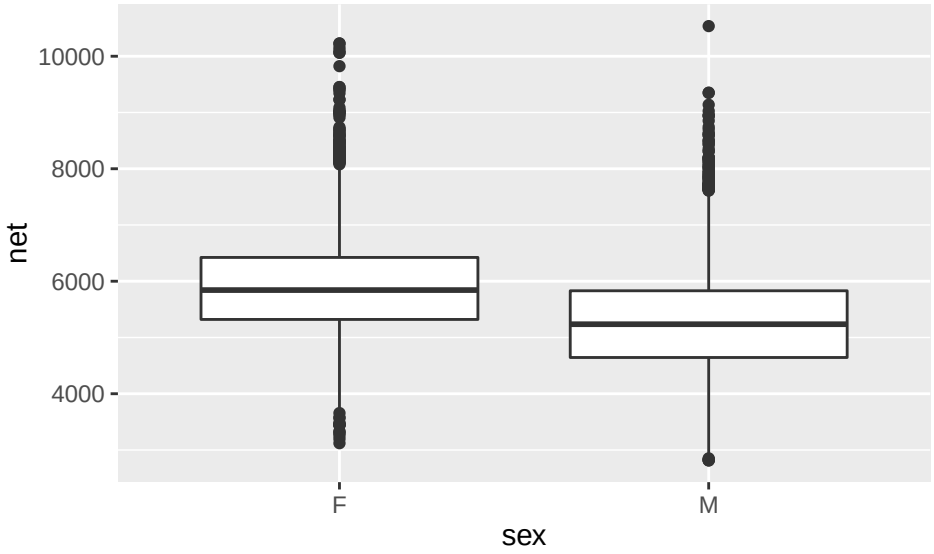
Often when a histogram is presented, the y-axis is labeled as "frequency" or "count" which is the number of observations that fall within a particular bin. However, it is often desirable to scale the y-axis so that if we were to sum up the area (height * width) then the total area would sum to 1. The rescaling that accomplishes this is

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

1.1.3 Bivariate - Categorical vs Continuous

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

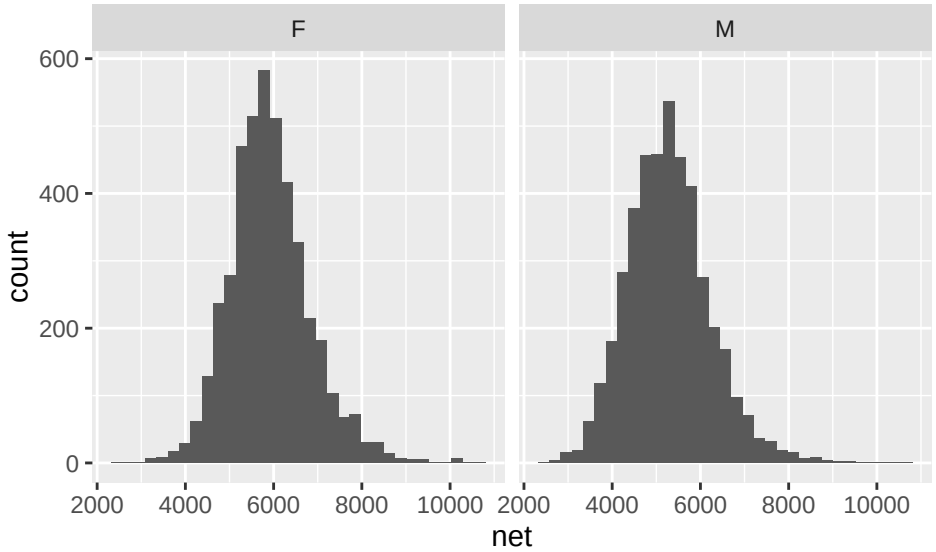
```
ggplot(TenMileRace, aes(x=sex, y=net)) + geom_boxplot()
```



In this graph, the edges of the box are defined by the 25% and 75% quantiles. That is to say, 25% of the data is to the below of the box, 50% of the data is in the box, and the final 25% of the data is to the above of the box. The dots are data points that traditionally considered outliers.[3]

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

```
ggplot(TenMileRace, aes(x=net)) +
  geom_histogram() +
  facet_grid( . ~ sex )  # side-by-side plots based on sex

## `stat_bin()` using `bins = 30`.  Pick better value with `binwidth`.
```

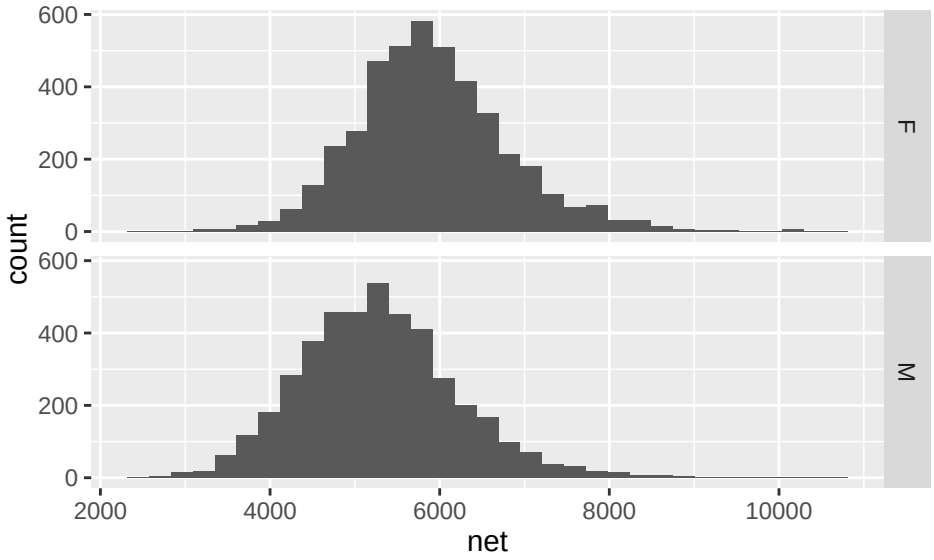


Orientation of graphs can certainly matter. In this case, it makes sense to *stack* the two graphs

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

to facilitate comparisons.

```
ggplot(TenMileRace, aes(x=net)) +
  geom_histogram() +
  facet_grid( sex ~ . )  # side-by-side plots based on sex

## `stat_bin()` using `bins = 30`.  Pick better value with `binwidth`.
```



1.1.4 Bivariate - Continuous vs Continuous

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

```
ggplot(TenMileRace, aes(x=age, y=net, color=sex)) +
  geom_point()
```

1.2 Measures of Centrality

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

Mean

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

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

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

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

This can easily be calculated in R by using the function `mean()`. We first extract the column we are interested in using the notation: `DataSet$ColumnName` where the `$` signifies grabbing the column.

```
mean( TenMileRace$net ) # Simplest way of doing this calculation

## [1] 5599.065

TenMileRace %>% summarise( mean(net) )  # using the dplyr package

##   mean(net)
## 1  5599.065
```

Median

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

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

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

```
TenMileRace %>% summarise( median(net) )

##   median(net)
## 1        5555
```

Mode

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

Examples

- If my father were to become bored with retirement and enroll in my STA 570 course, how would that affect the mean and median age of my 570 students?

  - The mean would move much more than the median. Suppose the class has 5 people right now, ages 21, 22, 23, 23, 24 and therefore the median is 23. When my father joins, the ages will be 21, 22, 23, 23, 24, 72 and the median will remain 23. However, the mean would move because we add in such a large outlier. Whenever we are dealing with skewed data, the mean is pulled toward the outlying observations.

- In 2010, the median NFL player salary was $770,000 while the mean salary was $1.9 million. Why the difference?

  - Because salary data is *skewed* superstar players that make huge salaries (in excess of 20 million) while the minimum salary for a rookie is $375,000. Financial data often reflects a highly skewed distribution and the median is often a better measure of centrality in these cases.

1.3 Measures of Variation

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

Range

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

```
max( TenMileRace$net ) - min( TenMileRace$net )

## [1] 7722

TenMileRace %>% summarise( range = max(net) - min(net) )

##   range
## 1  7722
```

Inter-Quartile Range

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

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

```
quantile( TenMileRace$net ) # this works

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

# I can't do the following because the quantile() function spits out 5 values, not 1
# TenMileRace %>% summarise( quantile(net) )
```

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

```
TenMileRace %>% summarise( IQR(net) )

##   IQR(net)
## 1     1219
```

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

Variance

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

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

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

There are some really good theoretical reasons to chose the square option[4], so we square the deviates and then find the average deviate size (approximately) and call that the **sample variance**.

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

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

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

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

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

[4]First, squared terms are easier to deal with compared to absolute values, but more importantly, the spread of the normal distribution is parameterized via squared distances from the mean. Because the normal distribution is so important, we've chosen to define the sample variance so it matches up with the natural spread parameter of the normal distribution.

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()`.

```
ToyData <- data.frame( x=c(3,6,4,8,2) )
ToyData %>% summarise( s2 = var(x) )

##    s2
## 1 5.8
```

For the larger `TenMileRace` data set, the variance is just as easily calculated.

```
TenMileRace %>% summarise( s2 = var(net) )

##         s2
## 1 940233.5
```

Standard Deviation

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

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

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

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

```
TenMileRace %>% summarise( s = sd(net) )

##          s
## 1 969.6564
```

Coefficient of Variation

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

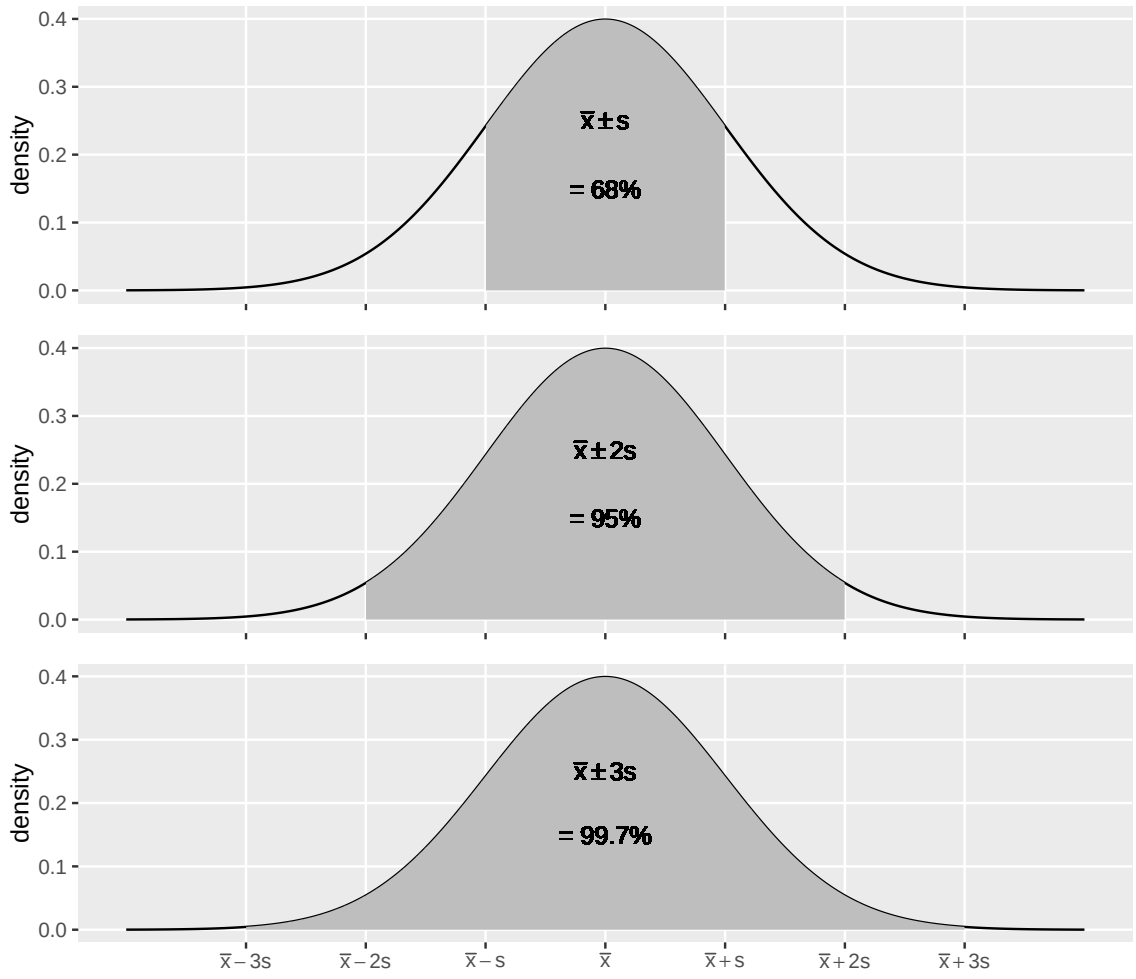
$$CV = \frac{s}{|\bar{x}|}$$

```
TenMileRace %>% summarise( s = sd(net),
                           xbar = mean(net),
                           cv = s / abs(xbar) )

##          s     xbar        cv
## 1 969.6564 5599.065 0.1731818

# For fun, lets calculate these same statistics but separated by sex...
TenMileRace %>%
  group_by(sex) %>%
  summarise( xbar = mean(net),
             s    = sd(net),
             cv   = s / abs(xbar) )

## Source: local data frame [2 x 4]
##
##      sex     xbar        s        cv
##   (fctr)    (dbl)    (dbl)     (dbl)
## 1      F 5916.398 902.1090 0.1524761
## 2      M 5280.702 929.9817 0.1761095
```

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%



1.4 Exercises

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

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

   (a) By hand, compute the mean and median for the 21 ratios, ignoring the type of feeder.

   (b) By hand, compute the mean and median separately for each type of feeder.

   (c) Using your results from parts (a) and (b), comment on the relative sensitivity of the mean and median to extreme values in a data set.

   (d) Which measure, mean or median, would you recommend as the most appropriate measure of the DDE to PCB level for both types of feeders? Explain your answer.

2. O&L 3.31. *Consumer Reports* in its June 1998 issue reports on the typical daily room rate at six luxury and nine budget hotels. The room rates are given in the following table.

Luxury	\$175	\$180	\$120	\$150	\$120	\$125			
Budget	\$50	\$50	\$49	\$45	\$36	\$45	\$50	\$50	\$40

   (a) By hand, compute the means and standard deviations of the room rates for each class of hotel.

   (b) Give a practical reason why luxury hotels might have higher variability than the budget hotels. *(Don't just say the standard deviation is higher because there is more spread in the data, but rather think about the Hotel Industry and why you might see greater price variability for upscale goods compared to budget items.)*

3. Use R to confirm your calculations in problem 1 (the pollution data). Show the code you used and the subsequent output. *It will often be convenient for me to give you code that generates a data frame instead of uploading an Excel file and having you read it in. The data can be generated using the following commands:*

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

# Print out some of the data to confirm what the column names are
head( PolutionRatios )

##   Ratio        Type
## 1 76.50 Terrestrial
## 2  6.03 Terrestrial
## 3  3.51 Terrestrial
## 4  9.96 Terrestrial
## 5  4.24 Terrestrial
## 6  7.74 Terrestrial
```

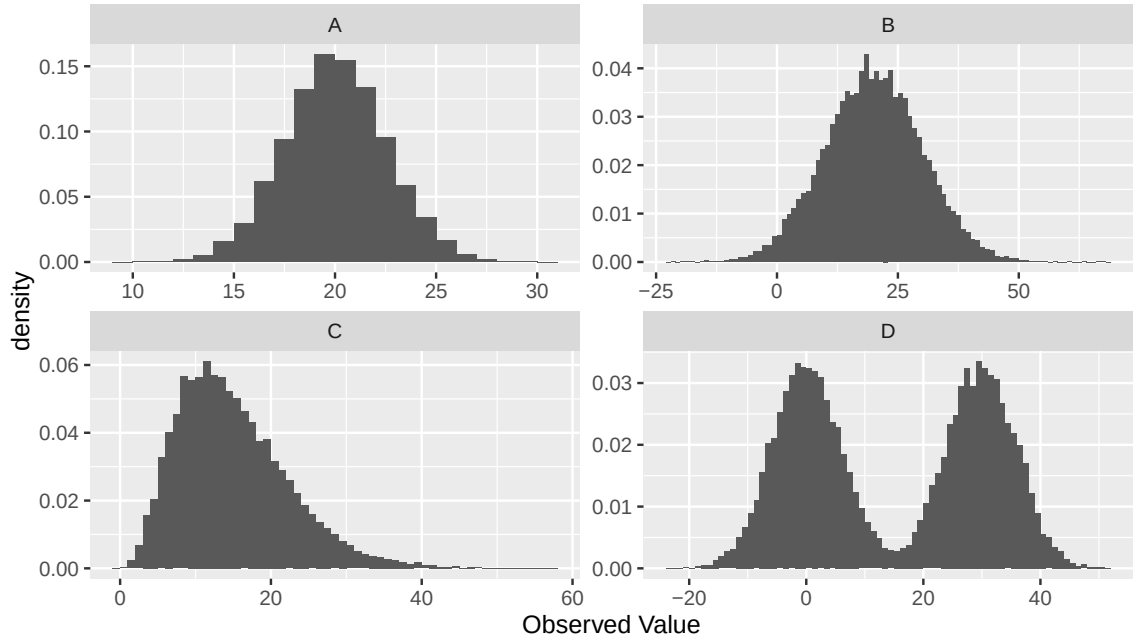
   Hint: for computing the means and medians for each type of feeder separately, there is a very convenient command

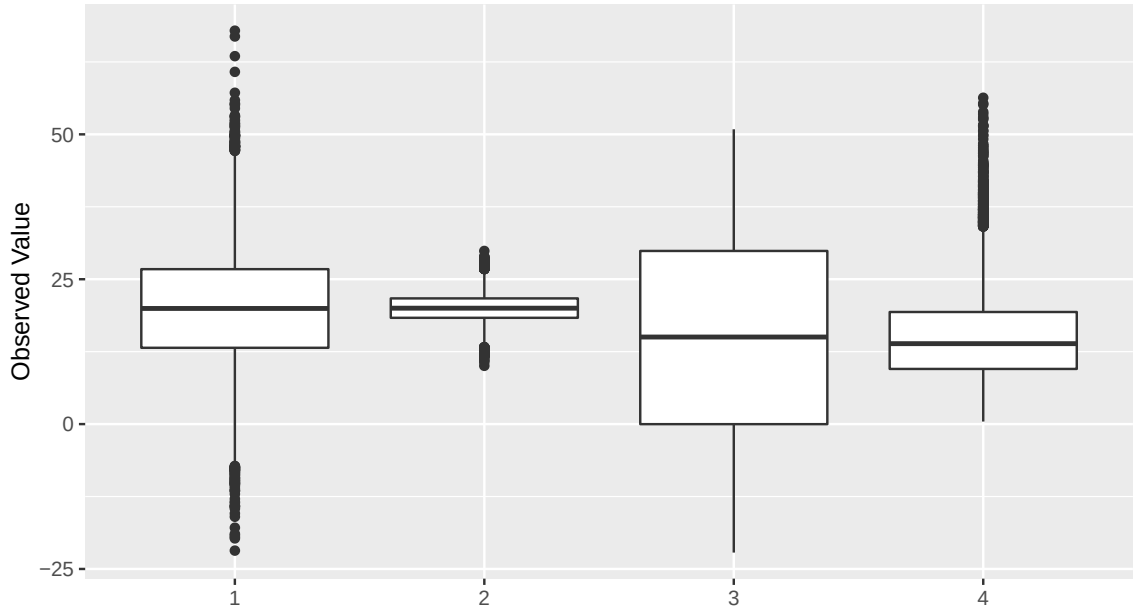
4. Use R to confirm your calculations in problem 2 (the hotel data). Show the code you used and the subsequent output. The data can be loaded into a data frame using the following commands Show the code you used and the subsequent output:

```
Hotels <- data.frame(
  Price = c(175, 180, 120, 150, 120, 125, 50, 50, 49, 45, 36, 45, 50, 50, 40),
  Type  = c( rep('Luxury',6),  rep('Budget', 9) ) )

# Print out some of the data to confirm what the column names are
head( Hotels )

##   Price   Type
## 1   175 Luxury
## 2   180 Luxury
## 3   120 Luxury
## 4   150 Luxury
## 5   120 Luxury
## 6   125 Luxury
```

5. For the hotel data, create side-by-side box-and-whisker plots to compare the prices.

6. Match the following histograms to the appropriate boxplot.





(a) Histogram A goes with boxplot _____

(b) Histogram B goes with boxplot _____

(c) Histogram C goes with boxplot _____

(d) Histogram D goes with boxplot _____

7. Twenty-five employees of a corporation have a mean salary of \$62,000 and the sample standard deviation of those salaries is \$15,000. If each employee receives a bonus of \$1,000, does the standard deviation of the salaries change? Explain your reasoning.

Chapter 2

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.

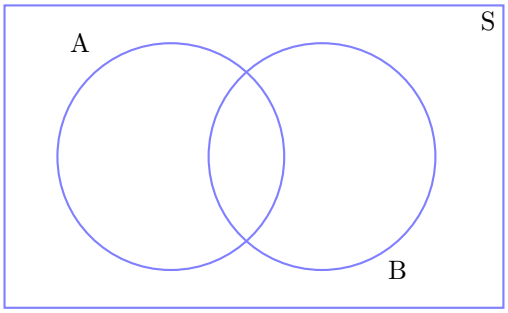
2.1 Introduction to Set Theory

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

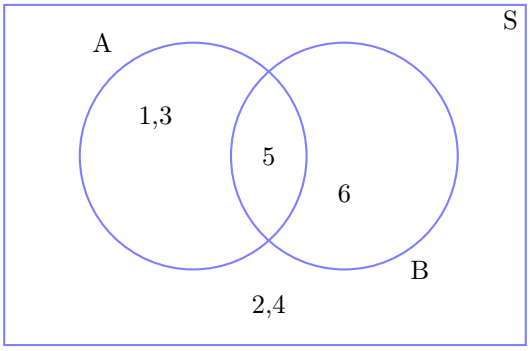
Events are properties of a particular outcome. For a coin flip, the event "Heads" would be the event that a heads was flipped. For the single roll of a six-sided die, a possible event might be that the result is even. For the NAU student, we might be interested in the event that the student is a biology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

2.1.1 Venn Diagrams

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



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



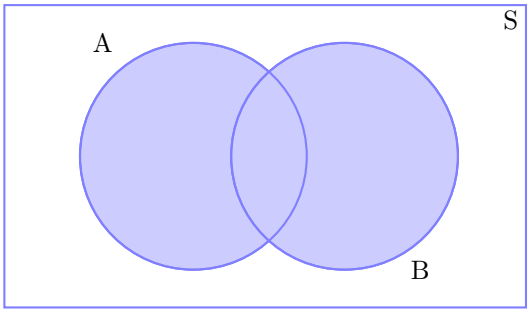
All of our possible events are present, and distributed amongst our possible events.

2.1.2 Composition of events

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

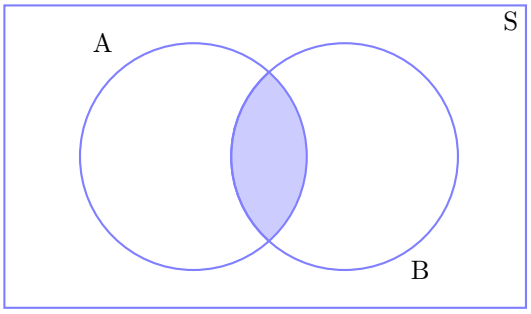
- Union: Denote the event that either A or B occurs as $A \cup B$.

$A \cup B$



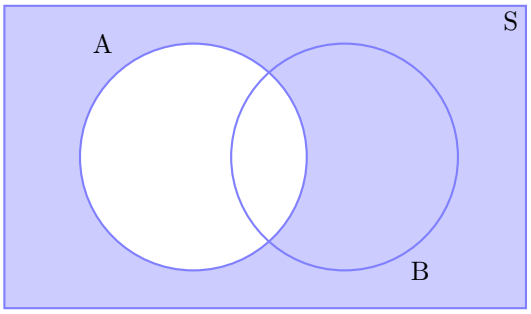
- Denote the event that both A and B occur as $A \cap B$

$A \cap B$



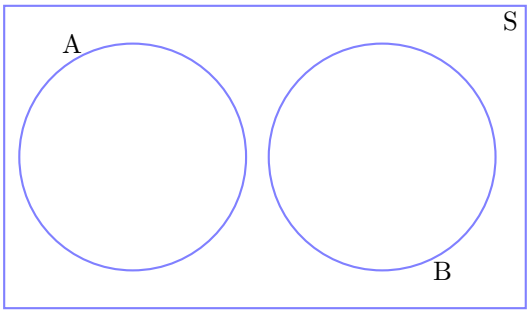
- Denote the event that A does not occur as $\bar{A}$ or A^C (different people use different notations)

$\bar{A}$ 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



2.2 Probability Rules

2.2.1 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.

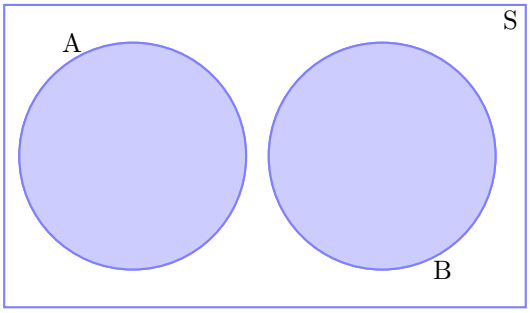
Definition 2. Probability is the proportion of times an event occurs in many repeated trials of a random phenomenon. In other words, probability is the long-term relative frequency.

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

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



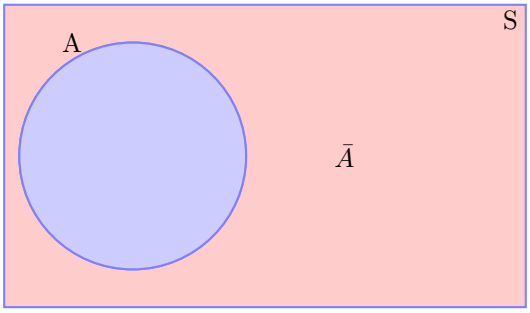
Example. Let S be the sum of two different colored dice. Suppose we are interested in $P(S \leq 4)$. Notice that the pair of dice can fall 36 different ways (6 ways for the first die and six for the second results in 6x6 possible outcomes, and each way has equal probability 1/36. Since the dice cannot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

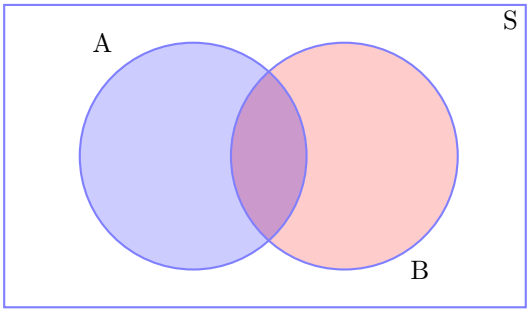
$$P(A) + P(\bar{A}) = 1$$



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.

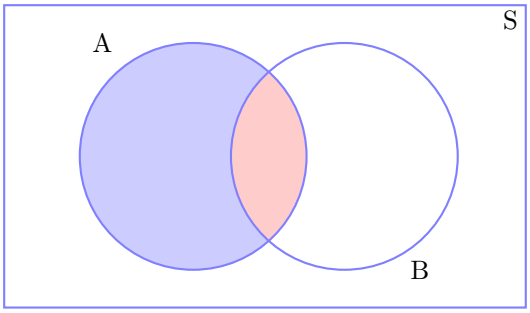
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$



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.

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



2.2.2 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

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

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.

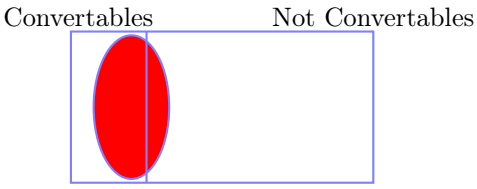
Fact 5. *If A and B are independent events, then*

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

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.

2.2.3 Summary of Probability Rules

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

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

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

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

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

2.3 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?

2.3.1 Introduction to Discrete Random Variables

The following facts hold for discrete random variables:

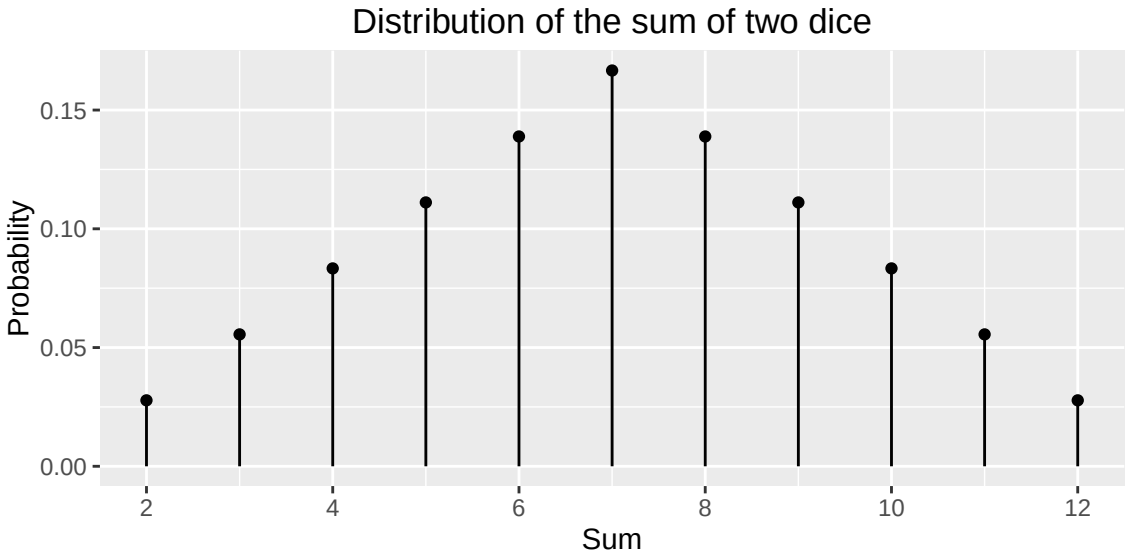
1. The probability associated with every value lies between 0 and 1

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

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

Expected Value

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



We often want to ask 'What is expected value of this distribution?' You might think about taking a really, really large number of samples from this distribution and then taking the mean of that *really really* big sample. 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.

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

Variance

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

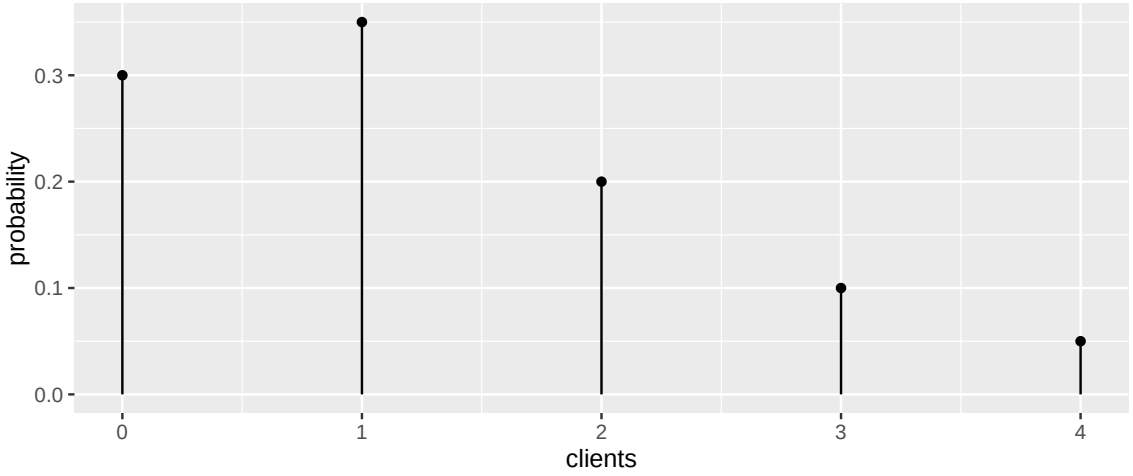
$$\begin{aligned}
\sigma^2 = V[S] &= \sum_{s=2}^{12} \left(s - \mu\right)^2 P(S = s) \\
&= (2-7)^2\left(\frac{1}{36}\right) + (3-7)^2\left(\frac{2}{36}\right) + \cdots + (12-7)^2\left(\frac{1}{36}\right) \\
&= \frac{35}{6} = 5.8\bar{3}
\end{aligned}$$

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

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

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

```
library(ggplot2) # graphing package

distr <- data.frame(    clients = c( 0,   1,    2,    3,    4   ),    # two columns
                     probability = c(0.3, 0.35, 0.20, 0.10, 0.05 ) )   #

ggplot(distr, aes(x=clients)) +                 # graph with clients as the x-axis
  geom_point(aes(y=probability)) +              # where the dots go
  geom_linerange(aes(ymax=probability, ymin=0)) # the vertical lines
```



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 and graph is often called a *probability mass function (pmf)* because it lists how the probability is spread across the possible values of the random variable. We might next ask ourselves what is the average number of clients per day? It looks like it ought to be between 1 and 2 clients per day.

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

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

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

What is the variance of this distribution?

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

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

2.4 Common Discrete Distributions

2.4.1 Binomial Distribution

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

The possible outcomes are

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

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

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

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

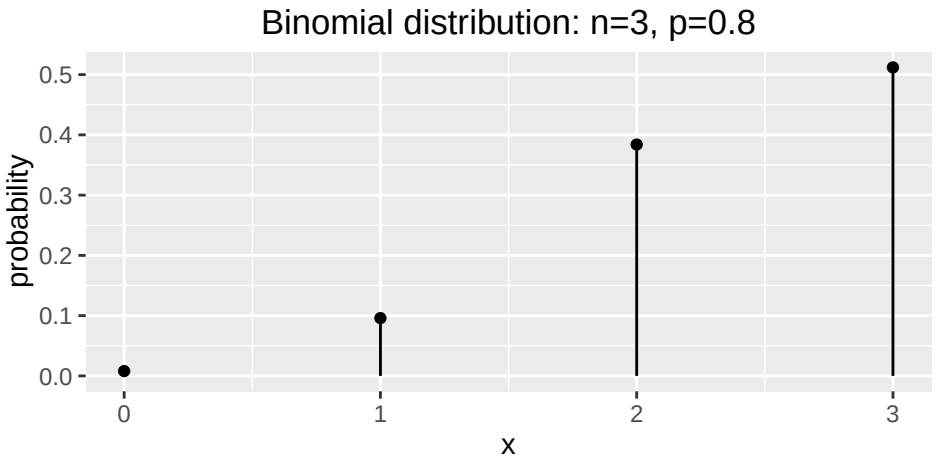
Recall that the probability mass function (pmf) describes how the probability is spread across the possible outcomes, and in this case, I can describe this via a nice formula. The pmf of a a binomial random variable X taken from n trials each with probability of success π is

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

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

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

```
library(ggplot2)
library(dplyr)
dist <- data.frame( x=0:3 ) %>%
  mutate(probability = dbinom(x, size=3, prob=0.8))
ggplot(dist, aes(x=x)) +
  geom_point(aes(y=probability)) +
  geom_linerange(aes(ymax=probability, ymin=0)) +
  ggtitle('Binomial distribution: n=3, p=0.8')
```



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

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

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:

```
# If    X ~ Binomial(n=3, pi=0.8)
# Then  P( X = 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

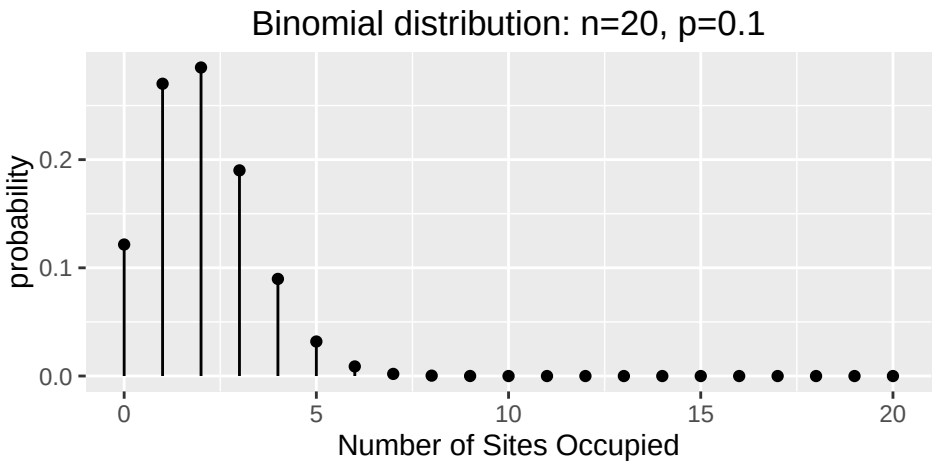
$$\begin{aligned} E[X] &= \sum_{x=0}^{n} x\, P(X=x) \\ &= \sum_{x=0}^{n} x\, \frac{n!}{x!\,(n-x)!}\pi^x\,(1-\pi)^{n-x} \\ &= \vdots \\ &= n\pi \end{aligned}$$

and the variance can be similarly calculated

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

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

```
dist <- data.frame( x=0:20 ) %>%
  mutate(probability = dbinom(x, size=20, prob=0.1))
ggplot(dist, aes(x=x)) +
  geom_point(aes(y=probability)) +
  geom_linerange(aes(ymax=probability, ymin=0)) +
  ggtitle('Binomial distribution: n=20, p=0.1') +
  xlab('Number of Sites Occupied')
```



Often we are interested in questions such as $P(X \leq 2)$ 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(X=0) + P(X=1) + P(X=2)
sum <- dbinom(0, size=20, prob=0.1) +
       dbinom(1, size=20, prob=0.1) +
       dbinom(2, size=20, prob=0.1)
sum

## [1] 0.6769268

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

## [1] 0.6769268
```

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

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

2.4.2 Poisson Distribution

A commonly used distribution for count data is the Poisson.

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

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

1. Two or more events do not occur at precisely the same time or in the same space
2. The occurrence of an event in a given period of time or region of space is independent of the occurrence of the event in a non overlapping period or region.
3. The expected number of events during one period or region, λ, is the same in all periods or regions of the same size.

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

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

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

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

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

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

What about the probability of finding at most 4?

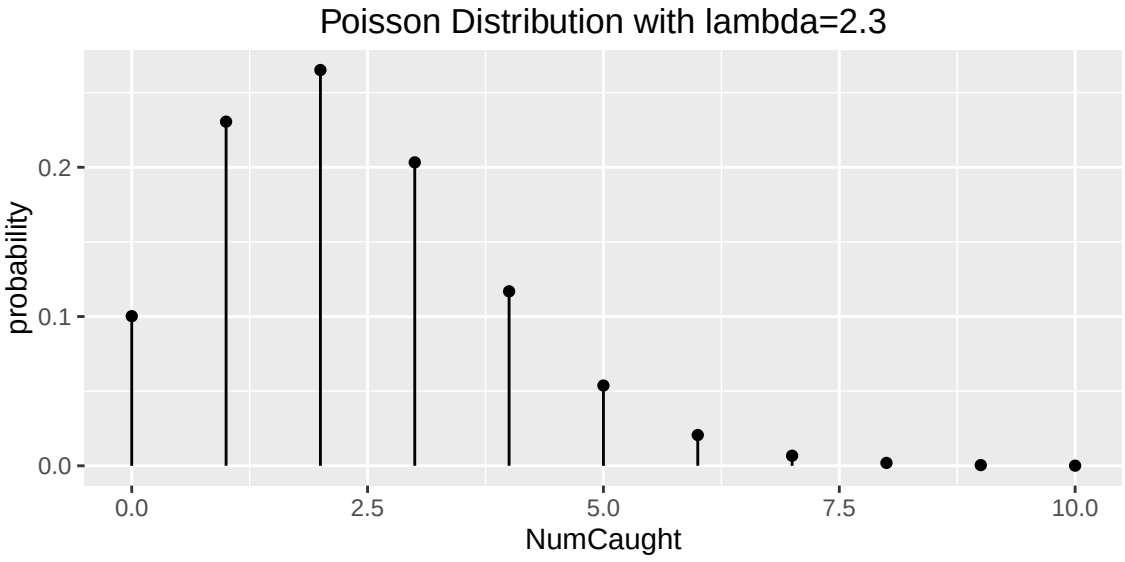
$$\begin{aligned} P(Y \leq 4) &= P(Y = 0) + P(Y = 1) + P(Y = 2) + P(Y = 3) + P(Y = 4) \\ &= 0.1003 + 0.2306 + 0.2652 + 0.2033 + 0.1169 \\ &= 0.9163 \end{aligned}$$

What about the probability of finding 5 or more?

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

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

```
dist <- data.frame( NumCaught = 0:10 ) %>%
  mutate( probability = dpois( NumCaught, lambda=2.3 ) )
ggplot(dist, aes(x=NumCaught)) +
  geom_point( aes(y=probability) ) +
  geom_linerange(aes( ymax=probability, ymin=0)) +
  ggtitle('Poisson Distribution with lambda=2.3')
```



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

## [1] 0.1169022

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

## [1] 0.9162493

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

## [1] 0.08375072
```

2.5 Continuous Random Variables

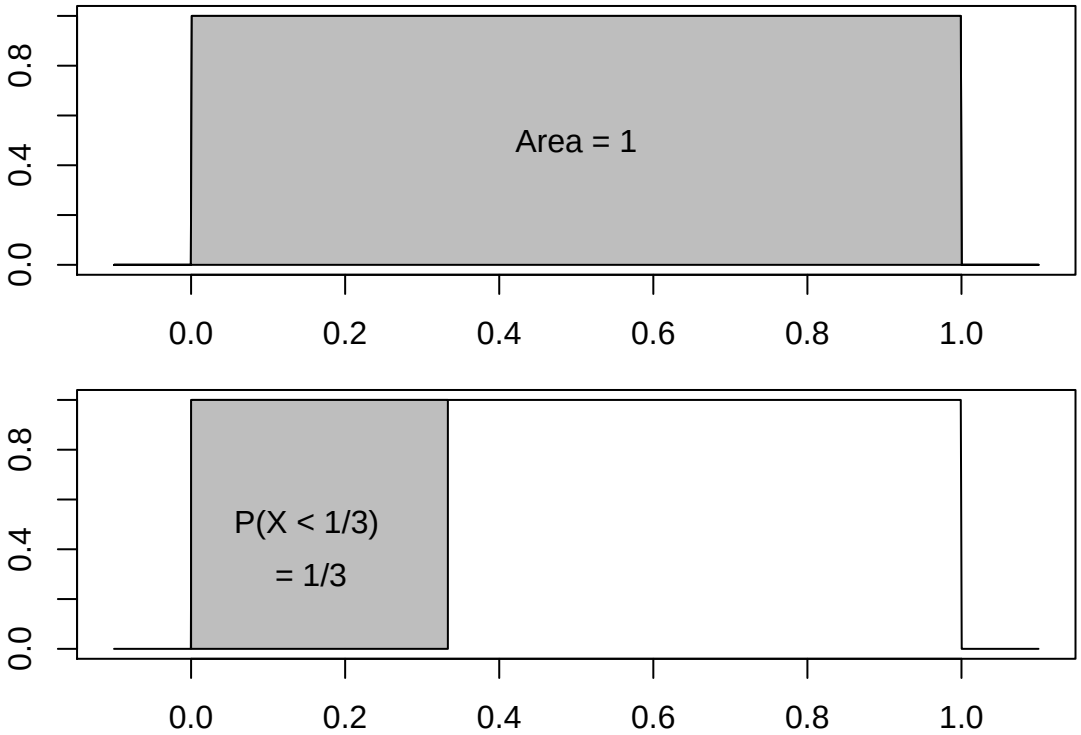
Finding the area under the curve of a particular density function $f(x)$ requires the use of calculus, but since this isn't a calculus course, we will resort to using R or tables of calculated values.

2.5.1 Uniform(0,1) Distribution

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

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

To make this distinction, we will define the distribution using a *probability density function* instead of the probability mass function. In the discrete case, we had to constrain the probability mass function to sum to 1. In the continuous case, we have to constrain the probability density function to integrate to 1.

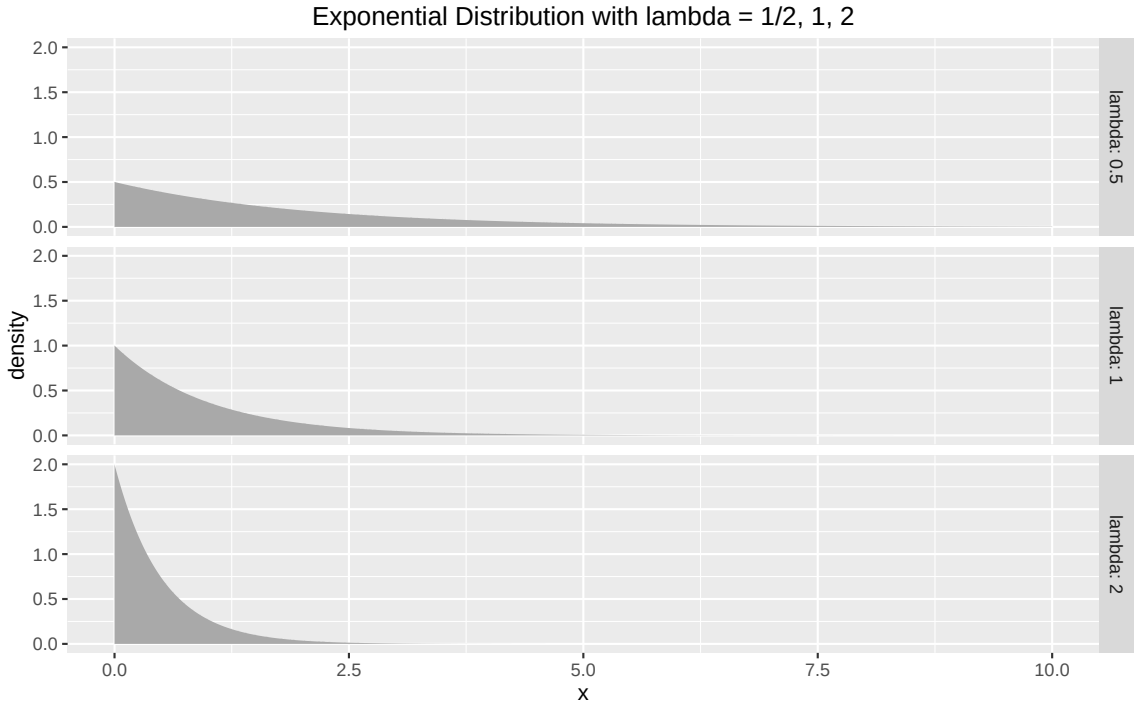


Finding the area under the curve of a particular density function $f(x)$ usually requires the use of calculus, but since this isn't a calculus course, we will resort to using R or tables of calculated values.

2.5.2 Exponential Distribution

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

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



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

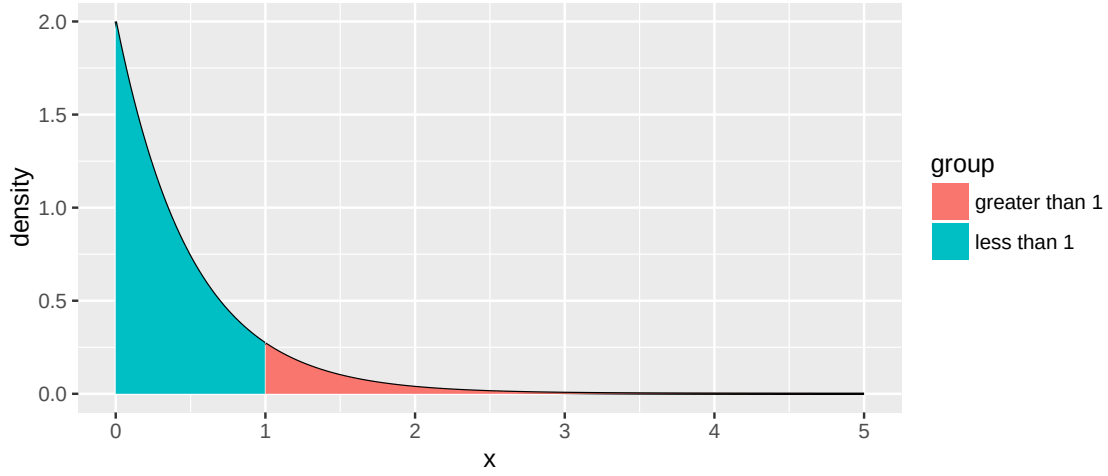
$$\begin{aligned} E[X] &= \int_0^\infty x\, f(x)\, dx \\ &= \vdots \\ &= \frac{1}{\lambda} \\ Var[X] &= \int_0^\infty (x - E[X])^2\, f(x)\, dx \\ &= \vdots \\ &= \frac{1}{\lambda^2} \end{aligned}$$

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

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

$P(X > 1) =$

```
data <- data.frame( x=seq(0,5,length=1000) ) %>%
    mutate(density = dexp(x, rate=2),
           group   = ifelse( x >= 1, 'greater than 1', 'less than 1'))
head(data)

##             x  density       group
## 1 0.000000000 2.000000 less than 1
## 2 0.005005005 1.980080 less than 1
## 3 0.010010010 1.960358 less than 1
## 4 0.015015015 1.940833 less than 1
## 5 0.020020020 1.921502 less than 1
## 6 0.025025025 1.902364 less than 1

ggplot(data, aes(x=x, y=density, fill=group)) +
  geom_line() +
  geom_area()
```



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

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

## [1] 0.1353353
```

2.5.3 Normal Distribution

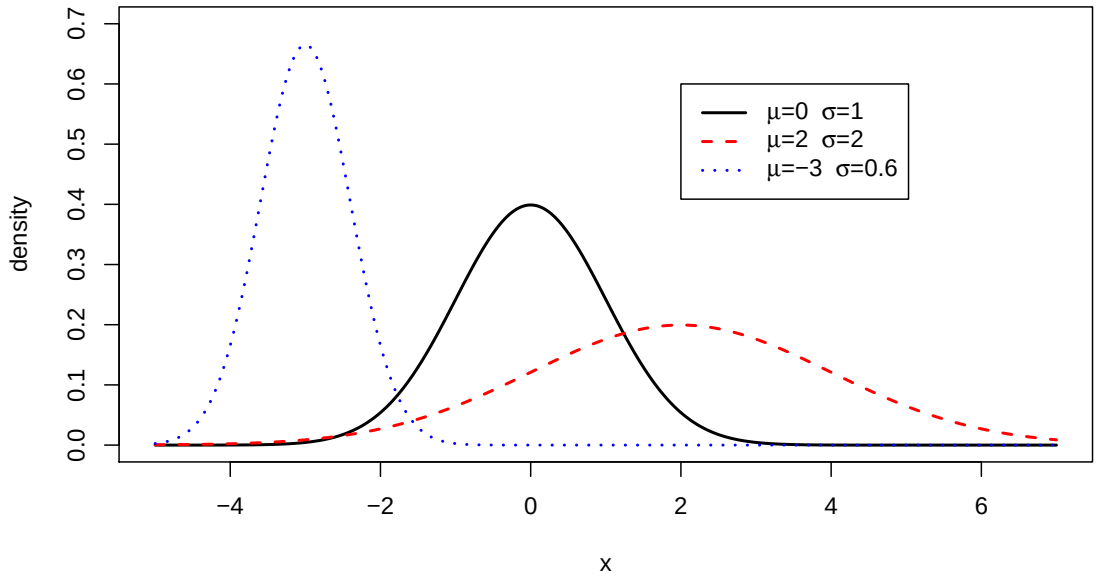
Undoubtably the most important distribution in statistics is the normal distribution. If my RV X is normally distributed with mean μ and standard deviation σ, its probability density function is given by

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

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

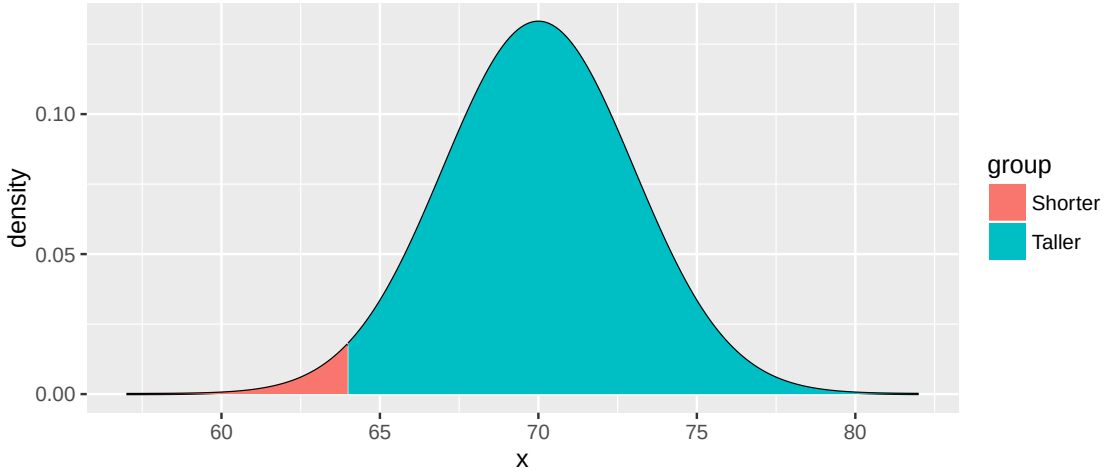
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

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



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

```
distr <- data.frame(x=seq(57,82,length=1000)) %>%
  mutate( density = dnorm(x, mean=70, sd=3),
          group = ifelse(x<=64, 'Shorter','Taller') )
ggplot(distr, aes(x=x, y=density, fill=group)) +
  geom_line() +
  geom_area()
```



Using R you can easily find this

```
pnorm(64, mean=70, sd=3)

## [1] 0.02275013
```

2.5.3.1 Standardizing

Before we had computers that could calculate these probabilities for any normal distribution, it was important to know how to convert a probability statement from an arbitrary $N\left(\mu, \sigma^2\right)$ distribution to a question about a *Stanard Normal* distribution, which is a normal distribution with mean 0 and standard deviation 1. If we have $X \sim N\left(\mu, \sigma^2\right)$, then

$$Z = \frac{X - \mu}{\sigma} \sim N\left(0, 1\right)$$

You might remember doing something similar in an undergraduate statistics course in order to use a table to look up some probability. From the height example, we calculate

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

note that this calculation shows that he is -2 standard deviations from the mean. Next we look at a table for $z = -2.00$. To do this we go down to the -2.0 row and over to the .00 column and find 0.0228. Only slightly over 2% of the adult male population is shorter!

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

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

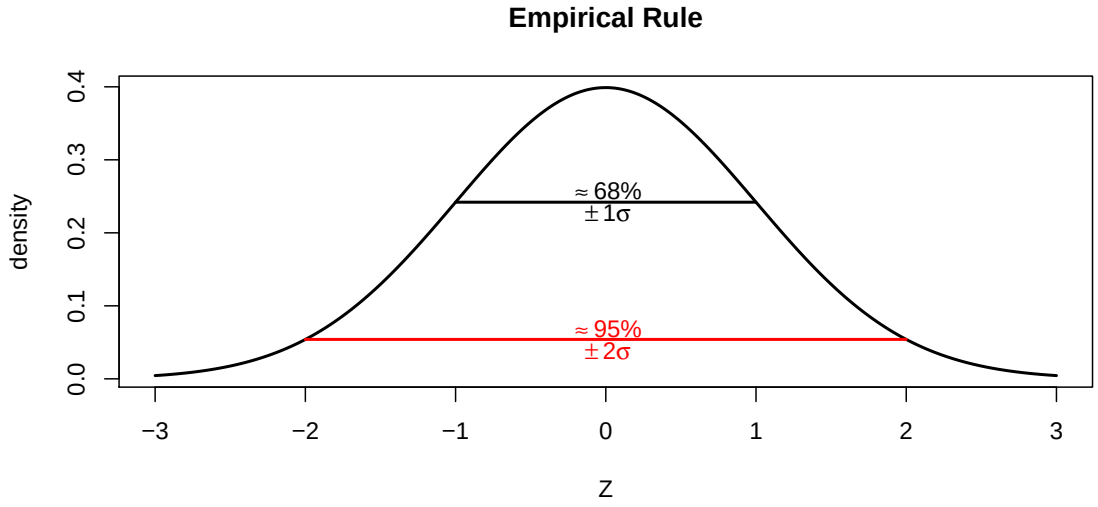
Alternatively we could use the quantile function for the normal distribution `(q-function)` in R and avoid the imprecision of using a table.

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

## [1] 0.8416212
```

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

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

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



2.6 R Comments

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

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

All the probability distributions available in R are accessed in exactly the same way, using a `d-function`, `p-function`, `q-function`, and `r-function`.

Function	Result	Example

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

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

2.7 Exercises

1. The population distribution of blood donors in the United States based on race/ethnicity and blood type as reported by the American Red Cross is given here:

	Blood Type				
Ethnicity	O	A	B	AB	Total
White	36%	32.2%	8.8%	3.2%	
Black	7%	2.9%	2.5%	0.5%	
Asian	1.7%	1.2%	1%	0.3%	
Other	1.5%	0.8%	0.3%	0.1%	
Total					

Notice that the numbers given in the table sum to 100%, so the data presented are the probability of a particular ethnicity and blood type.

(a) Fill in the column and row totals.

(b) What is the probability that a randomly selected donor will be Asian and have Type O blood? That is to say, given a donor is randomly selected from the list of all donors, what is the probability that the selected donor will Asian with Type O?

(c) What is the probability that a randomly selected donor is white? That is to say, given a donor is randomly selected from the list of *all donors*, what is the probability that the selected donor is white?

(d) What is the probability that a randomly selected donor has Type A blood? That is to say, given a donor is selected from the list of *all donors*, what is the probability that the selected donor has Type A blood?

(e) What is the probability that a white donor will have Type A blood? That is to say, given a donor is randomly selected from the list of *all the white donors,* what is the probability that the selected donor has Type A blood? (Notice we already know the donor is white because we restricted ourselves to that subset!)

(f) Is blood type and ethnicity independent? Justify your response mathematically using your responses from the previous answers.

2. For each of the following, mark if it is Continuous or Discrete.

(a) _____ Milliliters of tea drunk per day.

(b) _____ Different brands of soda drunk over the course of a year.

(c) _____ Number of days per week that you are on-campus for any amount of time.

(d) _____ Number of grizzly bears individuals genetically identified from a grid of hair traps in Glacier National Park.

3. For each scenario, state whether the event should be modeled via a binomial or Poisson distribution.

(a) _____ Number of M&Ms I eat per hour while grading homework

(b) _____ The number of mornings in the coming 7 days that I change my daughter's first diaper of the day.

(c) _____ The number of Manzanita bushes per 100 meters of trail.

4. During a road bike race, there is always a chance a crash will occur. Suppose the probability that at least one crash will occur in any race I'm in is $\pi = 0.2$ and that races are independent.

(a) What is the probability that the next two races I'm in will both have crashes?

(b) What is the probability that neither of my next two races will have a crash?

(c) What is the probability that at least one of the next two races have a crash?

5. My cats suffer from gastric distress due to eating house plants and the number of vomits per week that I have to clean up follows a Poisson distribution with rate $\lambda = 1.2$.

   (a) What is the probability that I don't have to clean up any vomits this coming week?

   (b) What is the probability that I must clean up 1 or more vomits?

   (c) If I wanted to measure this process with a rate per day, what rate should I use?

6. Suppose that the number of runners I see on a morning walk on the trails near my house has the following distribution (Notice I've never seen four or more runners on a morning walk):

y	0	1	2	3	4+
$p(y)$	0.45	0.25	0.20		0.0

   (a) What is the probability that I see 3 runners on a morning walk?

   (b) What is the expected number of runners that I will encounter?

   (c) What is the variance of the number of runners that I will encounter?

7. If $Z \sim N\left(\mu = 0, \sigma^2 = 1\right)$, find the following probabilities:

   (a) $P(Z < 1.58) =$

   (b) $P(Z = 1.58) =$

   (c) $P(Z > -.27) =$

   (d) $P(-1.97 < Z < 2.46) =$

8. Using the Standard Normal Table or the table functions in `R`, find z that makes the following statements true.

   (a) $P(Z < z) = .75$

   (b) $P(Z > z) = .4$

9. The amount of dry kibble that I feed my cats each morning can be well approximated by a normal distribution with mean $\mu = 200$ grams and standard deviation $\sigma = 30$ grams.

   (a) What is the probability that I fed my cats more than 250 grams of kibble this morning?

   (b) From my cats' perspective, more food is better. How much would I have to feed them for this morning to be among the top 10% of feedings?

Chapter 3

Confidence Intervals Using Bootstrapping

3.1 Theory of Bootstrapping

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

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

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

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

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

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

```
## Loading required package:  grid
```

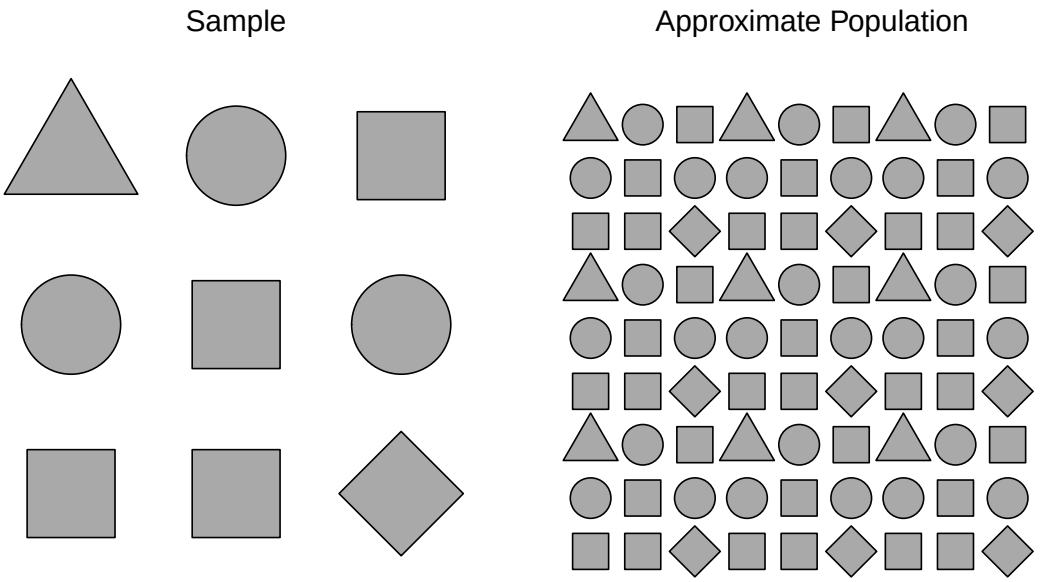


Figure 3.1.1: A possible sample from a population of shapes. Because 4/9 of our sample were squares, our best estimate is that the population is also approximately 4/9 squares. We can think of the approximated population as just many many copies of the observed sample data.

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

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

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

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

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

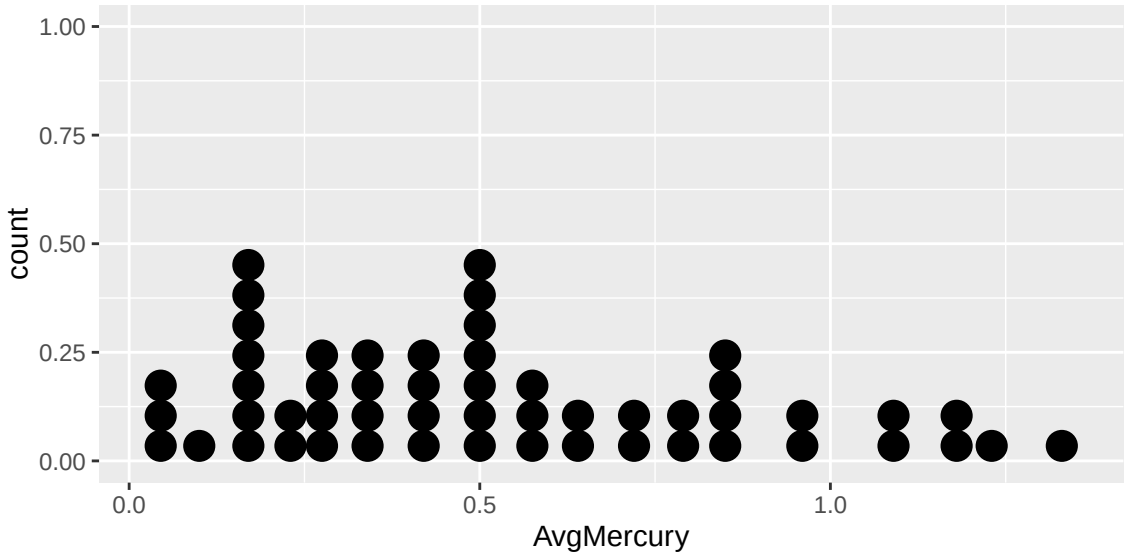
The sampling from the estimated population via sampling from the observed data is called *bootstrapping* because we are making no distributional assumptions about where the data came from, and the idiom "Pulling yourself up by your bootstraps" seemed appropriate.

Example: Mercury Levels in Fish from Florida Lakes

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

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

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

```
# make a nice picture... dot plots are very similar to histograms
# but in this case, my y-axis doen't make any sense.
ggplot(Lakes, aes(x=AvgMercury)) +
  geom_dotplot()

## `stat_bindot()` using `bins = 30`.  Pick better value with `binwidth`.
```



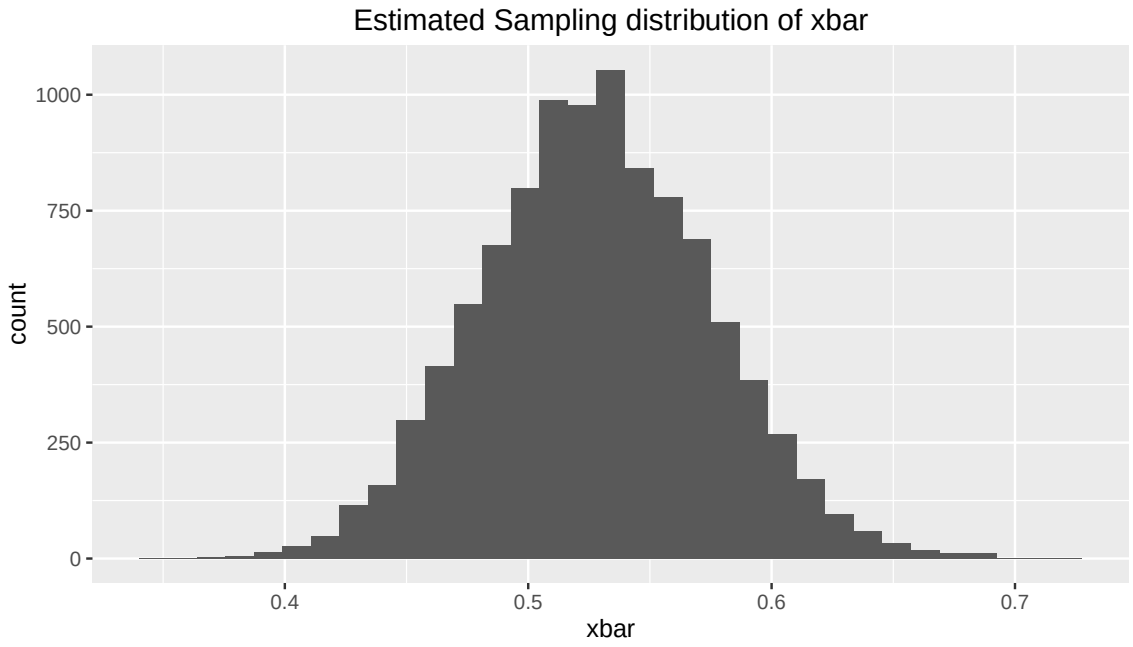
We can calculate mean average mercury level for the $n = 53$ lakes

```
Lakes %>% summarise(xbar = mean( AvgMercury ))

##        xbar
## 1 0.5271698
```

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

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

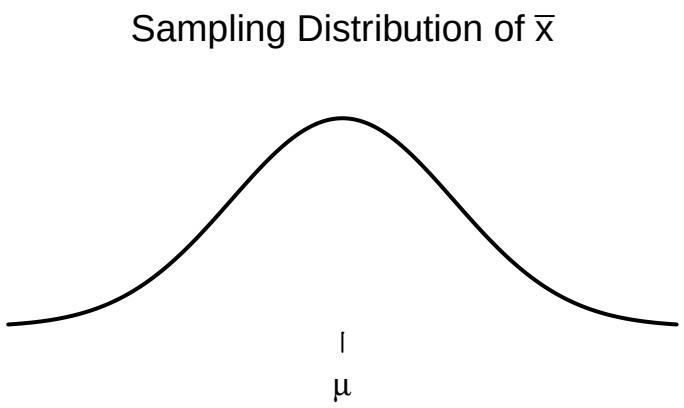
```
# create the sampling distribution of xbar
SamplingDist <- do(10000) * resample(Lakes) %>% summarise(xbar = mean(AvgMercury))

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

##        xbar
## 1 0.5154717
## 2 0.4864151
## 3 0.5456604
## 4 0.5252830
## 5 0.4920755
## 6 0.6128302

# show a histogram of the sampling distribution of xbar
ggplot(SamplingDist, aes(x=xbar)) +
  geom_histogram() +
  ggtitle('Estimated Sampling distribution of xbar' )

## `stat_bin()` using `bins = 30`.  Pick better value with `binwidth`.
```



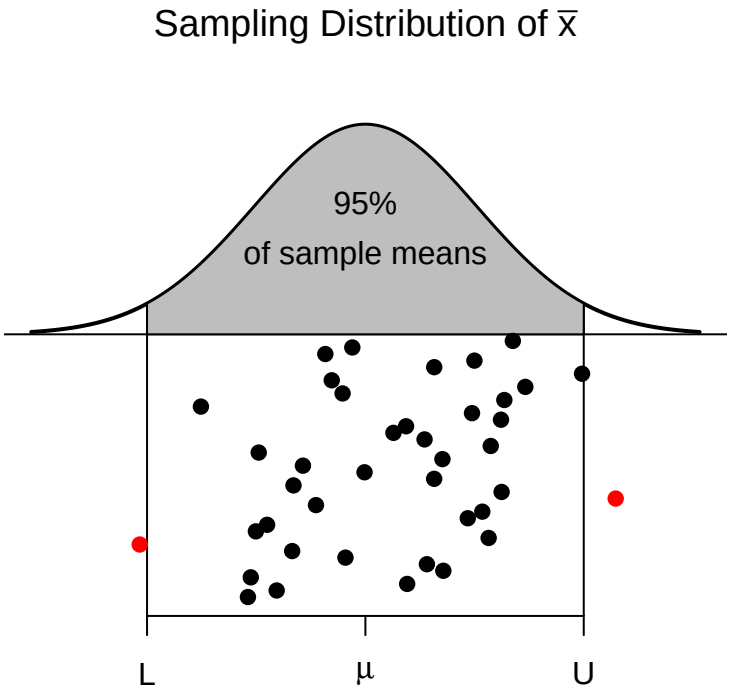
3.2 Using Quantiles of the Estimated Sampling Distributions to create a Confidence Interval

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

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

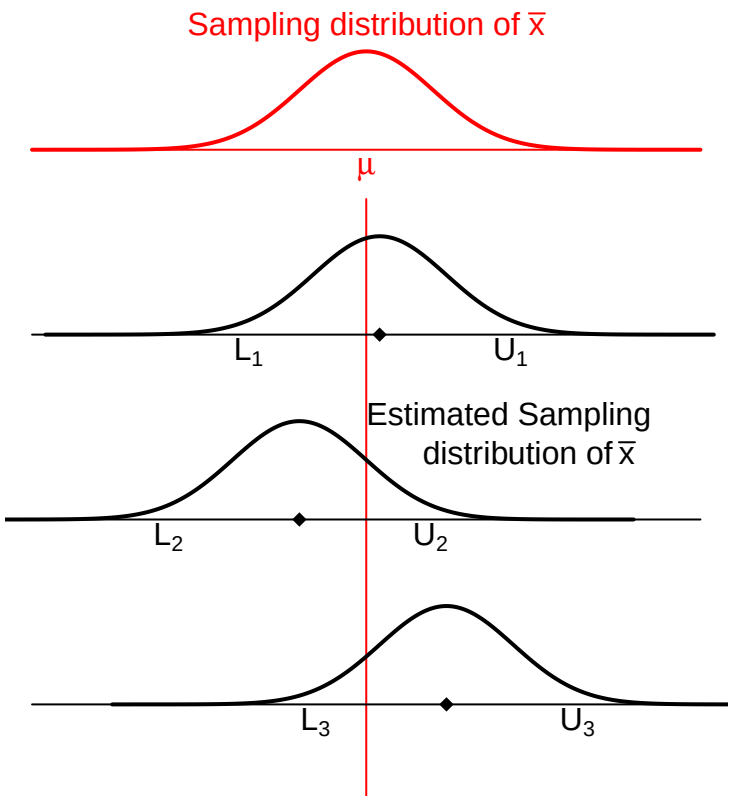


There are two points, (call them L and U) where for our given sample size and population we are sampling from, where we expect that 95% of the sample means to fall within. That is to say, L and U capture the middle 95% of the sampling distribution of $\bar{x}$.

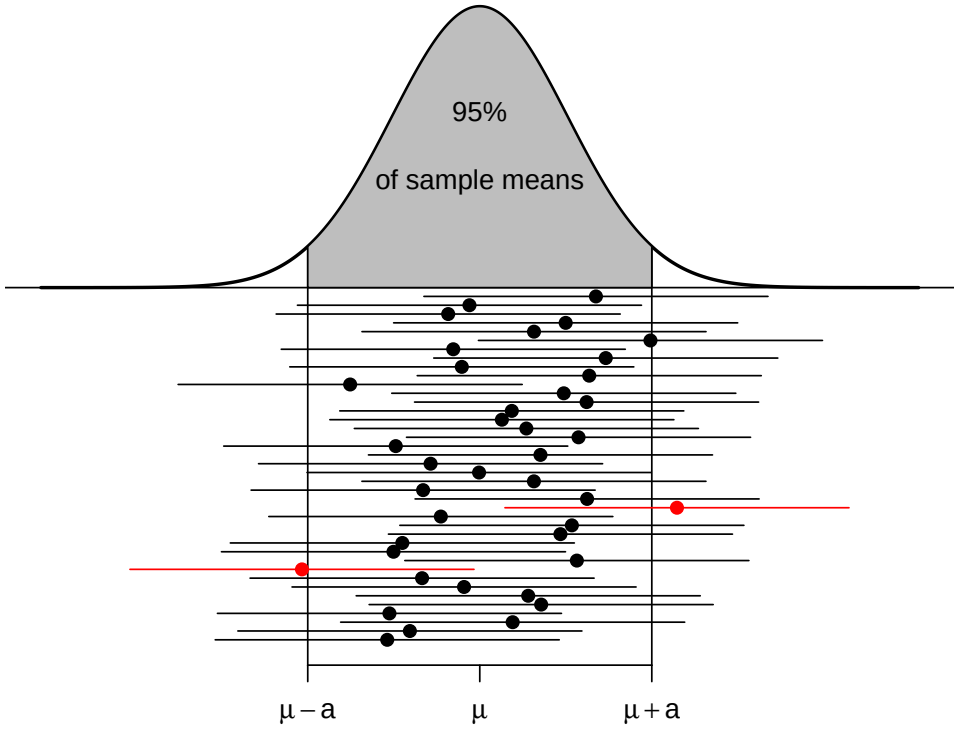


These sample means are randomly distributed about the population mean μ. Given our sample data and sample mean $\bar{x}$, we can examine how our *simulated* values of $\bar{x}^*$ vary about $\bar{x}$. I expect that these simulated sample means $\bar{x}^*$ should vary about $\bar{x}$ in the same way that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

and their associated sample means.



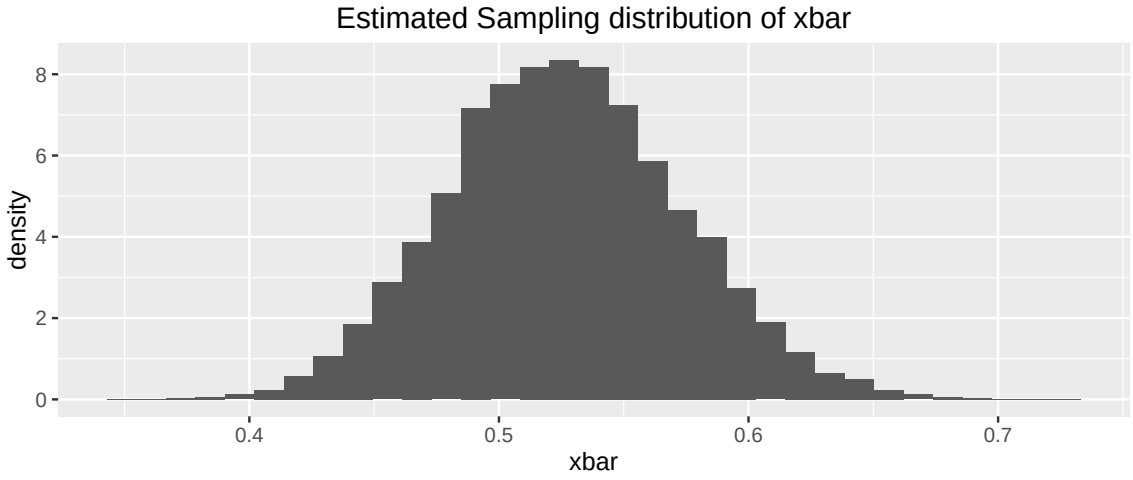
For each possible sample, we could consider creating the estimated sampling distribution of $\bar{X}$ and calculating the L and U values that capture the middle 95% of the estimated sampling distribution. Below are twenty samples, where we've calculated this interval for each sample.



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

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

```
# create the sampling distribution of xbar
SamplingDist <- do(10000) * resample(Lakes)%>%summarise(xbar=mean(AvgMercury))

# show a histogram of the sampling distribution of xbar
ggplot(SamplingDist, aes(x=xbar, y=..density..)) +
  geom_histogram() +
  ggtitle('Estimated Sampling distribution of xbar')

## `stat_bin()` using `bins = 30`.  Pick better value with `binwidth`.
```




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

##      2.5%     97.5%
## 0.4375472 0.6211368
```

There are several ways to interpret this interval.

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

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

3. The interval (0.44, 0.62) is the set of values of μ that are consistent with the observed data at the 0.05 threshold of statistical significance for a two-sided hypothesis test[2].

Example: Fuel Economy

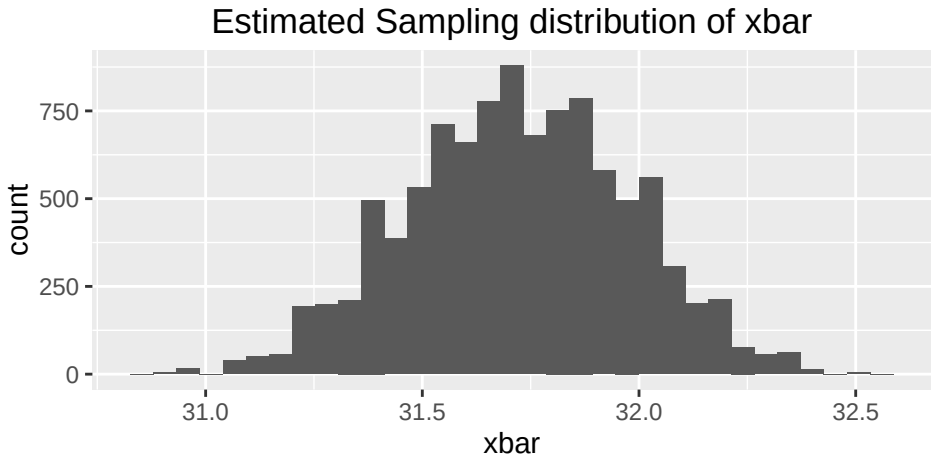
Suppose we have data regarding fuel economy of 5 new vehicles of the same make and model and we wish to test if the observed fuel economy is consistent with the advertised 31 mpg at highway speeds. We the data are

[2]See the chapters on hypothesis testing.

```
CarMPG <- data.frame( ID=1:5, mpg = c(31.8, 32.1, 32.5, 30.9, 31.3) )
CarMPG %>% summarise( xbar=mean(mpg) )

##    xbar
## 1 31.72
```

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

```
SamplingDist <- do(10000) * resample(CarMPG) %>% summarise(xbar=mean(mpg))
# show a histogram of the sampling distribution of xbar
ggplot(SamplingDist, aes(x=xbar)) +
  geom_histogram() +
  ggtitle('Estimated Sampling distribution of xbar')

## `stat_bin()` using `bins = 30`.  Pick better value with `binwidth`.
```




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

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

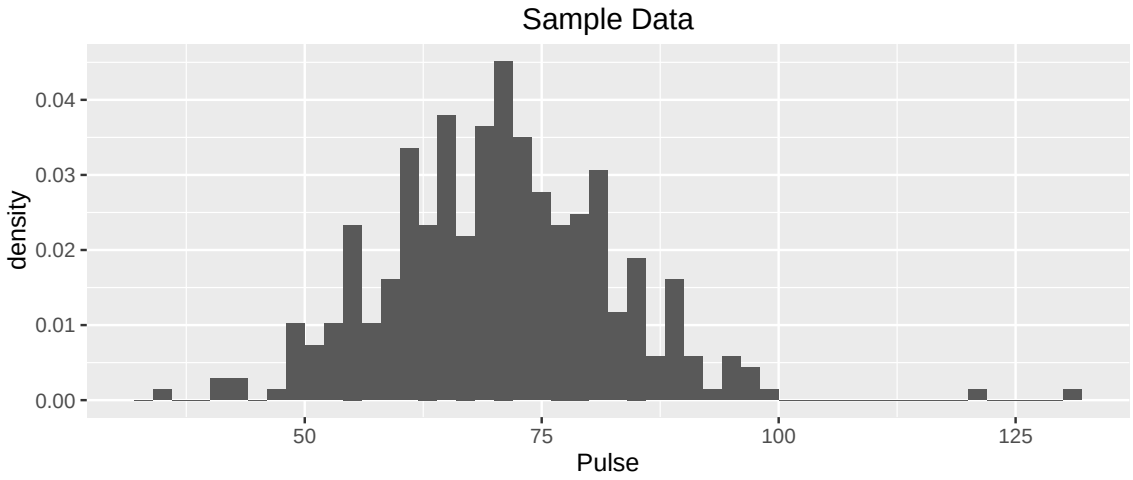
We see that the 95% confidence interval is (31.2, 32.2) and does not actually contain the advertised 31 mpg. However, I don't think we would object to a car manufacturer selling us a car that is *better* than advertised.

Example: Pulse Rate of College Students

In the package `Lock5Data`, the dataset `GPAGender` contains information taken from undergraduate students in an Introductory Statistics course. This is a convenience sample, but could be considered representative of students at that university. One of the covariates measured was the students pulse rate and we will use this to create a confidence interval for average pulse of students at that university.

First we'll look at the raw data.

```
library(Lock5Data)   # load the package
data(GPAGender)      # from the package, load the dataset

# Now a nice histogram
ggplot(GPAGender, aes(x=Pulse, y=..density..)) +
  geom_histogram(binwidth=2) +
  ggtitle('Sample Data')
```

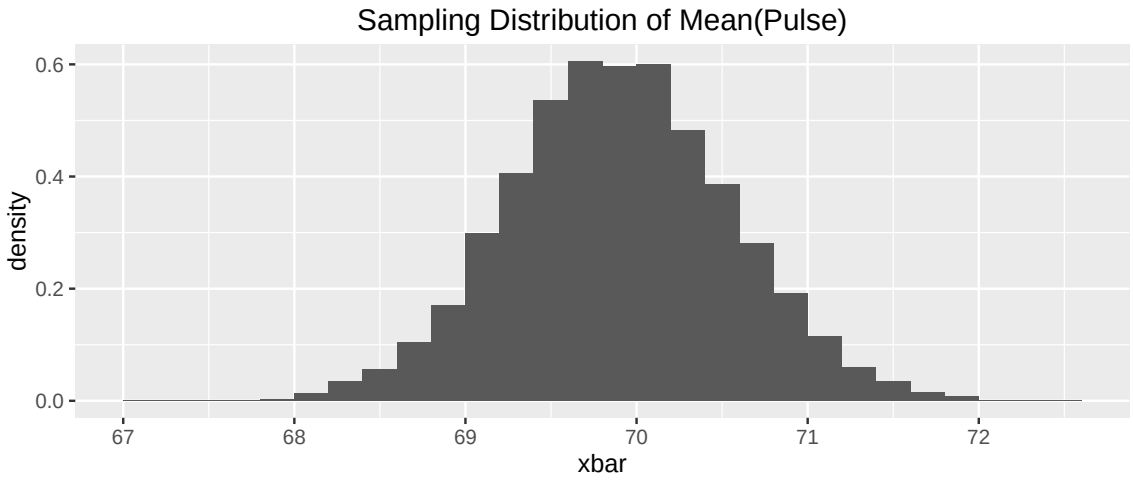


It is worth noting this was supposed to be measuring resting heart rates, but there are two students had extremely high pulse rates and six with extremely low rates. The two high values are approximately what you'd expect from someone currently engaged in moderate exercise and the low values are levels we'd expect from highly trained endurance athletes.

```
# Summary Statistics
GPAGender %>% summarise(xbar = mean(Pulse),
                        StdDev = sd(Pulse))

##       xbar   StdDev
## 1 69.90379 12.08569
```

So the sample mean is $\bar{x} = 69.9$ but how much should we expect our sample mean to vary from sample to sample when our sample size is $n = 343$ people? We'll estimate the sampling distribution of $\bar{X}$ using the bootstrap.

```
# Create the bootstrap replicates
SampDist <- do(10000) * {
  resample(GPAGender) %>% summarise(xbar = mean(Pulse))
}

ggplot(SampDist, aes(x=xbar, y=..density..)) +
  geom_histogram(binwidth=.2) +
  ggtitle('Sampling Distribution of Mean(Pulse)')
```



Just by sampling variability, we expect the sampling mean $\bar{X}$ to vary from approximately 68 to 72. The appropriate quantiles for a 95% bootstrap confidence interval are actually

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

##     2.5%    97.5%
## 68.64431 71.18076
```

3.3 Exercises

For several of these exercises, we will use data sets from the R package `Lock5Data`, which greatly contributed to the pedagogical approach of these notes. Install the package from CRAN using either the following R commands or using the RStudio point-and-click interface `Tools -> Install Packages...`.

1. Load the dataset `BodyTemp50` from the `Lock5Data` package. This is a dataset of 50 healthy adults. Unfortunately the documentation doesn't give how the data was collected, but for this problem we'll assume that it is a representative sample of healthy US adults.

   ```
   library(Lock5Data)
   data( BodyTemp50 )
   ?BodyTemp50
   ```

   One of the columns of this dataset is the `Pulse` of the 50 data, which is the number of heartbeats per minute.

   (a) Create a histogram of the observed pulse values. Comment on the graph and aspects of the graph that might be of scientific interest.

   (b) Calculate the sample mean $\bar{x}$ and sample standard deviation s of the pulses.

(c) Create a dataset of 10000 bootstrap replicates of $\bar{x}^*$.

(d) Create a histogram of the bootstrap replicates. Calculate the mean and standard deviation of this distribution. Notice that the standard deviation of the distribution is often called the *Standard Error* of $\bar{x}$ and we'll denote it as $\sigma_{\bar{x}}$.

(e) Using the bootstrap replicates, create a 95% confidence interval for μ, the average adult heart rate.

(f) Calculate the interval

$$(\bar{x} - 2 \cdot \hat{\sigma}_{\bar{x}}\,, \quad \bar{x} + 2 \cdot \hat{\sigma}_{\bar{x}})$$

and comment on its similarity to the interval you calculated in part (e).

2. Load the dataset `EmployedACS` from the `Lock5Data` package. This is a dataset drawn from American Community Survey results which is conducted monthly by the US Census Bureau and should be representative of US workers. The column `HoursWk` represents the number of hours worked per week.

(a) Create a histogram of the observed hours worked. Comment on the graph and aspects of the graph that might be of scientific interest.

(b) Calculate the sample mean $\bar{x}$ and sample standard deviation s of the worked hours per week.

(c) Create a dataset of 10000 bootstrap replicates of $\bar{x}^*$.

(d) Create a histogram of the bootstrap replicates. Calculate the mean and standard deviation of this distribution. Notice that the standard deviation of the distribution is often called the *Standard Error* of $\bar{x}$ and we'll denote it as $\sigma_{\bar{x}}$.

(e) Using the bootstrap replicates, create a 95% confidence interval for μ, the average worked hours per week.

(f) Calculate the interval

$$(\bar{x} - 2 \cdot \hat{\sigma}_{\bar{x}}\,, \quad \bar{x} + 2 \cdot \hat{\sigma}_{\bar{x}})$$

and comment on its similarity to the interval you calculated in part (e).

Chapter 4

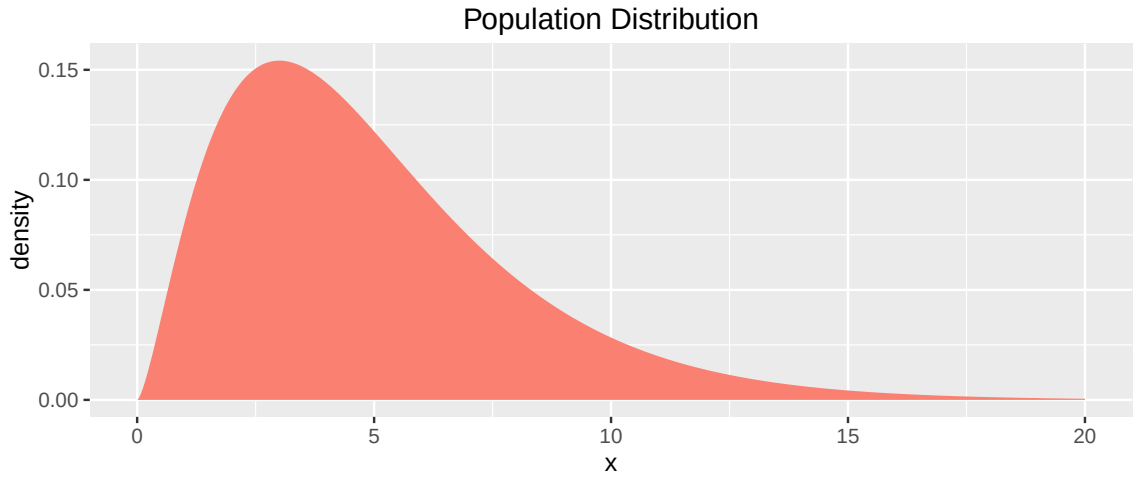
Sampling Distribution of $\bar{X}$

In the previous chapter, we using bootstrapping to estimate the sampling distribution of $\bar{X}$. We then used this bootstrap distribution to calculate a confidence interval for the population mean. Prior to the advent of modern computing, statisticians used a theoretical approximation known as the Central Limit Theorem (CLT). Even today, statistical procedures based on the CLT are widely used and often perform as well or better than the corresponding resampling technique.

4.1 Enlightening Example

Suppose we are sampling from a population that has a mean of $\mu = 5$ and is skewed. For this example, I'll use a Chi-squared distribution with parameter $\nu = 5$.

```
# load the ggplot2 and dplyr librarys... which I use constantly.
library(ggplot2)
library(dplyr)

# Population is a Chi-sq distribution with df=5
PopDist <- data.frame(x = seq(0,20,length=10000)) %>%
  mutate(density=dchisq(x,df=5))

ggplot(PopDist, aes(x=x, y=density)) +
  geom_area(fill='salmon') +
  ggtitle('Population Distribution')
```



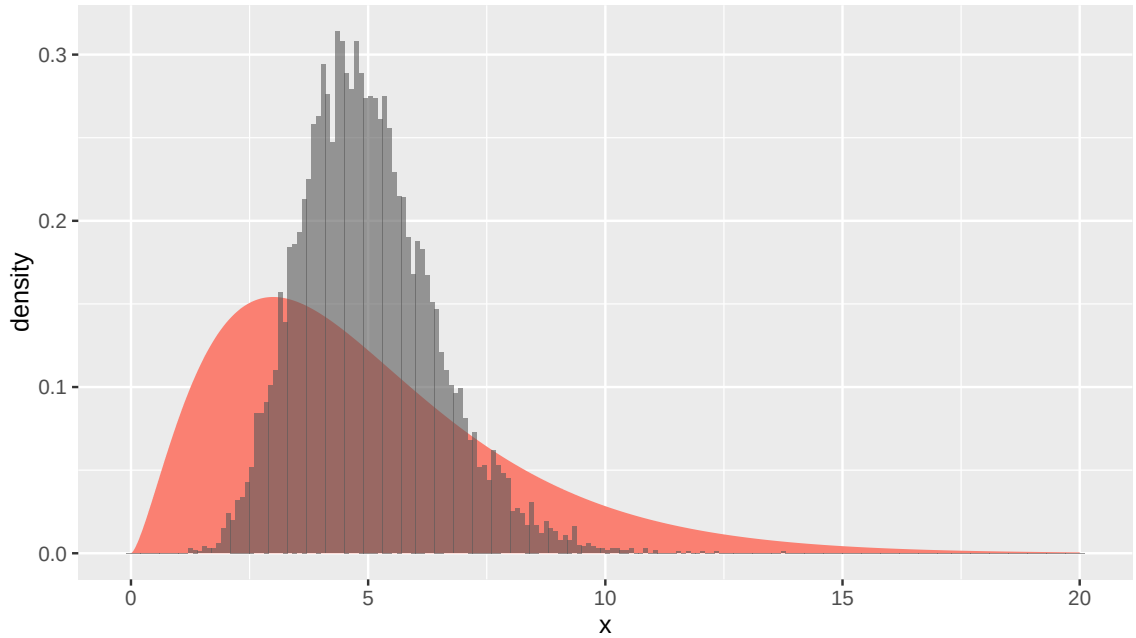
We want to estimate the mean μ and take a random sample of $n = 5$. Lets do this a few times and notice that the sample mean is never *exactly* 5, but is a bit off from that.

```
library(mosaic)  # For the do() loop mechanism
n <- 5  # Our Sample Size!
do(3) * {
  Sample.Data <- data.frame( x = rchisq(n,df=5) )
  Sample.Data %>% summarise( xbar = mean(x) )
}

##       xbar
## 1 6.252452
## 2 4.614849
## 3 3.398105
```

```
n <- 5
SampDist <- do(10000) * {
  Sample.Data <- data.frame( x = rchisq(n,df=5) )
  Sample.Data %>% summarise( xbar = mean(x)     )
}
```

We will compare the population distribution to the sampling distribution graphically.

```
ggplot() +
  geom_area(data=PopDist, aes(x=x, y=density),
            fill='salmon') +
   geom_histogram(data=SampDist, aes(x=xbar, y=..density..),
                  binwidth=.1,
                  alpha=.6)     # alpha is the opacity of the layer
```



From the histogram of the sample means, we notice three things:

- The sampling distribution of $\bar{X}$ is centered at the population mean μ.
- The sampling distribution of $\bar{X}$ has less spread than the population distribution.

- The sampling distribution of $\bar{X}$ is less skewed than the population distribution.

4.2 Mathematical details

4.2.1 Probability Rules for Expectations and Variances

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

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

Proving these results is relatively straight forward and is done in almost all introductory probability text books.

4.2.2 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

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

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.

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

If $X_i \overset{iid}{\sim} N\left(\mu, \sigma^2\right)$ then it is well known (and proven in most undergraduate probability classes) that $\bar{X}$ is also normally distributed with a mean and variance that were already established. That is

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

Notation: Because 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. Because 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?

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

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

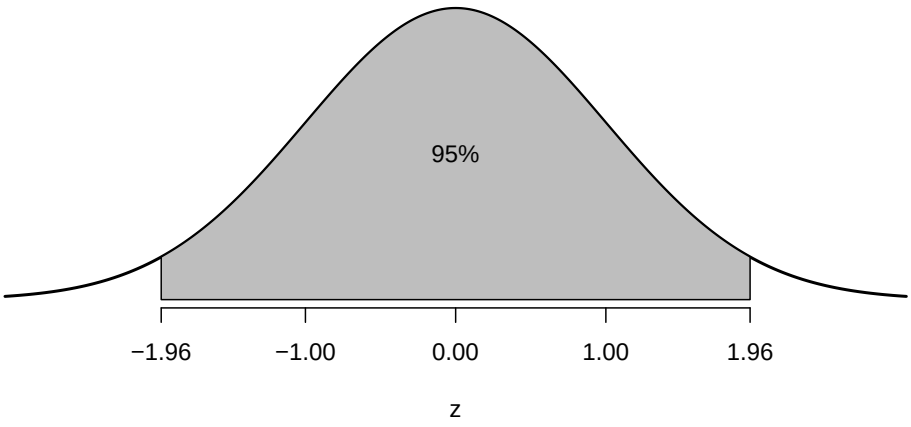
$$\begin{aligned} P\left(\bar{X} \geq 7\right) &= P\left(\frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} \geq \frac{7-6}{\sqrt{\frac{1}{3}}}\right) \\ &= P\left(Z \geq \sqrt{3}\right) \\ &= P\left(Z \geq 1.73\right) \\ &= 0.0418 \end{aligned}$$

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

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

4.4 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 the distribution of the sample mean if the data comes from a non-normal distribution? The answer is quite a lot![1]

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

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

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?

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

[1] Provided the distribution sample from has a non-infinite variance and we have a sufficient sample size.

```
# Answer obtained via simulation
SampDist <- do(10000) *{                                  # make 10,000
  Sample <- data.frame( x= rexp(n=40, rate=1/2 ) )        # simulated xbar
  Sample %>% summarise( xbar = mean( x )          )       # values
}
SampDist %>%                                              # What proportion of those
  mutate(Greater = ifelse(xbar >= 2.5, 1, 0)) %>%         # xbar values are
  summarise( ProportionGreater = mean(Greater) )          # greater than 2.5?

##   ProportionGreater
## 1            0.0642
```

4.5 Summary

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

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

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

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

4.6 Exercises

1. Suppose that the amount of fluid in a small can of soda can be well approximated by a Normal distribution. Let X be the amount of soda (in milliliters) in a single can and $X \sim N(\mu = 222, \ \sigma = 5)$.
   (a) $P(X > 230) =$
   (b) Suppose we take a random sample of 6 cans such that the six cans are independent. What is the expected value of the mean of those six cans? In other words, what is $E\left(\bar{X}\right)$?
   (c) What is $Var\left(\bar{X}\right)$? (Recall we denote this as $\sigma^2_{\bar{X}}$)
   (d) What is the standard deviation of $\bar{X}$? (Recall we denote this as $\sigma_{\bar{X}}$)
   (e) What is the probability that the sample mean will be greater than 230 ml? That is, find $P\left(\bar{X} > 230\right)$.

2. Suppose that the number of minutes that I spend waiting for my order at Big Foot BBQ can be well approximated by a Normal distribution with mean $\mu = 10$ minutes and standard deviation $\sigma = 1.5$ minutes.

   (a) Tonight I am planning on going to Big Foot BBQ. What is the probability I have to wait less than 9 minutes?

   (b) Over the next month, I'll visit Big Foot BBQ 5 times. What is the probability that the mean waiting time of those 5 visits is less than 9 minutes? (This assumes independence of visits but because I don't hit the same restaurant the same night each week, this assumption is probably ok.)

3. A bottling company uses a machine to fill bottles with a tasty beverage. The bottles are advertised to contain 300 milliliters (ml), but in reality the amount varies according to a normal distribution with mean $\mu = 298$ ml and standard deviation $\sigma = 3$ ml. (For this problem, we'll assume σ is known and carry out the calculations accordingly).

   (a) What is the probability that a randomly chosen bottle contains less than 296 ml?

   (b) Given a simple random sample of size $n = 6$ bottles, what is the probability that the sample mean is less than 296 ml?

   (c) What is the probability that a single bottle is filled within 1 ml of the true mean $\mu = 298$ ml? *Hint: Draw the distribution and shade in what probability you want... then convert that to a question about standard normals. To find the answer using a table or R, you need to look up two values and perform a subtraction.*

   (d) What is the probability that the mean of 10 randomly selected bottles is within 1 ml of the mean? What about a sample of 100?

   (e) If a sample of size $n = 50$ has a sample mean of $\bar{x} = 298$, should this be evidence that the filling machine is out of calibration? i.e., assuming the machine has a mean fill amount of $\mu = 300$ and $\sigma = 3$, what is $P\left(\bar{X} \leq 298\right)$?

Chapter 5

Confidence Intervals for μ

5.1 Asymptotic result, σ known

We know that our sample mean $\bar{x}$, should be close to the population mean μ. So when giving a region of values for μ that are consistent with the observed data, we would expect our CI formula to be something like $(\bar{x} - d,\ \bar{x} + d)$ for some value d. That value of d should be small if our sample size is big, representing our faith that a large amount of data should result in a statistic that is *very* close to the true value of μ. Recall that if our data $X_i \sim N\left(\mu, \sigma^2\right)$ or our sample size was large enough, then we know

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

or is approximately so. Doing a little re-arranging, we see that

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

So if we take the 0.025 and 0.975 quantiles of the normal distribution, which are $z_{0.025} = -1.96$ and $z_{0.975} = 1.96$, we could write

$$\begin{aligned} 0.95 &= P\left(-1.96 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq 1.96\right) \\ &= \\ &\quad P\left(-1.96\left(\frac{\sigma}{\sqrt{n}}\right) \leq \bar{X} - \mu \leq 1.96\left(\frac{\sigma}{\sqrt{n}}\right)\right) \\ &= P\left(\bar{X} - 1.96\left(\frac{\sigma}{\sqrt{n}}\right) \leq \mu \leq \bar{X} + 1.96\left(\frac{\sigma}{\sqrt{n}}\right)\right) \end{aligned}$$

Which suggests that a reasonable 95% Confidence Interval for μ is $\bar{x} \pm 1.96\left(\frac{\sigma}{\sqrt{n}}\right)$. In general for a $(1-\alpha)\cdot 100\%$ confidence interval, we would use the formula $\bar{x} \pm z_{1-\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$. Notice that I could write the formula using $z_{\alpha/2}$ instead of $z_{1-\alpha/2}$ because the normal distribution is symmetric about 0 and we are subtracting and adding the same quantity to $\bar{x}$.

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.

5.2 Confidence interval for μ assuming σ is unknown

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

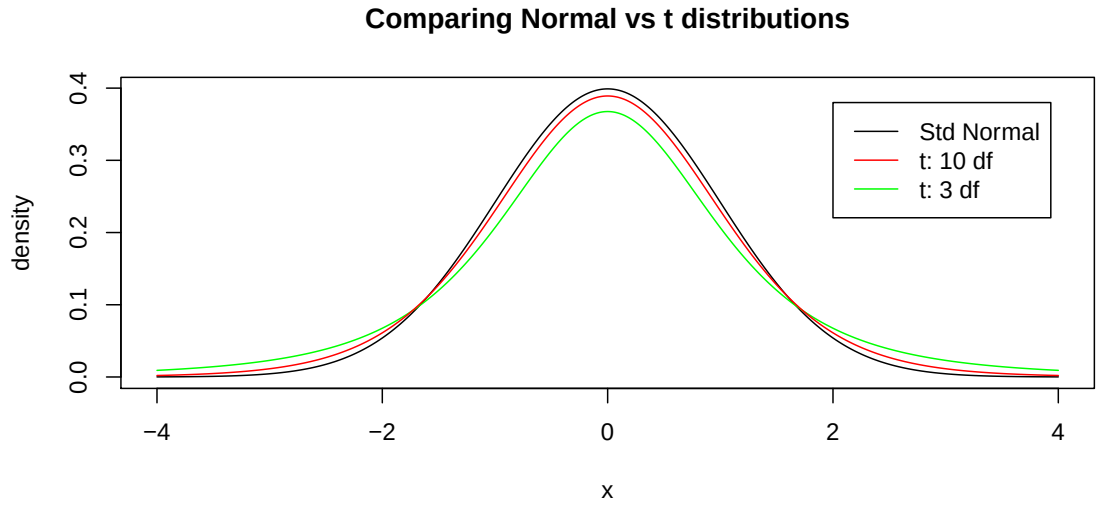
Our previous results about confidence intervals assumed that $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ (or is approximately so) and therefore

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

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

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

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



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

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

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

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

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

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

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

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

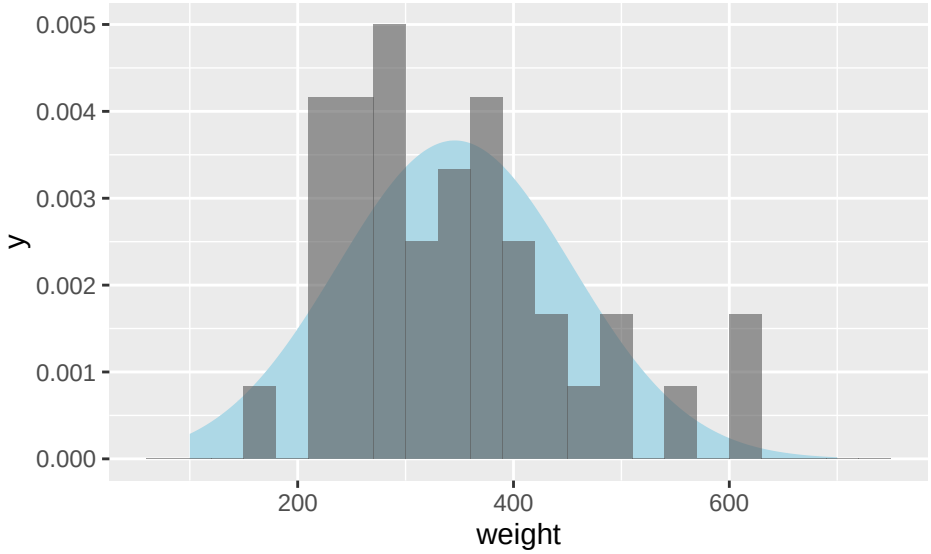
In our current problem, $\bar{x}$ is our estimate of μ and the estimated standard deviation (which is commonly called the *standard error*) is $s/\sqrt{n}$ and the appropriate degrees of freedom are $df = n-1$.

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

```
library(dplyr)    # data frame manipulation
library(ggplot2)  # pretty graphics
library(mosaic)   # for the "do" loop
bears <- data.frame(weight =
   c(306, 446, 276, 235, 295, 302, 374, 339, 624, 266,
     497, 384,  429, 497, 224, 157, 248, 349, 388, 391,
     266, 230, 621, 314, 344,  413, 267, 380, 225, 418,
     257, 466, 230, 548, 277, 354, 271, 369,  275, 272))

xbar <- mean(bears$weight)
s    <-   sd(bears$weight)
cbind(xbar, s)

##       xbar        s
## [1,] 345.6 108.8527
```

Notice that the data do not appear to come from a normal distribution, but a slightly heavier right tail. We'll plot the histogram of data along with a normal distribution with the same mean and standard deviation as our data.

```
normal.data <- data.frame(weight=seq(100,700,length=1000)) %>%
  mutate( y = dnorm(weight, mean=xbar, sd=s))
ggplot() +
  geom_area( data=normal.data, aes(x=weight, y=y), fill='light blue' ) +
  geom_histogram(data=bears, aes(x=weight, y=..density..),
                 binwidth=30, alpha=.6)
```



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

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

## [1] 2.022691
```

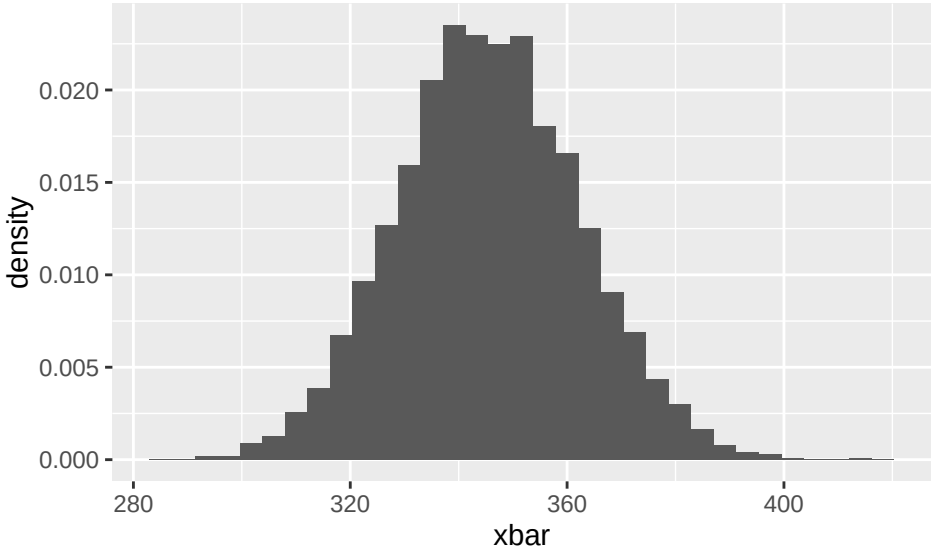
we find that $t_{n-1}^{1-\alpha/2} = 2.02$. Therefore the 95% confidence interval is

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

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

We can wonder how well this interval matches up with the interval we would have gotten if we had used the bootstrap method to create a confidence interval for μ. In this case, where the sample size n is relatively large, the Central Limit Theorem is certainly working and the distribution of the sample mean certainly looks fairly normal.

```
SampDist <- do(10000) * resample(bears) %>% summarise(xbar=mean(weight))
ggplot(SampDist, aes(x=xbar, y=..density..)) +
  geom_histogram()

## `stat_bin()` using `bins = 30`.  Pick better value with `binwidth`.
```



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

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

##     2.5%    97.5%
## 313.3744 379.1500
```

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.

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

```
qt( 1-.1/2, df=4)  # 1-(.1)/2 = 1-.05 = .95

## [1] 2.131847
```

$$\begin{aligned} 55 \quad &\pm \quad 2.13 \left(\frac{4}{\sqrt{5}} \right) \\ 55 \quad &\pm \quad 3.8 \end{aligned}$$

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?

5.3 Sample Size Selection

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

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

and therefore

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

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

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 $ME = 10$ pounds. Using prior collection data from the fall harvest (road side checks by game wardens), the researcher believes that $\sigma = 60$ lbs is a reasonable standard deviation number to use.

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

Notice that I don't bother using the t-distribution in this calculations because because I am assuming that σ is known. While this is a horrible assumption, the difference between using a t quantile instead of z quantile is small and what really matters is how good the estimate of σ is. As with many things, the quality of the input values is reflected in the quality of the output. Typically this sort of calculation is done with only a rough estimate of σ and therefore I would subsequently regard the resulting sample size n as an equally rough estimate.

5.4 Exercises

1. An experiment is conducted to examine the susceptibility of root stocks of a variety of lemon trees to a specific larva. Forty of the plants are subjected to the larvae and examined after a fixed period of time. The response of interest is the logarithm of the number of larvae per gram of of root stock. For these 40 plants, the sample mean is $\bar{x} = 11.2$ and the sample standard deviation is $s = 1.3$. Use these data to construct a 90% confidence interval for μ, the mean susceptibility of lemon tree root stocks from which the sample was taken.

2. A social worker is interested in estimating the average length of time spent outside of prison for first offenders who later commit a second crime and are sent to prison again. A random sample of $n = 100$ prison records in the count courthouse indicates that the average length of prison-free life between first and second offenses is 4.2 years, with a standard deviation of 1.1 years. Use this information to construct a 95% confidence interval for μ, the average time between first and second offenses for all prisoners on record in the county courthouse.

3. A biologist wishes to estimate the effect of an antibiotic on the growth of a particular bacterium by examining the number of colony forming units (CFUs) per plate of culture when

a fixed amount of antibiotic is applied. Previous experimentation with the antibiotic on this type of bacteria indicates that the standard deviation of CFUs is approximately 4. Using this information, determine the number of observations (i.e. cultures developed) necessary to calculate a 99% confidence interval with a half-width of 1.

4. In the R package `Lock5Data`, the dataset `FloridaLakes` contains information about the mercury content of fish in 53 Florida lakes. For this question, we'll be concerned with the average ppm of mercury in fish from those lakes which is encoded in the column `AvgMercury`.

   (a) Using the bootstrapping method, calculate a 90% confidence interval for μ, the average ppm of mercury in fish in all Florida lakes.

   (b) Using the asymptotic approximations discussed in this chapter, calculate a 90% confidence interval for μ, the average ppm of mercury in fish in all Florida lakes.

   (c) Comment on the similarity of these two intervals.

5. In the R package `Lock5Data`, the dataset `Cereal` contains nutrition information about a random sample of 30 cereals taken from an on-line nutrition information website (see the help file for the dataset to get the link). For this problem, we'll consider the column `Sugars` which records the grams of sugar per cup.

   (a) Using the bootstrapping method, calculate a 90% confidence interval for μ, the average grams of sugar per cup of all cereals listed on this website.

   (b) Using the asymptotic approximations discussed in this chapter, calculate a 90% confidence interval for μ, the average grams of sugar per cup of all cereals listed on this website.

   (c) Comment on the similarity of these two intervals.

   (d) We could easily write a little program (or pay an undergrad) to obtain the nutritional information about all the cereals on the website so the random sampling of 30 cereals is unnecessary. However, a bigger concern is that the website cereals aren't representative of cereals Americans eat. Why? For example, consider what would happen if we added 30 new cereals that were very nutritious but were never sold.

Chapter 6

Hypothesis Tests for the mean of a population

```
# packages we'll use in this chapter
library(mosaic)
library(dplyr)
library(ggplot2)
```

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

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

```
cows <- data.frame(
  weight = c(553, 466, 451, 421, 523, 517, 451, 510, 392, 466) )
cows %>% summarise( xbar=mean(weight), s=sd(weight) )

##   xbar        s
## 1  475 49.99556
```

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

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

Assuming[1] the true mean is 500, how likely is it to get a sample mean of 475 (or less)? One way to think about this is that we want a measure of how extreme the event is that we observed, and one way to do that is to calculate how much probability there is for events that are even more extreme.

To calculate how far into the tail our observed sample mean $\bar{x} = 475$ is by measuring the area of

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

the distribution that is farther into the tail than the observed value.

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

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

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

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

The above value is the actual value calculated using R

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

## [1] 0.07428219
```

but using tables typically found in intro statistics books, the most precise thing you would be able to say is

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

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

6.1 Writing Hypotheses

6.1.1 Null and alternative hypotheses

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

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

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

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

1. In approximately 300 BC, Eratosthenes[2] showed that the world was not flat[3] by measuring the different lengths of shadows of identical sticks in two cities that were 580 miles apart but lay on the same meridian (Alexandria is directly north of Aswan). His proposed alternative was that the Earth was a sphere. While his alternative is not technically true (it is actually an oblate spheroid that bulges at the equator), it was substantially better than the flat world hypothesis.

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

It is helpful to our understanding to label the different hypothesis, both the ones being tested and the different alternatives. We'll label the hypothesis being tested as H_0 which we often refer to as the "**null hypothesis**." The **alternative hypothesis**, which we'll denote H_a, should be the opposite of the null hypothesis. Had Eratosthenes known about modern scientific methods, he would have correctly considered H_0 : *the world is flat* verses H_a: *the world is not flat* and not incorrectly concluded that the world is a sphere[4]. Likewise Helmont should have considered the hypotheses H_0 : *plants only consume soil* versus the alternative H_a : *plants consume something besides soil.*

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

6.1.2 Error

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

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

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

[2]For more about Eratosthenes, start at his wikipedia page. http://en.wikipedia.org/wiki/Eratosthenes

[3]Carl Sagan has an excellent episode of *Cosmos* on this topic. https://www.youtube.com/watch?v=G8cbIWMv0rI

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

```
## `stat_bindot()` using `bins = 30`.  Pick better value with `binwidth`.
## `stat_bindot()` using `bins = 30`.  Pick better value with `binwidth`.
```

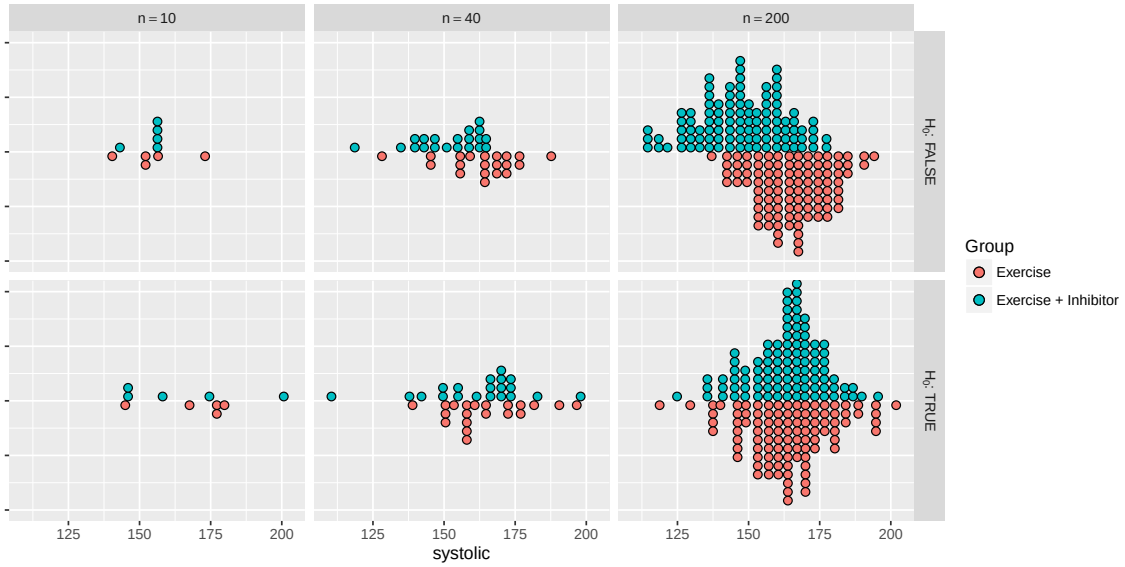


Figure 6.1.1: Comparing possible data assuming there is a difference between treatments versus no difference. In the top row of graphs, there is a difference between the *Exercise* and the *Exercise + Inhibitor* treatments. However, at small sample sizes, we can't tell if the observed difference is due to the difference in treatment or just random variation in the data. In the second row, there is no difference between the treatments.

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

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

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

and the alternative will be one of

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

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

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

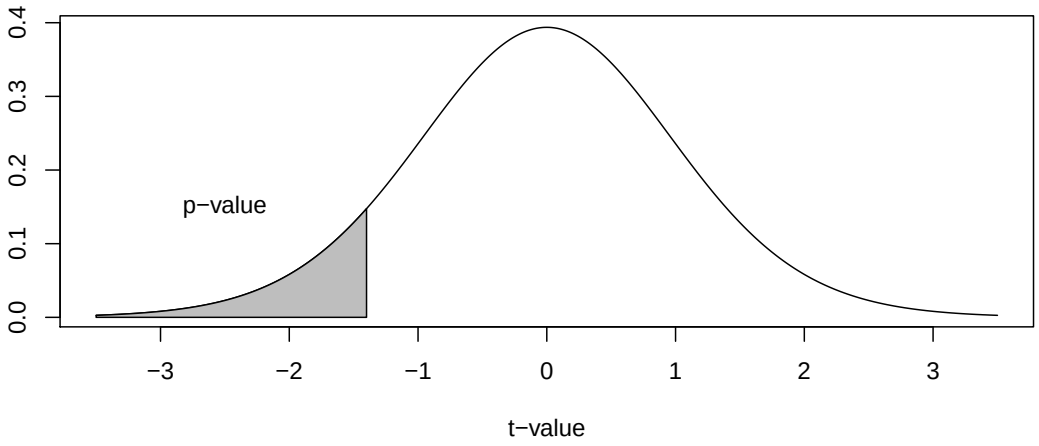
Example A light bulb company advertises that their bulbs last for 1000 hours. Consumers will be unhappy if the bulbs last less time, but will not mind if the bulbs last longer. Therefore *Consumer Reports* might perform a test and would consider the hypotheses

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

Suppose we perform an experiment with $n = 20$ lightbulbs and observe $\bar{x} = 980$ and $s = 64$ hours and therefore our test statistic is

$$\begin{aligned} t_{19} &= \frac{\bar{x} - \mu}{s/\sqrt{n}} \\ &= \frac{980 - 1000}{64/\sqrt{20}} \\ &= -1.40 \end{aligned}$$

Then the p-value would be



and we calculate

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

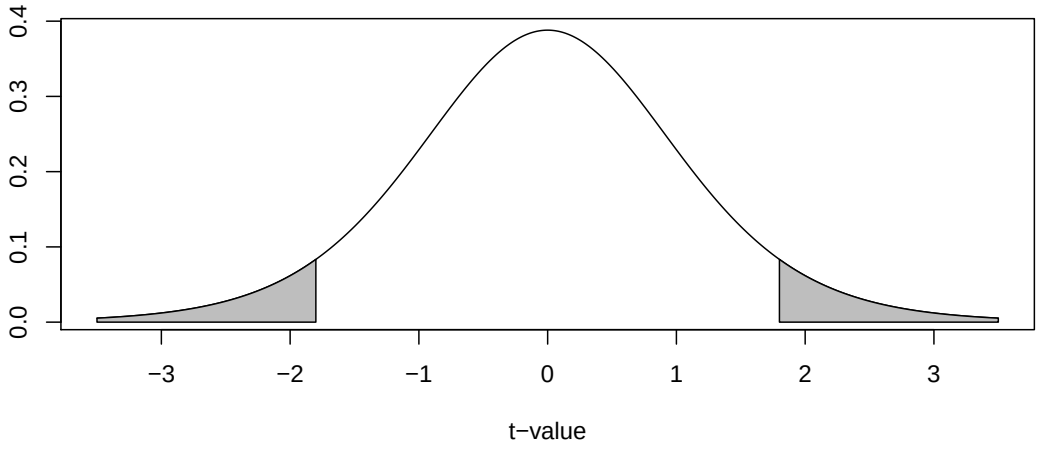
using R

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

## [1] 0.08881538
```

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

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

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



and we calculate

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

using the R commands

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

## [1] 0.1053907
```

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

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

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

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

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

Examples

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

   (a) State an appropriate null and alternative hypothesis.

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

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

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

(c) Calculate the p-value.

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

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

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

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

(a) State an appropriate null and alternative hypothesis.

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

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

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

(c) Calculate the p-value.

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

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

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

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

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

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

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

So the p-value is

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

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

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

and thus

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

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

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

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

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

6.1.3 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.[5] Most of the time you'll subtract from one $p^* = 1 - p^*$.

[5] 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.

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

6.1.4 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. Because the observed test statistic is less extreme than the cutoff value, we failed to reject the null hypothesis.

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

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

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

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

6.1.5 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

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

and calculate

```
cars <- data.frame(mpg = c(31.8, 32.1, 32.5, 30.9, 31.3))
cars %>% summarise(mean(mpg), sd(mpg))

##   mean(mpg)   sd(mpg)
## 1     31.72 0.6340347
```

The test statistic is:

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

The p-value is

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

and a 95% confidence interval is

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

```
t.test( cars$mpg, mu=31, alternative='two.sided' )

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

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

6.2 Type I and Type II Errors

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

There is an important philosophical debate about how much evidence do we need in order to reject the null hypothesis. My brother-in-law would have to have extremely strong evidence before

he stated the other rancher was wrong. Likewise, researchers needed solid evidence before concluding that Newton's Laws of Motion were incorrect.

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

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

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

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

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

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

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

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

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

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

6.2.1 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 $\hat{\sigma}$.

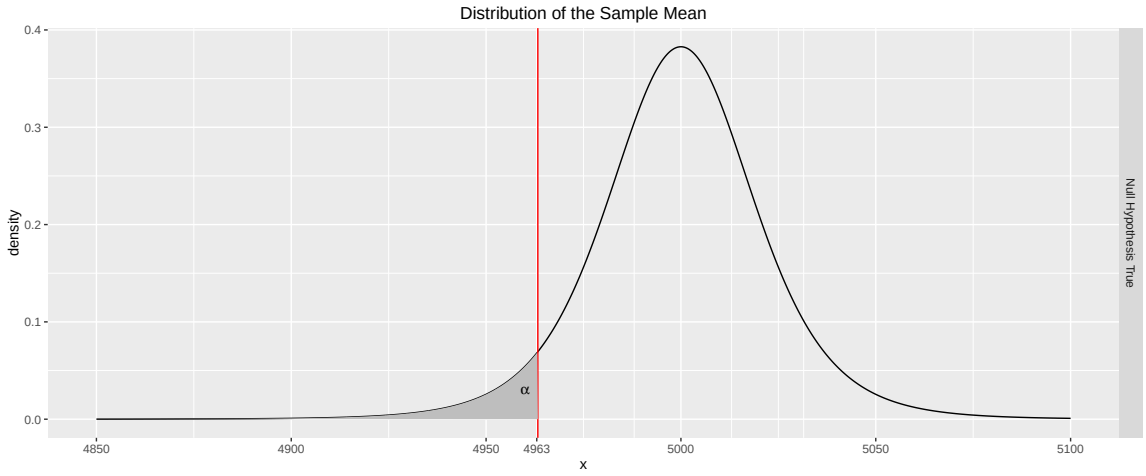
Example. Suppose that I work in Quality Control for a company that manufactures a type of rope. This rope is supposed to have a mean breaking strength of 5000 pounds and long experience with the process suggests that the standard deviation is approximately $s = 50$. As with many manufacturing processes,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

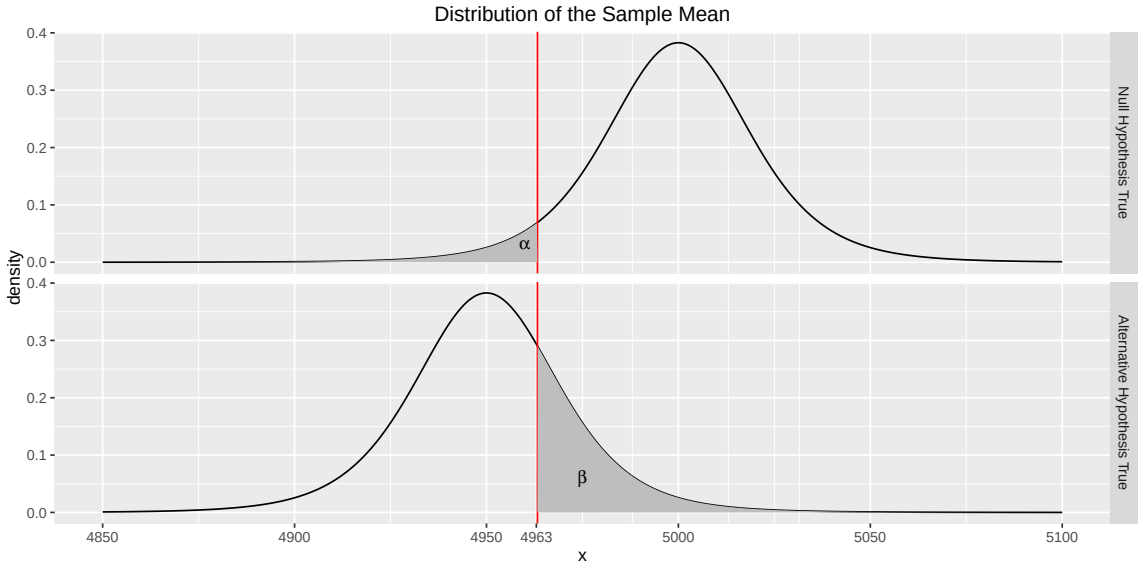
$$\begin{aligned} t_{crit} &= \frac{\bar{x}_{crit} - \mu_0}{\frac{s}{\sqrt{n}}} \\ t_{crit}\left(\frac{s}{\sqrt{n}}\right) + \mu_0 &= \bar{x}_{crit} \\ -1.943\left(\frac{50}{\sqrt{7}}\right) + 5000 &= \bar{x}_{crit} \\ 4963 &= \bar{x}_{crit} \end{aligned}$$

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 $\mu = 4950$.

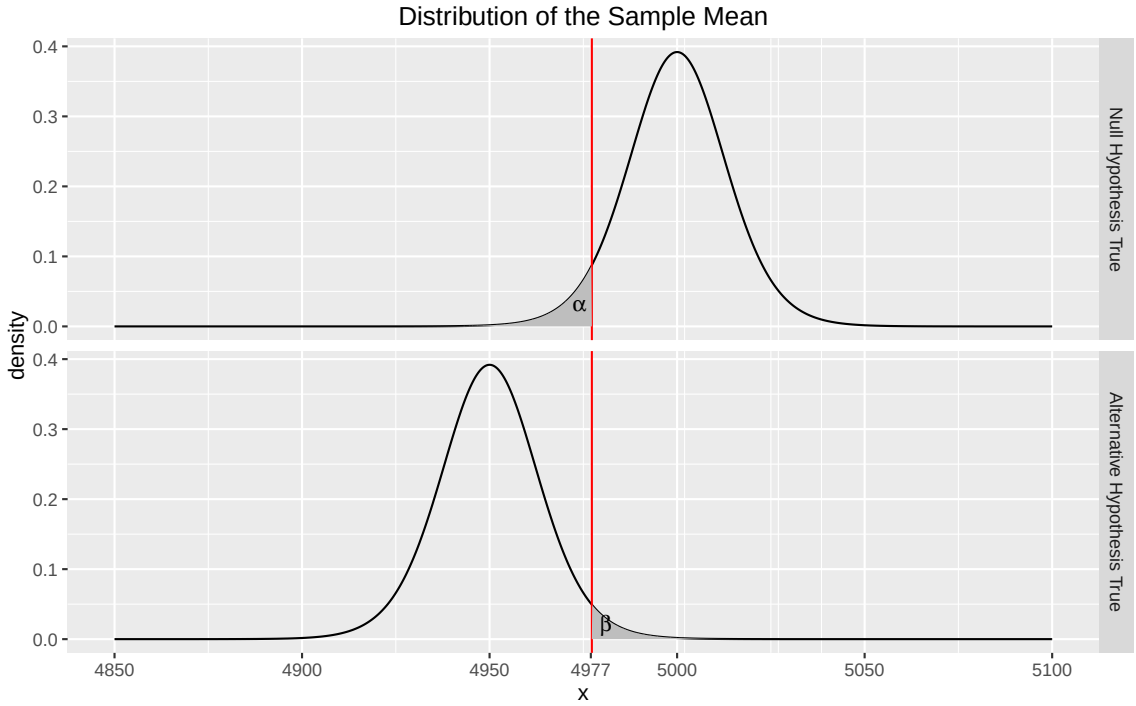


The the probability of a type II error is

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

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.[6] 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 $n = 15$.

[6]The power calculation should done using a t-distribution with non-centrality parameter instead of just shifting the distribution. The difference is slight, but is enough to cause our calculation to be slightly off.



Power calculations are relatively tedious to do by hand, but fortunately there are several very good resources for exploring how power and sample size interact. My favorite is a Java Applet web page maintained by Dr. Russ Lenth at http://www.stat.uiowa.edu/~rlenth/Power/. It will provide you a list of analysis to do the calculations for and the user is responsible for knowing that we are doing a one-sample t-test with a one-sided alternative.

Alternatively, we can do these calculations in R using the function `power.t.test()`.

Fundamentally there are five values that can be used and all power calculators will allow a user to input four of them and the calculator will calculate the fifth.

1. The difference δ from the hypothesized mean μ_0 that we wish to detect
2. The population standard deviation σ.
3. The significance level of the test α.
4. The power of the test $1 - \beta$.
5. The sample size n.

```
power.t.test(delta=50, sd=50, sig.level=0.05, n=7,
             type="one.sample", alternative="one.sided")

##
##      One-sample t test power calculation
##
##               n = 7
##           delta = 50
##              sd = 50
##       sig.level = 0.05
##           power = 0.7543959
##     alternative = one.sided
```

```
power.t.test(delta=50, sd=50, sig.level=0.05, power=0.95,
             type="one.sample", alternative="one.sided")

##
##      One-sample t test power calculation
##
##               n = 12.32052
##           delta = 50
##              sd = 50
##       sig.level = 0.05
##           power = 0.95
##     alternative = one.sided
```

The general process for selecting a sample size is to

1. Pick a α-level. Usually this is easy and people use $\alpha = 0.05$.
2. Come up with an estimate for the standard deviation σ. If you don't have an estimate, then a pilot study should be undertaken to get a rough idea what the variability is. Often this is the only good data that comes out of the first field season in a dissertation.
3. Decide how large of an effect is scientifically interesting.
4. Plug the results of steps 1-3 into a power calculator and see how large a study you need to achieve a power of 90% or 95%.

6.3 Exercises

1. One way the amount of sewage and industrial pollutants dumped into a body of water affects the health of the water is by reducing the amount of dissolved oxygen available for aquatic life. Over a 2-month period, 8 samples were taken from a river at a location 1 mile downstream from a sewage treatment plant. The amount of dissolved oxygen in the samples was determined and is reported in the following table. Current research suggests that the mean dissolved oxygen level must be at least 5.0 parts per million (ppm) for fish to survive.

Sample	1	2	3	4	5	6	7	8
Oxygen (ppm)	5.1	4.9	5.6	4.2	4.8	4.5	5.3	5.2

   (a) Use R to calculate the sample mean and standard deviation.

   (b) Using the asymptotic results and the quantities you calculated, create a 95% two-sided confidence interval for the mean dissolved oxygen level during the 2-month period. What assumption is being made for this calculation to be valid?

   (c) Calculate a 95% two-sided confidence interval using the bootstrap method. Examine the histogram of bootstrap means, does it appear normal? If so, what does that imply about the assumption you made in the calculation in the previous part?

   (d) Using the confidence interval calculated in part (b), do the data support the hypothesis that the mean dissolved oxygen level is equal to 5 ppm?

   (e) Perform a 1-sided hypothesis test that the mean oxygen level is less that 5 ppm with a significance level of $\alpha = 0.05$.

   (f) Use the function `t.test` in R to repeat the calculations you made in parts (b) and (e).

2. We are interested in investigating how accurate radon detectors sold to homeowners are. We take a randomly selection of $n = 12$ detectors and expose to 105 pico-curies per liter (pCi/l) of radon. The following values were given by the radon detectors. Do all of the following

calculations by hand (except for the calculations of the mean and standard deviation).

91.9	97.8	111.4	122.3	105.4	95.0
103.8	99.6	96.6	119.3	104.8	101.7

(a) Calculate a 90% confidence interval using the asymptotic method.

(b) State an appropriate null and alternative hypothesis for a two-sided t-test. Why is a two-sided test appropriate here?

(c) Calculate an appropriate test statistic.

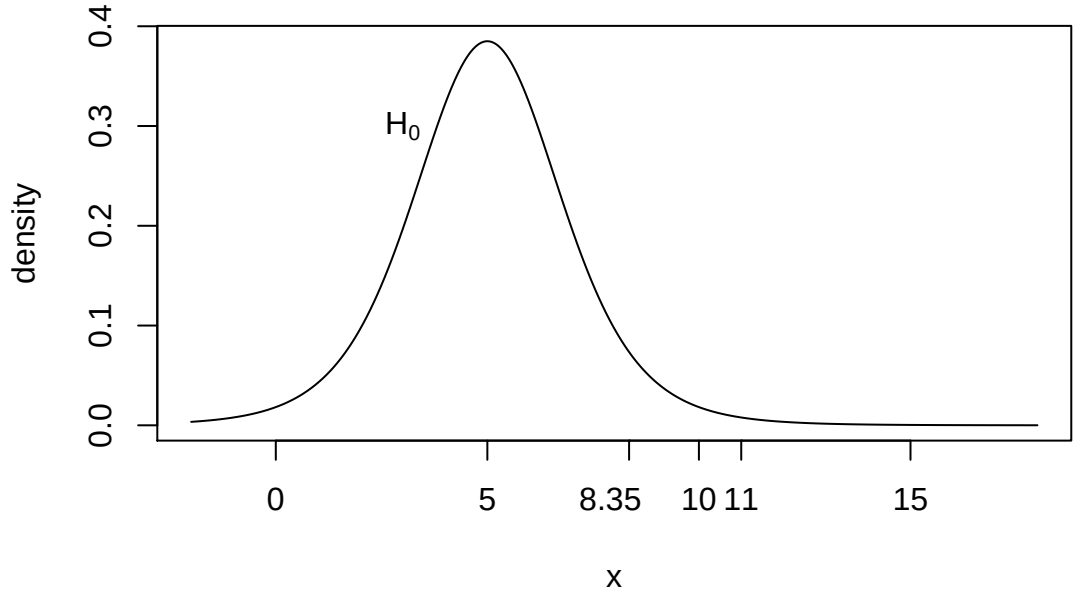
(d) Calculate a p-value.

(e) At an $\alpha = 0.10$ level, what is your conclusion. Be sure to state your conclusion in terms of the problem.

(f) Use the function `t.test()` to redo the the hand calculations you did in parts (a), (c), (d).

3. Given data such that $X_i \sim N\left(\mu, \sigma^2 = 5^2\right)$, the following graph shows the distribution of a sample mean of $n = 8$ observations under the null hypothesis $H_0 : \mu = 5$. We are interested in testing the alternative $H_a : \mu > 5$ at the $\alpha = 0.05$ level and therefore the cut off point for rejecting the null hypothesis is $t_{crit} = 1.895$ and $\bar{x}_{crit} = 1.895 * 5 + 5 = 8.35$.

(a) Add the plot of the distribution of the sample mean if $\mu = 11$ and denote which areas represent α, β, and the power in the figure below.



(b) Under the same alternative value of $\mu = 11$, find the probability of a Type II error. That is, calculate the value of $\beta = P\left(\bar{X} < 8.35 \,|\, \mu = 11\right)$.

4. A study is to be undertaken to study the effectiveness of connective tissue massage therapy on the range of motion of the hip joint for elderly clients. Practitioners think that a reasonable standard deviation of the differences (post - pre) would be $\sigma = 20$ degrees.

(a) Suppose an increase of 5 degrees in the range would be a clinically significant result. How large of a sample would be necessary if we wanted to control the Type I error rate by $\alpha = 0.1$ and the Type II error rate with $\beta = 0.1$ (therefore the power is $1-\beta = 0.90$)? Use the use the `power.t.test()` function available in the package `pwr` to find the necessary sample size.

(b) Suppose we were thought that only increases greater than 10 degrees were substantive. How large must our minimum sample size be in this case? What about for 15, 20, 25 and 30 degrees? Sketch a graph of n versus the difference to be detected and comment on how much larger a sample size must be to detect a difference half as small.

Chapter 7

Two-Sample Hypothesis Tests and Confidence Intervals

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

Example: For many years hormone (Estrogen and Progestin) replacement therapy's primary use for post-menopausal woman was to reduce the uncomfortable side-effects of menopause but it was thought to also reduced the rate of rate of breast cancer in post-menopausal women. This belief was the result of many observational studies where women who chose to take hormone replacement therapy also had reduced rates of breast cancer. The *lurking*[1] *variable* thing that the observational studies missed was that hormone therapy is relatively expensive and was taken by predominately women of a high socio- economic status. Those women tended to be more health conscious, lived in areas with less pollution, and were generally at a lower risk for developing breast cancer. Even when researchers realized that socio-economic status was *confounded*[2] with the therapy, they couldn't be sure which was the cause of the reduced breast cancer rates. To correctly test this, nearly 17,000 women underwent an experiment in which each women was randomly assigned to take either the treatment (E+P) or a placebo. The Women's Health Initiative (WHI) Estrogen plus Progestin Study[3] (E+P) was stopped on July 7, 2002 (after an average 5.6 years of follow-up) because of increased risks of cardiovascular disease and breast cancer in women taking active study pills, compared with those on placebo (inactive pills). The study showed that the overall risks exceeded the benefits, with women taking E+P at higher risk for heart disease, blood clots, stroke, and breast cancer, but at lower risk for fracture and colon cancer. Lurking variables such as income levels and education are correlated to overall health behaviors and with an increased use of hormone replacement therapy. By randomly assigning each woman to a treatment, the unidentified lurking variables were evenly spread across treatments and the dangers of hormone replacement therapy were revealed.

There is a fundamental difference between imposing treatments onto subjects versus taking a random sample from a population and observing relationships between variables. In general, designed

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

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

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

experiments allow us to determine cause-and-effect relationships while observational studies can only determine if variables are correlated. This difference in how the data is generated will result in different methods for generating a sampling distribution for a statistic of interest. In this chapter we will focus on experimental designs, though the same analyses are appropriate for observational studies.

7.1 Difference in means between two groups

Often researchers will obtain a group of subjects and them into two groups, provide different treatments to each, and the observe some response. The goal is to see if the two groups have different mean values, as this is the most common difference to be interested in.

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

Example: Finger Tapping and Caffeine

The effects of caffeine on the body have been well studied. In one experiment,[4] a group of male college students were trained in a particular tapping movement and to tap at a rapid rate. They were randomly divided into caffeine and non-caffeine groups and given approximately two cups of coffee (with either 200 mg of caffeine or none). After a 2-hour period, the students tapping rate was measured.

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

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

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

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

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

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

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

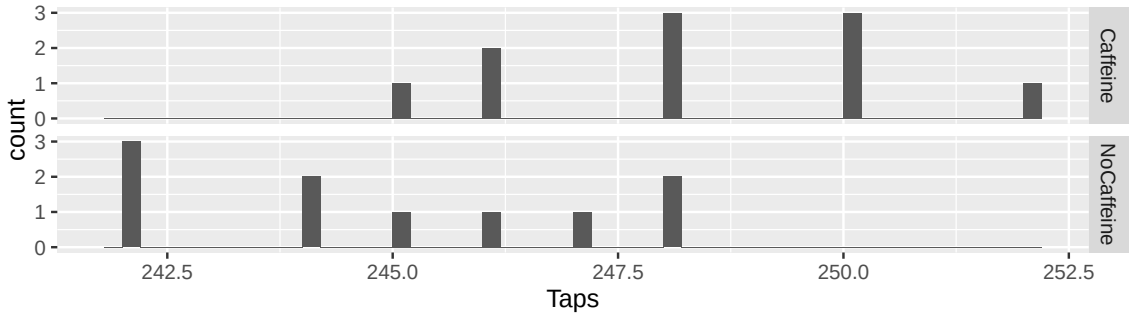
```
# Load all the libraries we'll need in this chapter
library(mosaic)    # for the do{}, shuffle(), resample() functions...
library(Lock5Data)
library(dplyr)
library(tidyr)     # spread() and gather()
library(ggplot2)
```

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

```
data(CaffeineTaps)   # load the data from the Lock5Data package
str(CaffeineTaps)

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

The first thing we should do is, as always, is graph the data.

```
ggplot(CaffeineTaps, aes(x=Taps)) +
  geom_histogram( binwidth=.2) +
  facet_grid( Group ~ . )  # two graphs stacked by Group (Caffeine vs non)
```



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

```
CaffeineTaps %>%
  group_by(Group) %>%  # group the summary stats by Treatment group
  summarise(xbar=mean(Taps), s=sd(Taps))

## Source: local data frame [2 x 3]
##
##        Group  xbar        s
##       (fctr) (dbl)    (dbl)
## 1   Caffeine 248.3 2.213594
## 2 NoCaffeine 244.8 2.394438
```

```
# No Caffeine  -  Caffeine
244.8 - 248.3

## [1] -3.5

CaffeineTaps %>% group_by(Group) %>%
  summarise(xbar=mean(Taps)) %>%
  summarise(d = diff(xbar))

## Source: local data frame [1 x 1]
##
##       d
##   (dbl)
## 1  -3.5
```

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

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

7.1.1 Inference via resampling

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

We can perform this shuffling with the following code:

```
# shuffle(): takes an input column and reorders it randomly
CaffeineTaps %>% mutate(ShuffledGroup = shuffle(Group))
```

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

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

```
CaffeineTaps %>%
  mutate( ShuffledGroup = shuffle(Group) ) %>%
  group_by( ShuffledGroup )  %>%
  summarise(xbar=mean(Taps)) %>%
  summarise(d.star = diff(xbar))
```

```
## Source: local data frame [1 x 1]
##
##   d.star
##    (dbl)
## 1    0.9
```

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

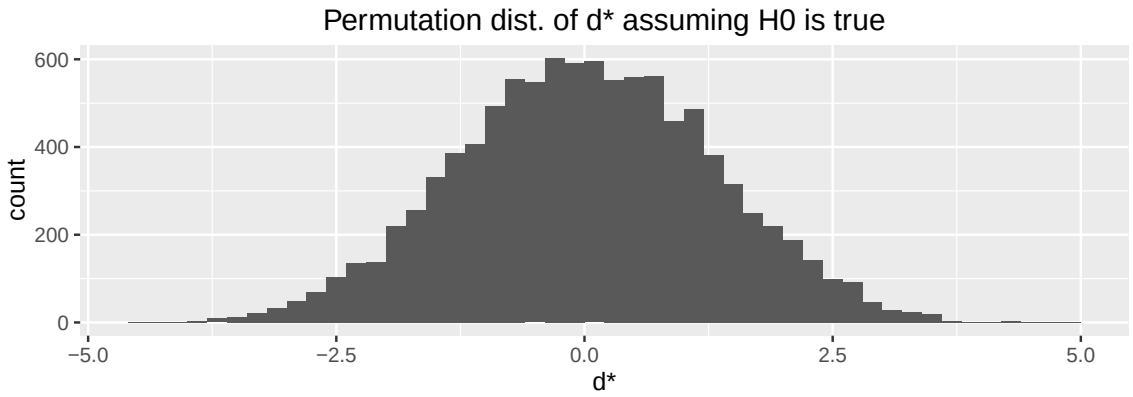
```
do(5) * {
  CaffeineTaps %>%
  mutate( ShuffledGroup = shuffle(Group) ) %>%
  group_by( ShuffledGroup )  %>%
  summarise(xbar=mean(Taps)) %>%
  summarise(d.star = diff(xbar))
}
```

```
##   d.star
## 1   -1.3
## 2   -0.1
## 3   -0.3
## 4    2.1
## 5    1.3
```

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

```
SamplingDist <- do(10000) * {
  CaffeineTaps %>%
  mutate( ShuffledGroup = shuffle(Group) ) %>%
  group_by( ShuffledGroup )  %>%
  summarise(xbar=mean(Taps)) %>%
  summarise(d.star = diff(xbar))
}
```

```
ggplot(SamplingDist, aes(x=d.star)) +
  geom_histogram(binwidth=.2) +
  ggtitle('Permutation dist. of d* assuming H0 is true') +
  xlab('d*')
```



We have almost no cases where the randomly assigned groups produced a difference as extreme as the actual observed difference of $d = -3.5$. We can calculate the percentage of the sampling distribution of the difference in means that is farther from zero

```
SamplingDist %>%
  mutate( MoreExtreme = ifelse( abs(d.star) >= 3.5, 1, 0)) %>%
  summarise( p.value1 = sum(MoreExtreme)/n(),        # these are all the
             p.value2 = mean(MoreExtreme),           # same calculation
             p.value3 = mean( abs(d.star) >= 3.5 )) # but more verbose

##   p.value1 p.value2 p.value3
## 1   0.0058   0.0058   0.0058
```

We see that only $58/10,000$ simulations of data produced assuming H_0 is true produced a d^* value more extreme than our observed difference in sample means so we can reject the null hypothesis $H_0 : \mu_{nc} - \mu_c = 0$ in favor of the alternative $H_a : \mu_{nc} - \mu_c \neq 0$ at an $\alpha = 0.05$ or any other reasonable α level.

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

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

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

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

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

```
SamplingDist %>%
  summarise( p.value = mean( d.star <= -3.5 ))

##   p.value
## 1  0.0028
```

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

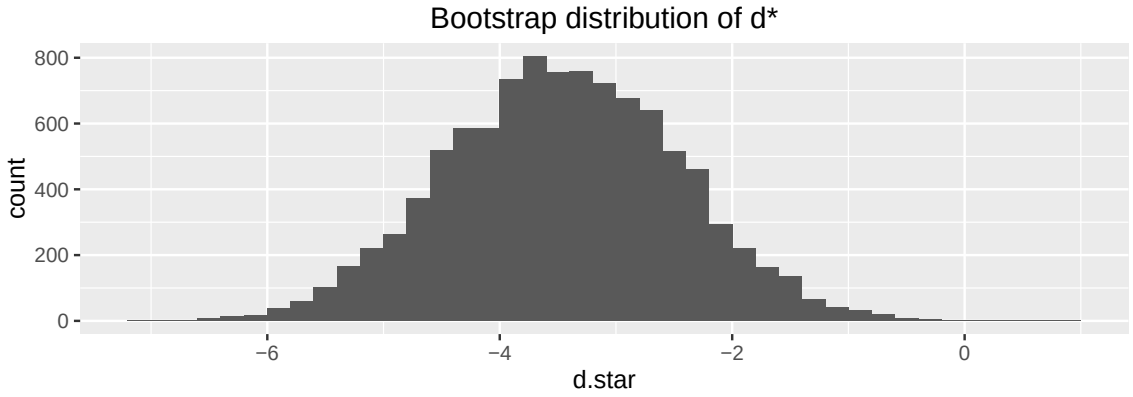
In general, we prefer to use a two-sided test because if the two-sided test leads us to reject the null hypothesis then so would the appropriate one-sided hypothesis[5]. Second, by using a two-sample test, it prevents us from from "tricking" ourselves when we don't know the which group should have a higher mean going into the experiment, but after seeing the data, thinking we should have known and using the less stringent test. Some statisticians go so far as to say that using a 1-sided test is outright fraudulent. Generally, we'll concentrate on two-sided tests as they are the most widely acceptable.

Notice that the corresponding confidence interval gives a similar inference.

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

```
BootDist <- do(10000)*{
  CaffeineTaps %>%
    group_by(Group) %>%
    resample() %>%
    summarise( xbar=mean(Taps) ) %>%
    summarise( d.star = diff(xbar)  )
}
```

```
ggplot(BootDist, aes(x=d.star)) +
  geom_histogram(binwidth=.2) +
  ggtitle('Bootstrap distribution of d*')
```




```
CI <- quantile( BootDist$d.star, probs=c(0.025, 0.975) )
CI
```

```
##      2.5%     97.5%
## -5.400000 -1.515152
```

Notice that the null hypothesis value, $\delta = 0$, is not a value supported by the data because 0 is not in the 95% confidence interval. A subtle point in the above bootstrap code is that I resampled each group separately. Because the experimental protocol was to have 10 in each group, then we want our simulated data sets should obey the same rule. Had I resampled first and then did the grouping, we might end up with 12 caffeinated and 8 un-caffeinated subjects, which is data that our experimental design couldn't have generated.

7.1.2 Inference via asymptotic results (unequal variance assumption)

Previously we've seen that the Central Limit Theorem gives us a way to estimate the distribution of the sample mean. So it should be reasonable to assume that for our two groups (1=NonCaffeine, 2=Caffeine),

$$\bar{X}_1 \dot{\sim} N\left(\mu_1, \sigma_1^2\right) \quad \text{and} \quad \bar{X}_2 \dot{\sim} N\left(\mu_2, \sigma_2^2\right)$$

It turns out that because $\bar{X}_C$ and $\bar{X}_{NC}$ both have approximately normal distributions, then the difference between them also does. This shouldn't be too surprising after looking at the permutation and bootstrap distributions of the d^* values.

So our hypothesis tests and confidence interval routine will follow a similar pattern as our one-sample tests, but we now need to figure out the correct standardization formula for the *difference in means*. The only difficulty will be figuring out what the appropriate standard deviation term $\hat{\sigma}_D$ should be.

Recall that if two random variables, A and B, are independent then

$$Var\left(A - B\right) = Var(A) + Var(B)$$

and therefore

$$\begin{aligned} Var(D) &= Var\left(\bar{X}_1 - \bar{X}_2\right) \\ &= Var\left(\bar{X}_1\right) + Var\left(\bar{X}_2\right) \\ &= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \end{aligned}$$

and finally we have

$$StdErr(D) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

and therefore my standardized value for the difference will be

$$\begin{aligned} t_{???} &= \frac{\text{estimate} - \text{hypothesized value}}{StdErr(\text{ estimate })} \\ &= \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \\ &= \frac{(-3.5) - 0}{\sqrt{\frac{2.39^2}{10} + \frac{2.21^2}{10}}} \\ &= -3.39 \end{aligned}$$

This is somewhat painful, but reasonable. The last question is what t-distribution should we compare this to? Previously we've used $df = n - 1$ but now we have *two* samples. So our degrees of freedom ought to be somewhere between $\min(n_1, n_2) - 2 = 8$ and $(n_1 + n_2) - 1 = 19$. There is no correct answer, but the best approximation to what it should be is called *Satterwaite's Approximation*.

$$df = \frac{(V_1 + V_2)^2}{\frac{V_1^2}{n_1 - 1} + \frac{V_2^2}{n_2 - 1}}$$

and

$$V_1 = \frac{s_1^2}{n_1} \text{ and } V_2 = \frac{s_2^2}{n_2}$$

So for our example we have

$$V_1 = \frac{2.39^2}{10} = 0.5712 \text{ and } V_2 = \frac{2.21^2}{10} = 0.4884$$

and

$$df = \frac{(0.5712 + 0.4884)^2}{\frac{(0.5712)^2}{9} + \frac{(0.4884)^2}{9}} = 17.89$$

So now we can compute our p-value as

$$p.value = P\left(T_{17.89} < -3.39\right)$$

```
pt(-3.39, df=17.89)
```

```
## [1] 0.00164277
```

In a similar fashion, we can calculate the confidence interval in our usual fashion

$$\text{Est} \ \pm \ t^{1-\alpha/2}_{???} \, \text{StdErr}\left(\text{ Est }\right)$$

$$(\bar{x}_1 - \bar{x}_2) \pm t^{1-\alpha/2}_{17.89} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$-3.5 \pm 2.10 \sqrt{\frac{2.39^2}{10} + \frac{2.21^2}{10}}$$

$$-3.5 \pm 2.16$$

$$(-5.66, \ -1.34)$$

It is probably fair to say that this is an ugly calculation to do by hand. Fortunately it isn't too hard to make R do these calculations for you. The function `t.test()` will accept two arguments, a vector of values from the first group and a vector from the second group.

```
Caffeine    <- CaffeineTaps$Taps[ 1:10]  # first 10 are Caffeine
NonCaffeine <- CaffeineTaps$Taps[11:20]  # last 10 are non-Caffeine

# Do the t-test
t.test( NonCaffeine, Caffeine )

##
##  Welch Two Sample t-test
##
## data:  NonCaffeine and Caffeine
## t = -3.3942, df = 17.89, p-value = 0.003255
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -5.667384 -1.332616
## sample estimates:
## mean of x mean of y
##     244.8     248.3
```

7.1.3 Inference via asymptotic results (equal variance assumption)

In the `CaffeineTaps` example, the standard deviations of each group are quite similar. Instead of thinking of the data as

$$\bar{X}_1 \dot{\sim} N\left(\mu_1, \, \sigma_1^2\right) \quad \text{and} \quad \bar{X}_2 \dot{\sim} N\left(\mu_2, \, \sigma_2^2\right)$$

we could consider the model where we assume that the variance term is the same for each sample.

$$\bar{X}_1 \dot{\sim} N\left(\mu_1, \, \sigma^2\right) \quad \text{and} \quad \bar{X}_2 \dot{\sim} N\left(\mu_2, \, \sigma^2\right)$$

First, we can estimate μ_1 and μ_2 with the appropriate sample means $\bar{x}_1$ and $\bar{x}_2$. Next we need to calculate an estimate of σ using all of the data. First recall the formula for the sample variance for one group was

$$s^2 = \frac{1}{n-1}\left[\sum_{j=1}^{n} \left(x_j - \bar{x}\right)^2\right]$$

In the case with two samples, we want a similar formula but it should take into account data from both sample groups. Define the notation x_{1j} to be the jth observation of group 1, and x_{2j}

to be the jth observation of group 2 and in general x_{ij} as the jth observation from group i. We want to subtract each observation from the its appropriate sample mean and then, because we had to estimate two means, we need to subtract two degrees of freedom from the denominator.

$$\begin{aligned} s^2_{pooled} &= \frac{1}{n_1+n_2-2}\left[\sum_{j=1}^{n_1}(x_{1j}-\bar{x}_1)^2+\sum_{j=1}^{n_2}(x_{2j}-\bar{x}_2)^2\right] \\ &= \frac{1}{n_1+n_2-2}\left[\sum_{j=1}^{n_1}e_{1j}^2+\sum_{j=1}^{n_2}e_{2j}^2\right] \\ &= \frac{1}{n_1+n_2-2}\left[\sum_{i=1}^{2}\sum_{j=1}^{n_i}e_{ij}^2\right] \end{aligned}$$

where $\bar{x}_1$ and $\bar{x}_2$ are the sample means and $e_{ij} = x_{ij} - \bar{x}_i$ is the residual error of the i, j observation. A computationally convenient formula for this same quantity is

$$s^2_{pooled} = \frac{1}{n_1+n_2-2}\left[(n_1-1)\,s_1^2+(n_2-1)\,s_2^2\right]$$

Finally we notice that this pooled estimate of the variance term σ^2 has $n_1 + n_2 - 2$ degrees of freedom. One benefit of the pooled procedure is that we don't have to mess with the Satterthwaite's approximate degrees of freedom.

Recall our test statistic in the unequal variance case was

$$t_{???} = \frac{(\bar{x}_1-\bar{x}_2)-0}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$$

but in the equal variance case, we will use the pooled estimate of the variance term s^2_{pooled} instead of s_1^2 and s_2^2. So our test statistic becomes

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

where we have $StdErr\left(\bar{X}_1-\bar{X}_2\right) = s_{pooled}\sqrt{(1/n_1)+(1/n_2)}$. For the `CaffeineTaps` data, this results in the following analysis for

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

First we have to calculate the summary statistics (along with the pooled σ_{pooled}).

```
CaffeineTaps %>%
  group_by(Group) %>%
  summarise(xbar.i = mean(Taps),    # sample mean for each group
            s2.i   = var(Taps),     # sample variances for each group
            s.i    = sd(Taps),      # sample standard deviations for each group
            n.i    = n()       )    # sample sizes for each group

## Source: local data frame [2 x 5]
##
##        Group xbar.i     s2.i      s.i   n.i
##       (fctr)  (dbl)    (dbl)    (dbl) (int)
## 1   Caffeine  248.3 4.900000 2.213594    10
## 2 NoCaffeine  244.8 5.733333 2.394438    10

CaffeineTaps %>%
  group_by(Group) %>%
  summarize(  n.i = n(),
              s2.i = var(Taps) ) %>%
  summarize( s2.p = sum( (n.i-1)*s2.i ) / ( sum(n.i)-2 ),
             s.p  = sqrt(s2.p) )

## Source: local data frame [1 x 2]
##
##       s2.p     s.p
##      (dbl)   (dbl)
## 1 5.316667 2.30579
```

Next we can calculate

$$t_{18} = \frac{(244.8 - 248.3) - 0}{2.31\sqrt{\frac{1}{10} + \frac{1}{10}}} = -3.39$$

```
p.value <- 2 * pt(-3.39, df=18)    # 2-sided test, so multiply by 2
p.value

## [1] 0.003262969
```

The associated 95% confidence interval is

$$(\bar{x}_1 - \bar{x}_2) \pm t^{1-\alpha/2}_{n_1+n_2-2} \left(s_{pool}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right)$$

```
qt( .975, df=18 )

## [1] 2.100922
```

$$-3.5 \pm 2.10 \left(2.31\sqrt{\frac{1}{10} + \frac{1}{10}} \right)$$

$$-3.5 \pm 2.17$$

$$(-5.67,\ -1.33)$$

This p-value and 95% confidence interval are quite similar to the values we got in the case where we assumed unequal variances.

As usual, these calculations are pretty annoying to do by hand and we wish to instead do them using R. Again the function `t.test()` will do the annoying calculations for us.

```
Caffeine    <- CaffeineTaps$Taps[ 1:10]  # first 10 are Caffeine
NonCaffeine <- CaffeineTaps$Taps[11:20]  # last 10 are non-Caffeine

# Do the t-test
t.test( NonCaffeine, Caffeine, var.equal=TRUE )

##
##  Two Sample t-test
##
## data:  NonCaffeine and Caffeine
## t = -3.3942, df = 18, p-value = 0.003233
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -5.66643 -1.33357
## sample estimates:
## mean of x mean of y
##     244.8     248.3
```

```
# 99% confidence interval instead
t.test( NonCaffeine, Caffeine, var.equal=TRUE, conf.level=.99 )

##
##  Two Sample t-test
##
## data:  NonCaffeine and Caffeine
## t = -3.3942, df = 18, p-value = 0.003233
## alternative hypothesis: true difference in means is not equal to 0
## 99 percent confidence interval:
##  -6.4681918 -0.5318082
## sample estimates:
## mean of x mean of y
##     244.8     248.3
```

Example - Does drinking beer increase your attractiveness to mosquitos?

In places in the country substantial mosquito populations, the question of whether drinking beer causes the drinker to be more attractive to the mosquitoes than drinking something else has plagued campers. To answer such a question, researchers[6] conducted a study to determine if drinking beer attracts more mosquitoes than drinking water. Of $n = 43$ subjects, $n_b = 25$ drank a liter beer and $n_w = 18$ drank a liter of water and mosquitoes were caught in traps as they approached the different subjects. The critical part of this study is that the treatment (beer or water) was randomly assigned to each subject.

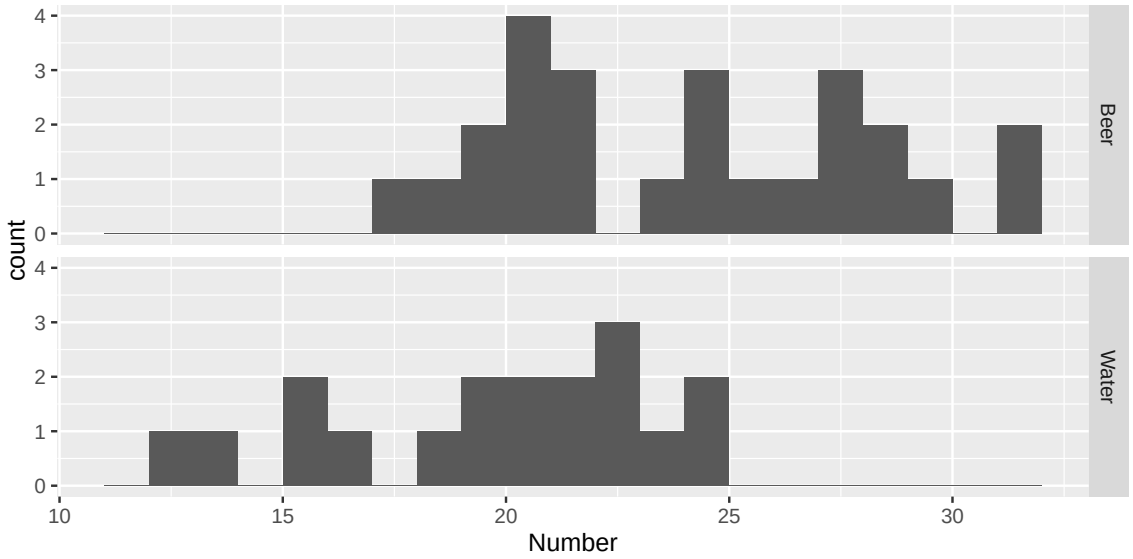
For this study, we want to test

$$H_0 : \delta = 0 \qquad \text{vs} \qquad H_a : \delta < 0$$

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

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

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

# Plot the data
ggplot(Mosquitoes, aes(x=Number)) +
  geom_histogram(binwidth=1) +
  facet_grid( Treat ~ . )
```



For this experiment and the summary statistic that captures the difference we are trying to understand is

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

where $\bar{x}_w$ is the sample mean number of mosquitoes attracted by the water group and $\bar{x}_b$ is the sample mean number of mosquitoes attracted by the beer group. Because of the order we chose for the subtraction, a negative value for d is supportive of the alternative hypothesis that mosquitoes are more attracted to beer drinkers.

```
Mosquitoes %>% group_by(Treat) %>%
  summarise(xbar.i = mean(Number),
            s2.i   = var(Number),
            s.i    = sd(Number),
            n.i    = n())

## Source: local data frame [2 x 5]
##
##    Treat   xbar.i     s2.i      s.i   n.i
##   (fctr)    (dbl)    (dbl)    (dbl) (int)
## 1   Beer 23.60000 17.08333 4.133199    25
## 2  Water 19.22222 13.47712 3.671120    18
```

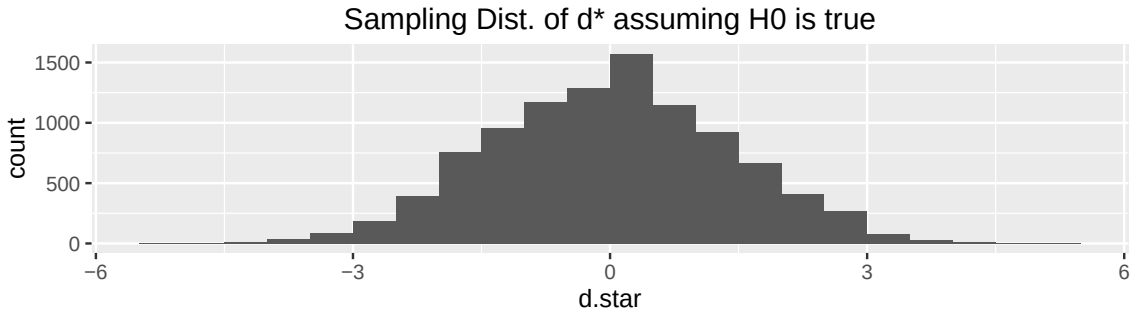
Here we see that our statistic of interest is

$$\begin{aligned} d &= \bar{x}_w - \bar{x}_b \\ &= 19.22 - 23.6 \\ &= -4.37\bar{7} \end{aligned}$$

The hypothesis test and confidence interval to see if this is statistically significant evidence to conclude that beer increases attractiveness to mosquitos is as follows. First we perform the hypothesis test by creating the sampling distribution of d^* assuming H_0 is true by repeatedly shuffling the group labels and calculating differences.

```
SamplingDist <- do(10000) *{
  Mosquitoes %>%
    mutate(ShuffledTreat = shuffle(Treat)) %>%
    group_by( ShuffledTreat )              %>%
    summarise( xbar.i = mean(Number) )     %>%
    summarise( d.star = diff(xbar.i) )
}
```

```
ggplot(SamplingDist, aes(x=d.star)) +
  geom_histogram(binwidth=.5) +
  ggtitle('Sampling Dist. of d* assuming H0 is true')
```

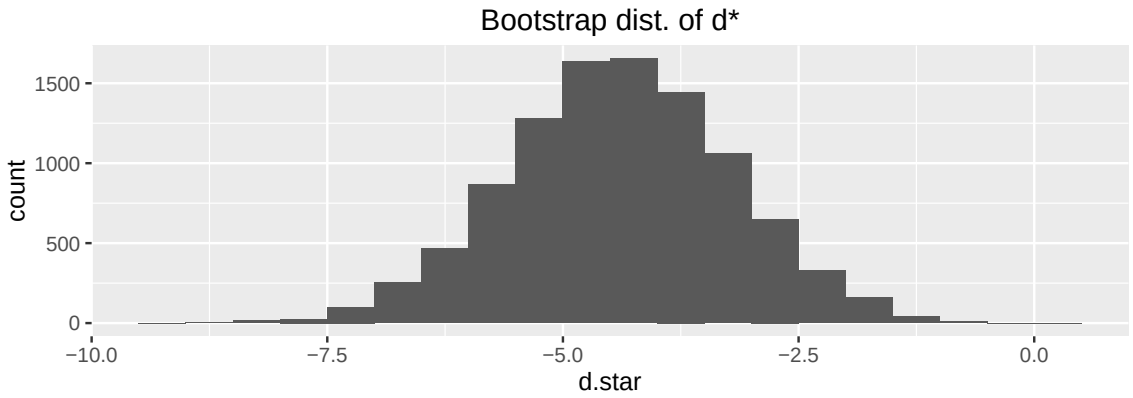



```
p.value <- SamplingDist %>%
  summarise( mean( d.star <= -4.377 ))
p.value

##   mean(d.star <= -4.377)
## 1                  2e-04
```

The associated confidence interval (lets do a 90% confidence level), is created via bootstrapping.

```
BootDist <- do(10000)*{
  Mosquitoes %>%
    group_by(Treat)                      %>%
    resample()                           %>%
    summarise( xbar.i = mean(Number) )   %>%
    summarise( d.star = diff(xbar.i) )
}
```

```
ggplot(BootDist, aes(x=d.star)) +
  geom_histogram(binwidth=.5) +
  ggtitle('Bootstrap dist. of d*')
```




```
quantile( BootDist$d.star, probs=c(.05, .95))
```

```
##        5%       95%
## -6.336262 -2.449065
```

The calculated p-value is extremely small (p = 2e-04 $\leq \alpha = 0.10$) and the associated two-sided 90% confidence interval doesn't contain 0, so we can conclude that the choice of drink does cause a change in attractiveness to mosquitoes.

If we wanted to perform the same analysis using asymptotic methods we could do the calculations by hand, or just use R.

```
# the package mosaic augments the t.test() function to
# input the data in a more convenient fashion.
t.test( Number ~ Treat, data=Mosquitoes,
        var.equal=TRUE,
        conf.level=0.90)
```

```
##
##  Two Sample t-test
##
## data:  Number by Treat
## t = 3.587, df = 41, p-value = 0.0008831
## alternative hypothesis: true difference in means is not equal to 0
## 90 percent confidence interval:
##  2.323889 6.431666
## sample estimates:
##  mean in group Beer mean in group Water
##            23.60000            19.22222
```

Notice that we didn't specify the order for the subtraction and the `mosaic:t.test()` function did the subtraction in the opposite order than we did. The p-value is slightly different but doesn't change the resulting inference.