

Introduction to Statistical Methodology

Derek L. Sonderegger

2017-01-20

Contents

1	Summary Statistics and Graphing	5
1.1	Graphical summaries of data	6
1.2	Measures of Centrality	12
1.3	Measures of Variation	13
1.4	Exercises	17
2	Probability	21
2.1	Introduction to Set Theory	21
2.2	Probability Rules	24
2.3	Discrete Random Variables	28
2.4	Common Discrete Distributions	31
2.5	Continuous Random Variables	37
2.6	Exercises	43
3	Confidence Intervals via Bootstrapping	45
3.1	Theory of Bootstrapping	45
3.2	Quantile-based Confidence Intervals	48
3.3	Exercises	53

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.

```
# Every chapter, we will load all the libraries we will use at the beginning  
# of the chapter. If you  
library(mosaicData) # library of datasets we'll use  
library(ggplot2)    # graphing functions  
library(dplyr)      # data summary tools
```

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.

```
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. 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." 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.

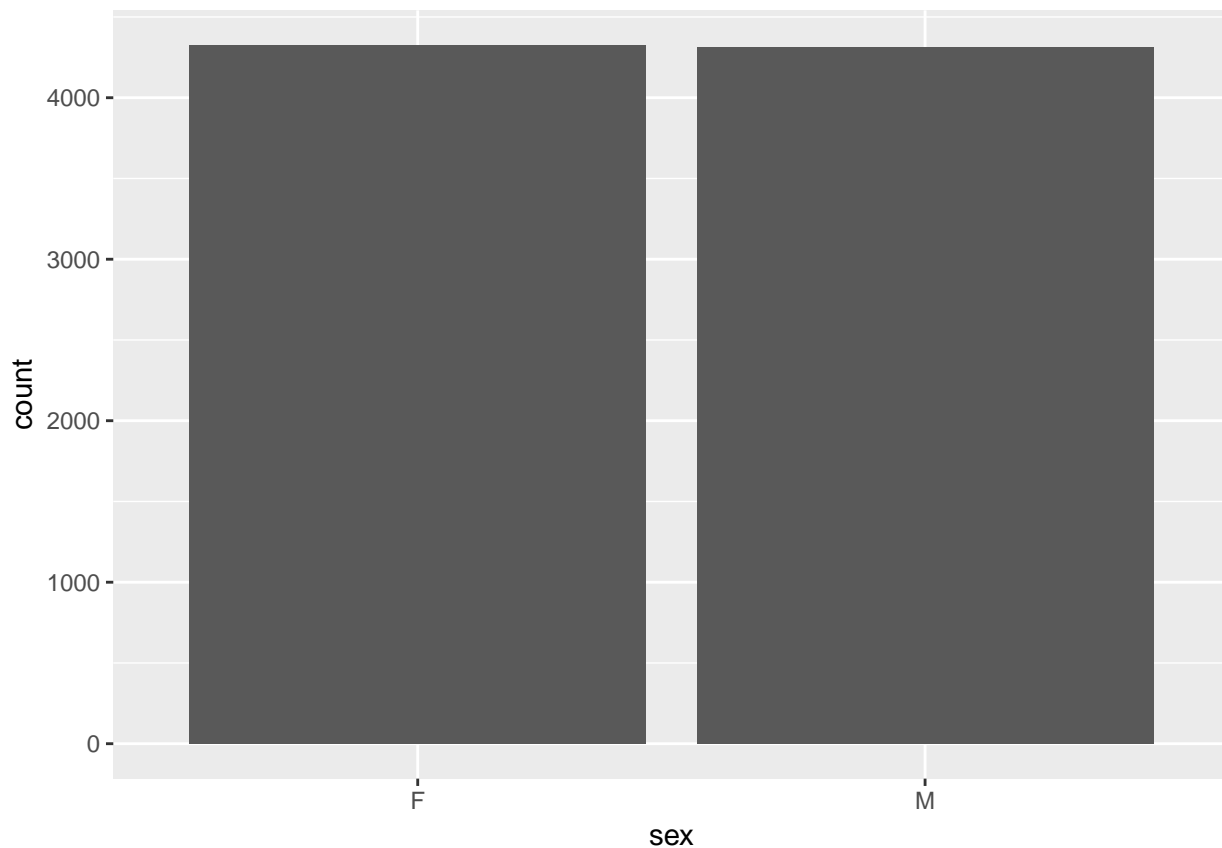
To decide if a data attribute is discrete or continuous, I often ask "Does a fraction of a value make sense?" If so, then the data is continuous.

1.1 Graphical summaries of data

1.1.1 Univariate - Categorical

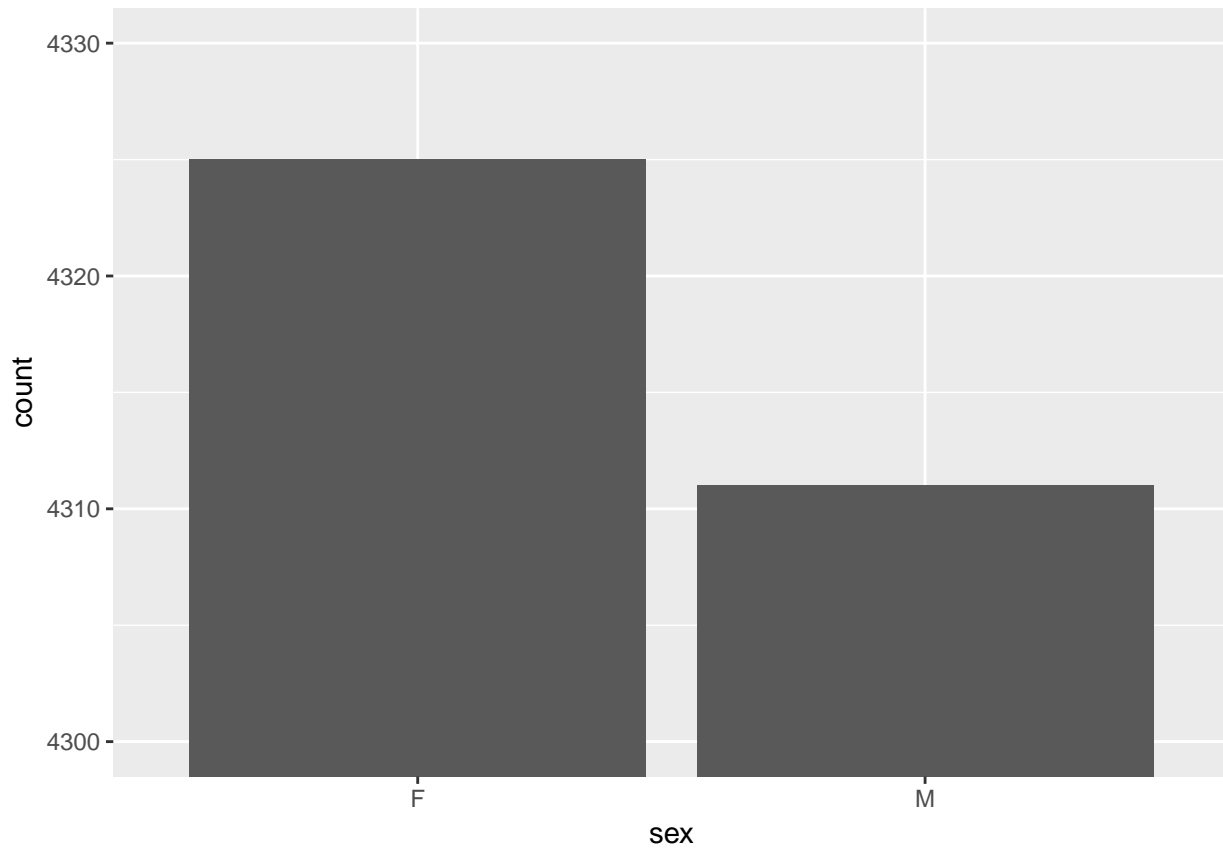
If we have univariate data about a number of groups, often the best way to display it is using barplots. They have the advantage over pie-charts that groups are easily compared.

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



One thing that can be misleading is if the zero on the y-axis is removed. In the following graph it looks like there are twice as many female runners as male until you examine the y-axis closely. In general, the following is a very misleading graph.

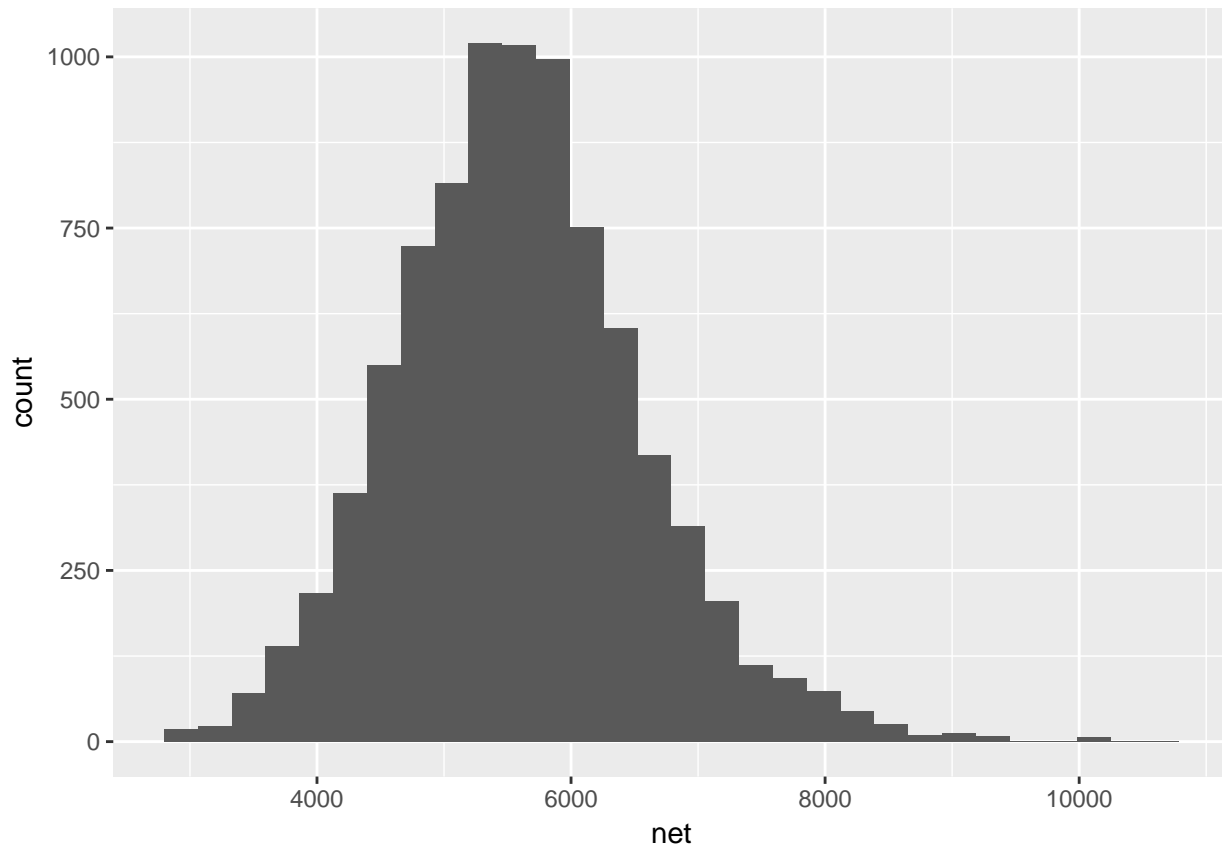
```
ggplot(TenMileRace, aes(x=sex)) +  
  geom_bar() +  
  coord_cartesian(ylim = c(4300, 4330))
```



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.

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



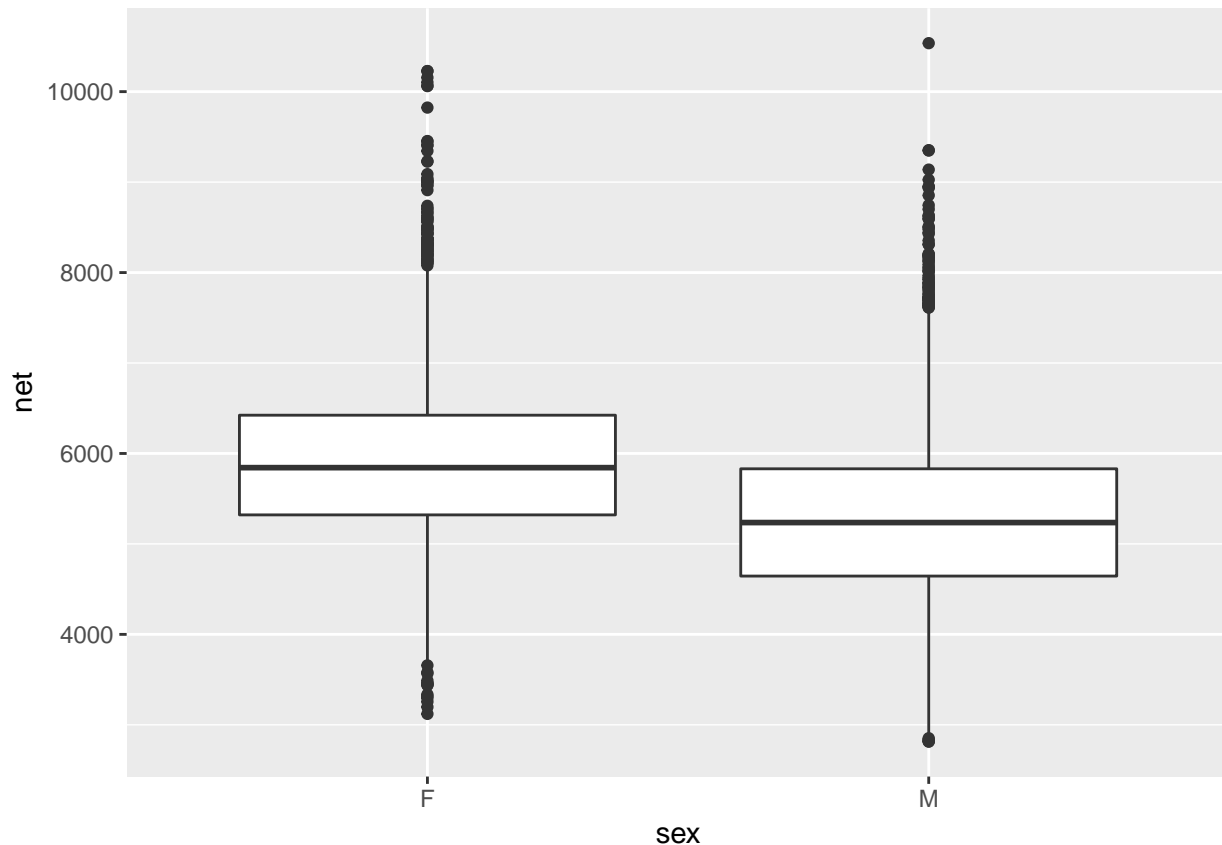
Often when a histogram is presented, the y-axis is labeled as “frequency” or “count” which is the number of observations that fall within a particular bin. However, it is often desirable to scale the y-axis so that if we were to sum up the area (*height*width*) then the total area would sum to 1. The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

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

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.

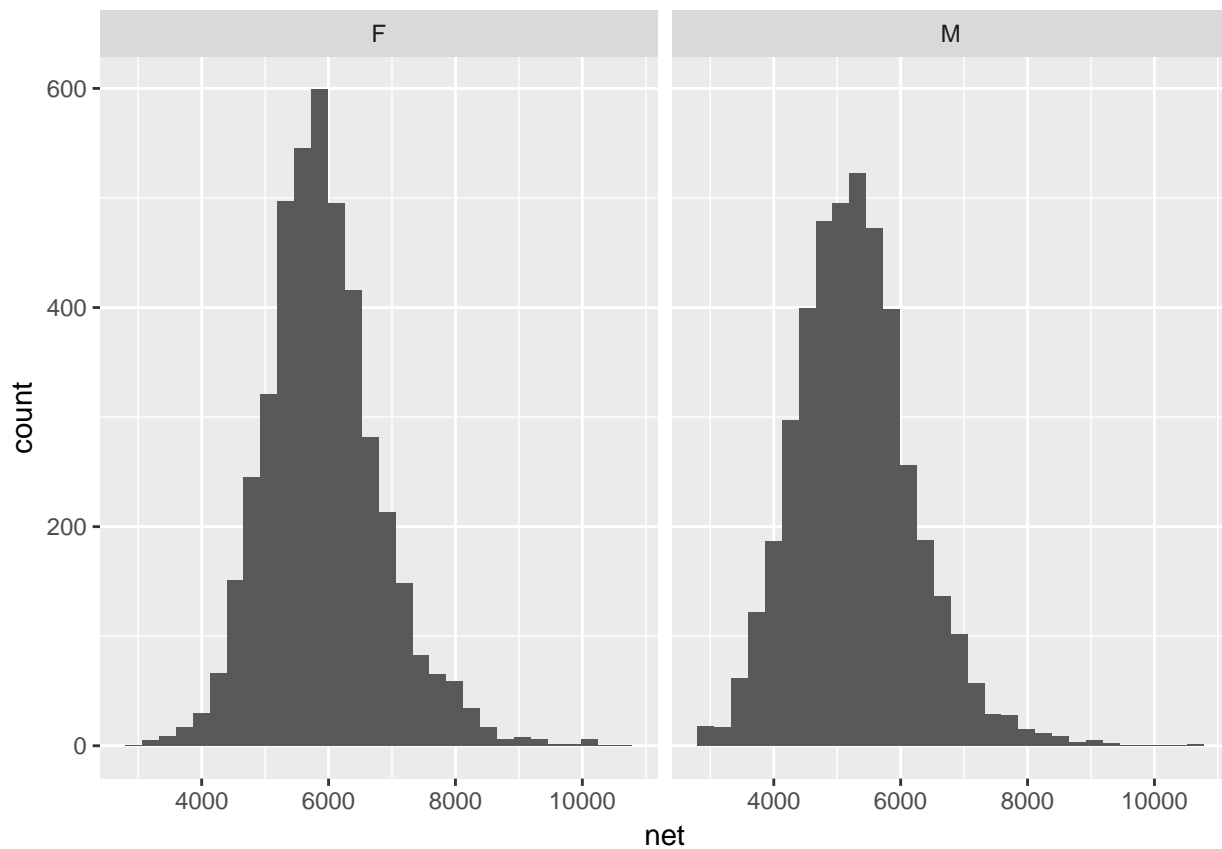
```
ggplot(TenMileRace, aes(x=sex, y=net)) + geom_boxplot()
```

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. Define the Inter-Quartile Range (IQR) as the length of the box. Then any observation more than $1.5 \times \text{IQR}$ from the box is considered an outlier.

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.

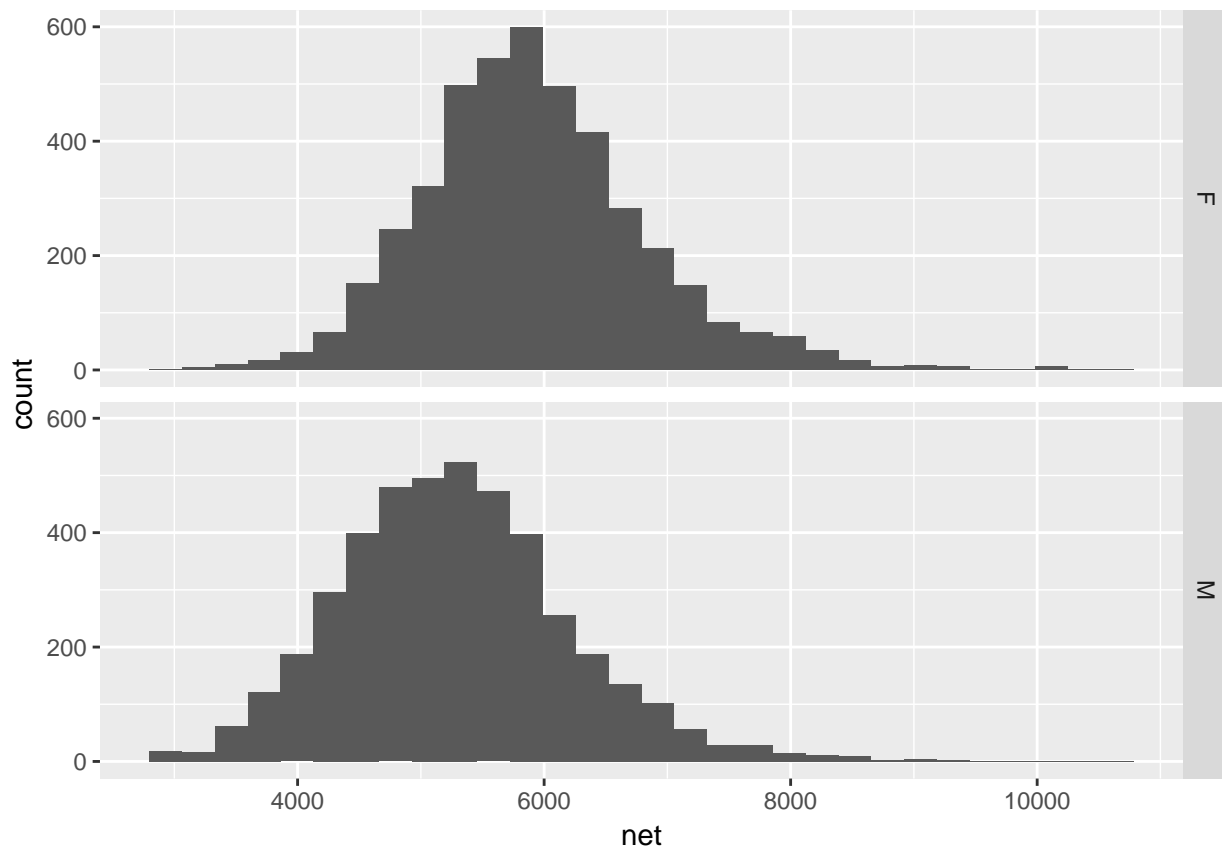
```
ggplot(TenMileRace, aes(x=net)) +
  geom_histogram() +
  facet_grid( . ~ sex ) # side-by-side plots based on sex
```



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

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

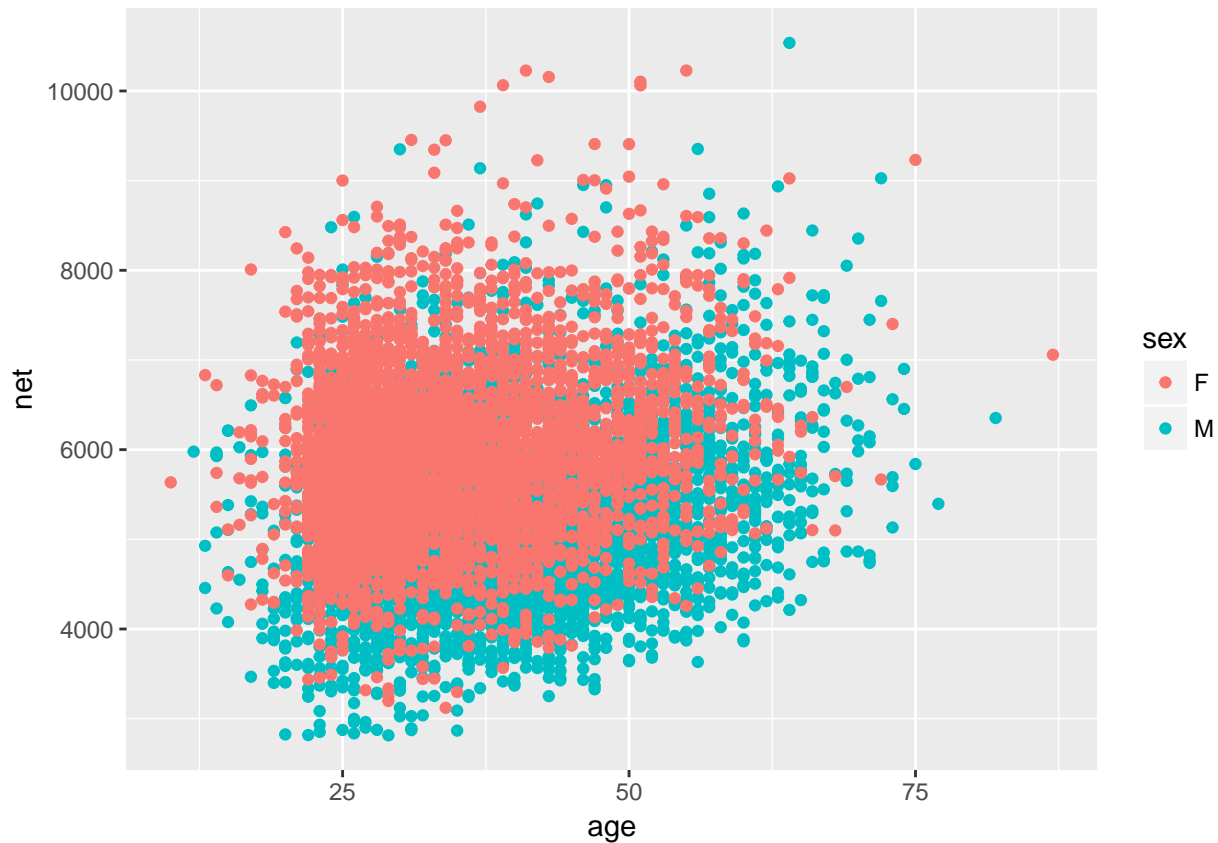
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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.

1.2.1 Mean

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

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

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

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

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
```

1.2.2 Median

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

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

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

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

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

1.2.3 Mode

This is the observation value with the most number of occurrences. This measure of “center” is not often used

1.2.4 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.

1.3.1 Range

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

```
#vmax( TenMileRace$net ) - min( TenMileRace$net )
TenMileRace %>% summarise( range = max(net) - min(net) )
```

```
## range
## 1 7722
```

1.3.2 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.

```
# IQR( TenMileRace$net )
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.

1.3.3 Variance

One way to measure the spread of a distribution is to ask “what is the typical distance of an observation to the mean?” We could define the i th 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.

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

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. 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. 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 because 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 that quantity, you only need $n - 1$ data points and can then figure out the last one, we have used one degree of freedom on estimating the mean and we need to adjust the formula accordingly.

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

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

and the Mean Squared Error as

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

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

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

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

and so the sample variance is

$$s^2 = \frac{SSE}{n-1} = \frac{23.2}{(n-1)} = \frac{23.2}{4} = 5.8$$

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) )
# var( ToyData$x )
```

```
ToyData %>% summarise( s2 = var(x) )
```

```
##      s2
## 1 5.8
```

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

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

```
##      s2
## 1 940233.5
```

1.3.4 Standard Deviation

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

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

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

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

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

```
##      s
## 1 969.6564
```

1.3.5 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) )
```

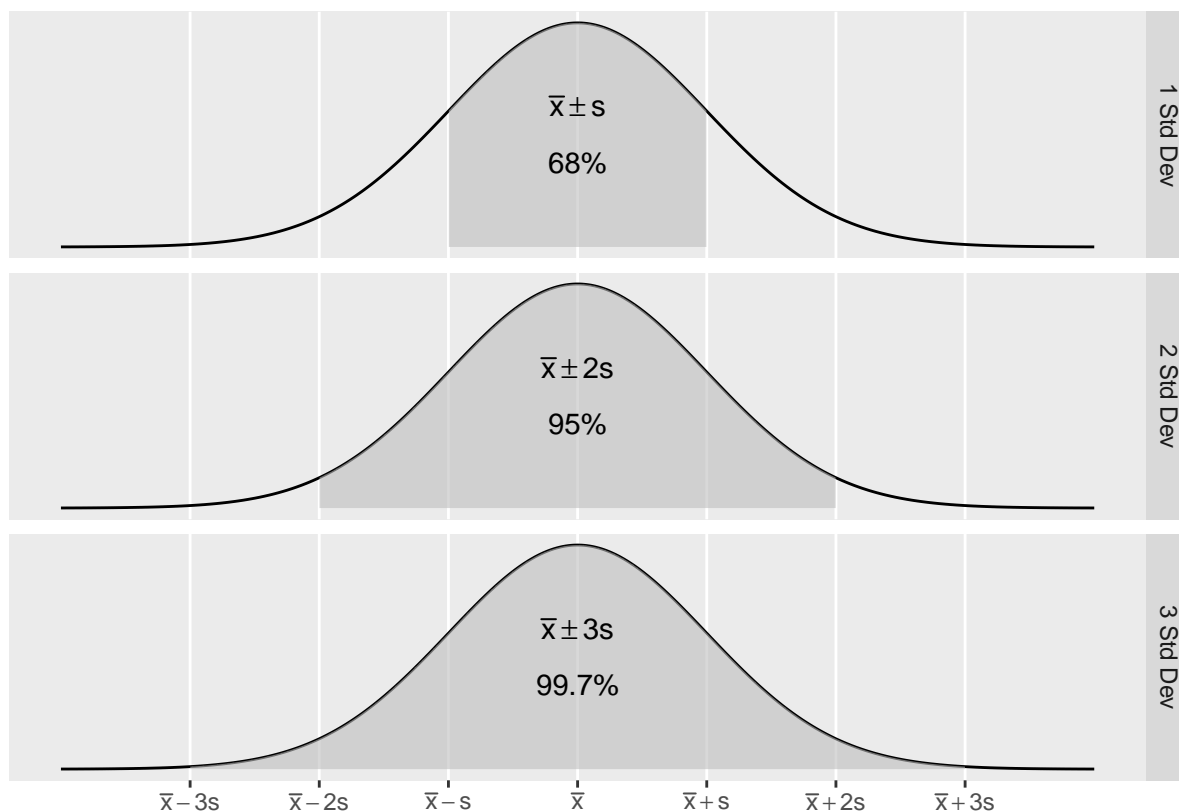


```
## # A tibble: 2 × 4
##   sex    xbar      s      cv
##   <fctr> <dbl> <dbl> <dbl>
## 1     F 5916.398 902.1090 0.1524761
## 2     M 5280.702 929.9817 0.1761095
```

1.3.6 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 use 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.

Source Type	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

- By hand, compute the mean and median separately for each type of feeder.
 - Using your results from part (a), comment on the relative sensitivity of the mean and median to extreme values in a data set.
 - 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.

Hotel Type	Nightly Rate
Luxury	\$175, \$180, \$120, \$150, \$120, \$125
Budget	\$50, \$50, \$49, \$45, \$36, \$45, \$50, \$50, \$40

- By hand, compute the means and standard deviations of the room rates for each class of hotel.
 - 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, the `group_by()` command we demonstrated earlier in the chapter is convenient.

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)) )

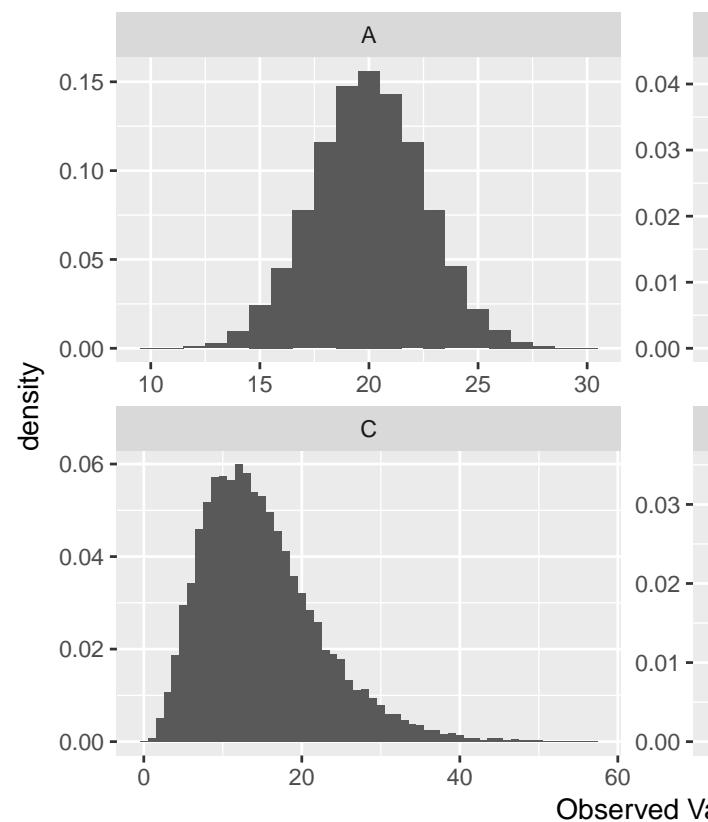
# Print out some of the data to confirm what the column names are

```

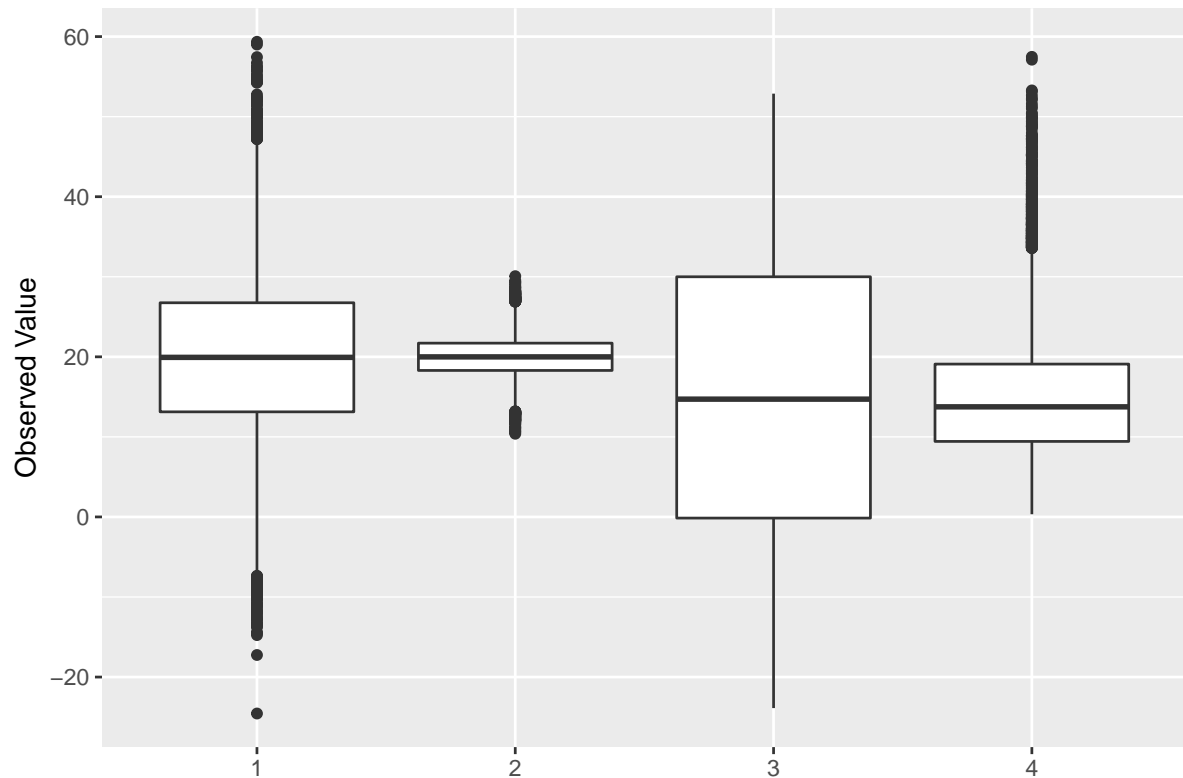
```
head( Hotels )
```

```
##   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

```
# Every chapter, we will load all the librarys we will use at the beginning  
# of the chapter.  
library(ggplot2)    # graphing functions  
library(dplyr)      # data summary tools
```

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.

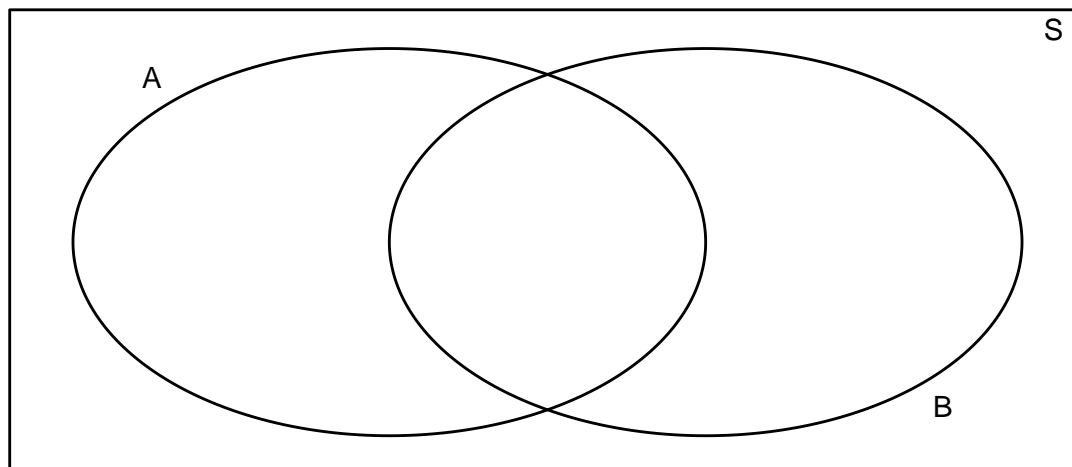
2.1 Introduction to Set Theory

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

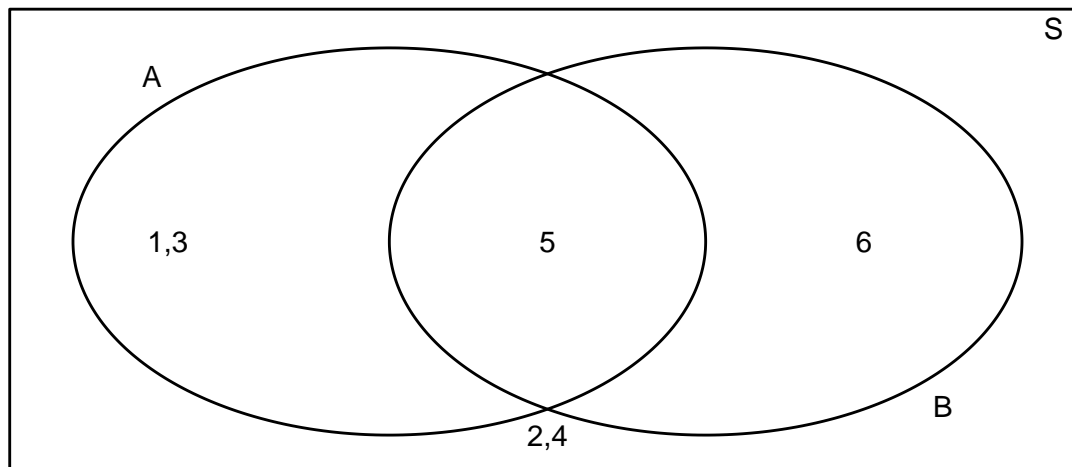
Events are properties of a particular outcome. For a coin flip, the event “Heads” would be the event that a heads was flipped. For the single roll of a six-sided die, a possible event might be that the result is even. For the NAU student, we might be interested in the event that the student is a biology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

1.1.1 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 $S = \{1, 2, 3, 4, 5, 6\}$. Suppose I then define events $A = \text{roll is odd}$ and $B = \text{roll is 5 or greater}$. In this case our picture is:

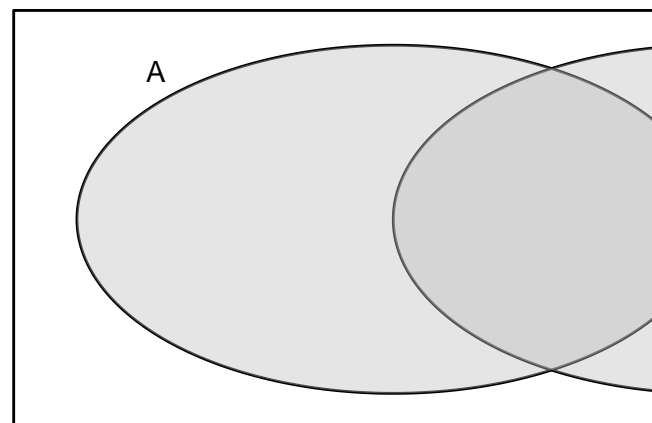


All of our possible events are present, and distributed amongst our possible events.

1.1.2 Composition of events

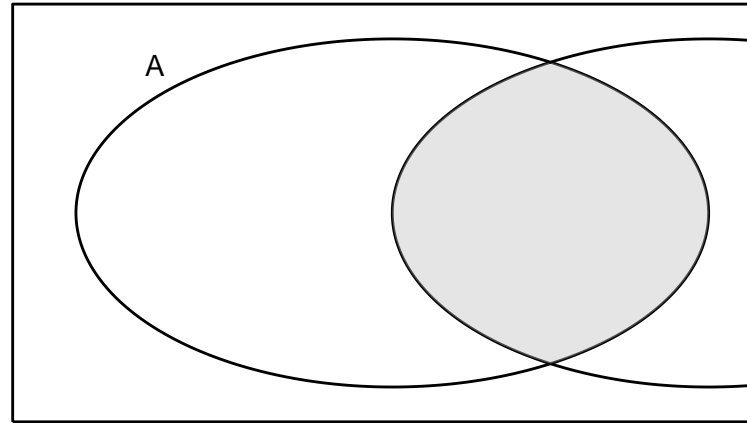
I am often interested in discussing the composition of two events and we give the common set operations below.

$A \cup B$

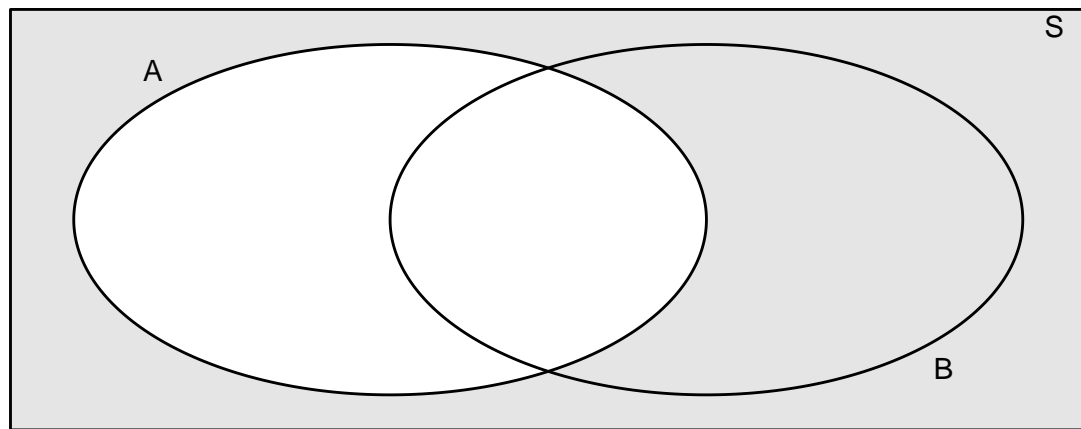


- Union: Denote the event that either A or B occurs as $A \cup B$.

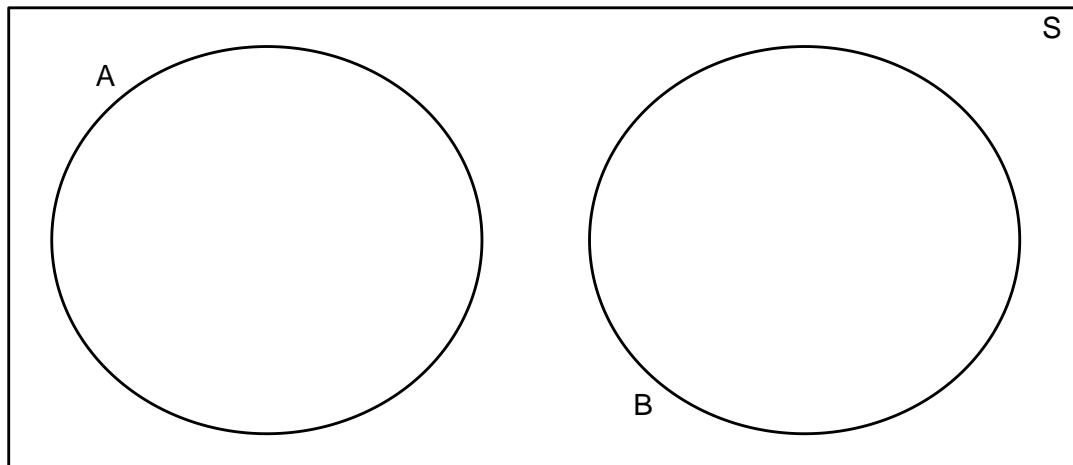
$$A \cap B$$



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



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.

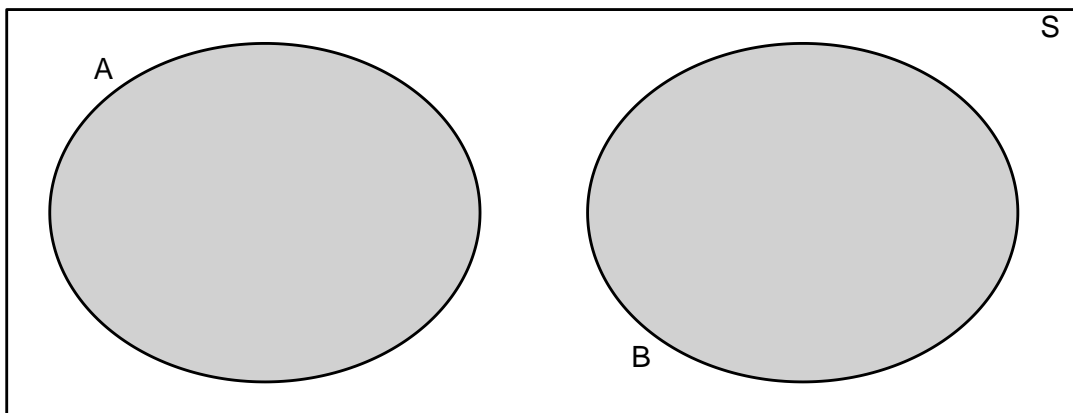
Definition 2. Probability is the proportion of times an event occurs in many repeated trials of a random phenomenon. In other words, probability is the long-term relative frequency.

Fact. For any event A the probability of the event $P(A)$ satisfies $0 \leq P(A) \leq 1$ because 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 represented by the area in the circle A .

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

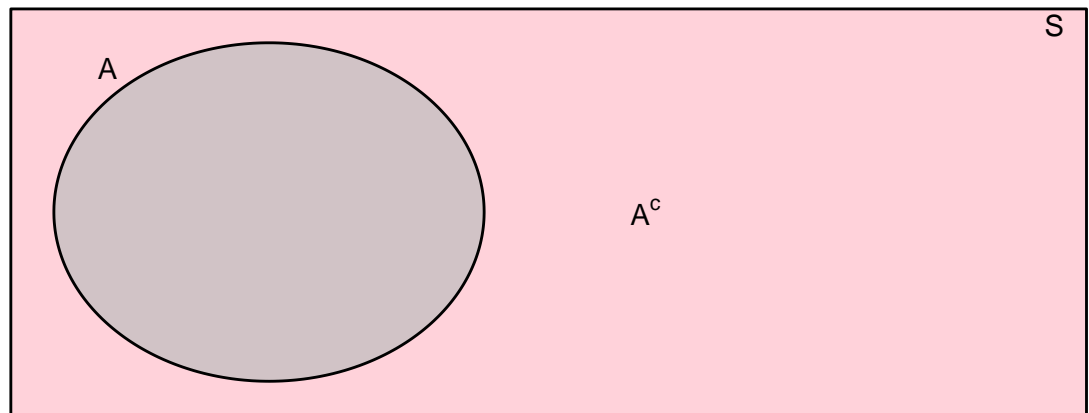
$$P(A \cup B) = P(A) + P(B)$$



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

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

$$P(A) + P(A^c) = 1$$



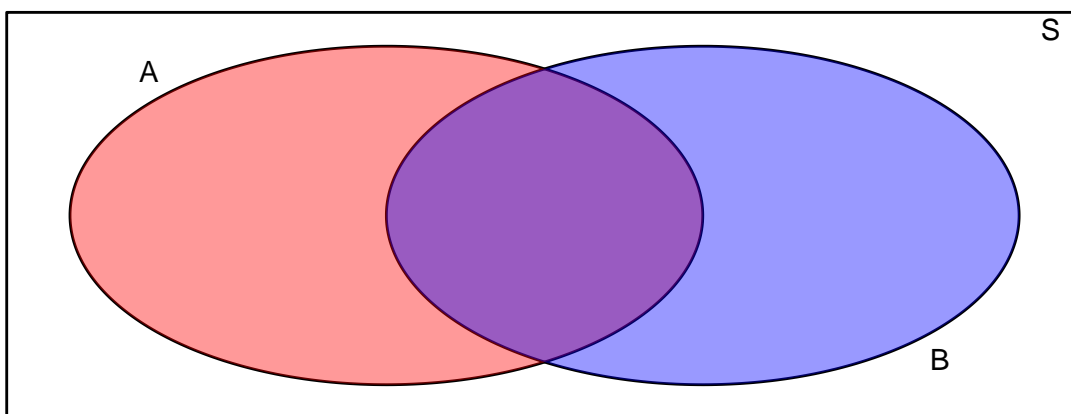
Fact. $P(A) + P(\bar{A}) = 1$

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.

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

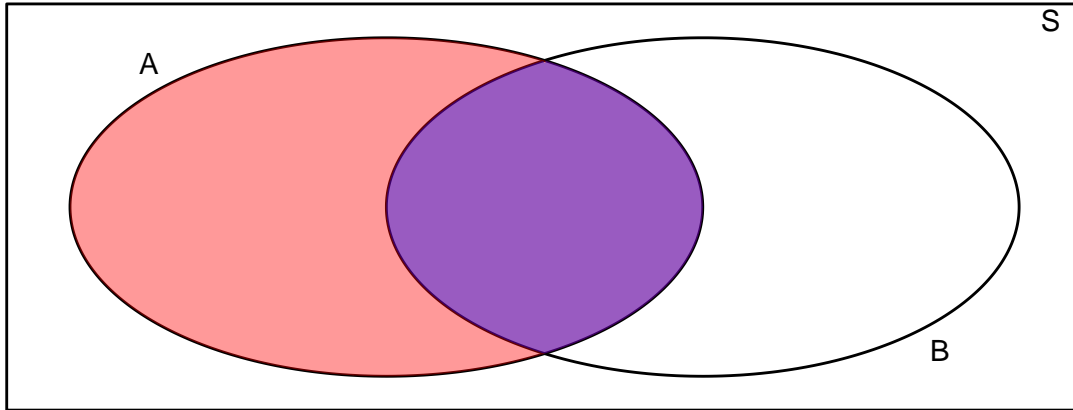
The reason behind this fact is that if there is if A and B are not disjoint, then some area is added twice when I calculate $P(A) + P(B)$. To account for this, I simply subtract off the area that was double counted.

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



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

$$P(A) = P(A \cap B^c) + P(A \cap B)$$



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

2.2.2 Conditional Probability

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

	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.

	Fire	Auto	Other	
Fraudulent	6%	1%	3%	10%
non-Fraudulent	14%	29%	47%	90%
	20%	30%	50%	100%

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

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

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

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

which can also be rearranged to show

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

Because 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.

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

Definition 4. Two events A and B are said to be independent if $P(A|B) = P(A)$ 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 $P(A \cap B) = P(A|B)P(B) = P(A)P(B)$.

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 $P(\text{Convertible}) = 0.10$. Of the non-convertibles, red is not unheard of but it isn't common either. So suppose $P(\text{Red} | \text{NonConvertible}) = 0.15$. However red is an extremely popular color for convertibles so let $P(\text{Red} | \text{Convertible}) = 0.60$.

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

	Convertible	Not Convertible	
Red			
Not Red			
	10%	90%	100%

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

$$\begin{aligned} P(\text{Red} \cap \text{Convertible}) &= P(\text{Red} | \text{Convertible}) P(\text{Convertible}) \\ &= 0.60 * 0.10 \\ &= 0.06 \end{aligned}$$

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

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

So when I ask for $P(\text{red} | \text{convertible})$, I am narrowing my space of cars to consider only convertibles. While the 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 $P(\text{Red}) = 0.195 \neq 0.60 = P(\text{Red} | \text{Convertible})$ then the events *Red* and *Convertible* are not independent.

2.2.3 Summary of Probability Rules

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

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

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

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

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

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?

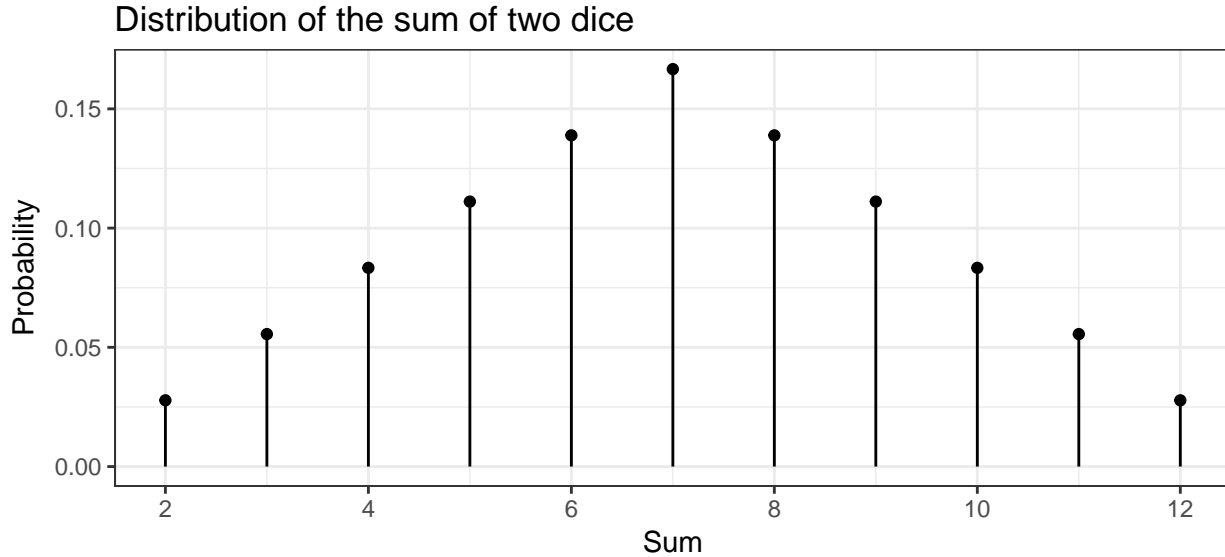
2.3.1 Introduction to Discrete Random Variables

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

2.3.1.1 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.

$$\mu = E[S] = \sum_{\text{possible } s} s \cdot P(S = s)$$

In this case, we have that

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

2.3.1.2 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.

$$\sigma^2 = V[S] = \sum_{\text{possible } s} (s - \mu)^2 \cdot P(S = s)$$

which in this example can be calculated as

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

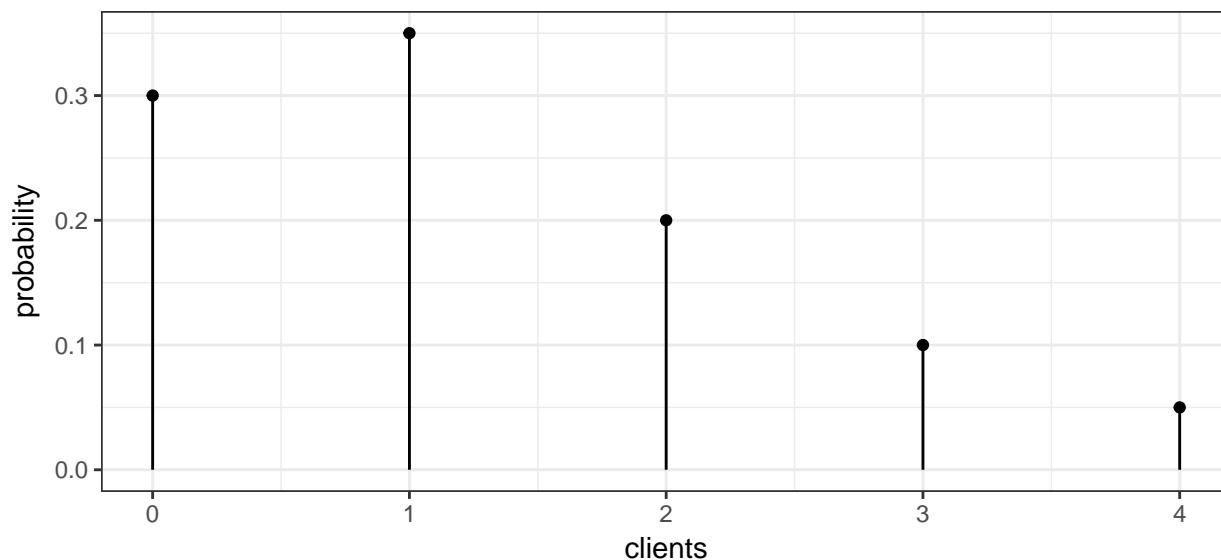
We could interpret the expectation as the sample mean of an infinitely large sample, and the variance as the sample variance of the same infinitely large sample. These are two very important numbers that describe the distribution.

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

Number of Clients	0	1	2	3	4
Frequency/Probability	0.30	0.35	0.20	0.10	0.05

```
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
  theme_bw()                                                    # set background color...
```



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?

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

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 with a critter as C_i and denote the event that the trap is empty as E_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_1, E_2, E_3
C_1, E_2, E_3
E_1, C_2, E_3
E_1, E_2, C_3

Outcome
C_1, C_2, E_3
C_1, E_2, C_3
E_1, C_2, C_3
C_1, C_2, C_3

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

$$P(E_1 \cap C_2 \cap E_3) = P(E_1)P(C_2)P(E_3) = (1 - 0.8)0.8(1 - 0.8) = 0.032$$

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_1, E_2, E_3	0.008	$S = 0$	0.008
C_1, E_2, E_3	0.032	$S = 1$	$3(0.032) = 0.096$
E_1, C_2, E_3	0.032		
E_1, E_2, C_3	0.032		
C_1, C_2, E_3	0.128	$S = 2$	$3(0.128) = 0.384$
C_1, E_2, C_3	0.128		
E_1, C_2, C_3	0.128		
C_1, C_2, C_3	0.512	$S = 3$	0.512

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

S Outcome	Probability
$S = 0$	0.008
$S = 1$	0.096
$S = 2$	0.384
$S = 3$	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 binomial random variable X taken from n trials each with probability of success π is

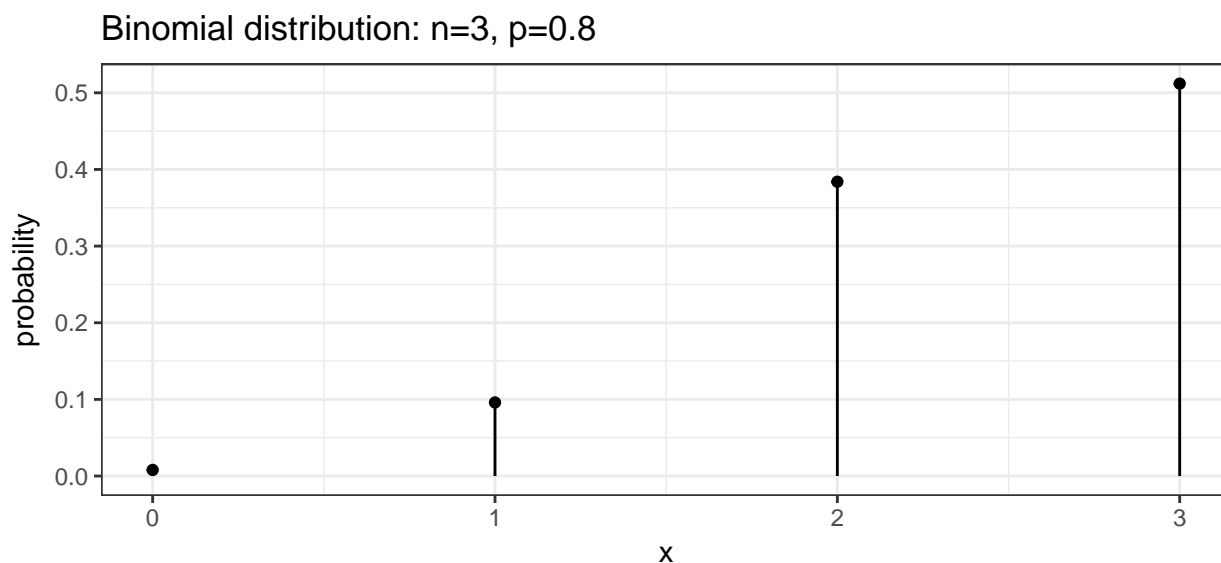
$$P(X = x) = \underbrace{\frac{n!}{x!(n-x)!}}_{\text{orderings}} \underbrace{\pi^x}_{y \text{ successes}} \underbrace{(1-\pi)^{n-x}}_{n-y \text{ failures}}$$

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:

```
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') +
  theme_bw()
```



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

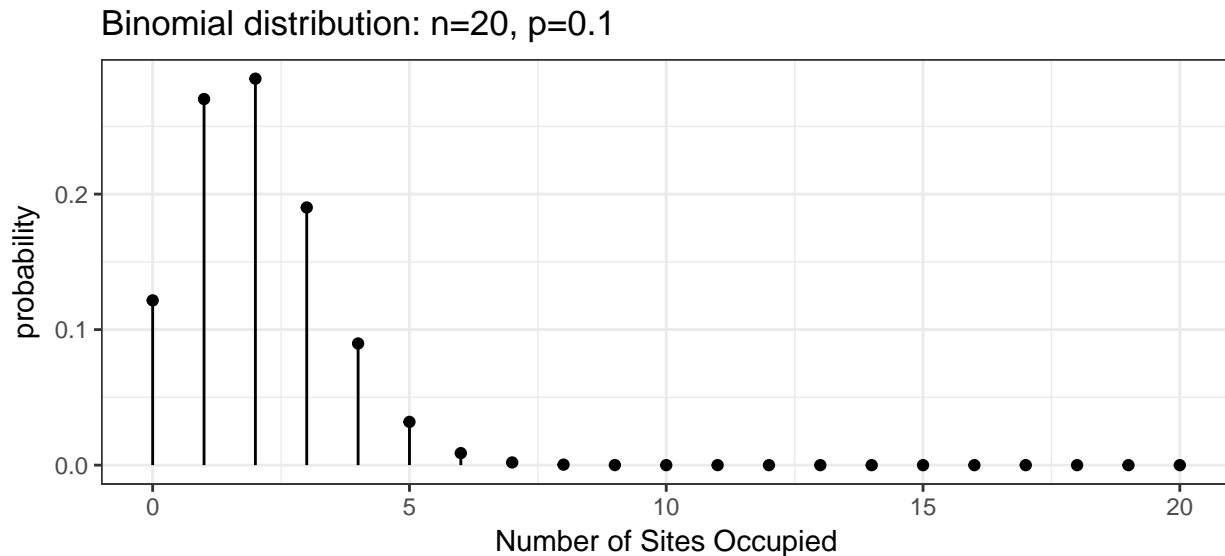
$$\begin{aligned}
 V[X] &= \sum_{x=0}^n (x - E[X])^2 P(X = x | n, \pi) \\
 &= \sum_{x=0}^n (x - E[X])^2 \frac{n!}{x! (n-x)!} \pi^x (1-\pi)^{n-x} \\
 &= \vdots \\
 &= n\pi(1-\pi)
 \end{aligned}$$

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(ymin=0, ymax=probability)) +
  ggtitle('Binomial distribution: n=20, p=0.1') +
  xlab('Number of Sites Occupied') +
  theme_bw()
```



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...

$$E[Y] = \lambda$$

$$Var[Y] = \lambda$$

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?

$$P(Y = 4) = \frac{2.3^4 e^{-2.3}}{4!} = 0.1169$$

What about the probability of finding at most 4?

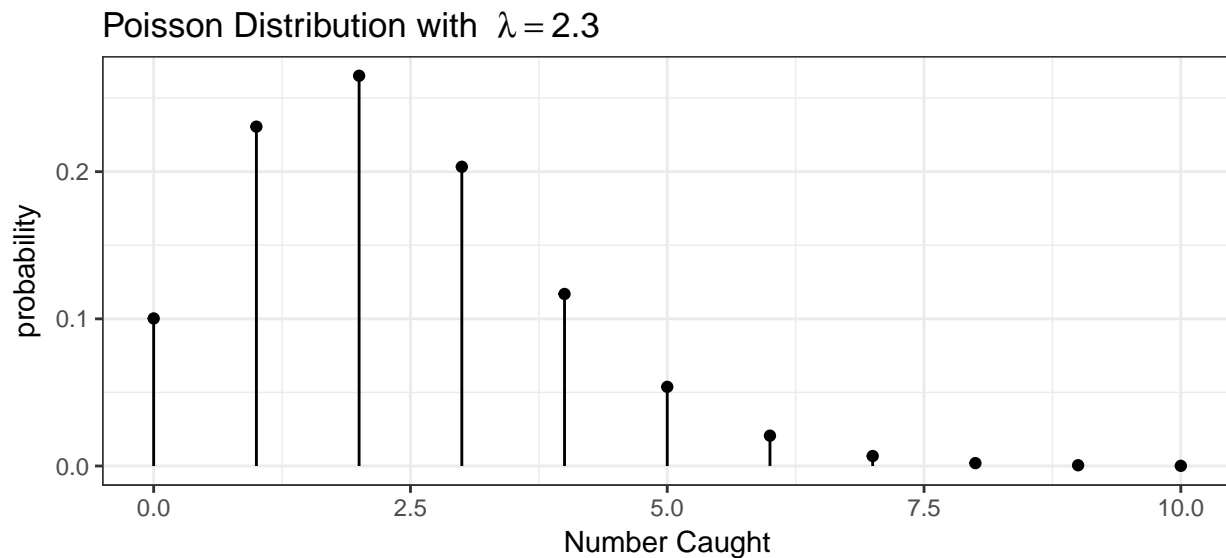
$$\begin{aligned} P(Y \leq 4) &= P(Y = 0) + P(Y = 1) + P(Y = 2) + P(Y = 3) + P(Y = 4) \\ &= 0.1003 + 0.2306 + 0.2652 + 0.2033 + 0.1169 \\ &= 0.9163 \end{aligned}$$

What about the probability of finding 5 or more?

$$P(Y \geq 5) = 1 - P(Y \leq 4) = 1 - 0.9163 = 0.0837$$

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(expression(paste('Poisson Distribution with ', lambda == 2.3))) +
  labs(x='Number Caught') +
  theme_bw()
```



```
# 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

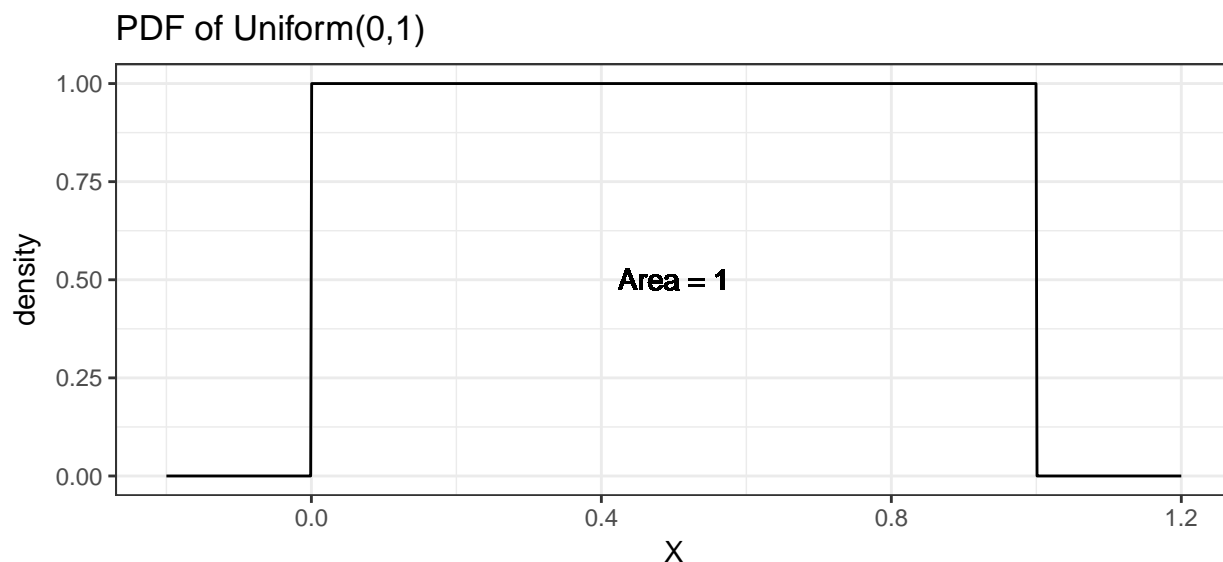
Continuous random variables can take on an (uncountably) infinite number of values, and this results in a few obnoxious mathematical differences between how we handle continuous and discrete random variables. In particular, the probability that a continuous random variable X will take on a particular value will be zero, so we will be interested in finding the probability that the random variable is in some interval instead. Wherever we had a summation, \sum , we will instead have an integral, but because many students haven't had calculus, we will resort to using R or tables of calculated values.

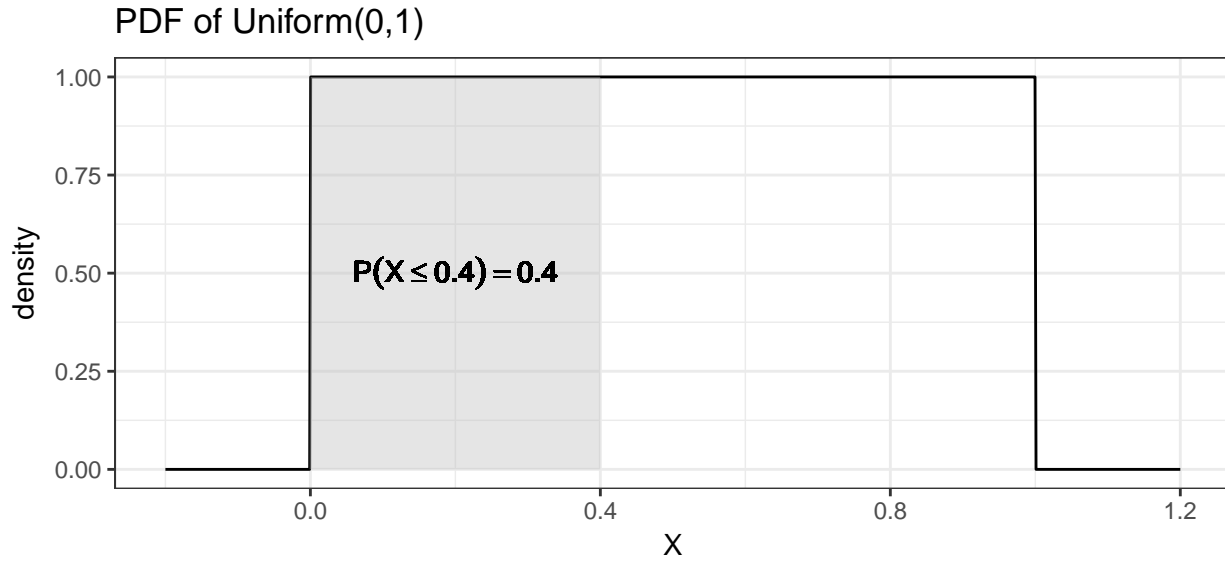
2.5.1 Uniform(0,1) Distribution

Suppose you wish to draw a random number between 0 and 1 and any two intervals of equal size should have the same probability of the value being in them. 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 (pdf) 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 probability density function is

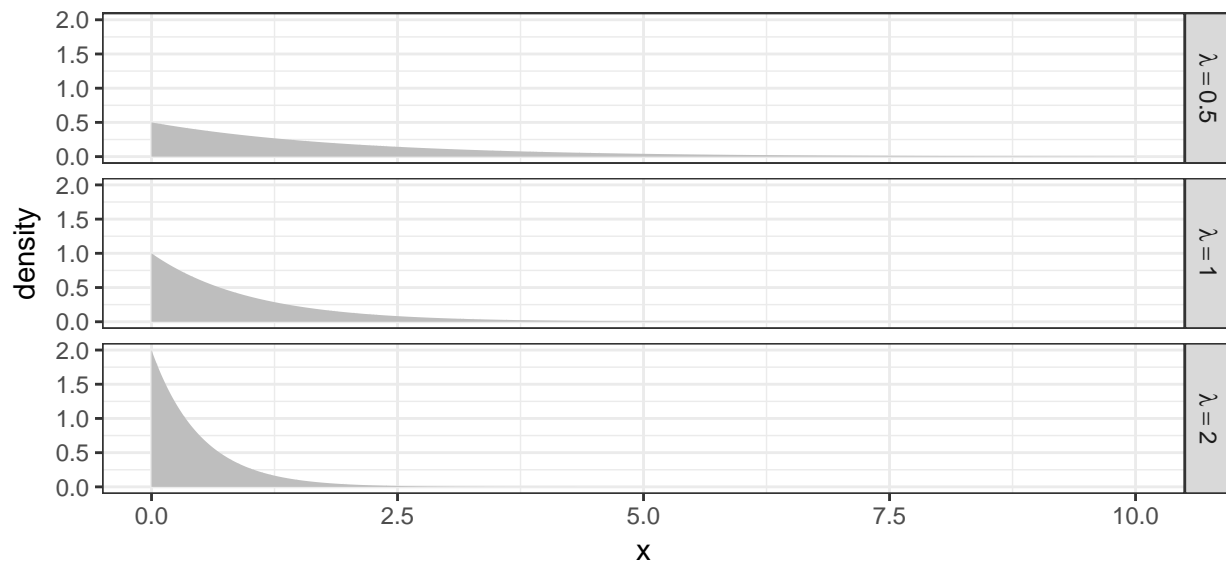
$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \text{ and } \lambda > 0 \\ 0 & \text{otherwise} \end{cases}$$

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

$$\mu = E[X] = \int_0^{\infty} x f(x) dx = \cdots = \frac{1}{\lambda}$$

$$\sigma^2 = Var[X] = \int_0^{\infty} (x - \mu)^2 f(x) dx = \cdots = \frac{1}{\lambda^2}$$

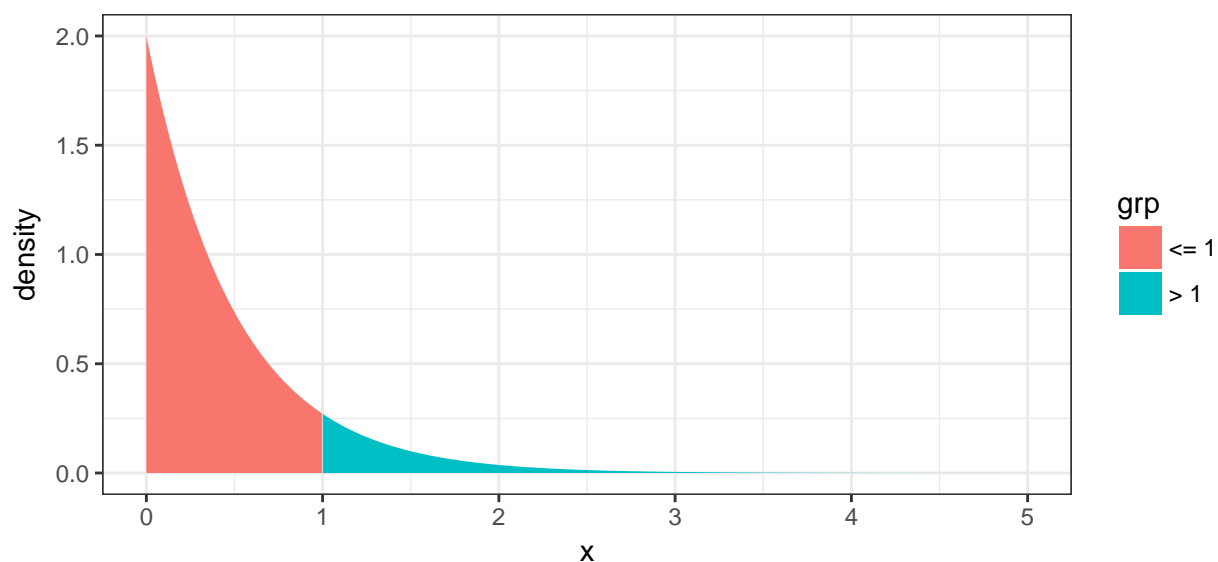
Because the exponential distribution is defined by the rate of occurrence of an event, increasing that rate decreases the time between events. Furthermore because the rate of occurrence cannot be negative, we restrict $\lambda > 0$.



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), lambda = 2) %>%
  mutate(y=dexp(x, rate = lambda),
         grp = ifelse( x > 1, '> 1', '<= 1'))
ggplot(data, aes(x=x, y=y, fill=grp)) +
  geom_area() +
  labs(y='density') +
  theme_bw()
```



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

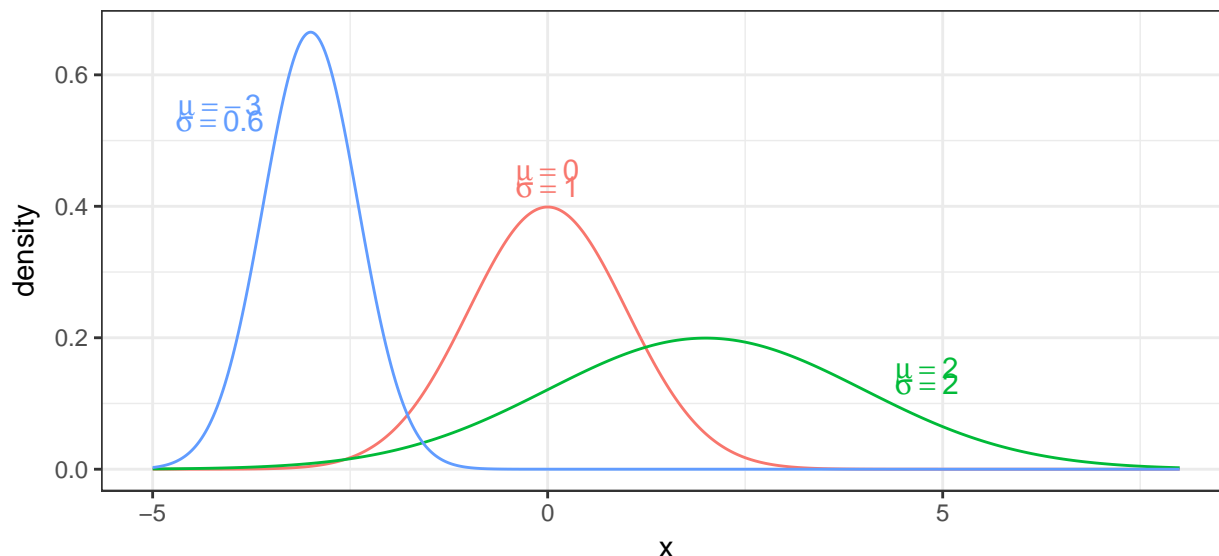
Undoubtably the most important distribution in statistics is the normal distribution. If my RV X is normally distributed with mean μ and standard deviation σ , its probability density function is given by

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

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

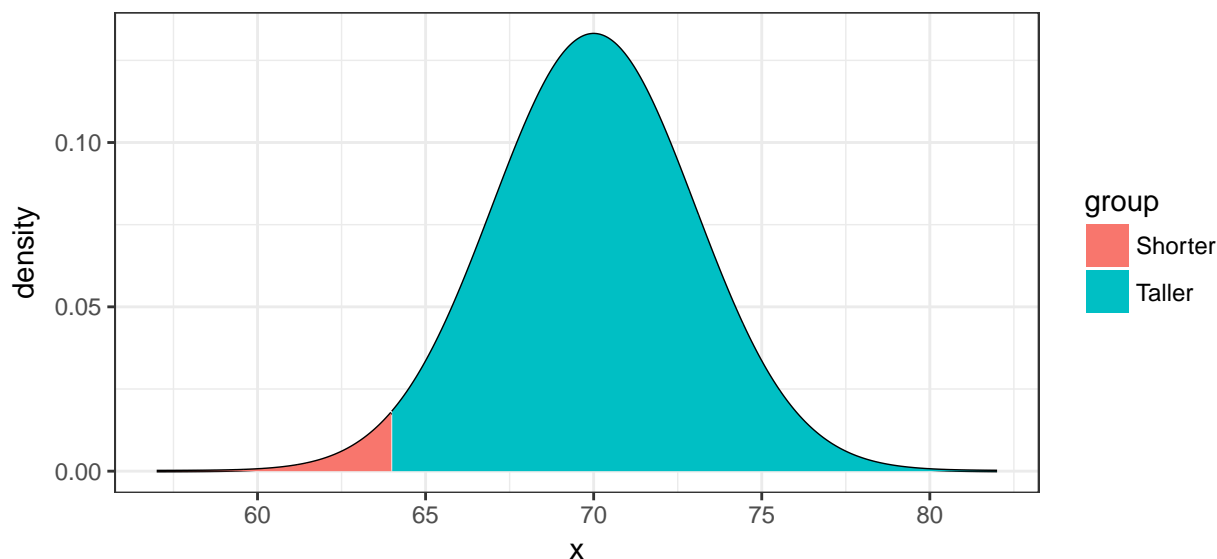
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left[-\frac{1}{2} \left(\frac{x - \mu}{\sigma} \right)^2 \right]$$

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() +
  theme_bw()
```

Using R you can easily find this

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

```
## [1] 0.02275013
```

2.5.4 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 $N(\mu, \sigma^2)$ distribution to a question about a Standard Normal distribution, which is a normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$. If we have

$$X \sim N(\mu, \sigma^2)$$

then

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

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 person be to be taller than 80% of the rest of the adult 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

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

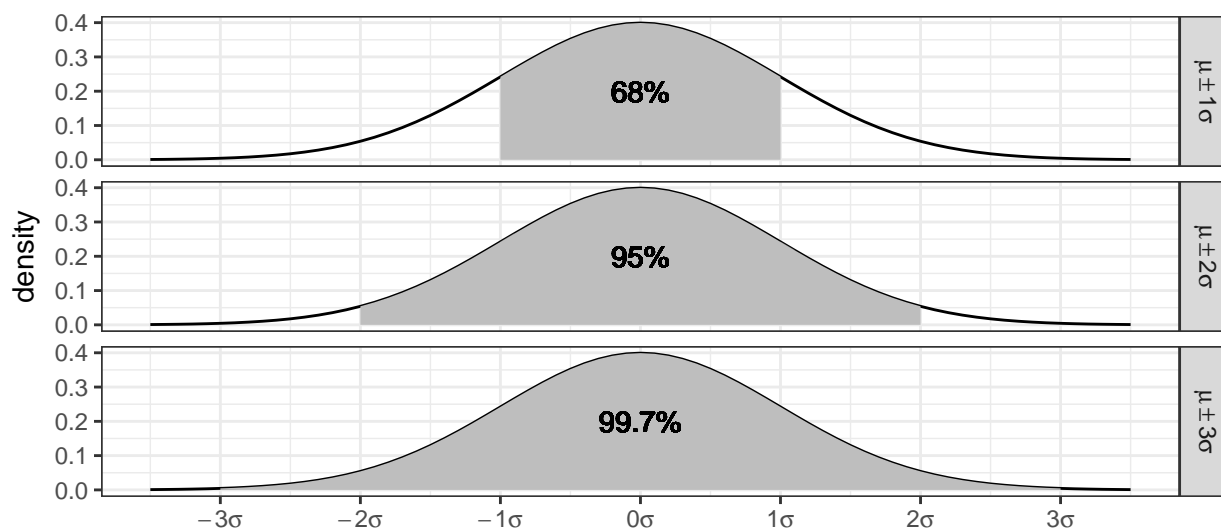
Alternatively we could use the quantile function for the normal distribution (q-function) in R and avoid the imprecision of using a table.

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

```
## [1] 0.8416212
```

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

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



1.6 R Comments

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

Distribution	Stem	Parameters	Parameter Interpretation
Binomial	<code>binom</code>	<code>size prob</code>	Number of Trials Probability of Success (per Trial)
Exponential	<code>exp</code>	<code>rate</code>	Mean of the distribution
Normal	<code>norm</code>	<code>mean=0 sd=1</code>	Center of the distribution Standard deviation
Uniform	<code>unif</code>	<code>min=0 max=1</code>	Minimum of the distribution Maximum of the distribution

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

Function	Result
d-function(x)	The height of the probability distribution/density at x
p-function(x)	$P(X \leq x)$
q-function(q)	x such that $P(X \leq x) = q$
r-function(n)	n random observations from the distribution

2.6 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:

	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%	
					100%

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 son'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.
- What is the probability that the next two races I'm in will both have crashes?
 - What is the probability that neither of my next two races will have a crash?
 - 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 $\lambda = 1.2$ pukers per week.
- What is the probability that I don't have to clean up any vomits this coming week?
 - What is the probability that I must clean up 1 or more vomits?
 - 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+
Probability	0.45	0.25	0.20		0.0

- What is the probability that I see 3 runners on a morning walk?
 - What is the expected number of runners that I will encounter?
 - What is the variance of the number of runners that I will encounter?
7. If $Z \sim N(\mu = 0, \sigma^2 = 1)$, find the following probabilities:
- $P(Z < 1.58) =$
 - $P(Z = 1.58) =$
 - $P(Z > -.27) =$
 - $P(-1.97 < Z < 2.46) =$
8. Using the Standard Normal Table or the table functions in R, find z that makes the following statements true.
- $P(Z < z) = .75$
 - $P(Z > z) = .4$
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.
- What is the probability that I fed my cats more than 250 grams of kibble this morning?
 - 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 via Bootstrapping

```
# Every chapter, we will load all the libraries we will use at the beginning  
# of the chapter.  
library(mosaic)      # for the resample, shuffle, and do functions  
library(ggplot2)     # graphing functions  
library(dplyr)       # data summary tools
```

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 .

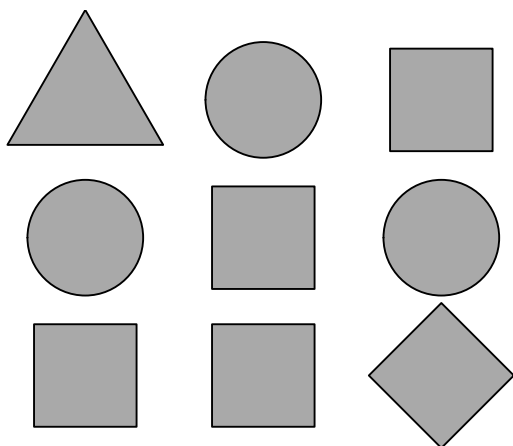
Next suppose we wish to ask if a particular value for μ , say μ_0 , is consistent with our observed data? We know that \bar{x} will vary from sample to sample, but we have no idea how much it will vary between samples. However, if we could understand how much \bar{x} varied sample to sample, we could 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.

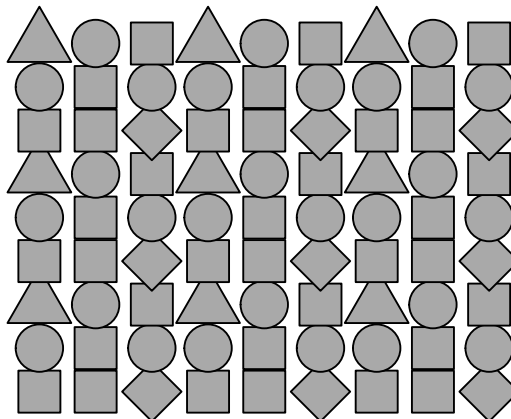
Suppose we were to sample from a population of shapes, and we observed 4/9 of the sample were squares, 3/9 were circles, and a triangle and a diamond. Then our best guess of what the population that we sampled from was a population with 4/9 squares, 3/9 circles, and 1/9 of triangles and diamonds.

```
multiplot(sample.plot, population.plot, cols=2)
```

Sample



Approximate Population



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

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

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

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

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

```
##      name orig.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.

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.

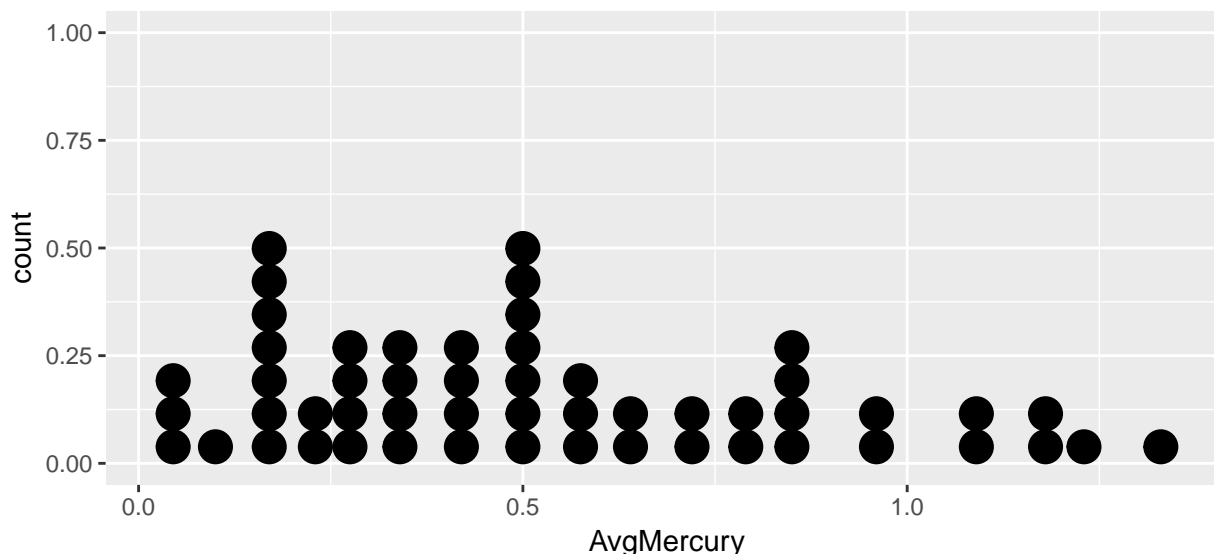
Example: Mercury Levels in Fish from Florida Lakes

A data set provided by the Lock⁵ introductory statistics 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.

```
# 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 doesn'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 an estimate of 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))

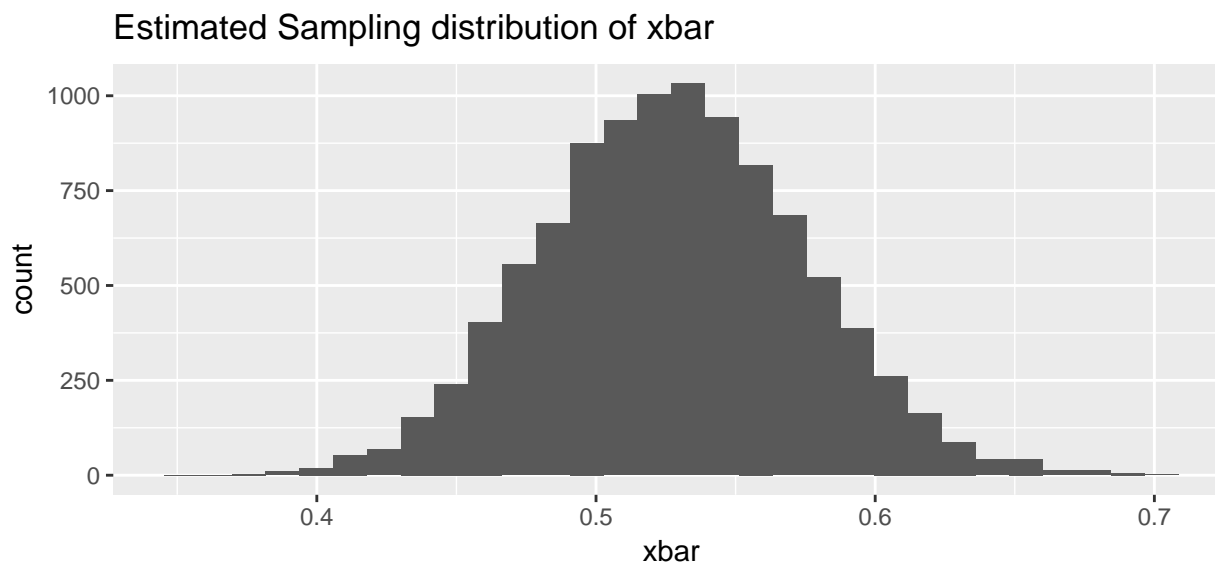
# what columns does the data frame "SamplingDist" have?
head(SamplingDist)

##          xbar
## 1 0.5111321
## 2 0.5043396
```

```
## 3 0.5252830
## 4 0.5362264
## 5 0.4992453
## 6 0.5960377
```

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

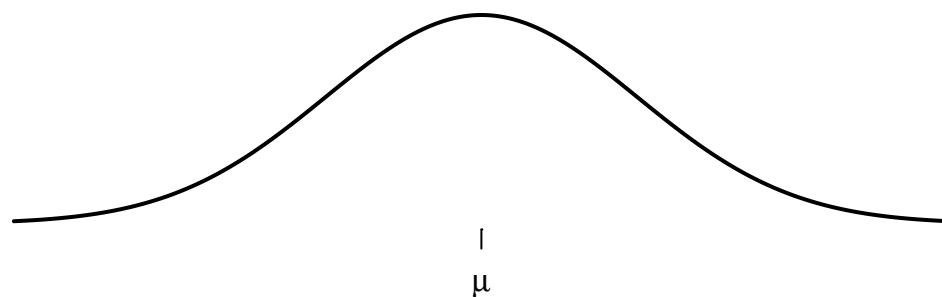
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



3.2 Quantile-based Confidence Intervals

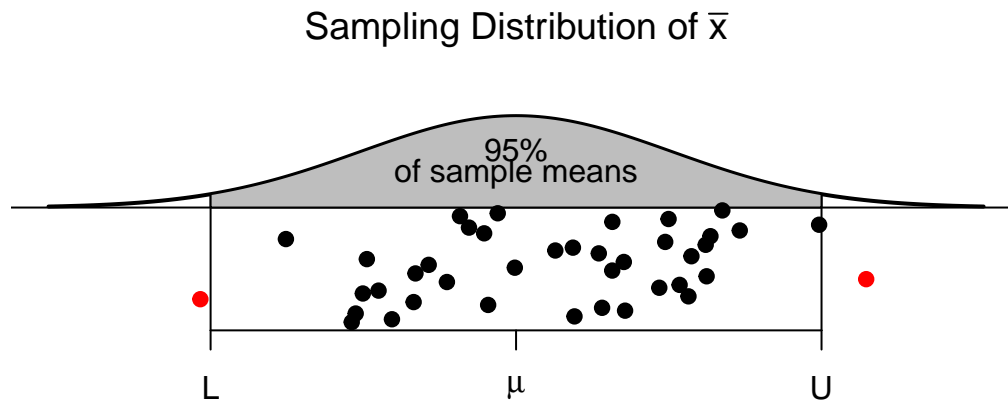
In many cases we have seen, the sampling distribution of a statistic is centered on the parameter we are interested in estimating and is symmetric about that parameter. There are actually several ways to create a confidence interval from the estimated sampling distribution. The method presented here is called the “percentile” method and works when the sampling distribution is symmetric and the estimator we are using is unbiased. 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.

Sampling Distribution of \bar{x}

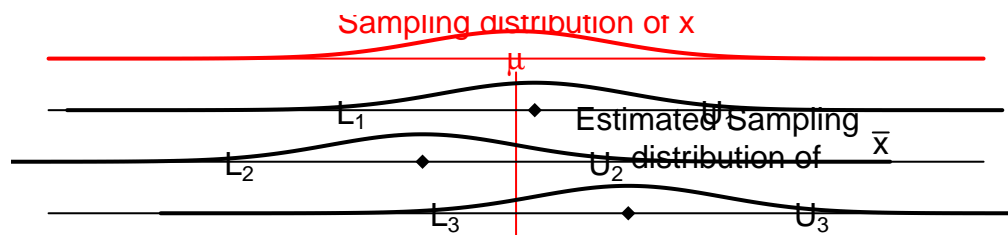


There are two points, (call them L and U) where for our given sample size and population we are sampling from, where we expect that 95% of the sample means to fall within. That is to say, L and U capture the

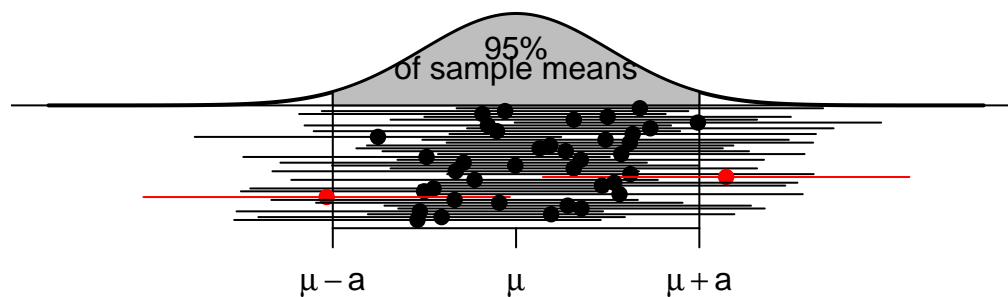
middle 95% of the sampling distribution of \bar{x} .



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.

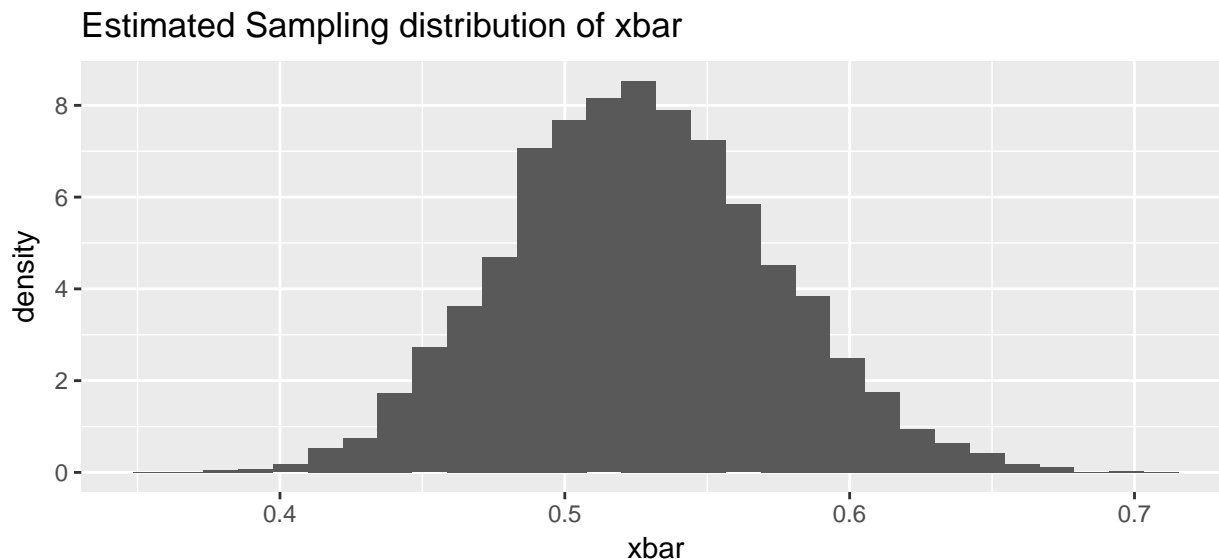
After the sample is taken and the interval is calculated, the numbers lower and upper bounds of the confidence interval are fixed. Because μ is a constant value and the confidence interval is fixed, nothing is changing. To distinguish between a future random event and the fixed (but unknown) outcome 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

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

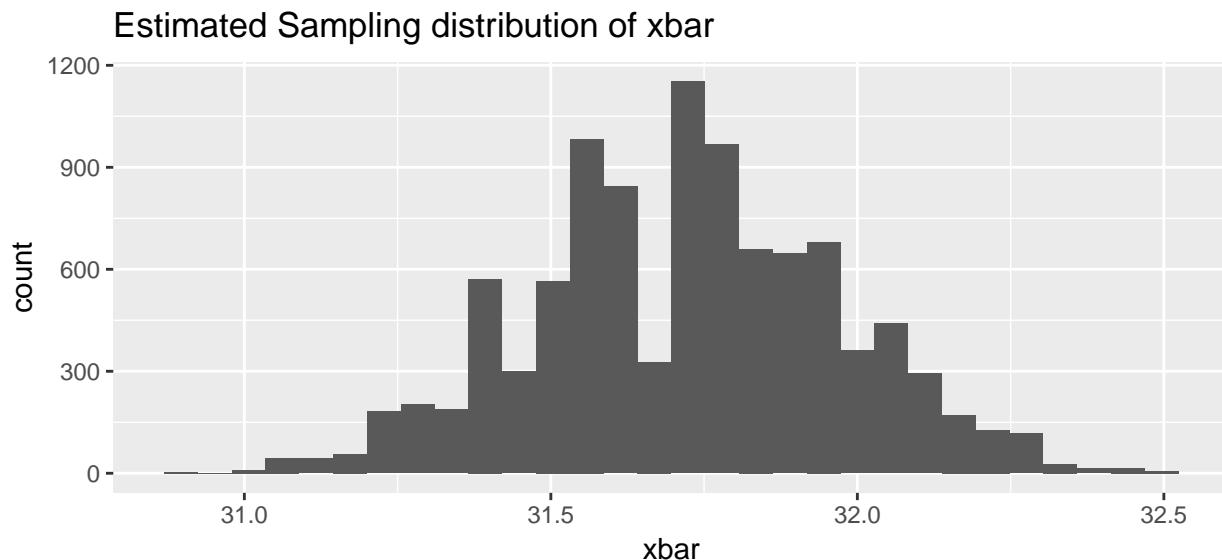
```
CarMPG <- data.frame( ID=1:5, mpg = c(31.8, 32.1, 32.5, 30.9, 31.3) )
CarMPG %>% summarise( xbar=mean(mpg) )
```

```
##      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%
## 31.24 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.

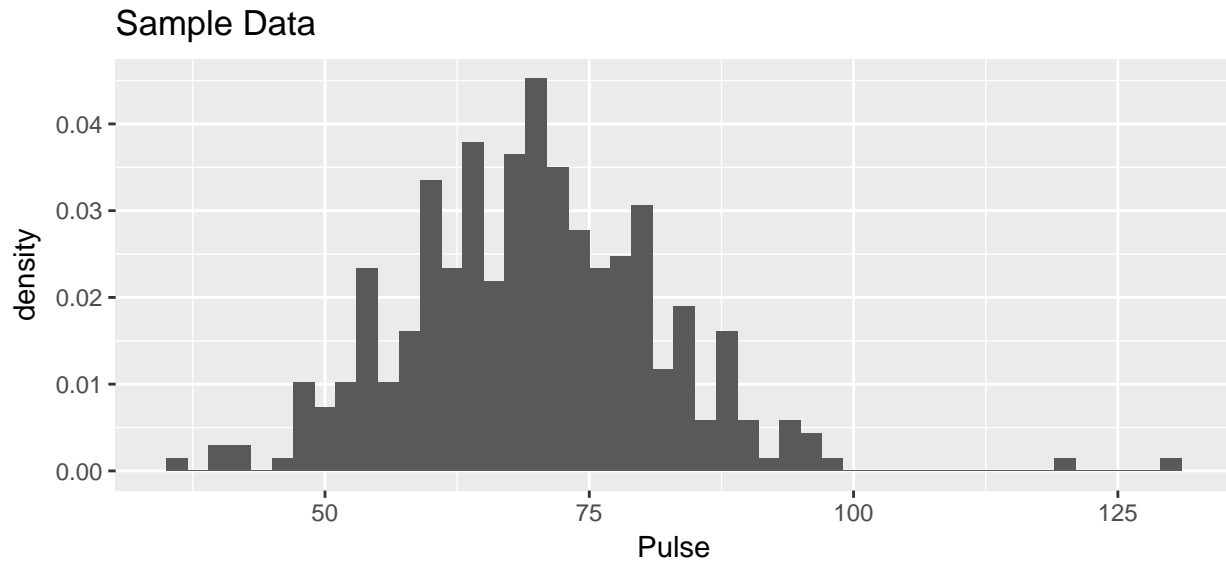
Example: Pulse Rate of College Students

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.

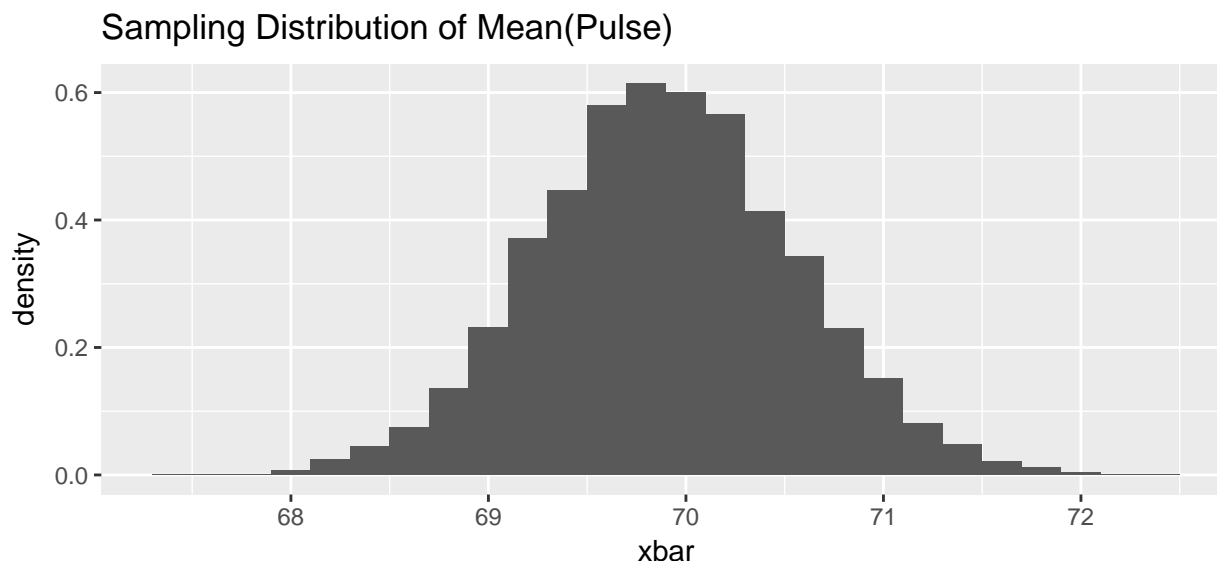
```
# Summary Statistics
GPAGender %>% summarise(xbar = mean(Pulse),
                        StdDev = sd(Pulse))
```

```
##      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%
## 68.65015 71.18958
```

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 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 points, 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

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

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

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

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

Bibliography