Chapter 1 Describing Data

1.1 Tidy Data

It doesn't take a data scientist to know that there isn't a single, defined format for data to be organized and stored. More often that not, we'll have to adjust and manage the data to get it into a format that is suitable for visualizations and analysis. One step in this process is generally called **cleaning**, and involves tasks such as "cleaning up" missing or incorrect values, column names, inconsistencies in the data, etc. There's also ususally some amount of **wrangling**. Data wrangling is the steps of creating new variables, reshaping the data, joining multiple datasets into one, etc. While this course won't require you to do either of these tasks in depth, it is good to know that they're part of everyday work for a data scientist.

What should good data look like though? How does one know that their data is ready to be analyzed and visalized? The answer is that the dataset should be **tidy**.

> "Tidy data is a standard way of mapping the meaning of a dataset to its structure. A dataset is messy or tidy depending on how rows, columns and tables are matched up with observations, variables and types."
> — Hadley Wickham

To better explain what this means and give it some context, consider some example datasets:

(Note: much of the following comes from "R for Data Science" by Garrett Grolemund and Hadley Wickham and the *Tidy Data* paper written by Hadley Wickham and published in the Journal of Statistical Software.)

You can represent the same underlying data in multiple ways. The example below shows the same data organised in four different ways. Each dataset shows the same values of four variables country, year, population, and cases, but each dataset organises the values in a different way.

Table 1

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

Table 2

country	year	type	count
Afghanistan	1999	cases	745
Afghanistan	1999	population	19987071
Afghanistan	2000	cases	2666
Afghanistan	2000	population	20595360
Brazil	1999	cases	37737
Brazil	1999	population	172006362
Brazil	2000	cases	80488
Brazil	2000	population	174504898
China	1999	cases	212258
China	1999	population	1272915272
China	2000	cases	213766
China	2000	population	1280428583

Table 3

country	year	rate
Afghanistan	1999	745/19987071
Afghanistan	2000	2666/20595360
Brazil	1999	37737/172006362
Brazil	2000	80488/174504898
China	1999	212258/1272915272
China	2000	213766/1280428583

Spread across two tables

Table 4a: `cases`

country	1999	2000
Afghanistan	745	2666
Brazil	37737	80488
China	212258	213766

Table 4b: `population`

country	1999	2000
Afghanistan	19987071	20595360
Brazil	172006362	174504898
China	1272915272	1280428583

These are all representations of the same underlying data, but they are not equally easy to use.

There are three interrelated rules which make a dataset tidy:

1. Each variable forms a column.
2. Each observation forms a row.
3. Each value must have its own cell.

Figure 1.1 shows the rules visually.

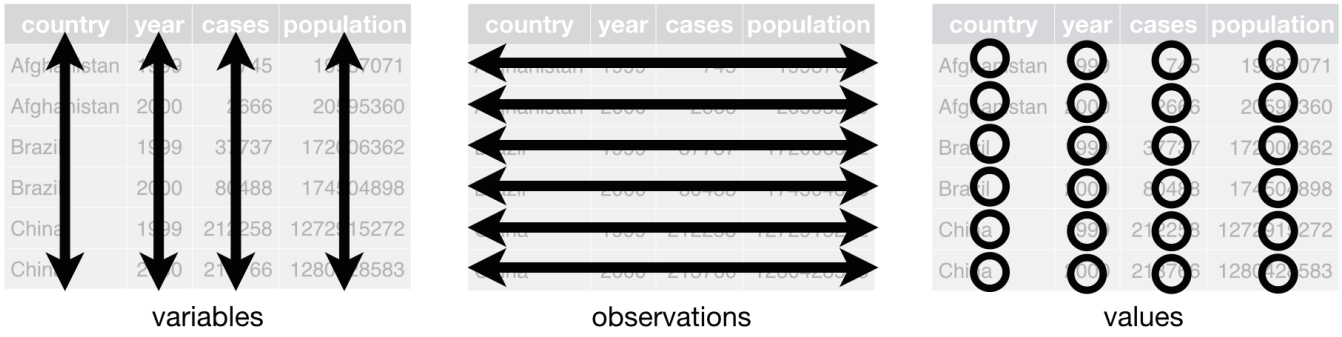


Figure 1.1: Following three rules makes a dataset tidy: variables are in columns, observations are in rows, and values are in cells.

These three rules are interrelated because it's impossible to only satisfy two of the three. That interrelationship leads to an even simpler set of practical instructions:

1. Put each dataset in a tibble. (Don't worry if you've never heard of a tibble before. For now, just think of it as a dataset that follows the rules of being tidy. Read here for more information if you're curious. Tibbles will have a lot more meaning if you ever program in the R language.)

2. Put each variable in a column.

In this example, only `table1` is tidy. It's the only representation where each column is a variable.

Why ensure that your data is tidy?:

There's a general advantage to picking one consistent way of storing data. If you have a consistent data structure, it's easier to learn the tools that work with it because they have an underlying uniformity.

For further reading about how to tidy data and more examples see here. Note: This page discusses tidying data in relation to the R programming language.

We will be using Google Sheets for this course to store and manipulate data. Google Sheets doesn't require data to be stored in a tidy format, but you should choose to keep your data tidy. Your life will be much better for it.

1.2 Course Tools

For this course, we will use both Google Sheets and Plotly Chart Studio.

1.2.1 Using Google Sheets

With Google Sheets, you can create and edit spreadsheets directly in your web browser—no special software is required. Multiple people can work simultaneously, you can see people's changes as they make them, and every change is saved automatically. A more extensive guide can be found in Tools, but you can refer to their cheat sheet here.

1.2.2 Using Tableau for visualizations

We will be using Tableau to create visualizations. A reference guide to some of its features can be found in 6.2.

1.3 Numerical Summaries

When data is numeric in nature, it is often helpful to look at summaries of the data, rather than try and take in the data as a whole. There are more summary statistics than are shown here in this book, but this book will cover the most commonly used numerical summaries. This chapter is split into **measures of center** and **measures of spread**.

1.3.1 Measures of Center

Think about conversations you may have had about data. That may be the average score on an exam, the average miles per gallon on a tank of gas, or the median starting salary for a given degree or job. Note that all of these values are just that, values. These values don't tell us anything about the **spread** of the data, but they tell us about the likely values of the distribution. Understanding the spread as well can be very useful (see Data vs. Summaries for more on that conversation) but understanding the center by itself has plenty of use. When buying a car, you probably don't feel the need to ask about the standard deviation of the car's reported gas mileage becasue the average alone is probably enough for you to make a decision.

There are more measures of center than are listed here, but these are arguably the most common and useful for general purpose uses.

The material below on numerical and graphical summaries is almost entirely from the Statistics Notebook.

Mean

1 Quantitative Variable

Formula:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- Although the formula looks complicated, all it states is "add all the data values up and divide by the total number of values."

Symbols in the Formula

- $\bar{x}$ is read "x-bar" and is the symbol typically used for the **sample mean**, the mean computed on a *sample* of data from a population.
- Σ, the capital Greek letter "sigma," is the symbol used to imply "add all of the data values up."
- The x_i's are the data values. The i in the x_i is stated to go from $i = 1$ all the way up to n. In other words, data value 1 is represented by x_1, data value 2: x_2, . . ., up through the last data value x_n. In general, we just write x_i.
- n represents the sample size, or number of data values.

Explanation:

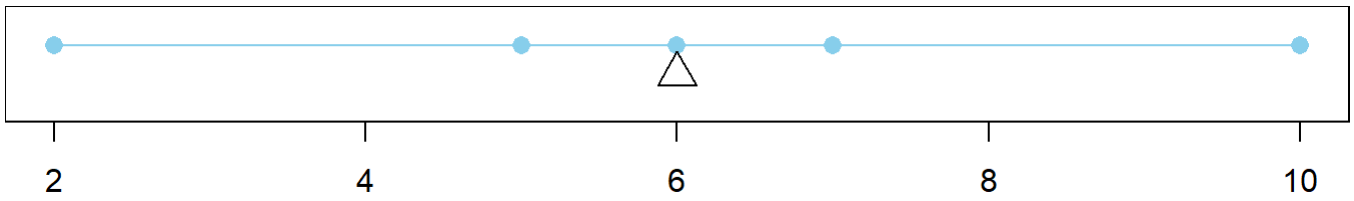
- The “balance point” or “center of mass” of quantitative data. It is calculated by taking the numerical sum of the values divided by the number of values. Typically used in tandem with the standard deviation. Most appropriate for describing the most typical values for relatively normally distributed data. Influenced by outliers, so it is not appropriate for describing strongly skewed data.

Physical Interpretation

- The mean is sometimes described as the “balance point” of the data. The following example will demonstrate.
- Say there are $n = 5$ data points with the following values.
- $x_1 = 2$
- $x_2 = 5$
- $x_3 = 6$
- $x_4 = 7$
- $x_5 = 10$
- The sample mean is calculated as follows.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{2 + 5 + 6 + 7 + 10}{5} = 6$$

- If these values were plotted, and an “infinitely thin bar” connected the points, then the bar would “balance” at the mean (the triangle) as shown below.



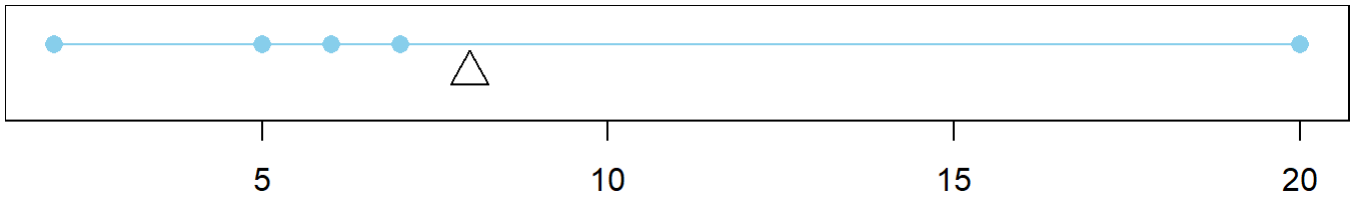
Middle of the Deviations

- The above plot demonstrates that there are equal, but opposite, “sums of deviations” to either side of the mean. Note that a deviation is defined as the distance from the mean to a given point. Thus, x_1 has a deviation of -4 from the mean, x_2 a deviation of -1, x_3 a deviation of 0, x_4 a deviation of 1, and x_5 a deviation of 4. To the left there is a sum of deviations equal to -5 and on the right, a sum of deviations equal to 5. This can be verified to hold for any scenario.

Effect of Outliers

- The mean can be strongly influenced by *outliers*, points that deviate abnormally from the mean. This is shown below by changing x_5 to be 20. Note that the deviation of x_5 is 12, and the sum of deviations to the left of the mean ($\bar{x} = 8$) is $-1 + -2 + -3 + -6 = -12$.
- The mean of the altered data
- $x_1 = 2$
- $x_2 = 5$
- $x_3 = 6$
- $x_4 = 7$
- $x_5 = 20$

is now $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{2+5+6+7+20}{5} = 8$.



Population Mean

- When **all** of the data from a population is available, the **population mean** is calculated instead of the sample mean. The mathematical formula for the **population mean** is the same as the formula for the sample mean, but is written with slightly different notation.

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

- Notice that the symbol for the population mean is μ, pronounced “mew,” another Greek letter. The only other difference between the two formulas is that the sample mean uses a sample of data, denoted by n, while the population mean uses all the population data, denoted by N .

Median

1 Quantitative Variable

Formula:

- The mathematical formula used to compute the median of data depends on whether n, the number of data points in the sample, is even or odd.
- If n is even, then there is no “middle” data point, so the middle two values are averaged.

$$\text{Median} = \frac{x_{(n/2)} + x_{(n/2+1)}}{2}$$

If n is odd, then the middle data point is the median.

$$\text{Median} = x_{((n+1)/2)}$$

Symbols in the Formula

- There is no generally accepted symbol for the median. Sometimes a capital M or even lower-case m is used, but generally the word median is just written out.
- $x_{(n/2)}$ represents the data value that is in the $(n/2)^{th}$ position in the ordered list of values. It only exists when n is even.
- $x_{(n/2+1)}$ represents the data value that immediately follows the $(n/2)^{th}$ value in the ordered list of values. It only exists when n is even.
- $x_{((n+1)/2)}$ represents the data value that is in the $((n+1)/2)^{th}$ position in the ordered list of values. It only exists when n is odd.
- n represents the sample size, or number of data values in the sample.

Explanation

The “middle data point,” i.e., the 50^{th} percentile. Half of the data is below the median and half is above the median. Typically used in tandem with the five-number summary to describe skewed data because it is not heavily influenced by outliers, i.e., it is *robust*. Can also be used with normally distributed data, but the mean and standard deviation are more useful measures in such cases.

Population Median

When **all** of the data from a population is available, the **population median** is calculated by the above formulas with the slight change that N, the total number of data values in the population, instead of n, the number of values in the sample, is used.

If N is even, then there is no “middle” data point, so the middle two values are averaged.

$$\text{Median} = \frac{x_{(N/2)} + x_{(N/2+1)}}{2}$$

If N is odd, then the middle data point is the median.

$$\text{Median} = x_{((N+1)/2)}$$

Physical Interpretation

The median is the 50^{th} percentile of the data.

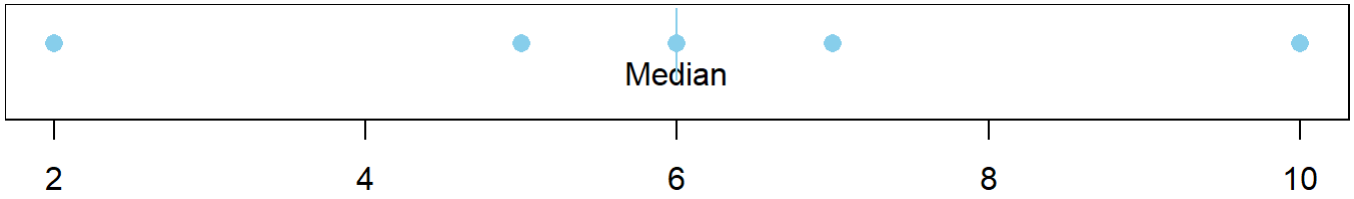
Say there are $n = 5$ data points in the sample with the following values.

- $x_1 = 2$
- $x_2 = 5$
- $x_3 = 6$
- $x_4 = 7$
- $x_5 = 10$

The sample median is calculated as follows. Note that $n = 5$ is odd.

$$\text{Median} = x_{((n+1)/2)} = x_{((5+1)/2)} = x_{(3)} = 6$$

When these values are plotted it is clear that exactly 50% of the data (excluding the median) is to either side of the median.

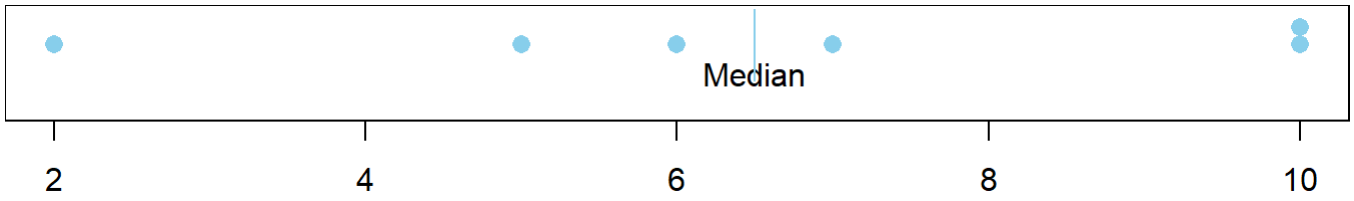


Second Example

Say there was a sixth value in the data set equal to 10, so that $n = 6$ is even.

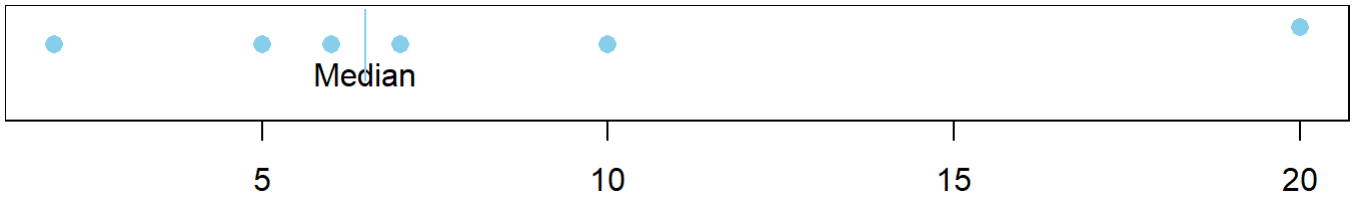
- $x_1 = 2$
- $x_2 = 5$
- $x_3 = 6$
- $x_4 = 7$
- $x_5 = 10$
- $x_6 = 10$

$$\text{Median} = \frac{x_{(n/2)} + x_{(n/2+1)}}{2} = \frac{x_{(6/2)} + x_{(6/2+1)}}{2} = \frac{x_{(3)} + x_{(4)}}{2} = \frac{6+7}{2} = 6.5$$



Effect of Outliers

The median is not greatly influenced by *outliers*. It is said to be *robust*. This is shown below by changing x_6 to be 20, which does not change the value of the median.



Mode

1 Quantitative or Qualitative Variable

The most commonly occurring value. There may be more than one mode.

Example:

A set of data might contain the following values: 5, 7, 8, 8, 8, 11, 15, 16, 17, 17, 20.

The mode would be 8, becuase it occured most often in the data.

If another 17 were added, then the data would be: 5, 7, 8, 8, 8, 11, 15, 16, 17, 17, 17, 20.

The new modes would be 8 and 17 because they’re tied for the highest number of occurences.

Percentile

1 Quantitative Variable

The percent of data that is equal to or less than a given data point. Useful for describing the relative position of a data point within a data set. If the percentile is close to 100, then the observation is one of the largest. If it is close to zero, then the observation is one of the smallest.

An example may help this make more sense. Imagine a very long street with houses on one side. The houses increase in value from left to right. At the left end of the street is a small cardboard box with a leaky roof. Next door is a slightly larger cardboard box that does not leak. The houses eventually get larger and more valuable. The rightmost house on the street is a huge mansion.

Notice that if there was a fence between each house, it would take 99 fences to separate the houses.

house 1 | house 2 | … | house 99 | house 100

The home values are representative of data. If we have a list of data, sorted in increasing order, and we want to divide it into 100 equal groups, we only need 99 dividers (like fences) to divide up the data. The first divider is as large or larger than 1% of the data. The second divider is as large or larger than 2% of the data, and so on. The last divider, the 99th, is the value that is as large or larger than 99% of the data. These “dividers” (i.e. the fences) are called percentiles. A percentile is a number such that a specified percentage of the data are at or below this number. For example, the 99th percentile is a number such that 99% of the data are at or below this value. As another example, half (50%) of the data lie at or below the 50th percentile. The word “percent” means “÷100.” This can help you remember that the percentiles divide the data into 100 equal groups.

Quartiles are special percentiles. The word “quartile” is from the Latin quartus, which means “fourth.” The quartiles divide the data into four equal groups. The quartiles correspond to specific percentiles. The first quartile, Q1, is the 25th percentile. The second quartile, Q2, is the same as the 50th percentile or the median. The third quartile, Q3, is equivalent to the 75th percentile.

Understanding the five-number summary will help perecentiles and quartiles have more meaning.

1.3.2 Measures of Spread

Quartiles (five-number summary)

25th, 50th, 75th *and* 100th *Percentiles*

1 Quantitative Variable

Good for describing the spread of data, typically for skewed distributions. There are four quartiles. They make up the ***five-number summary*** when combined with the minimum. The second quartile is the median (50th percentile) and the fourth quartile is the maximum (100th percentile). The first quartile (Q_1 or lower quartile) and third quartile (Q_3 or upper quartile) show the spread of the "middle 50%" of the data, which is often called the ***interquartile range***. Comparing the interquartile range to the minimum and maximum shows how the possible values spread out around the more probable values.

Standard Deviation

1 Quantitative Variable

Formula:

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

Data often varies. The values are not all the same. To capture, or measure how much data varies with a single number is difficult. There are a few different ideas on how to do it, but by far the most used measurement of the variability in data is the standard deviation.

The first idea in measuring the variability in data is that there must be a reference point. Something from which everything varies. The most widely accepted reference point is the **mean**.

A **deviation** is defined as the distance an observation lies from the reference point, the mean. This distance is obtained by subtraction in the order $x_i - \bar{x}$, where x_i is the data point value and $\bar{x}$ is the mean of the data. There are thus n deviations because there are n data points.

Unfortunately, because of the order of subtraction in obtaining deviations, the average deviation will always work out to be zero. This is because the mean by nature splits the deviations evenly.

One solution would be to take the absolute value of the deviations and obtain what is known as the "absolute mean deviation." This is sometimes done, but a far more attractive choice (to mathematicians and statisticians) is to square each deviation. You'll have to trust us that this is the better choice.

Squaring a deviation results in the expression $(x_i - \bar{x})^2$. **SQUARE**

Summing up all of the squared deviations results in the expression $\sum_{i=1}^{n}(x_i - \bar{x})^2$.

Dividing the sum of the squared deviations by n would seem like an appropriate thing to do. Experience (and some fantastic statistical theory!) demonstrated that this is wrong. Dividing by $n - 1$, the *degrees of freedom* is right. **MEAN**

To undo the squaring of the deviations, the final results are square rooted. **ROOT**

The end result is the beautiful formula for s, the standard deviation! (At least the symbol for standard deviation is a simple s.) It is also know as the **ROOT-MEAN-SQUARED ERROR**. Error is another word for deviation.

The *standard deviation* is thus the representative deviation of all deviations in a given data set. It is never negative and only zero if all values are the same in a data set. Larger values of s imply the data is highly variable, very spread out or very inconsistent. Smaller values mean the data is consistent and not as variable.

Population Standard Deviation

When **all** of the data from a population is available, the **population standard deviation** σ (the lower-case Greek letter “sigma”) is calculated by the following formula.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

Note that N is the number of data points in the full population. In this formula the denominator is actually N and the deviations are calculated as the distance each data point is from the population mean μ.

An Example

Say there are five data points given by

- $x_1 = 2$
- $x_2 = 5$
- $x_3 = 6$
- $x_4 = 7$
- $x_5 = 10$

The mean of these values is $\bar{x} = 6$.

The five deviations are

- $(x_1 - \bar{x}) = (2 - 6) = -4$
- $(x_2 - \bar{x}) = (5 - 6) = -1$
- $(x_3 - \bar{x}) = (6 - 6) = 0$
- $(x_4 - \bar{x}) = (7 - 6) = 1$
- $(x_5 - \bar{x}) = (10 - 6) = 4$

The squared deviations are

- $(x_1 - \bar{x})^2 = (2 - 6)^2 = (-4)^2 = 16$
- $(x_2 - \bar{x})^2 = (5 - 6)^2 = (-1)^2 = 1$
- $(x_3 - \bar{x})^2 = (6 - 6)^2 = (0)^2 = 0$
- $(x_4 - \bar{x})^2 = (7 - 6)^2 = (1)^2 = 1$
- $(x_5 - \bar{x})^2 = (10 - 6)^2 = (4)^2 = 16$

The sum of the squared deviations is

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = 16 + 1 + 0 + 1 + 16 = 34$$

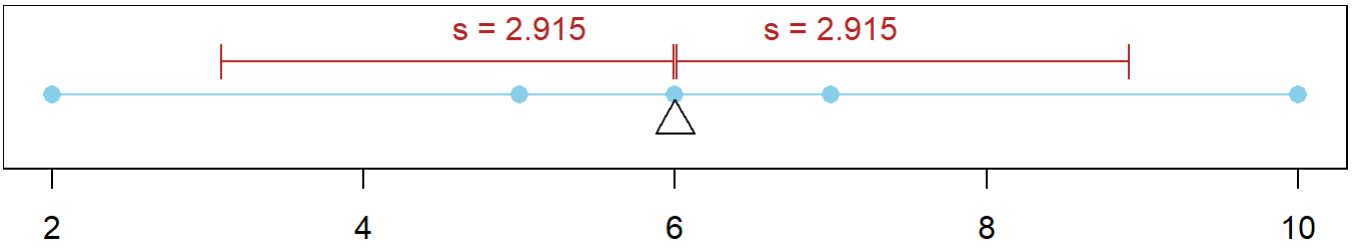
Dividing this by the degrees of freedom, $n - 1$, gives

$$\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1} = \frac{34}{5 - 1} = \frac{34}{4} = 8.5$$

Finally, s is obtained by taking the square root

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1}} = \sqrt{8.5} \approx 2.915$$

The red lines below show how the standard deviation represents all deviations in this data set. Recall that the magnitudes of the individual deviations were $4, 1, 0, 1$, and 4. The representative deviation is 2.915.

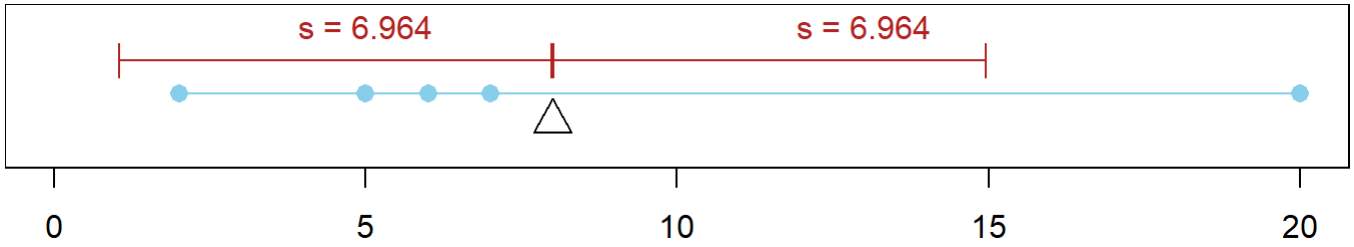


Effect of Outliers

Like the mean, the standard deviation is influenced by outliers. This is shown below by changing x_5 to be 20. Note that the deviation of x_5 is now 12 (instead of 4 like it was previously) and that the mean is now 8. The standard deviation of the altered data

- $x_1 = 2$
- $x_2 = 5$
- $x_3 = 6$
- $x_4 = 7$
- $x_5 = 20$

is now $s \approx 6.964$. Not very "representative" of all the deviations. It is biased towards the largest deviation. It is important to be aware of outliers when reporting the standard deviation s.



Variance

1 Quantitative Variable

Formula:

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

Notice that the formula is just the formula for standard deviation without the square root. This is becasue variance is just standard deviation squared. Great theoretical properties, but seldom used when describing data. Difficult to interpret in context of data because it is in squared units. The standard deviation is typically used instead because it is in the original units and is thus easier to interpret.

1.4 Graphical Summaries

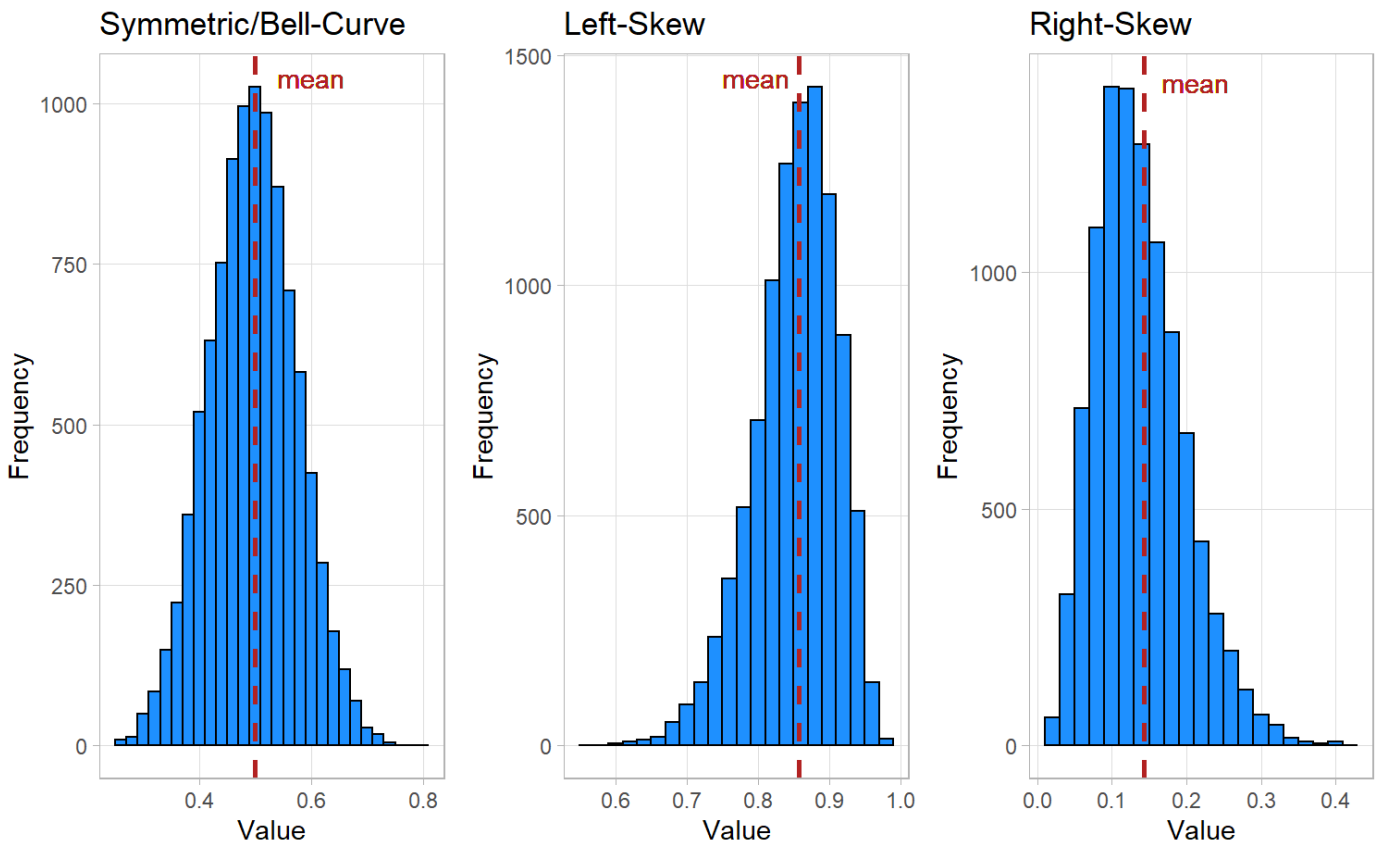
Section 6.2.4 provides links to information about using Tableau to create some of the different types of graphics we will be discussing and using in this cousrse. This section will talk about these graphical summaries more generally, as well as their relevance in statistics and analysis.

Histograms

1 Quantitative Variable

Histograms are great for showing the distribution of data for a single quantitative variable when the sample size is large. Dotplots are a good alternative for smaller sample sizes. Histograms are generally either symmetric/bell-shaped, left-skewed, or right-skewed.

Histograms group data that are close to each other into "bins" (the vertical bars in the plot). The height of a bin is determined by the number of data points that are contained within the bin.



Here is an excellent visualizer for histograms, with a walkthrough of how they're created. You should take some time to interact with and undersand it.

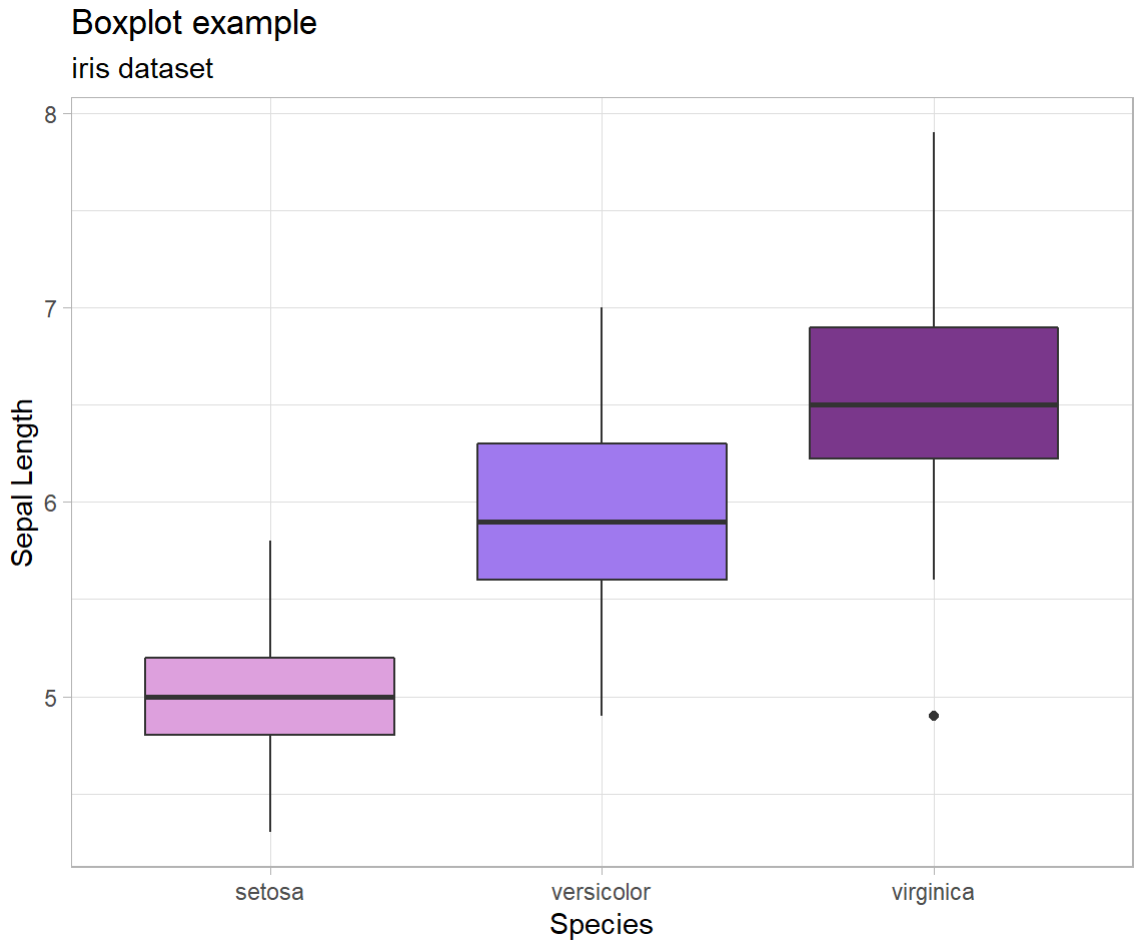
Boxplots

1 Quantitative Variable | 2+ Groups

Graphical depiction of the five-number summary. Great for comparing the distributions of data across several groups or categories. Provides a quick visual understanding of the location of the median as well as the range of the data. Can be useful in showing outliers. Sample size should be larger than at least five, or computing the *five*-number summary is not very meaningful. Side-by-side dotplots are a good alternative for smaller sample sizes.

How Boxplots are Made

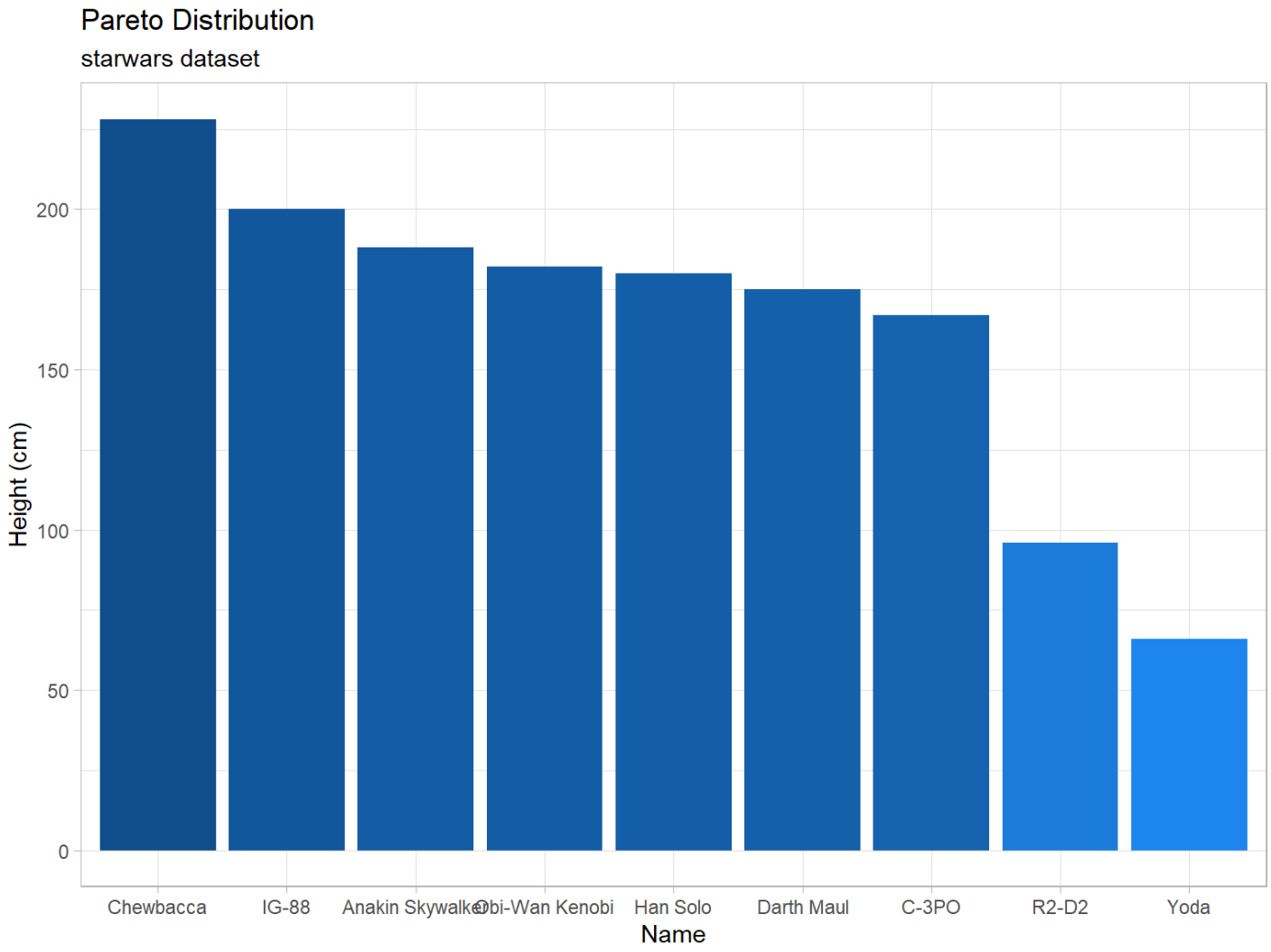
1. The five-number summary is computed.
2. A box is drawn with one edge located at the first quartile and the opposite edge located at the third quartile.
3. This box is then divided into two boxes by placing another line inside the box at the location of the median.
4. The maximum value and minimum value are marked on the plot.
5. Whiskers are drawn from the first quartile out towards the minimum and from the third quartile out towards the maximum.
6. If the minimum or maximum is too far away, then the whisker is ended early.
7. Any points beyond the line ending the whisker are marked on the plot as dots. This helps identify possible outliers in the data.

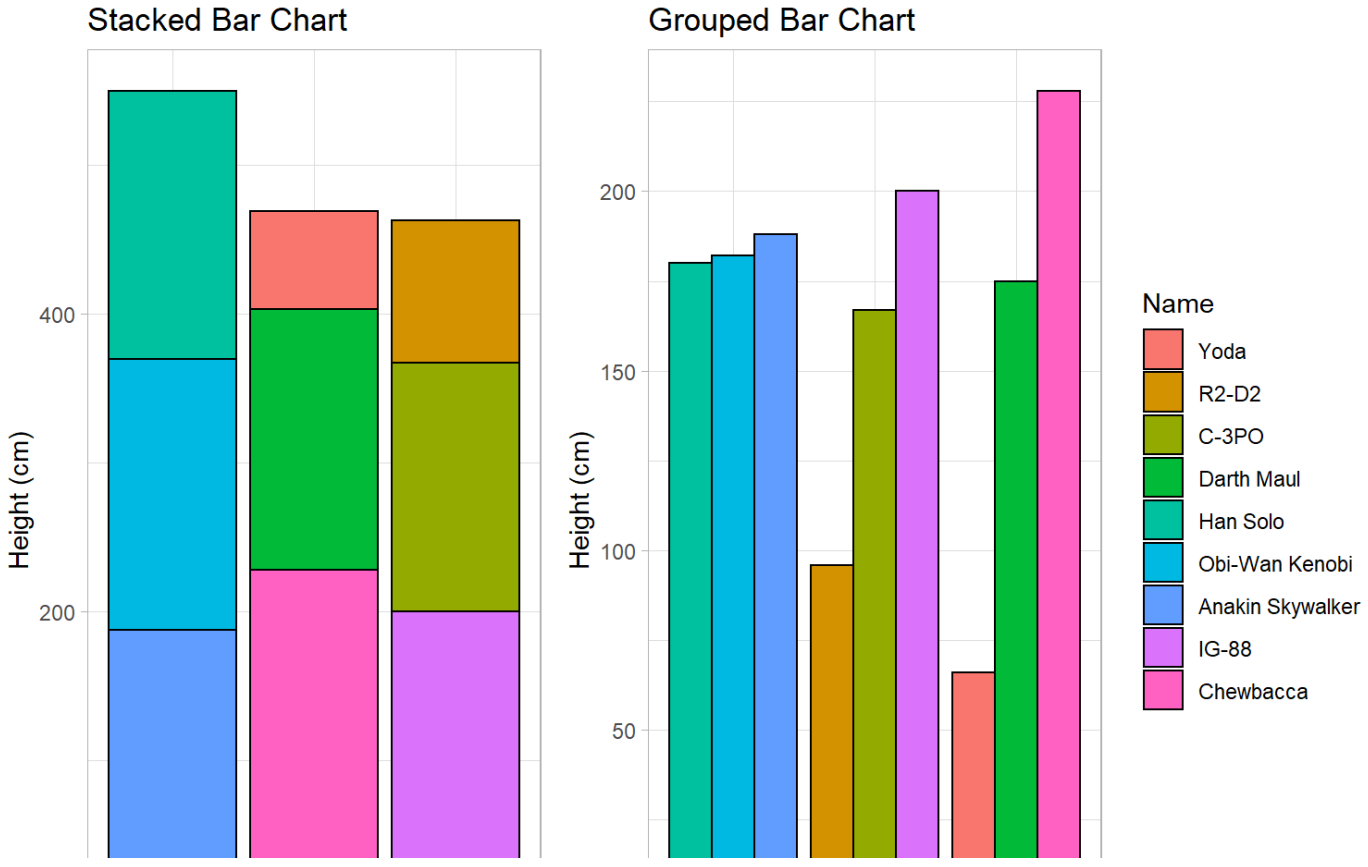


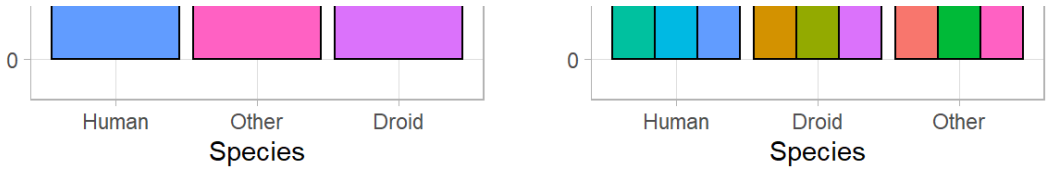
Bar Charts

1 (or 2) Qualitative Variable(s)

Depicts the number of occurrances for each category, or *level*, of the qualitative variable. Similar to a histogram, but there is no natural way to order the bars. Thus the white-space between each bar. It is called a *Pareto* chart if the bars are ordered from tallest to shortest. Grouped and stacked bar charts are often used to display information for two qualitative variables simultaneously.


Pareto Distribution
starwars dataset
Height (cm)
200
150
100
50
0
Chewbacca
IG-88
Anakin Skywalker
Obi-Wan Kenobi
Han Solo
Darth Maul
C-3PO
R2-D2
Yoda
Name


Stacked Bar Chart
Height (cm)
400
200
Grouped Bar Chart
Height (cm)
200
150
100
50
Name
Yoda
R2-D2
C-3PO
Darth Maul
Han Solo
Obi-Wan Kenobi
Anakin Skywalker
IG-88
Chewbacca



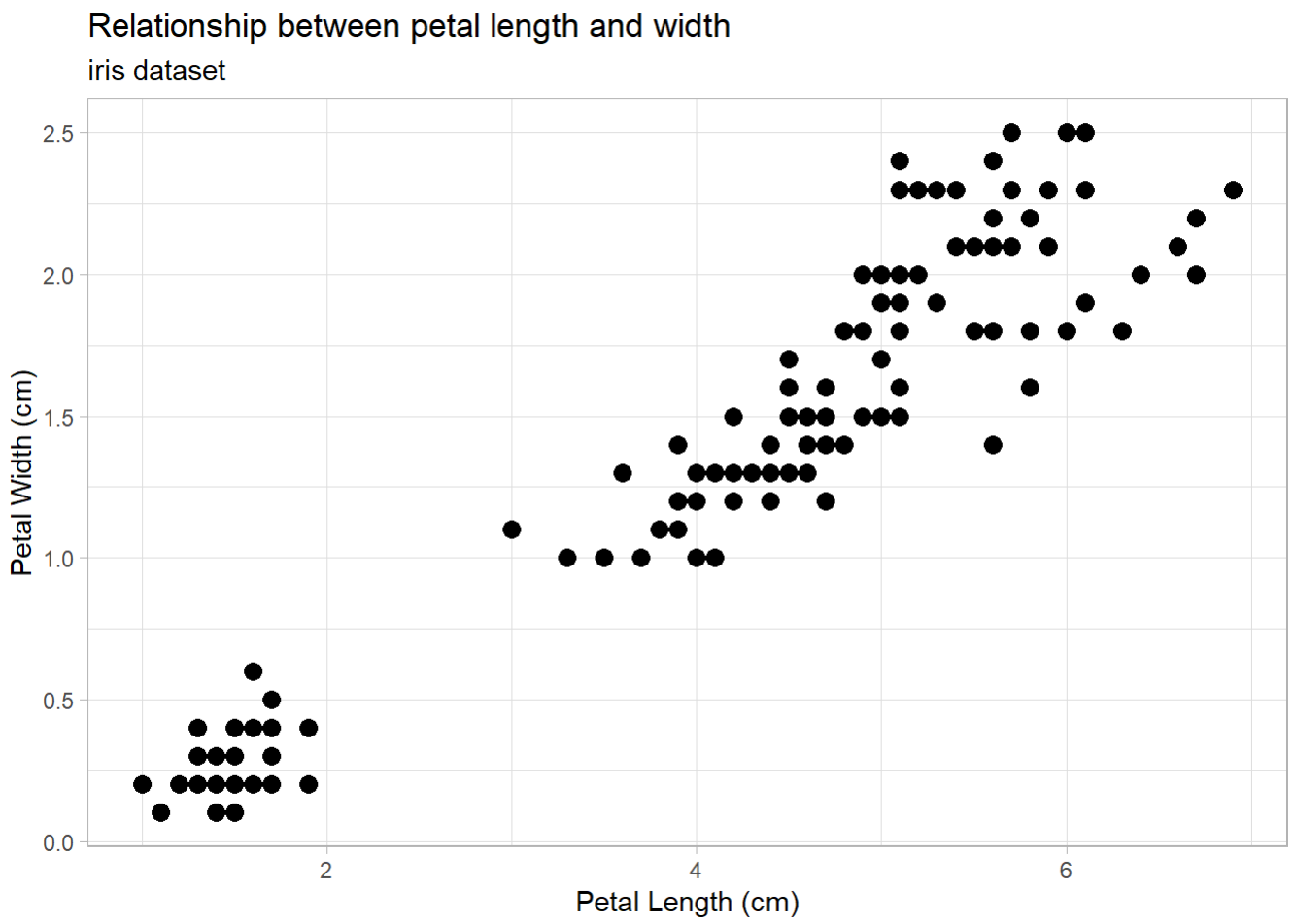
Note that stacked and grouped bar charts each come with their own advantages and disadvantages. Stacked bar charts are an effective way to compare the groups as a whole, while leaving some ability to compare the individual components of the group. Grouped bar charts are excellent for comparing the individuals both within and across groups, while also distinguishing the different groups that the individuals belong to. If you look at the stacked bar chart, you will find that it is hard to tell who is taller between Obi-Wan Kenobi and Darth Maul. You will find that this question is easy to answer though on the grouped plot, Obi-Wan is clearly taller, even if it is only by a handful of centimeters. On the other hand, it is difficult to tell from the grouped bar chart whether the Droid or Other species has the greatest total height. This becomes very easy to answer though using the stacked plot, the Other species has a greater total height.

Stacked bar charts are good for comparing groups as a whole, grouped bar charts are good for comparing individuals within and across groups. Both have their uses, but the decision of which to use should be a deliberate one.

Scatterplots

2 Qualitative Variables

Depicts the actual values of the data points, which are (x,y) pairs. Works well for small or large sample sizes. Visualizes well the correlation between the two variables. Should be used in linear regression contexts whenever possible.

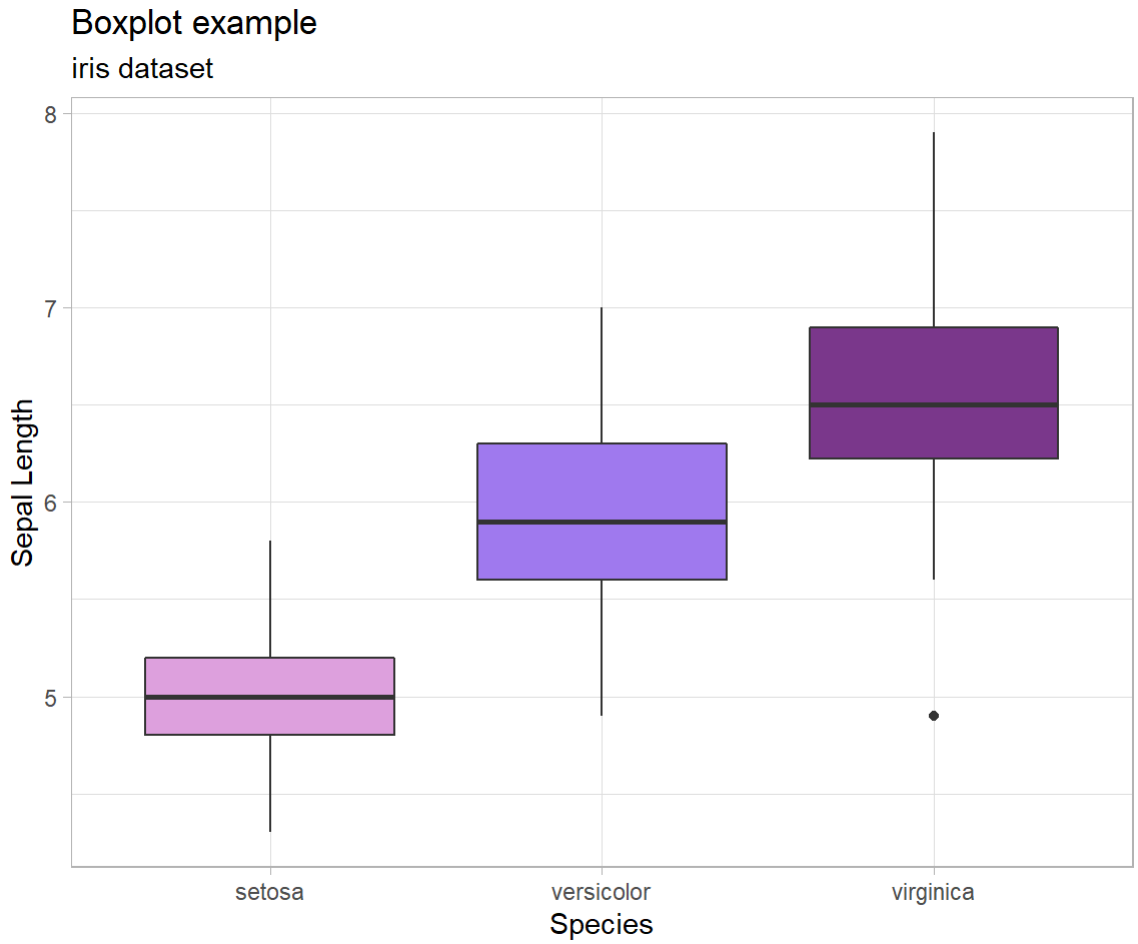


A lot of what you learned in your algebra classes comes into play with scatter plots and linear regression. A positive slope (like in the iris plot above) means that as the x-variable increases the y-variable also increases. A negative slope would mean that an increase in x would correspond with a decrease in y (example: an increase of mileage on a car corresponds with a decrease in its value). We must be careful here though to recognize the difference between **correlation** and **causation**. The iris plot is a good example of correlation. There is an obvious trend between petal length and petal width; when petal length is large we can be confident that the petal width will also be large. **This doesn't mean that a long petal causes a wide petal**, but simply that iris flowers with long petals will usually also have wide petals. There is no causation between the two, just correlation. On the other hand, consider the relationship between how much you use your cell phone and it's battery. The more you use your phone, the lower it's battery level will be, and this is an example of a causal relationship, or causation. Additionally, this is a negative relationship. As one variable increases, the other decreases.

1.5 Data vs. Summaries

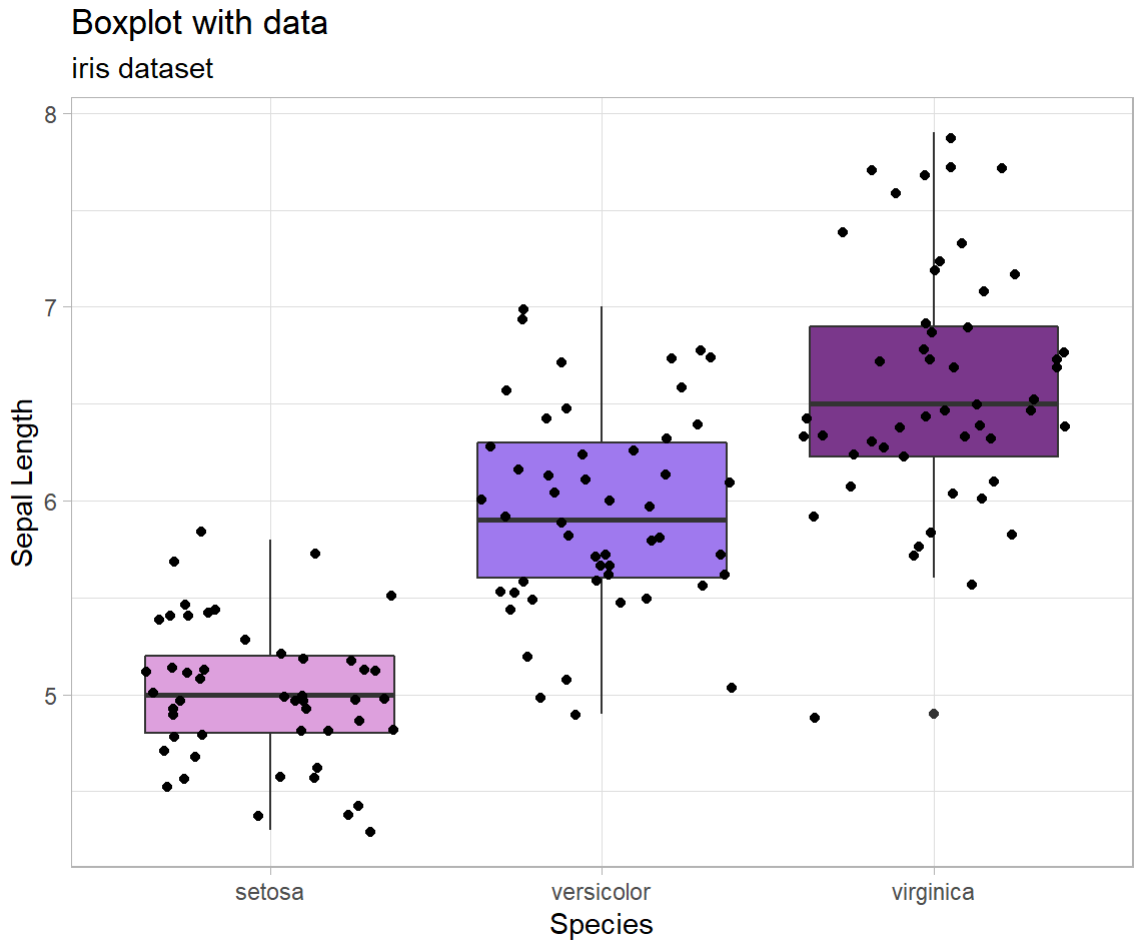
Note that there is a very important difference between *data* and *data summaries*. Data summaries can give us insight into data, but they are not data. Consider the idea of a mean. The mean is generally a good measure of the center of data (at least when the data is non-skewed) and is therefore able to give us some insight into the data. We have to be careful though to remember what the mean doesn't gell us about the data. It doesn't tell us the spread, (see variance and standard deviation) the number of observations that were used to produce it, or really anything about the distrubtion of the data.

As another example, consider again the boxplot for the iris dataset from above:



From this graphic, we are able to understand the five-number summary of each type of iris - and from there make comparisons and even decisions. That said, there is still a lot we don't know. For example, how many observations are there in the dataset? Is it the same number in each group? When there are very few observations, boxplots can be very deceptive. In this case, it turns out that there are 50 observations in each group so the boxplot works well here, but that's not something we could know just from the graphic.

See how the effect of the graphic changes once the data is added into the graphic:



Notice that with the data laid on top of the boxplots we can see that there is indeed more variance in the virginica group than the setosa, as we saw just from the boxplot. Having an understanding of the data itself allows data summaries to have power in decision making.

Chapter 2 Introduction to the Normal Distribution and Z-Scores

2.1 The Normal Distribution

2.2 Z-Scores

Source: MATH 221 Textbook

2.2.1 Introduction to Z-scores

In Ghana, the mean height of young adult women is normally distributed with mean 159.0 cm and standard deviation 4.9 cm. (Monden & Smits, 2009) Serwa, a female BYU-Idaho student from Ghana, is 169.0 cm tall. Her height is $169.0 - 159.0 = 10$ cm greater than the mean. When compared to the standard deviation, she is about two standard deviations ($\approx 2 \times 4.9$ cm) taller than the mean.

The heights of men are also normally distributed. The mean height of young adult men in Brazil is 173.0 cm ("Oramento," 2010), and the standard deviation for the population is 6.3 cm. (Castilho & Lahr, 2001) A Brazilian BYU-Idaho student, Gustavo, is 182.5 cm tall. Compared to other Brazilians, he is taller than the mean height of Brazilian men.

When we examined the heights of Serwa and Gustavo, we compared their height to the standard deviation. If we look carefully at the steps we did, we subtracted each individual's height from the mean height for people of the same gender and nationality.

2.2.2 Computing Z-scores

This shows how much taller or shorter the person is than the mean height. In order to compare the height difference to the standard deviation, we divide the difference by the standard deviation. This gives the number of standard deviations the individual is above or below the mean.

For example, Serwa's height is 169.0 cm. If we subtract the mean and divide by the standard deviation, we get

$$z = \frac{169.0 - 159.0}{4.9} = 2.041$$

We call this number a z-score. The z-score for a data value tells how many standard deviations away from the mean the observation lies. If the z-score is positive, then the observed value lies above the mean. A negative z-score implies that the value was below the mean.

We compute the z-score for Gustavo's height similarly, and obtain

$$z = \frac{182.5 - 173.0}{6.3} = 1.508$$

Gustavo's z-score is 1.508. As noted above, this is about one-and-a-half standard deviations above the mean. In general, if an observation x is taken from a random process with mean μ and standard deviation σ, then the z-score is

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

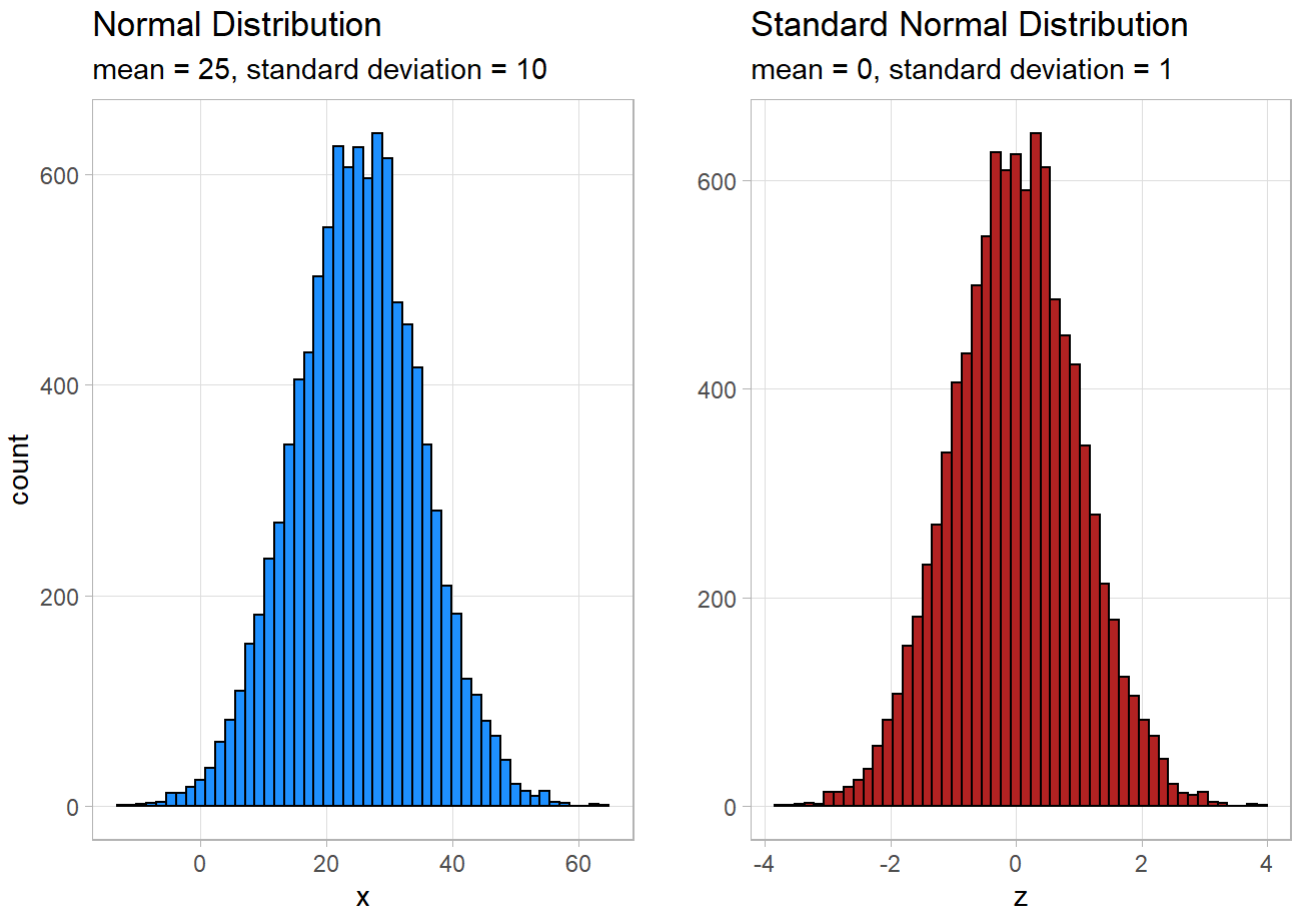
The z-score can be computed for data from any distribution, but it is most commonly applied to normally distributed data.

2.3 Standard Normal Distributions

Source: MATH 221 Textbook (excluding graphic)

A standard normal distribution is a normal distribution with a mean of 0 and a standard deviation of 1. We call it a *standard* normal because the values are *standardized*. No matter what the distribution is about (heights, temperatures, etc.), if the distribution is a standard normal then they are all on the same scale. This is useful when comparing distributions because differences in units or scales no longer need to be considered in the comparison.

As shown below, standardizing values doesn't alter the shape of the distribution. The counts in each bin change very little, and any discrepancies are likely to be as a result of rounding error as much as anything.



variable	mean	standard deviation
x	24.98	9.98
z	0	1

2.4 Rules of Thumb

2.4.1 68-95-99.7% Rule

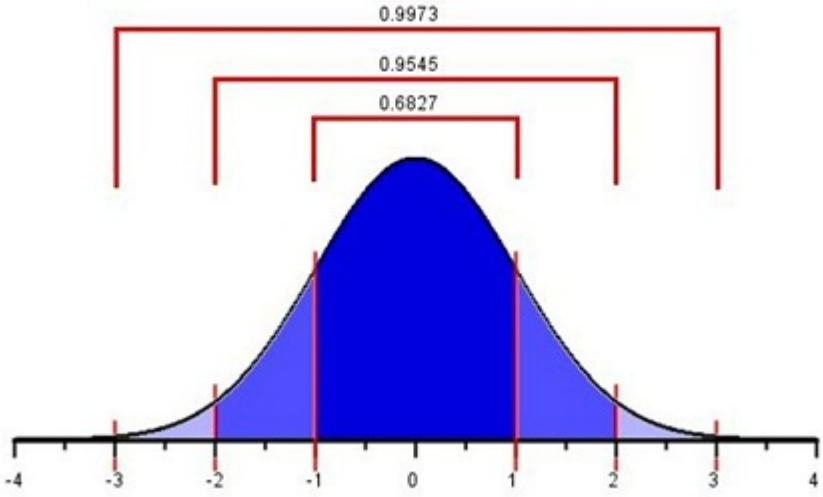
Source: MATH 221 Textbook

Heights of women (or men) in a particular population follow a normal distribution. Knowing this, think about the following statement:

Most people's heights are close to the mean. A few are very tall or very short.

While this may be true, suppose we would like to make a more precise statement than this. To do so, we can use the following rule of thumb:

For any bell-shaped distribution:
- 68% of the data will lie within 1 standard deviation of the mean,
- 95% of the data will lie within 2 standard deviations of the mean, and
- 99.7% of the data will lie within 3 standard deviations of the mean.



68-95-99.7% Rule

This is called the *68-95-99.7% Rule for Bell-shaped Distributions*. Some statistics books refer to this as the Empirical Rule.

Approximately 68% of the observations from a bell-shaped distribution will be between the values of $\mu - \sigma$ and $\mu + \sigma$. Consider the heights of young adult women in Ghana. We expect that about 68% of Ghanian women have a height between the values of

$$\mu - \sigma = 159.0 - 4.9 = 154.1 \text{ cm}$$

and

$$\mu + \sigma = 159.0 + 4.9 = 163.9 \text{ cm}.$$

So, if a female is chosen at random from all the young adult women in Ghana, about 68% of those chosen will have a height between 154.1 and 163.9 cm. Similarly, 95% of the women's heights will be between the values of

$$\mu - 2\sigma = 159.0 - 2(4.9) = 149.2 \text{ cm}$$

and

$$\mu + 2\sigma = 159.0 + 2(4.9) = 168.8 \text{ cm}.$$

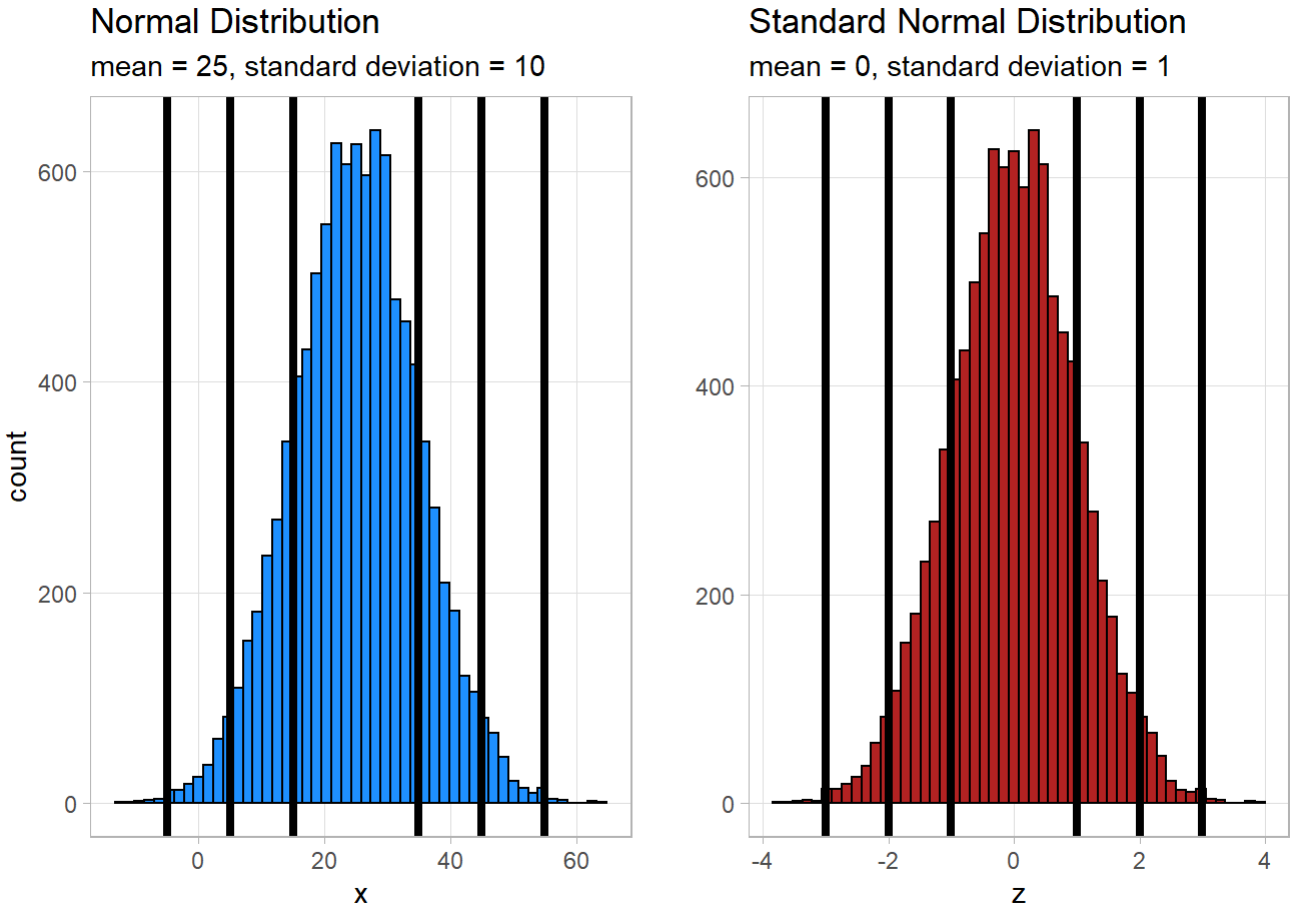
Finally, 99.7% of the women's heights will be between

$$\mu - 3\sigma = 159.0 - 3(4.9) = 144.3 \text{ cm}$$

and

$$\mu + 3\sigma = 159.0 + 3(4.9) = 173.7 \text{ cm}.$$

Consider again the plots from above, this time with 1, 2, and 3 standard deviations from the mean marked.



As you can see, the spread of the data is the same whether the data has been standardized or not. The 68-95-99.7% Rule still holds true because even though the scale has changed the relative distance of the data from the mean has not.

Unusual Events

If a z-score is extreme (either a large positive number or a large negative number), then that suggests that that observed value is very far from the mean. The 68-95-99.7% rule states that 95% of the observed data values will be within two standard deviations of the mean. This means that 5% of the observations will be more than 2 standard deviations away from the mean (either to the left or to the right).

We define an **unusual observation** to be something that happens less than 5% of the time. For normally distributed data, we determine if an observation is unusual based on its z-score. We call an observation unusual if $z < -2$ or if $z > 2$. In other words, we will call an event unusual if the absolute value of its z-score is greater than 2.

2.4.2 Standard Deviation Estimation

While the concept of the standard deviation isn't terribly complicated, computing it by hand isn't particularly convenient. A general rule of thumb for estimating it is to subract the miniumum from the maximum, and then divide that difference by either 4 or 6.

$$\frac{Maximum - Minimum}{4 \text{ or } 6}$$

If accuracy is important, it is recommended that you actually compute the standard deviation using software or the formula here.

See this website for a more detailed explanation on why this trick is a valid method, but it follows the principle that nearly all of the data should fall within 2 (or 3) standard deviations on either side of the mean, or within 4 (or 6) total standard deviations.

Chapter 3 Probability, Sampling, and Confidence Intervals

3.1 Probability

Principles of probability are essential to statistics. It is through probability that we understand how likely events are, which then allows us to make data-driven decisions. This course and textbook aren't sufficient to gain an in-depth understanding of probability, but a few of the basics will be covered.

3.1.1 Probability Notation

Source: MATH 221 Textbook

You may already have a good understanding of the basics of probability. It is worth noting that there is a special notation used to denote probabilities. The probability that an event, x, will occur is written $P(x)$. As an example, the probability that you will roll a 6 on a die can be written as

P (Roll a 6 on a die)= $\frac{1}{6}$

3.1.2 Rules of Probability

Source: MATH 221 Textbook

Probabilities follow patterns, called **probability distributions,** or distributions, for short. There are three rules that a probability distribution must follow.

The three rules of probability are:

- **Rule 1**: The probability of an event X is a number between 0 and 1.

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

- **Rule 2**: If you list all the outcomes of an experiment (such as rolling a die) the probability that one of these outcomes will occur is 1. In other words, the sum of the probabilities of all the possible outcomes of any experiment is 1.

$$\sum P(X) = 1$$

- **Rule 3**: (Complement Rule) The probability that an event X will not occur is 1 minus the probability that it will occur.

$$P(\text{not } X) = 1 - P(X)$$

You may have noticed that the Complement Rule is just a combination of the first two rules.

3.2 Sampling from a Population

Very rarely do we have access to an entire population for one reason or another (too large, not enough resources, etc.), so we are left with taking samples from the population that should be representative of that population. If we had access to the entire population then we wouldn't need to do statistical tests or analysis, we would just make observations on the whole population. The goal of statistical analysis is to determine if what we see in a sample is likely to occur in the population. For example, if we observe a common trend among 100 BYU-Idaho can we assume that that trend will hold for all BYU-Idaho students? Or as a made-up larger scale example, assume that 1000 Toyota Camrys are tested and 1% of them are found to have a defect in the braking system. Can Toyota assume that 1% of all Toyta Camrys will have the same defect? Through statistical analysis we are able to obtain answers to these questions.

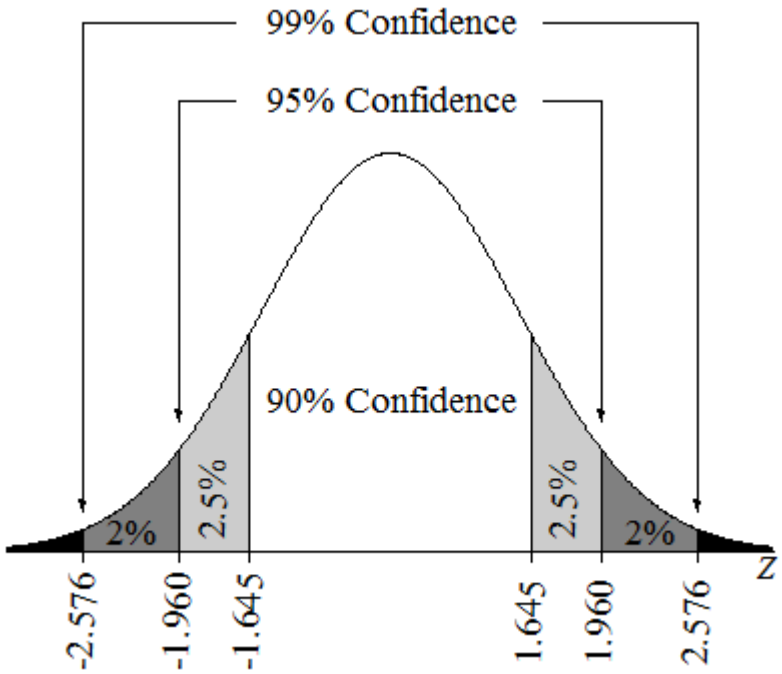
Hopefully this helps you see the importance of sampling. Even if we aren't able to observe every member of a population, through proper sampling and statistical analysis we are able to gain insight into the population as a whole. For those insights to be valid, however, the sampling must be done in a statistically correct way. Not just any sample can be taken, so methods for sampling have been developed. Some of the more common methods are shown below.

Source: MATH 221 Textbook

- There are many sampling methods used to obtain a **sample** from a **population**:
    - A **simple random sample (SRS)** is a random selection taken from a population

- A **systematic sample** is every k^{th} item in the population, beginning at a random starting point
- A **cluster sample** is all items in one or more randomly selected clusters, or blocks
- A **stratified sample** divides data into similar groups and an **SRS** is taken from each group
- A **convenience sample** is one easily obtained in a less-than-systematic way and should be avoided whenever possible

3.3 Confidence Intervals



90%, 95%, and 99% Confidence Intervals

In statistics, we usually don't know the exact values of the population parameters so we use statistical methods to approximate them. For example, we might take a random sample from the population, calculate the sample mean, and use that value as an estimate for the population mean. The sample mean is an example of a **point estimator** because it is a single value, or a point. There are also **interval estimators**, which instead offer a range, or interval, of values which are likely to contain the value we are trying to estimate. Arguably the most common interval estimator is the **confidence interval**.

3.3.1 Definition and Interpretation

First, a confidence interval is always associated with some percentage, or a probability. For example, a 95% confidence interval. Second, we find confidence intervals for values. Perhaps the most common confidence interval is a **95% confidence interval for the mean**. The correct interpretation of this 95% confidence interval would be: “We are 95% confident that the true mean lies within the lower and upper bounds of the confidence interval.”

Notice that with this interpretation we aren’t saying anything about the exact value of the population mean, we are simply giving a range of values that the mean is very likely to lie in.

Example

Source: MATH 221 Textbook

Consider the 95% confidence interval for the true mean of 25 rolls of a fair die. We find the 95% confidence interval to be: (2.37,3.71). When we interpret this confidence interval, we say, “We are 95% confident that the true mean is between 2.37 and 3.71.”

The word, “confident” implies that if we repeated this process many, many times, 95% of the confidence intervals we would get would contain the true mean μ. It does not imply anything about whether or not one specific confidence interval will contain the true mean.

We do not say that “there is a 95% probability (or chance) that the true mean is between 2.37 and 3.71.” The probability that the true mean μ is between 2.37 and 3.71 is either 1 or 0.

3.3.2 Finding a Confidence Interval

More often than not, confidence intervals will be 95% confidence intervals. Think back to the **68-95-99.7% Rule for Bell-Curves** from last chapter, especially note the 95. Assuming the data is approximately normally distributed, then approximately 95% of the data lies within two standard deviations of the mean, so by computing the values that are two standard deviations away from the mean on either side we compute the 95% confidence interval; with the upper bound of the CI being the mean plus two standard deviations ($\mu + 2\sigma$), and the lower bound being the mean minus two standard deviations ($\mu - 2\sigma$).

Another way to think of this is to say that if the data is approximately normally distributed then the true mean will lie within the 95% confidence interval approximately 95% of the time, or within two standard deviations of the sample mean 95% of the time.

For a step-by-step example of finding a confidence interval and a real-world example, see How to Determine the Confidence Interval for a Population Proportion.

Chapter 4 Hypothesis Tests

4.1 Hypothesis Testing

Source: MATH 221 Textbook

Whenever sample data is used to infer a characteristic of a population, it is called making an inference. **Inferential statistics** represents a collection of methods that can be used to make inferences about a population.

This foundational assumption is called the null hypothesis. The **null hypothesis** is a statement about the population that represents the status quo, conventional wisdom, or what is generally accepted as true. Using the made-up Toyota Camry example from the last chapter, the null hypothesis is:

H_0: 1% of all Toyota Camrys have a defect in the braking system.

The purpose of a statistical study or experiment is to see if there is sufficient evidence against the null hypothesis. If there is sufficient evidence, we reject the null hypothesis. If the null hypothesis is rejected, it is rejected in favor of another statement about the population: the **alternative hypothesis**. In our example, let's assume Toyota wants to know if more than 1% of the population has a deffective braking system. The alternative hypothesis would then be:

H_a: More than 1% of all Toyota Camrys have a defect in the braking system

Notice that both the null and alternative hypotheses are statements about the population, not just the sample being tested. Again, the goal is to determine if what is observed in the sample can be assumed to be true in the population.

There is a formal procedure for testing the null and alternative hypotheses, called a **hypothesis test**. In a hypothesis test, the null hypothesis is always assumed to be true. If there is sufficient evidence against the null hypothesis, it is rejected. The evidence against the null hypothesis is assessed using a number called the P-value. The **P-value** is the probability of obtaining a result (called a test statistic) at least as extreme as the one you calculated, assuming the null hypothesis is true. We reject the null hypothesis if the P-value is small, say less than 0.05. If we

assume that only 1% of Camrys have the defective braking system, the P-value is the probability of observing a number of defective cars that is as large or larger than that which was observed in the test sample.

For this example (again, which is completely made up), the P-value was determined to be 0.68. Assuming the null hypothesis is true, the probability of observing defects in the braking system at least as often as was observed in the test sample was 0.68. This is a very large value. So, it is not surprising to have observed the number of the defects in the sample in this case. The probability that these differences could occur due to chance is very high. The conclusion is that the null hypothesis should not be rejected.

If the P-value is low, the null hypothesis is rejected. If this probability is large, the null hypothesis is not rejected.

Hypothesis tests sometimes lead accidentally to incorrect conclusions because we use data from samples (as opposed to data from entire populations). When random samples are selected, some of the samples will contain disproportionately few or many cars with defects, just by chance.

Think about drawing marbles from a container in which most of the marbles are white and a few are red. Each marble represents a Toyota Camry, and the red marbles represent Camrys with defective brakes.

If you choose a random sample of the marbles in the jar, you might get all the red marbles in your sample, just by chance. This might lead you to conclude that there are many red marbles in the container, which is false. This is like Toyota rejecting the null hypothesis when it is true, because their sample contains more Camrys with bad brake systems than it should—just due to chance.

Likewise, when drawing marbles from your container, you might select none of the red marbles. This may lead you to conclude that there are no red marbles in the container, or very few, which is false. This is like Toyota failing to reject the null hypothesis when it is false, because their sample contains fewer cars with bad brakes than it should—again, just due to chance.

Notice that if you draw only one marble, it will be either white or red, and you will be in one of the situations discussed in the previous two paragraphs. On the other hand, if you draw a larger sample, say 40, the chances are you will get a pretty good idea of the proportion of red to white marbles. Certainly better than if you only draw one marble. This emphasizes the roll of sample size in making inference; in general, the larger the sample size the better we understand the population.

Such errors are no one's fault; they are an inherent part of hypothesis testing. They make it impossible for us to be certain of the conclusions we draw using the statistical process. The thing to remember is that if we carry out the process correctly, our results are correct often enough to be very useful.

For more information about hypothesis testing, look at the Making Inference section of the Statistics Notebook.

This course will focus on Chi-Square testing, but there are several other types of hypothesis tests available to statisticians depending on the circumstances of their test and data. More information about some of them can be found under the "Making Inference" tab in the Statistics Notebook. The course *Intermediate Statistics* (MATH 325) covers these different tests in more depth.

4.2 Test for Two Proportions

Source: MATH 221 Textbook

The ability to taste the chemical Phenylthiocarbamide (PTC) is hereditary. Some people can taste it, while others cannot. Even though the ability to taste PTC was observed in all age, race, and sex groups, this does not address the issue about whether men or women are more likely to be able taste PTC.

Further exploration of the PTC data allows us to investigate if there is a difference in the proportion of men and women who can taste PTC. The following contingency table summarizes Elise Johnson's results:

Can Taste PTC?	Female	Male	Total
No	15	14	29
Yes	51	38	89
Total	66	52	118

Researchers want to know if the ability to taste PTC is a sex-linked trait. This can be summarized in the following research question: **Is there a difference in the proportion of men and the proportion of women who can taste PTC?** The hypothesis is that there is no difference in the

the true proportion of men who can taste PTC compared to the true proportion of women who can taste PTC.

$$H_0 : \text{There is no difference in the proportion of PTC tasters for men and women}$$

$$H_1 : \text{There is a difference in the proportion of PTC tasters for men and women}$$

A sample of 66 females and 52 males were provided with PTC strips and asked to indicate if they could taste the chemical or not. (This research was approved by the BYU-Idaho Institutional Review Board.)

When working with categorical data, it is natural to summarize the data by computing proportions. If someone has the ability to taste PTC, we will call this a success. The sample proportion is defined as the number of successes observed divided by the total number of observations. For the females, the proportion of the sample who could taste the PTC was:

$$\overbrace{\hat{p}_1}^{\text{Proportion of females PTC tasters}} = \frac{x_1}{n_1} = \frac{51}{66}$$

This is approximately 77.3% of the people who were surveyed. For the males, the proportion who could taste PTC was:

$$\overbrace{\hat{p}_2}^{\text{Proportion of male PTC tasters}} = \frac{x_2}{n_2} = \frac{38}{52}$$

This works out to be about 73.1%.

Recall the hypothesis described earlier, in simpler terms, it can be read as

$$H_0 : p_1 = p_2$$

$$H_1 : p_1 \neq p_2$$

If the null hypothesis is true, then the proportion of females who can taste PTC is the same as the proportion of males who can taste PTC.

The test statistic is a z, and is given by:

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2)}}}$$

In our case:

$$z = \frac{(\frac{51}{66} - \frac{38}{118}) - (0)}{\sqrt{\frac{89}{118}(1 - \frac{89}{118})(\frac{1}{66} + \frac{1}{52)})}}$$

Assume we use $\alpha = 0.05$ to help us make our decision of whether to reject or accept the null hypothesis.

After plugging in our values, we end up with $z = 0.526$. When converted to a P-value, we get $\text{P-value} = 0.599 > 0.05 = \alpha$. **We fail to reject the null hypothesis**. In English we say, there is insufficient evidence to suggest that the true proportion of males who can taste PTC is different from the true proportion of females who can taste PTC.

Men and women appear to be able to taste PTC in equal proportions. There is not enough evidence to say that one gender is able to taste PTC more than the other. It appears that the ability to taste PTC is not a sex-linked trait.

4.3 Chi-Squared Test of Independence

People often wonder whether two things influence each other. For example, people seek chiropractic care for different reasons. We may want to know if those reasons are different for Europeans than for Americans or Australians. This question can be expressed as “Do reasons for seeking chiropractic care depend on the location in which one lives?”

This question has only two possible answers: “yes” and “no.” The answer “no” can be written as “Motivations for seeking chiropractic care and one’s location are independent.” (The statistical meaning of “independent” is too technical to give here. However, for now, you can think of it as meaning that the two variables are not associated in any way. For example, neither variable depends on the other.) Writing the answer “no” this way allows us to use it as the null hypothesis of a test. We can write the alternative hypothesis by expressing the answer “yes” as “Motivations for seeking chirporactic care and one’s location are not independent.” (Reasons for wording it this way will be given after you’ve been through the entire hypothesis test.)

When we have our observed counts in hand, software will calculate the counts we should expect to see, if the null hypothesis is true. We call these the “expected counts.” The software will then subtract the observed counts from the expected counts and combine these differences to create a single number that we can use to get a P-value. That single number is called the χ2 test statistic. (Note that χ is a Greek letter, and its name is “ki”, as in “kite”. The symbol χ2 should be pronounced “ki squared,” but many people pronounce it “ki-square.”)

4.3.1 Assumptions

The following requirements must be met in order to conduct a χ2 test of independence:

- You must use simple random sampling to obtain a sample from a single population.
- Each expected count must be greater than or equal to 5. Let's walk through the rest of the chiropractic care example.

A study was conducted to determine why patients seek chiropractic care. Patients were classified based on their location and their motivation for seeking treatment. Using descriptions developed by Green and Krueter, patients were asked which of the five reasons led them to seek chiropractic care :

- Wellness: defined as optimizing health among the self-identified healthy
- Preventive health: defined as preventing illness among the self-identified healthy
- At risk: defined as preventing illness among the currently healthy who are at heightened risk to develop a specific condition
- Sick role: defined as getting well among those self-perceived as ill with an emphasis on therapist-directed treatment
- Self care: defined as getting well among those self-perceived as ill favoring the use of self vs. therapist directed strategies The data from the study are summarized in the following contingency table :

Location	Wellness	Preventive Health	At Risk	Sick Role	Self Care	Total
Europe	23	28	59	77	95	282
Australia	71	59	83	68	188	469
United States	90	76	65	82	252	565
Total	184	163	207	227	535	1316

The research question was whether people's motivation for seeking chiropractic care was independent of their location: Europe, Australia, or the United States. The hypothesis test used to address this question was the chi-squared (χ2) test of independence. (Recall that the Greek letter χ is pronounced, "ki" as in "kite.")

The null and alternative hypotheses for this chi-squared test of independence are:

$$H_0 : \text{The location and the motivation for seeking treatment are independent}$$

$$H_1 : \text{The location and the motivation for seeking treatment are not independent}$$

When the Test statistic ($\chi^2 = 49.743$) is calculated, we get a p-value that is essentially 0, which is lower than our $\alpha = 0.05$ and thus we **reject the null hypothesis**.

4.3.2 Interpretation

If the null hypothesis is true, then the interpretation is simple, the two variables are independent. End of story. However, when the null hypothesis is rejected and the alternative is concluded, it becomes interesting to interpret the results because all we know now is that the two variables are somehow associated.

One way to interpret the results is to consider the individual values of

$$\frac{(O_i - E_i)^2}{E_i}$$

which, when square-rooted are sometimes called the Pearson residuals.

$$\sqrt{\frac{(O_i - E_i)^2}{E_i}} = \frac{(O_i - E_i)}{E_i}$$

The Pearson residuals allow a quick understanding of which observed counts are responsible for the χ2 statistic being large. They also show the direction in which the observed counts differ from the expected counts.

4.4 Chi-Square Goodness of Fit Test

The chi-square goodness of fit test is shown below.

$$\sum \frac{(O_i - E_i)^2}{E_i}$$

The purpose of this test is to show whether a group of observations follows an expected distribution.

4.4.1 Hypothesis

For a chi-square goodness of fit test, the hypotheses take the following form.

- Ho: The data are consistent with a specified distribution.
- Ha: The data are not consistent with a specified distribution.

These are then to be compared to an α (alpha) of your choosing (0.01, 0.05, 0.1)

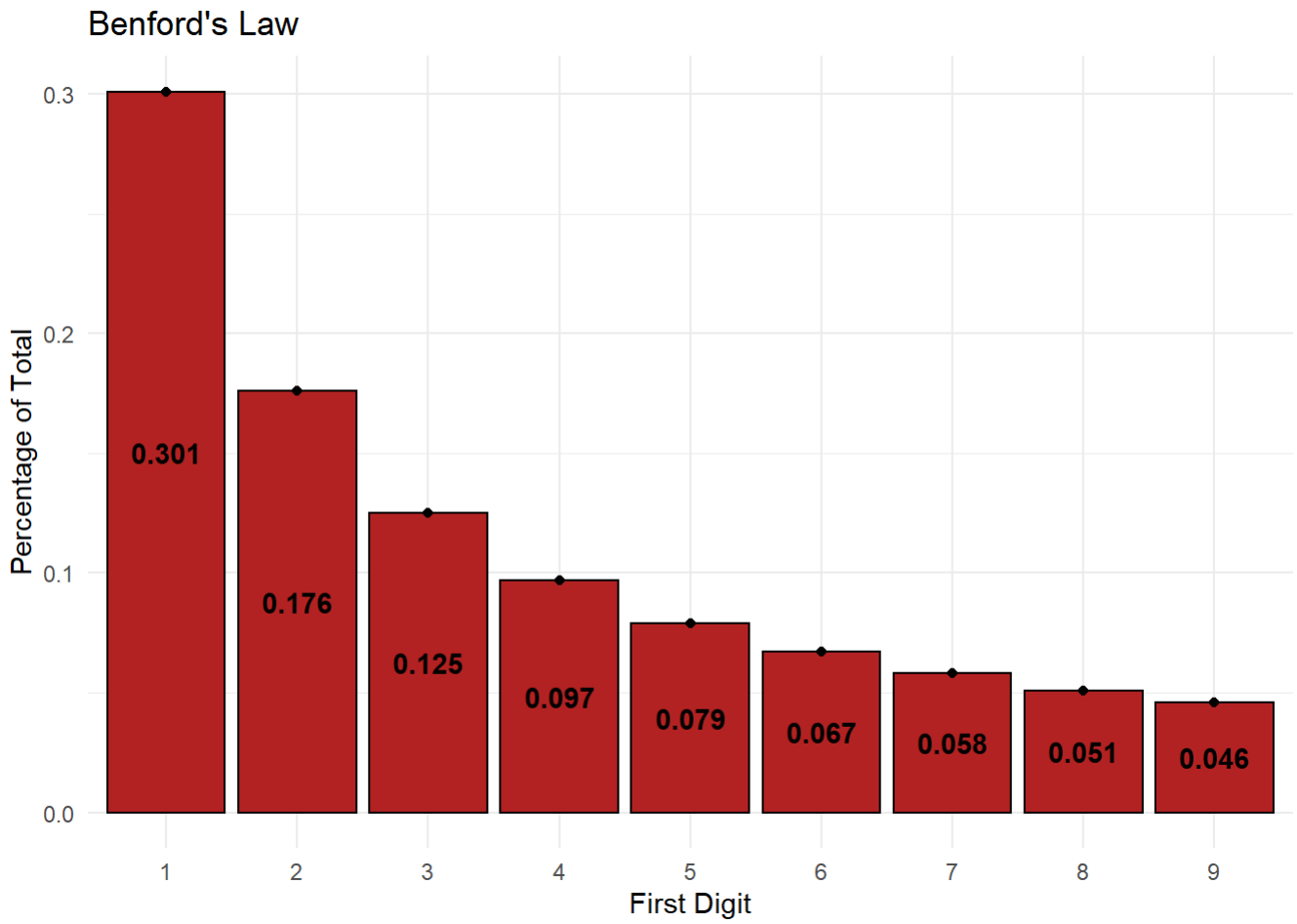
4.4.2 Assumptions

The chi-square goodness of fit test is appropriate when the following conditions are met:

- The data are from a simple random sample
- The groups that are being looked at is categorical, i.e. Qualitative
- There are at least five expected observations per group.

4.4.3 Benford's Law

Let's take a look at Benford's law as an example of how to use the chi-square goodness of fit test. Benford's law states that the first digits of a random group of numbers always follows a certain pattern.

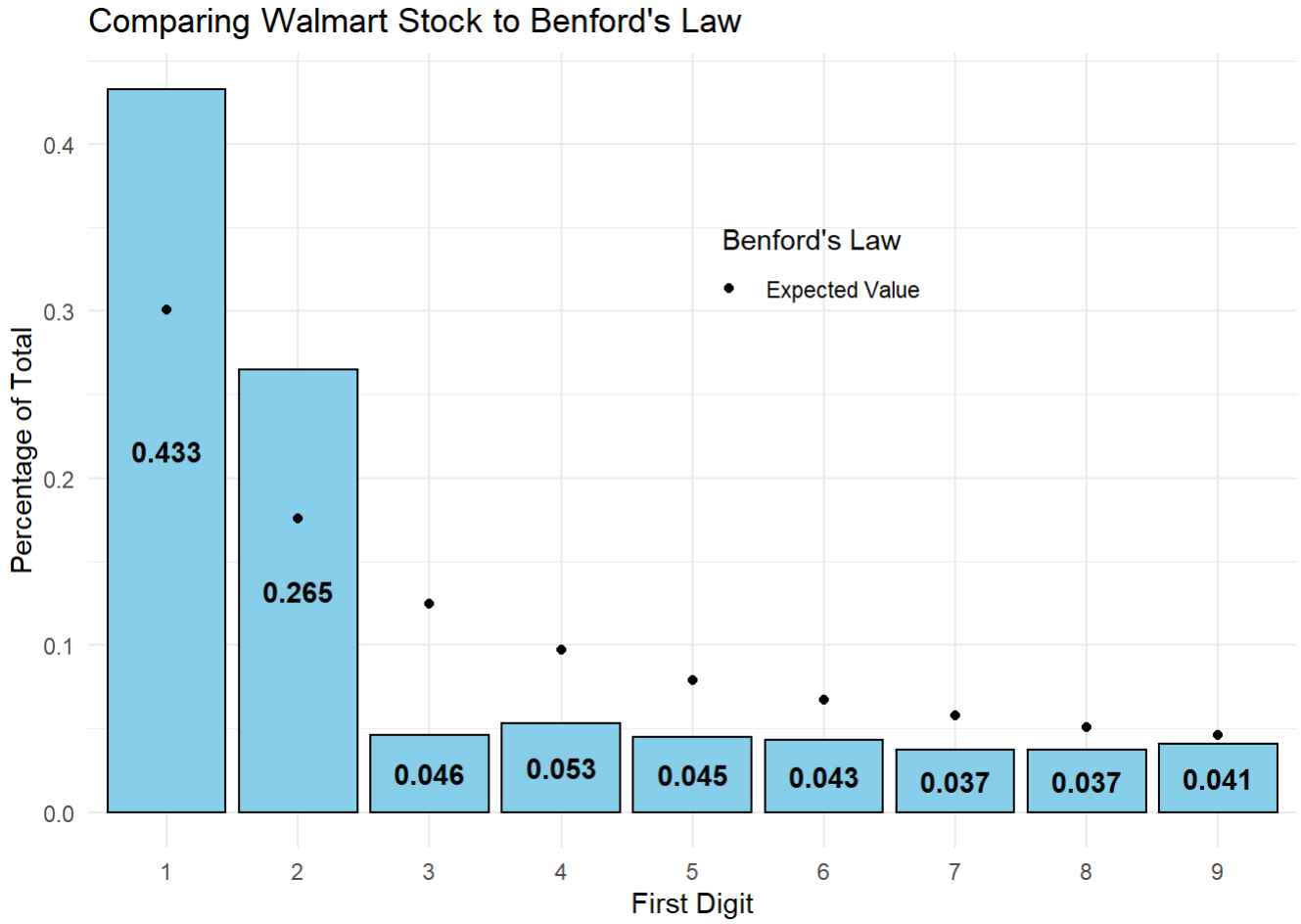


We can assume that if I were to take a set of numbers and pull their first digits from that set then we should see this same distribution or something close to it. This is the purpose of the chi-square goodness of fit test.

Example

Let's take a look at the distribution of the first digits of the Walmart stock. This is a good test set because there are a lot of records of what the stock has been over many years, and we also know that it's truly random. From there we will randomly subset all of walmart data and we will

perform a goodness of fit test.



As we can se it looks pretty similar to Benford's Law, but lets take a look at the actual p-value calculated from the goodness of fit test. The first thing we need to do is find the expected counts of what Benford's Law says the distribution should be. To do this, we multiply Benford's distribution percentages by the total.

301, *176*, *125*, *97*, *79*, *67*, *58*, *51* and *46*

This should be the expected count, now let's calculate the chi-squared value based on the formula at the beginning of this page. In layman's - The observed values minus the expected values all squared / divided by the expected values, all summed.

$$\sum \frac{(O_i - E_i)^2}{E_i}$$

The chi-square value is *208*

We can now calculate the the p-value by using a chi-square test. The p-value that is produced is: *0.00000000000000000000000000000000000000347*

As wee can see the p-value is less than our α of .05, therefore we fail to reject and we can say that the random sample of numbers from the Walmart stock follows Benford's Law.

Chapter 5 What is Data Science?

5.1 What is a Data Scientist?

Data science is an exciting field with a lot of expected growth and interesting opportunities. But what exactly is a data scientist? What do they do?

One writer gave some insight into the day-to-day life of a data scientist in **this article**.

Some questions you may want to ask yourself as you read this article that might help you see how a data scientist working in industry thinks and works, and how you would fit in to this kind of a job:

- What questions is the writer asking himself?
- What questions is he asking others? (colleagues, employers, etc)
    - Think of this perhaps as what types of questions is he asking, rather than the literal questions. What information is he trying to gain through the questions? Why those questions? What's good about them?
- What surprised you about the day-to-day life of a data scientist?
- What interests or excites you?
- What are you not looking forward to, or what would you not enjoy if this was your job?
- What aspects of the job that he describes do you feel like you need more information about?
- What are the skills (both hard and soft) that the data scientist possesses?

To be an effective data scientist, you must become a problem solver. Everything you will do in your career will be about solving some kind of problem using data. This usually requires learning to think about problems in an organized way. This article discusses the practice of **structured thinking**:

The art of structured thinking and analyzing

Some things you may want to ponder:
- How would you define structured thinking?
- How could past projects or work that you've done have been helped through more structured

thinking?
- How can you demonstrate to potential employers that you've developed a structured thinking approach to your work?

Another article to help show how data scientists work and think:
What I do when I get a new data set as told through tweets

5.2 Languages of Data Science

Being a good data scientist requires a lot more than just being able to write code well, but not being able to code well is a sign of a poor data scientist. Currently, three programming languages drive the data science community. If you want to argue that you are a data scientist, you need to be proficient in at least one and able to use all three.[1]

- **R:** - A successor to the S language with it's first beta release in 2000. Heavily used by trained statisticians and researchers. Thanks to RStudio (established in 2010), data scientists also use this software for their work. (ref)
- **Python:** - Version 2.0 was released in 2000, with version 3.0 arriving in 2008. Pandas is the foundation for data science in Python, and it started development in 2008. Python is heavily used by software engineers as well. (ref1 and ref2)
- **SQL:** - Has been around much longer. In the early 1970s, IBM implemented the language. Oracle created the first commercially available implementation. It is built to handle relational databases but has also been leveraged for other big data database constructs. IT departments heavily use SQL. (ref)

5.2.1 R for data science

- Why you should learn R first for data science
- Why Learn R? 10 Handy Reasons to Learn R programming Language
- R for Data Science Introduction

BYUI students can take MATH 325 to be introduced to R for statistics and MATH 335 to learn R for data wrangling and visualization.

5.2.2 Python for data science

- Advantages of Learning Python for Data Science
- WHY SHOULD YOU LEARN PYTHON FOR DATA SCIENCE?
- A Beginner's Guide to Python for Data Science

BYU-I students can take CSE 110 to be introduced to Python and CSE 250 to be introduced to Python for data science.

5.2.3 SQL for data science

- Is SQL needed to be a data scientist?
- Why do you need to learn SQL?

BYU-I students can take CIT 111 or CIT 225 to be introduced to SQL.

5.3 Requesting and Communicating Data

It's important to remember that most of the people that you'll interact with in your career won't be data scientists, and may not have any experience working with data in the way that a data scientist does. You may be the only "data person" on a team, and will need to communicate with your teammates about your work and present it in a way that they will understand. Read **this article again** for an example of this.

Similarly, you will need to get access to the data that you will be working with. Usually that will come from people who either aren't familiar with your project, aren't data people, or both. Without good work doesn't happen without good data. There's an art to requesting data from and communicating with people unfamiliar with what you are doing. Part of getting good at doing so can only come through time and practice, but there are things you can do from the get-go. Listed below are links to some articles that offer some good advice:

- **How to get the right data? Trying asking for it.**
- **How to ask for datasets**

Here is a real example of some of the pains of requesting data that you should be prepared to handle in your career. Note the actions that the data scientist took to ensure that he had the data that he needed to solve the client's problem. - What questions did he have to ask? - What data was important to him? What didn't matter? Why? - What principles can you learn from this about requesting and communicating data?

Fundamental Statistical Concepts in Presenting Data

Apart from being a good demonstration of what it's like to acquire data in real-world data science work, this is also an excellent example of great data science work in solving a client's problem.

5.4 Marketing Yourself as a Data Scientist

While data science is a rapidly growing field with a lot of opportunities for employment, it's also very competetive. Simply having a degree isn't enough to land the best jobs, you will need to be able to show employers that you're capable of meeting their data needs. Three of the best tools for accomplishing this are your resume, personal Github repository, and LinkedIn profile.

5.4.1 Resumes

In many cases, your resume will be the first thing that a potential employer will see about you. It should showcase your skills, contributions to past projects, and value as a data scientist. Here are some resources to help guide you through gearing your resume geared towards data science jobs, as well as some tools for helping you make a resume in general.

Guides to data science resumes

- How to Write a Great Data Science Resume
- How to Build an Effective Data Science Resume?
- How to Write the Perfect Data Scientist Resume

Resume builder websites

- cvmaker.com
- resume.com

- kickresume.com
- BYU-I Resume support
- Zety Resume Templates

5.4.2 Github

Github has become a staple in the software world for collaborative work, and data scientists also make great use of it. It's a great way to show potential employers the work that you've done in the past so that they can get see a concrete example of some of your technical abilities. Use it as a place to store your work for projects, both professional (where appropriate - don't share anything sensitive in a public place) and personal. Do work that isn't required by work or school to show that data science work is something that you enjoy and post it to your personal Github repository.

If you are unfamiliar with Github, here is a good place to start:
What Is GitHub, and What Is It Used For?

This article explains the kind of material that you should share on your Github repo, and how you can use Github as a tool to present your work and talent:
What do job-seeking developers need in their GitHub?

5.4.3 Linkedin

You've probably heard of LinkedIn before, it's a popular social media platform designed to help people connect with potential employers and other people in a professional setting. Many recruiters will use it as a tool for finding potential employees to fill openings in companies, so it's worth your time to build up a good LinkedIn profile. It's a place where you can post your resume, experience, skills, a link to your Github repostiory, and establish a professional presence as a data scientist. You should also use it as a networking tool to connect with potential employers and interact with professionals already in the industry who can offer advice and further connections and opportunities.

Here is a guide to developing your LinkedIn footprint, as well as how to use it as a tool to further your professional pursuits:
The Complete Data Science LinkedIn Profile Guide

5.4.4 Resources for Finding and Landing a Job

Data science job postings

- **Indeed**
  - Indeed: Data Science
  - Indeed: Analyst
  - Indeed: Statistician
  - Indeed: Data Analysis
  - Indeed: Data Visualization

- **Glassdoor**
  - glassdoor: Data Scientist Intern
  - glassdoor: Data Visualization

- **Chegg Internships**
  - Data Science Internships

Interview tips

109 Data Science Interview Questions and Answers

1. Knowing a language doesn't make you a data scientist, just like knowing English doesn't make you a poet. You will also need to have analytics and visualization capabilities.↩

Chapter 6 Tools

This chapter serves as a reference page for the tools that we will use in this course.

6.1 Google Sheets

- Cheat sheet

Part 1

6.1.1 Create or import files

- 1.1 Create a new file
- 1.2 Import and convert existing files

6.1.2 Add content to your spreadsheet

- 2.1 Enter and edit your data
- 2.2 Customize your spreadsheet
- 2.3 Work with rows, columns, and cells
- 2.4 Work with multiple sheets

6.1.3 Share and collaborate on files

- 3.1 Share files in Drive, Docs, Sheets, or Slides
- 3.2 Unshare files in Drive, Docs, Sheets, or Slides
- 3.3 Add comments and replies in Drive, Docs, Sheets, or Slides
- 3.4 Suggest edits in Docs 3.5 Chat with people directly in Docs, Sheets, or Slides

6.1.4 Print and download files

- 4.1 Print your file
- 4.2 Download versions in other formats
- 4.3 Make a copy
- 4.4 Email a copy as an attachment

6.1.5 Access your calendar, notes, and tasks

- 5.1 Open your Google Calendar and events
- 5.2 Open notes in Google Keep
- 5.3 Open your to-do lists in Google Tasks
- 5.4 Get add-ons

Part 2 - More on Google Sheets

6.1.6 Access Sheets

- 1.1 Get Sheets on your devices
- 1.2 (Optional) Add multiple Google Accounts
- 1.3 Create a browser bookmark
- 1.4 Add a Sheets desktop shortcut (Windows only)
- 1.5 Work offline (Chrome only)

6.1.7 Sheets and Excel best practices

- 2.1 Work with Excel files in Drive
- 2.2 Use Excel and Sheets together
- 2.3 Edit Excel files in Sheets
- 2.4 Import Excel data into Sheets
- 2.5 Convert Excel files to Sheets
- 2.6 Share a copy of a Sheets file in Excel format

6.1.8 Manage data in Sheets

- 3.1 Perform basic operations
- 3.2 Search for data
- 3.3 See changes to data
- 3.4 Restrict data sharing
- 3.5 Use keyboard shortcuts

6.1.9 Analyze data in Sheets

- 4.1 Add charts
- 4.2 Get automatic charts
- 4.3 Add charts to Docs and Slides
- 4.4 Functions in Sheets and Excel
- 4.5 Add pivot tables
- 4.6 Get automatic pivot tables

6.1.10 Export spreadsheets

- 7.1 Print spreadsheets
- 7.2 Download in different formats
- 7.3 Make a copy
- 7.4 Email a copy

6.1.11 Get Sheets productivity tips

- 8.1 Import data from Forms
- 8.2 Save time with templates
- 8.3 Find out if someone changes a spreadsheet

6.2 Using Tableau for Visualizations

6.2.1 Background

Tableau is heavily used among business analysts and data scientists. As we use Tableau, we will not be covering all the powerful data connections and dashboard tools in the software in our class. We will be focusing on creating different data views for visual analytics in this class. The links below will be used during the semester.

6.2.2 Registration and Download

Tableau provides free access for college students to their professional tools. They also have a Tableau public option for all users. We will be using their training videos and online guides to help us get up to speed with the tool. You will need to create a `sign up` through the link at the top right of their website to see some of the videos we will use during the semester.

1. Download Tableau Desktop and Tableau Prep here
2. Select each product download link to get started. When prompted, enter your school email address for Business Email and enter the name of your school for Organization.
3. Activate with your product key found on Canvas.

Students can continue using Tableau after the class is over by individually requesting their one-year license through the Tableau for Students program.

6.2.3 Training materials

Getting started

- Tutorial: Get Started with Tableau Desktop
  - Step 1: Connect to your data and Connection to Google Sheets
  - Step 2: Drag and drop to take a first look
  - Step 3: Focus your results
- Use Show Me to Start a View
- Understanding pill types
- Granularity, Aggregation, and Details

A little more introduction depth

Once you are logged in, the first 12 minutes of this getting started video will be the most useful. Here are some other excellent guides.

- Buidling data Views from Scratch and the Getting started with Visual Analytics 6-minute video.
- Additional features of Drag and drop fields
- Measure Values and Measure Names

Saving and exporting your work

- Export Views and Workbooks
- Print Views
- Save your Work

Editing your graphic

Part 1

- Using Shelves and Cards
- Changing the mark or geometry of the graphic
- Editing the marks or geometry
- Reference Lines, Bands, Distributions, and Boxes

Part 2

- Sorting visualization information and a 6-minute video on sorting
- Filters
- Editing tooltips
- Add annotations
- Grouping categories

6.2.4 Primary charts

We will focus on the first few charts in this section.

- Build a Histogram and Create Bins from a Continuous Measure
- Build a Line Chart
- Build a Bar Chart
- Build a Box Plot

- Build a Scatter Plot
- Visualize Benford’s Law