Objectives

Displaying data and distributions with graphs

- Variables
- Types of variables (CIS p40-41)
- Distribution of a variable
- Bar graphs for categorical variables (CIS p42)
- Histograms for quantitative variables (CIS p43-46)
- Interpreting histograms (CIS p56-57)
- Scatterplots
- Time series plots

Further reading:
https://www.openintro.org/stat/textbook.php?stat_book=os (from now on abbreviated to OS3) Section 1.6.3
Topics: Populations and Samples

- **Learning objectives**
  - Understand what population, sample and random variables are.
  - Know if a random variable is binary, categorical or numerical (continuous or discrete)
  - Understand what a distribution of a variable is and that a histogram is a good way of depicting the distribution.
  - Know how to plot a relative frequency histogram in Statcrunch.
  - Know the main features of a histogram of a variable will look like.
Samples, Populations and Variables
Definitions

- A **population** is group of individuals of interest. They are not necessarily humans. The population is usually comprised of millions/billions or infinite number of individual. The population is generally never observed.
  - Example: Individuals can be individual people, companies, animals or any object of interest.

- A **sample** is a subset of the population. In statistics we usually assume it is an extremely tiny proportion of the population. The ratio of the sample size to the population size has absolutely no influence on inference. Typically this ratio will be close to zero.

- A **variable** is any characteristic of an individual/case. A variable **varies** (thus the name) among individuals.
  - Example: If an individual is a person, then variables of interest can be this persons height, blood pressure, ethnicity, mother tongue etc.

Further reading: [https://www.openintro.org/stat/textbook.php?stat_book=os Section 1.6 (numerical data)](https://www.openintro.org/stat/textbook.php?stat_book=os Section 1.6 (numerical data))
Two types of variables

- Variables can be either quantitative...

- Usually numerical values, where arithmetic operations, such as adding and averaging, make sense (adding bus numbers together makes no sense!).
  - **Example:** How tall you are, your age, your blood iron level, the number of credit cards you own.
  - If the numerical variable can only take integers (such as 1, 2, 3) it is called *numerical discrete*.
  - If it can take “any” number within a range it is called *numerical continuous* (such as weights, heights etc).
Two types of variables

- **Categorical.**
  
  Something that falls into one of several categories (if it is categorical there is no ordering in the categories). What can be counted is the count or proportion of cases in each category.

  **Example:** Your blood type (A, B, AB, O), your ethnicity, who you would vote for in an election.

  If there are only two categories, we call it a **binary variable**.

  One can always give a categorical variable a numerical labelling. For example, one can label your favorite color as Green = 1, Blue = 2 Orange = 3. However, the labelling can change Green = 2, Blue = 3 Orange = 1, without any information being lost. Moreover, the “average” favorite color in a class has no meaning. The number is used in this example as a label and has no interpretation as a number. Despite “appearing” numerical it is categorical.
Students at Texas A&M were randomly sampled and asked four questions. What type of variable do we associate to each answer:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Their Major</td>
<td>Their GPA</td>
<td>Distance of home from Campus</td>
<td>Number of Siblings</td>
</tr>
</tbody>
</table>

Answer:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Numerical continuous</td>
<td>Categorical</td>
<td>Numerical discrete</td>
<td>Numerical continuous</td>
</tr>
<tr>
<td>B</td>
<td>Numerical continuous</td>
<td>Categorical</td>
<td>Numerical discrete</td>
<td>Numerical discrete</td>
</tr>
<tr>
<td>C</td>
<td>Numerical discrete</td>
<td>Categorical</td>
<td>Categorical</td>
<td>Numerical continuous</td>
</tr>
<tr>
<td>D</td>
<td>Categorical</td>
<td>Numerical continuous</td>
<td>Numerical continuous</td>
<td>Numerical Discrete</td>
</tr>
<tr>
<td>E</td>
<td>Categorical</td>
<td>Numerical continuous</td>
<td>Numerical continuous</td>
<td>Numerical continuous</td>
</tr>
</tbody>
</table>
261 (most are verified customers) submitted a review for this fitness trackers. This is a sample from all customers who have bought such a fitness tracker.

The variable of interest is the score from 1-5 that an individual gives the tracker. The variable is **numerical discrete**.

We need to be cautious about analyzing reviews, as samples like this tend to be biased. The people who submit reviews tend to love or hate the product. Recently, companies have been trying to coax people to write reviews to reduce the bias. In this course, we will be analyzing this type of data, but we need to careful about the conclusions draw, since data sampled this way tends to be biased.
Statcrunch: Inputting data

- Transferring data into Statcrunch
- It is fun to analyze Amazon scores in Statcrunch.
  - Find a product of interest whose scores you want to input into Statcrunch.
  - I will use the tracker on the previous slide.
  - Transform the percentages into numbers:
    - 5 star: 261*0.72 = 188
    - 4 star: 261*0.08 = 21 etc.
  - These number have to be put into the Statcrunch spreadsheet. The easiest, but most boring method is simply to type the column of 188 5s, then below that 21 4’s and so forth (there are other methods too).
  - Once you have one long column, where the number of rows is equal to the number of reviewers. Save the data by Data -> Save.
Consider the example of EU referendum (in the UK):

- The variable of interest is how each voter would vote. The choice was leave or remain. This is a **binary variable**.
- The population of interest are all voters in the UK (about 44 million). The `parameter’ (quantity) within this population of interest is the proportion of voters who want to Remain within the EU or leave the EU.
- Before election day, we did not know this proportion. But opinion polls were conducted to estimate it.
- Here us a snap shot of the polls over a two week period before the vote:

<table>
<thead>
<tr>
<th>Date</th>
<th>Polling Organization</th>
<th>Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>22 June</td>
<td>55%</td>
<td>46%</td>
</tr>
<tr>
<td>20-22 June</td>
<td>49%</td>
<td>46%</td>
</tr>
<tr>
<td>20-22 June</td>
<td>44%</td>
<td>46%</td>
</tr>
<tr>
<td>17-22 June</td>
<td>48%</td>
<td>42%</td>
</tr>
<tr>
<td>16-22 June</td>
<td>41%</td>
<td>43%</td>
</tr>
<tr>
<td>20 June</td>
<td>45%</td>
<td>44%</td>
</tr>
<tr>
<td>18-19 June</td>
<td>42%</td>
<td>44%</td>
</tr>
<tr>
<td>16-19 June</td>
<td>53%</td>
<td>46%</td>
</tr>
<tr>
<td>17-18 June</td>
<td>45%</td>
<td>42%</td>
</tr>
<tr>
<td>16-17 June</td>
<td>44%</td>
<td>43%</td>
</tr>
<tr>
<td>14-17 June</td>
<td>44%</td>
<td>44%</td>
</tr>
<tr>
<td>16 June</td>
<td>42%</td>
<td>44%</td>
</tr>
<tr>
<td>10-16 June</td>
<td>53%</td>
<td>47%</td>
</tr>
<tr>
<td>10-15 June</td>
<td>45%</td>
<td>55%</td>
</tr>
<tr>
<td>11-14 June</td>
<td>43%</td>
<td>49%</td>
</tr>
<tr>
<td>12-13 June</td>
<td>39%</td>
<td>46%</td>
</tr>
<tr>
<td>10-13 June</td>
<td>44%</td>
<td>49%</td>
</tr>
</tbody>
</table>

All official campaigning suspended until 16 June after the fatal shooting of Jo Cox
On the day of the referendum the Leave camp got 51.89% and won the vote.  

However, in the days before the vote it different polls gave different results.  

- One of the reasons for the different outcomes is the closeness of the vote.  
- When the differences are so small, it is the sample proportions can swing both ways.  
- This also explains why an interval of possible values is more useful then just one number.
Population and sample: Example 3

- In a medical check-up, often, a blood sample is taken to check whether, for example, your iron level is at a healthy level.
- The variable of interest is the concentration of iron in a blood sample. This is **numerical continuous**.
- The concentration of iron will fluctuate over the samples. Several factors can determine the concentration of iron in a given sample:
  - Food eaten that day
  - Bowel movements
  - Time of menstrual cycle.
- The population in this case are all the blood samples that can be taken over a period of time for that individual.
- The parameter of interest is the average of these blood samples (which in reality is impossible to take).
- Given a few blood samples, the doctor will be able to **estimate** the mean iron level in the body.
The population of interest are calves (it is impossible to observe all of them). However, we do have a small subsample, 44 calves monitored biweekly from birth to 7 weeks.

This data set has a whole bunch of different interesting variables:

- Weights are **numerical continuous** (though since the are rounded to the nearest pound it “feels” numerical discrete).
- Treatment the calf is given is **categorical**.

Questions that are of interest:

- How do calf weights evolve over a period of time.
- Focusing on a single calf can be misleading, an individuals weight can be highly erratic.
- It makes sense to investigate how the mean weight of calves changes over time.
- However, we only have an 44 calves. Statistics allows us to make inference (confident statements) about the population of calves based on just the 44 we observe.
The above is a snapshot from the calf data.

- Tag # corresponds to an individual.
- Igg and TSP correspond to treatments that each calf was given. These are categorical variables.
- Wt 1. Corresponds to the weight at week 1. This a numerical variable (continuous or discrete?).
- The data for each individual/case is called an observation.

What the data looks like in Statcrunch:

<table>
<thead>
<tr>
<th>Row</th>
<th>Tag #</th>
<th>TRT</th>
<th>Igg</th>
<th>TSP</th>
<th>Wt 0</th>
<th>wt 0.5</th>
<th>Wt 1</th>
<th>wt 1.5</th>
<th>Wt 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>30</td>
<td>A</td>
<td>C</td>
<td>5.6</td>
<td>106.5</td>
<td>103</td>
<td>104</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>34</td>
<td>B</td>
<td>F</td>
<td>4</td>
<td>86.5</td>
<td>79</td>
<td>74</td>
<td>69.5</td>
<td>73</td>
</tr>
<tr>
<td>3</td>
<td>42</td>
<td>A</td>
<td>C</td>
<td>5.1</td>
<td>79</td>
<td>76</td>
<td>71</td>
<td>71</td>
<td>70</td>
</tr>
<tr>
<td>4</td>
<td>43</td>
<td>D</td>
<td>P</td>
<td>4.3</td>
<td>85.5</td>
<td>82</td>
<td>81</td>
<td>80</td>
<td>83.55</td>
</tr>
<tr>
<td>5</td>
<td>44</td>
<td>C</td>
<td>F</td>
<td>4.2</td>
<td>90</td>
<td>83</td>
<td>82</td>
<td>82</td>
<td>80</td>
</tr>
<tr>
<td>6</td>
<td>45</td>
<td>D</td>
<td>F</td>
<td>4.2</td>
<td>97</td>
<td>89</td>
<td>89</td>
<td>85</td>
<td>87.5</td>
</tr>
<tr>
<td>7</td>
<td>46</td>
<td>D</td>
<td>P</td>
<td>4.7</td>
<td>91</td>
<td>91</td>
<td>89</td>
<td>87.5</td>
<td>86</td>
</tr>
<tr>
<td>8</td>
<td>48</td>
<td>A</td>
<td>F</td>
<td>4.5</td>
<td>89.5</td>
<td>87</td>
<td>82</td>
<td>84.5</td>
<td>85</td>
</tr>
<tr>
<td>9</td>
<td>49</td>
<td>C</td>
<td>F</td>
<td>4.4</td>
<td>85</td>
<td>82</td>
<td>83</td>
<td>80</td>
<td>81.5</td>
</tr>
</tbody>
</table>
Population and sample: Example 5

Old Faithful is a famous geyser in Yellowstone national park. It has been erupting water over thousands of years.

- Over 272 consecutive eruptions were recorded. Both the duration of an eruption and the waiting time between eruption were recorded.
- These as numerical continuous variables.
- The population of interest are all the eruptions over a period of time.

<table>
<thead>
<tr>
<th>var1</th>
<th>eruptions</th>
<th>waiting</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.6</td>
<td>79</td>
</tr>
<tr>
<td>2</td>
<td>1.8</td>
<td>54</td>
</tr>
<tr>
<td>3</td>
<td>3.333</td>
<td>74</td>
</tr>
<tr>
<td>4</td>
<td>2.283</td>
<td>62</td>
</tr>
<tr>
<td>5</td>
<td>4.533</td>
<td>85</td>
</tr>
<tr>
<td>6</td>
<td>2.883</td>
<td>55</td>
</tr>
<tr>
<td>7</td>
<td>4.7</td>
<td>88</td>
</tr>
<tr>
<td>8</td>
<td>3.6</td>
<td>85</td>
</tr>
<tr>
<td>9</td>
<td>1.95</td>
<td>51</td>
</tr>
</tbody>
</table>
Distribution of a variable

- We previously defined a variable as a quantity in a population that we are interested in.
  - **Example**: Heights of people, weights of people, favorite colour.
- A variable by definition is `random’ in the sense that it can vary from individual to individual.
- However, it is not so random in the sense that it can take any values (the values are all over the place). In general certain outcomes are more likely than others.
  - Example 1: It is more likely that adult heights lie between 5 to 6 feet rather than 3 to 4 feet.
  - Example 2: The total number of M&Ms in a bag tend to be between 9-13.
- The frequency of a variable to take different outcomes is known as the distribution of the variable.
- The distribution tells us which outcomes are more or less likely.
The distribution of a variable is rarely known. However, we can get some idea of what it `looks' like by analysing the distribution of a sample.

Of course, a whole bunch of numbers is very hard to comprehend. The best course of action is to plot the distribution.

Various plotting tools can be used:
- Frequency Histograms (most popular)
- Pie chart (good for a quick look)

The distribution of each variable will have certain characteristics, which we should look out for when we make a plot.

The next few slides will be on making distribution plots based on the data.

Always make plots. A picture can convey far more information than words. I use this picture to convince students to reduce on plastic: https://www.stat.tamu.edu/~suhasini/plastic.html
Ways to plot categorical data

When the variable is categorical, the categories (recall a categorical variable is where ordering does not matter) in the graph can be ordered any way we want (alphabetical, by increasing value, by personal preference, etc.)

- **Bar graphs**
  Each category is represented by a bar with length equal to a count or percent (starting from the x-axis).

- **Pie charts**
  The slices represent the parts of one whole.

More reading:
http://onlinestatbook.com/2/graphing_distributions/graphing_qualitative.html
Plotting numerical data

- Histograms (for plotting the distribution of quantitative data)
  
  A **histogram** breaks the *range of values* of a variable into “bins” (intervals) and displays the count or percent of the observations that fall into each bin.

  The histogram and its close relative the density will be widely used in this class. Further reading:
  
Histograms

The range of values that a variable can take is divided into equal size intervals (bins).

The histogram shows the number of individual data points that fall in each interval.

This is a histogram for the weight of 44 newborn calves. The weights are in pounds. Each bin has length 5 pounds and the number of calves in each bin is counted.

The y-axis gives the number of occurrences in each bin. A more useful measure is to plot the proportion on the y-axis rather than the number. This gives rise to the relative frequency histogram.
Relative Frequency Histograms

- A relative frequency histogram is when there is a percentage of the total on the y-axis rather than just the count.
- For example, in the previous example the y-axis would be the percentage of calves rather than the number of calves.
- The shape of a relative frequency histogram is identical to the shape of a regular histogram, the only difference is that the y-axis gives you the percentage (or chance) compared to the total rather than the numbers.
- It is useful for reading off information:
  - Example: In the previous example, we see that $11 + 5 + 1 = 17$ new born calves have a weight less than 90 pounds. This corresponds to $100 \times \frac{17}{44} = 38.6\%$ of the sample.
  - Proportions often convey more information than raw numbers.
Amazon gives the relative frequency histogram rating for all its products.

If you rotate the plot you get the Amazon plot.
Statcrunch: loading data

- Download onto your computer the data set of interest (know the folder where the data lies, usually it will be in the Download folder).
- Log into Statcrunch. To upload the data into Statcrunch:
  - Data -> Load -> From file -> on my computer
  - You will see a new window
  - **Choose File** selects the data set
  - The **delimiter** determines how columns in the file are separated. Usually whitespace or comma is fine.
  - Click **Load File** (at bottom of page)
Statcrunch: Making a plot

To make a plot:
- Graph -> Histogram
- You will get a pop down menu like the one on the right.
- Select the **column(s)** you want plotted
- Select **Type** (I usually use relative frequency).
- Using Bins Width you can select the length of bin. There is usually a default chosen by the software.
- We discuss later how bin widths can influence the information the plot conveys.
- If you check Marker, then it will include the location of the average in the plot.
Most common distribution shapes

- A distribution is **symmetric** if the right and left sides of the histogram are approximately mirror images of each other.

- A distribution is **skewed to the right** if the right side of the histogram (side with larger values) extends much farther out than the left side. It is **skewed to the left** if the left side of the histogram extends much farther out than the right side.

- Not all histograms have a simple overall shape. It can be difficult to identify the shape when the sample size is small..
Examples of Shapes

- Distributions which tend to be nearly symmetric are:
  - Heights of a particular gender.
  - Lengths of bird bills.
  - Other biological lengths.

- Distributions which tend to be skewed:
  - RIGHT SKEWED: The price of houses. Many which are moderately priced, but the expensive ones can be extremely expensive (right skewed). In addition house prices are not usually below zero dollars.
  - LEFT SKEWED: The gestation period of baby. Common knowledge tells us that the due data is at 40 weeks. But the baby rarely arrives on the due data, and unless it is induced the gestation period is random. However, a baby born several weeks earlier can still survive but most babies cannot survive beyond 43 weeks in the womb.

- Distribution with multiple modes (many peaks):
  - Height of adult humans (male and females populations put together).
- **Distribution with multiple modes:**
  - The number of M&Ms in a bag (several modes because the of the different types).
  - Anything where several subpopulation, each with their own characteristics are put together.

- **Distributions which tend to be flat (no modes):**
  - Numbers in a lottery
  - The outcome of a die.
  - The chance of all outcomes for flat distributions are equally likely. One outcome cannot be favoured over another (for example in a die 4 occurring more often than a 6). Such distribution are very rare.
Question Time

Which characteristic below best describes the distribution (histogram) for the length of adult feet?

A. Bimodal.
B. Symmetric and Bell shaped.
1. Temperature in College Station in July (in Fahrenheit)
2. The time a student takes to complete a midterm (in minutes).
3. The result of an 8 sided die.
4. The number of cars a student owns in College Station

<table>
<thead>
<tr>
<th></th>
<th>Midterm</th>
<th>Cars</th>
<th>Temp</th>
<th>Die.</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Midterm</td>
<td>Die</td>
<td>Temp</td>
<td>Car</td>
</tr>
<tr>
<td>B</td>
<td>Temp</td>
<td>Die</td>
<td></td>
<td>Midterm</td>
</tr>
</tbody>
</table>
As you can see from the previous question, different variables will have different distributions.

You **cannot change** the distribution. It is what it is.

There is a common misconception that as the sample size increases (for example the number of people surveyed), the distribution will become increasingly symmetric and bell shaped. This is **not true**.

Just look at the examples on the previous slide, the distribution for the number of cars owned by a student cannot become more symmetric (this would the same proportion of students do and do not own a car, which is highly implausible).

Remember most distributions are not symmetric and unimodal.

For a summary of distribution features see OS3 Section 1.6.3
Influence of bin width on histograms

- We plot a histogram of the M&M data set using 4 different bin widths.
- Different bin widths give you a different perception of the data.
Role of bin width on the distribution

- The length of a bin (the block) is known as a bin width.
- When we make a histogram it is from a sample of the population. This means that we do not know what the true distribution will look like.
- Our aim is to choose a bin width which will give a good representation of the true unknown population histogram.
- As illustrated on the previous slide the shape of the histogram that we do plot depends on the bin width that we use.
  - If the bin width is too small, the histogram will appear too spikey with too many modes (peaks). This is probably not a feature of the distribution, but due to the fact that the sample is too small to have data in all the intervals.
  - If the bin width is too large, the histogram will appear too flat. This will smooth out features such as modes that are in the distribution.
- There is usually no clear cut solution on how to choose the bin width, some guide lines are given on the next slide.
Here are three different plots of the same Old Faithful Geyser data set.

Which plot communicates the most information about the data

(A) Top

(B) Middle

(C) Bottom

http://www.easypolls.net/poll.html?p=59adc9b0e4b0f0b62d08c6bc
A. Eruption times tend to be every 50-65 minutes or 70-90 minutes.
B. Eruption times can be anywhere between 40 to 100 minutes and all times are *equally* likely.
C. Eruptions time are unimodal with the average time about 80 minutes.
Outliers

An important kind of observation is an **outlier**. Outliers are values that lie *far* outside the overall pattern of a distribution. Always look for outliers and try to explain them. *Do not delete them!*

The overall pattern is fairly symmetrical, but 2 states have exceptional values. Alaska and Florida have unusual representations of elderly in their populations.

The most extreme values are outliers only if they stand far away from the main body of values.
Other plotting tools
Scatterplots (relationships)

In a scatterplot, one axis is used to represent each of the variables, and the data are plotted as points on the graph.

Old Faithful Geyser data

Each dot is the length of one eruption and the length of time until the next eruption.

We see that there appears to be a positive relationship between the duration between eruptions and the time of each eruption.

https://www.nps.gov/features/yellowstone/exhibits/eruption/prediction/predict8.htm

See OS3, Section 1.6.1
We can use this plot to predict the time until the next eruption.

How would you do that?
Average global temperatures

- On the right we plot the global average temperature against year.
- We observe a steady increase, though it is unclear whether this is linear.
Objectives

Describing distributions with numbers

- Measures of center: mean, median OS3 Section 1.6.2
- Mean versus median
- Measure of spread: standard deviation and the IQR OS3 Section 1.6.4
- The Boxplot
- Changing the unit of measurement
- The z-transform
Topics: Summary statistics

- Learning Targets:
  - Know what a mean and median is.
  - Know what a Quartile, IQR and standard deviation is.
  - Know where to roughly place the mean and standard deviation on a histogram.
  - Understand the effect outliers will have an all of the above.
  - Understand how linear transformations of data will effect mean, standard deviations, and quartiles.
  - Know the z-transform and how it can be used to make comparisons between different data sets.
  - Know that the mean and standard deviation of z-transformed data is zero and one.
Numerical summaries of Data

- Summarizing the data with a few numbers can simplify comparisons between samples.

- Two simple ways to describe the data is:
  - A number which describes its center.
  - A number which describes its spread.

- These numbers will not give a true description of the distribution of the data.
  - For example, using these numbers we cannot identify whether it is uni-modal/bi-modal etc.

- But it does give us a useful summary of the data.

- The notions we describe now will play a very important role in the rest of this course.
Measures of center
Measure of center 1: the mean

The sample mean or arithmetic average

To calculate the average, or sample mean, add all values, then divide by the number of cases. It is the “center of mass” of the histogram. We often denote it with the symbol \( \bar{x} \) (say “x bar”).

Sum of heights is 1598.3 and \( n = 25 \). Dividing by 25 gives 63.9 inches.

\[
\bar{x} = \frac{1598.3}{25} = 63.9
\]
Measure of center 2: the median

The median ($M$) is the midpoint of a distribution—the number such that half of the observations are smaller and half are larger.

1. Sort observations by size.
   $n = \text{number of observations}$

2.a. If $n$ is odd, the median is observation $(n+1)/2$ down the list

   $\leftarrow n = 25$
   $(n+1)/2 = 26/2 = 13$
   Median = 3.4

2.b. If $n$ is even, the median is the mean of the two middle observations.

   $n = 24 \Rightarrow$
   $n/2 = 12$
   Median = $(3.3+3.4)/2 = 3.35$
Comparing the sample mean and the median

If the distribution is symmetrical then the mean and median are the same. The median is a measure of center that is resistant to skewness and outliers. The mean is not. If there is a skew in the data the mean tends to be pulled towards the side of the skew (long tail).

Mean is center of balance
Median splits the area in half

Mean and median for a symmetric distribution

Mean and median for skewed distributions

Left skewed

Right skewed
Impact of skewed data

Symmetric distribution…

\[ n = 25 \]

Disease X: \[ \bar{x} = 3.39 \]
\[ M = 3.35 \]

Mean and median are about the same.

… and a right-skewed distribution

\[ n = 25 \]

Multiple myeloma: \[ \bar{x} = 3.34 \]
\[ M = 2.32 \]

The mean is pulled in the direction of the skewness.
When summarizing data often we give the mean and median. We saw in the examples above that outliers can have too much of an influence on the average, whereas it will not have a large influence on the median.

Example: What is the average of the sequence: 0,1,1,2,2,100?
Answer: 17.66.

Example: What is the median of the sequence: 0,1,1,2,2,100?
Answer: (1+2)/2=1.5.

Which number do you think describes best the ‘center’ of this data.

Here we see that the average is ‘pulled’ towards 100, even though most of the data takes quite small values.

For this example the mean does not give a good idea of the center.
Example: Mean and Median

- Comparing means and medians. Statcrunch: Applets -> Mean/SD vs. Median/IQR (you can load your own data by using Data Table) and press compute (move the balls around and see what happens to the Mean and Median).

Here we plot the reviews of the Smart tracker. We see that a few bad reviews can push down the average but has no influence on the median.

It is useful to state both both the average and median in a report.
Demonstration 2

- You can do the same thing in Statcrunch: Applets -> Sampling Distributions (press compute) -> customize the plot by left clicking over the plot and tracing the plot you want. Observe how the mean and median change.

- Place the cursor over the distribution to change it. Watch how the mean and median change too.
One of the 5 star ratings turns out to be misclassified, it is reclassified as a 1 star rating? What happens to the mean and the median?

- (A) The mean decreases and the median stays the same.
- (B) The mean increases and the median stays the same.
- (C) The mean stays the same and the median decreases.
What should you use, when, and why?

Mean or median?

Currently taxation is an important issue in the news. In particular the impact that taxation will have on the `typical family’. How we define `typical’ depends on its application:

- **Mean**: Although income is likely to be right-skewed, the city government wants to know about the total tax base (the total income of the population), this is proportional to average income. Thus the mean is useful here (since mean income = total income divided by number of people).

- **Median**: On the other hand the sociologist is interested in a “typical” family of “middle” income. She uses the median to measure this, to lessen the impact on the value of her summary statistic due to extreme incomes.

- In a good data summary both the mean and median are given, and it left to the reader to decide which is the most representative.
Statcrunch: Obtaining summary statistics

- Load the data into Statcrunch.

- To obtain the summary statistics:
  - Go to Stat -> Summary Stat -> Columns,
  - A drop down menu appears. In select column(s) choose which variables you want to summarize (you can choose several by highlighting and pressing tab at the same time).
  - Click Compute!

- You will obtain several different summary statistics, which look like this:
Measures of spread
Measure of spread 1: the quartiles

The **first quartile**, $Q_1$, is the value in the sample that has 25% of the data at or below it (≡ it is the median of the lower half of the sorted data, excluding $M$).

The **third quartile**, $Q_3$, is the value in the sample that has 75% of the data at or below it (≡ it is the median of the upper half of the sorted data, excluding $M$).

The **IQR (interquartile range)** is the difference between the third and first quartile. It tells is where 50% of lie.

<table>
<thead>
<tr>
<th>Value</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>11</td>
<td>4</td>
</tr>
<tr>
<td>12</td>
<td>5</td>
</tr>
<tr>
<td>13</td>
<td>3.4</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>2</td>
</tr>
<tr>
<td>16</td>
<td>3</td>
</tr>
<tr>
<td>17</td>
<td>4</td>
</tr>
<tr>
<td>18</td>
<td>5</td>
</tr>
<tr>
<td>19</td>
<td>6</td>
</tr>
<tr>
<td>20</td>
<td>7</td>
</tr>
<tr>
<td>21</td>
<td>1</td>
</tr>
<tr>
<td>22</td>
<td>2</td>
</tr>
<tr>
<td>23</td>
<td>3</td>
</tr>
<tr>
<td>24</td>
<td>4</td>
</tr>
<tr>
<td>25</td>
<td>5</td>
</tr>
</tbody>
</table>

$Q_1=$ first quartile = 2.2

$M=$ median = 3.4

$Q_3=$ third quartile = 4.35
Largest = max = 6.1

$Q_3 =$ third quartile = 4.35

$M =$ median = 3.4

$Q_1 =$ first quartile = 2.2

Smallest = min = 0.6

Interquartile range $Q_3 - Q_1 = 4.35 - 2.2 = 2.15$
Boxplot and histograms

Boxplots often give information about the underlying histogram of the data.

The histogram of a symmetric distribution.

The histogram and boxplot of a right skewed distribution.
For the Old Faithful Data the Boxplot is not very informative. It does not inform about multiple modes.
The weight of 45 calves are tracked from birth to 7 weeks (weights are taken every half a week). The data is plotted below. What does this plot suggest?

(A) The plot suggests that the distribution of weights at each week is symmetric.
(B) There appears to be small drop in the weights early on (week 0-1.5), then the weights starts to increase.
(C) There seems to an increase in the spread of weights over time.
(D) [A] and [B]  
(E) [B] and [C].
(5) The median of a data set is 6, the first quartile is 4 and the third quartile is 6. What can we say about this data set (making a boxplot can help)?

(A) 25% of the data set is the same. (B) 50% of the data set is the same.
(C) All of the data set is the same. (D) The standard deviation is zero.
(E) C and D.
Identifying suspected outliers using quartiles and the Boxplot

- The boxplot is often used to identify outliers.

- IQR: The interquartile range is the difference between the first and third Quartile (Q3 – Q1). The IQR is a measure of spread 50% of the sample. The more spread out the majority of the sample, the larger the IQR.

- An observation is typically identified as a suspected outlier if it lies $1.5 \times \text{IQR}$ above the third quartile or $1.5 \times \text{IQR}$ below the first quartile.
Statisticians are obsessed with averages. The most common measure of spread is the standard deviation which is the average spread about the mean:

1. First calculate the **deviations**.
   
   deviation = value \( - \bar{x} \)

2. Then calculate the **variance** \( s^2 \).
   
   \( s^2 = \frac{\text{sum of squared deviations}}{n - 1} \)

3. Finally, take the square root to get the **standard deviation** \( s \).
   
   \( s = \sqrt{s^2} \)
Measure of spread 2: A calculation

The standard deviation “s” is used to describe the variation around the mean. Like the mean, it is not resistant to skew or outliers.

Numerical Example:

- We observe: 1,3,4,5,8
- The sample mean is 4.2
- Deviation of data from mean is -3.2, -1.2, -0.2, 0.8, 3.8
- Squared deviation is 10.24, 1.44, 0.04, 0.64, 14.44
- Average of squares is: 6.7.

1. First calculate the deviations.
   \[ \text{deviation} = \text{value} - \bar{x} \]

2. Then calculate the variance \( s^2 \).
   \[ s^2 = \frac{\text{sum of squared deviations}}{n - 1} \]

3. Finally, take the square root to get the standard deviation \( s \).
   \[ s = \sqrt{s^2} \]
Measure of spread 3: Invariance to shift

The standard deviation “s” is invariant to shifts. We add 10 the previous sample, the mean shifts by 10 but the standard deviation is the same.

Numerical Example:

- We observe: 10, 13, 14, 15, 18
- The sample mean is 10+4.2=14.2
- Deviation of data from sample mean is -3.2 -1.2 -0.2 0.8 3.8
- Squared deviation is 10.24 1.44 0.04 0.64 14.44
- Average of squares is: 6.7.
- 6.7 is too large it almost covers the entire data set. This is because we squared the deviation. We standardize by taking the root $s = 2.58 = \sqrt{6.7}$

1. First calculate the deviations.
   \[
   \text{deviation} = \text{value} - \bar{x}
   \]

2. Then calculate the variance $s^2$.
   \[
   s^2 = \frac{\text{sum of squared deviations}}{n-1}
   \]

3. Finally, take the square root to get the standard deviation $s$.
   \[
   s = \sqrt{s^2}
   \]
You observe the sample 1,1,16,8,9,10,14,19.

The average (sample mean) is **9.75** and the standard deviation is **6.54**

What happens to the average and standard deviation if we subtract 5 from each observation?

A. The mean is 4.75 and the standard deviation is 1.54
B. The mean is 9.75 and the standard deviation is 6.54
C. The mean is 4.75 and the standard deviation is 6.54
Variance and Standard Deviation

- Why do we square the deviations?
  - The sum of the non-squared deviations from the mean is always zero.
  - The sum of the squared deviations of any set of observations from their mean is the smallest that the sum of squared deviations from any number can possibly be.

- Why do we emphasize the standard deviation rather than the variance?
  - $s$ has the same unit of measurement as the original observations, and so is a natural measure of spread.
  - $s$ is the best measure of spread for Normal distributions.

- Why do we average by dividing by $n - 1$ rather than $n$ in calculating the variance (a statistical quirk)?
  - The sum of the deviations is always zero, so only $n - 1$ of the squared deviations can vary freely.

This is a summary of one of the samples take in class.

The average of all 20 heights is 66 inches. The average of the averages is also 66 inches.

The standard deviation of all heights is about 4 inches.

The standard deviation of the averages is about 1.7 inches.

The averages “tend” to be closer to the mean than individual heights.
Example 1: IQR vs Std deviation

- In Statcrunch go to Applets -> Mean/SD vs. Median/IQR, tick the boxes standard deviation and IQR and press computer. Move the balls around and compare the IQR with the standard deviation.

Observe how an outlier can change the standard deviation (just like the mean), but the IQR is not effected.
Example 2: IQR vs Std. deviation

- Here we illustrate further differences between the two measures of spread.

- If all the values in the data set take the same value, then IQR = 0 and standard deviation = 0. The standard deviation is only zero if all the values are the same.

- If most of the values in the data set are the same (but not all), in this case the standard deviation cannot be zero but the IQR = 0.
Properties of the Standard Deviation

- Usually, $s > 0$.
  
  $s = 0$ only when all observations have the same value and there is no spread. Eg. The data is 1, 1, 1, 1, 1. The sample mean is 1. The difference about the mean is 0, 0, 0, 0, 0. Thus $s = 0$.

- $s$ has the same units of measurement as the original observations.

- $s$ measures spread about the mean.

- **Rule of Thumb** `Most` of the data is within two standard deviations of the mean:

  $$\bar{x} - 2s \text{ and } \bar{x} + 2s.$$  

  The more standard deviations it is from the mean, the more `extreme` it is. We calculate the number of standard deviations using the z-transform (we do this in the next few slides).

- $s$ is not resistant to outliers or skewness. That is, a few extreme values can change $s$ considerably.
Disease X:

Symmetric distribution...

\[ n = 25 \]
\[ \bar{x} = 3.39 \]
\[ s = 1.48 \]

Data mostly within 1 st. dev. of the mean and nearly all within 2 st. dev.

... and a right-skewed distribution

Multiple myeloma:

\[ n = 25 \]
\[ \bar{x} = 3.34 \]
\[ s = 3.17 \]

Data to the left of the mean are more bunched together than data to the right of the mean.
Question Time

The weights of 44 calves are followed from week 0 to week 8. The plot below gives the **frequency histogram** for the weights of calves at week 7 (note a bin width of pounds was used). Calculate the proportion of calves (in **this sample of 44 calves** whose weight lies within **two** standard deviations of the mean.

Mean = 131 and standard deviation = 18.

(A) 95%  (B) 100%  (C) 44%  (D) 68%  (E) 92%

http://www.easypolls.net/poll.html?p=59add323e4b0f0b62d08c730
Changing the unit of measurement

Variables can be recorded in different units of measurement. Most often, one measurement unit is a **linear transformation** of another measurement unit: \( x_{\text{new}} = a + bx_{\text{old}} \).

Temperatures can be expressed in degrees Fahrenheit or degrees Celsius.\[\text{Temp}^{\text{Fahrenheit}} = 32 + 1.8* \text{Temp}^{\text{Celsius}} \Rightarrow a + bx.\]

Linear transformations do not change the basic shape of a distribution (skew, symmetry, multimodal). But they do change the measures of center and spread:

- Multiplying each observation by a positive number \( b \) multiplies both measures of center (mean, median) and spread \((s, IQR)\) by \( b \).
- Adding the same number \( a \) (positive or negative) to each observation adds \( a \) to measures of center (mean, median) and to quartiles but it does not change measures of spread \((s, IQR)\).
The new mean and standard deviation in a linear transformation

- Consider the transformation Fahrenheit = 32 + 1.8 \times Celsius.
  - The new mean after transformation is = 32 + 1.8 \times \text{old mean}.
  - The standard deviation after transformation is 
    = 1.8 \times \text{old standard deviation}.

- In general if the transformation is Y = a + bX
  - The new mean after transformation is = a + b \times \text{old mean}.
  - The standard deviation after transformation is 
    = |b| \times \text{old standard deviation} (remember to use the absolute, positive, of b, the standard deviation can never be negative).

To see why this is true, look at the real Fahrenheit to Celsius example given in:
http://onlinestatbook.com/2/summarizing_distributions/transformations.html
Examples

- **Examples 1**: 5 students are 19, 19, 20, 21, 22 years old. Their mean age is 20.2 years and their standard deviation is 1.3.

  - **Question**: In 10 years time what will be their mean and standard deviation?
  - **Answer**: $a = 10$ and $b = 1$. Their mean age will be $20.2 + 10 = 30.2$ but the amount of variation remains the same, $1.3 \times 1 = 1.3$ (the data has just shifted to the right).

- **Examples 2**: 5 children are 0.5, 1.5, 2, 3.2, 3.8 years old. The mean age is 2.2 years and their standard deviation is 1.32.

  - **Question**: Suppose we convert years into months, what is their age and standard deviation?
  - **Answer**: $a = 0$ and $b = 12$. There are 12 months in a year. So the mean age in months is $2.2 \times 12 = 26.4$ and the standard deviation is $1.32 \times 12 = 15.84$.

  Because the units have changed (from years to months), it `looks' like the mean and standard have increased. Of course, this is not the case, it is simply that a different unit of measurement has been used.
Suppose we observe the data 1, 3, 4, 5, 6. It has the sample mean 3.8 and sample standard deviation 1.92.

Question: The data is transformed with the transformation \( y = -2x \) i.e. -12, -10, -8, -6, -2

What is the new sample mean and standard deviation?

A. The new mean is \(|2| \times 3.8 = 7.6 \) and the new standard deviation is \(|2| \times 1.92 = 3.84\)

B. The new mean is \(-2 \times 3.8 = -7.6\) and the new standard deviation is \(|2| \times 1.92 = 3.84\)

C. The new mean is \(|2| \times 3.8 = 7.6\) and the new standard deviation is \(-2 \times 1.92 = -3.84\)
What unit to choose?

- As illustrated on the previous slide there are so many variables which can be measured with different units of measurements. This makes comparisons extremely difficult:
  - Temperatures: Do we use Fahrenheit or Celsius?
  - Length: Do we use Miles, meters or kilometers?
  - Weight: Do we use Pounds, or kilograms?

- It is clear that there is a large amount of ambiguity on which unit of measurement one should use for these and many other situations.
  - In general we choose the unit of measurement which gives the least number of zeros. Why: it is easier to say $0.5$ million than $500,000$ (in reference to a baseball salary)
  - The **units of measurement should always be given** whenever data is collected.


- In the Mars Mission, the British thought the Americans would use American standard Units. The Americans used International standard units. This discrepancy meant errors in the calculation and the orbiter was lost.
We should just choose one unit of measurement to make comparisons.

In statistics, there is a simple solution for making comparisons, which avoids selecting a particular unit of measurement.

This is by making a **z-transform**, which is described next. This transformation is **free of units of measurement**.
The z-transform/z-score

This is a “relative” distance, that takes into account the variation of the distribution.

OS3: Section 3.1 (page 130)
Comparing test scores

- Below is the distribution of test scores for two exams

- The mean score in both exams is roughly the same (about 33.5).
- A student scores 28 in both exams. Which exam did the student do worse?
  - For both exams the student scores about 5 points less than the mean. However, it is clear that the student did worse in one. Why?
Relative distance/z-score/z-transform

- In both exams the student scores 5.5 points **less** than the mean.
- -5.5 (we use a negative since it is less than the average) does not adequately describe the performance of the student.
- A better measure would be one which also took into account the spread of grades.
- The larger the “spread” the closer 28 point is to the mean 33.5.
- A distance measure which takes into account spread is often called a relative distance, or in statistics the z-score or z-transform (I usually call this a z-transform).

\[
\text{z-transform} = \frac{\text{data} - \text{mean}}{\text{st.dev}}
\]

- **Exam 1**  
  \[
  \text{Score 1} = \frac{28 - 34}{10.34} = -0.58
  \]

- **Exam 2**  
  \[
  \text{Score 1} = \frac{28 - 33.2}{5.88} = -0.88
  \]
\[
\text{z-transform} = \frac{\text{data} - \text{mean}}{\text{st.dev}}
\]

- Despite the student getting the same score in both exams and the mean score in both exams being the same, the student in performed relatively worse in Exam 2 as compared with Exam 1.

- By comparing z-transforms, z-scores we can determine which observations are more “unusual”. Determining “unusual” is very important in statistics.

- A student scores 45 in both exams. In which exam did the student do better relative to the class?
Example: Often we want to compare temperatures in different regions of the World. For example, the temperatures between Manchester (UK) and College Station (USA). It does not make much sense comparing their raw temperature. College Station is a lot warmer during the summer months. On Sept 15th in Manchester it was 14C whereas in CS it was 87F (30C).

But we can make a comparison against what is `normal’ in that part of the world. For example, during September on average it is 18 degrees Celsius in Manchester with standard deviation 3C, in College station on average it is 90 (32C) Fahrenheit with standard deviation 10F.

Which place experienced the more “unusual” weather on Sept 15th?

Manchester is 4C less than usual. CS is 3F less (about 2C less than usual).

But these numbers do not take into account the regional variability of the temperatures. If there is barely any variation, a drop of just 2F is unusual!
The z-transform of temperatures

\[ z\text{-transform} = \frac{\text{data} - \text{mean}}{\text{st.dev}} \]

- The z-transform of the Manchester temperature
  \[ \text{Manchester} = \frac{14 - 18}{3} = -1.33 \]

- The z-transform of the College Station temperature
  \[ \text{CS} = \frac{87 - 90}{10} = -0.33 \]

- The z-transforms show that the weather in Manchester on Sept 15th was more unusually cold than the weather in College Station.

- The z-transform is said to \textit{standardize data} as it is free of units of measurements. It is centered about zero and has standard deviation one.
The z-transform and the standard deviation

\[ z\text{-transform} = \frac{\text{data} - \text{mean}}{\text{st.dev}} \]

- The z-transform measures the number of standard deviations the data is from the mean.
- If the z-transform is **negative** it is to the **left** of the mean.
- If the z-transform is **positive** it is to the **right** of the mean.
- The larger the z-transform the further it is from the mean; the more extreme the data is relative to the mean.
The mean heights of students is 68 inches and the standard deviation is 4 inches.

A student is 62 inches, how many standard deviation are they from the mean?

A. The z-transforms is -1.5, they are 1.5 standard deviations to the LEFT of the mean.
B. The z-transforms is -1.5, they are 1.5 standard deviations to the RIGHT of the mean.
C. The z-transform is -6, they are 6 standard deviations to the LEFT of the mean.

http://www.easypolls.net/poll.html?p=59b81f77e4b08651d47502d1
Male heights have a mean 68 inches and standard deviation 4 inches.

Female heights have a mean 65 inches and a standard deviation 2 inches.

**Question** Suppose Jane (female) is 68 inches tall and Peter (male) is 72 inches tall. Who is more more “extreme/unusual” from the mean relative to their gender?

A. Jane
B. Peter

[http://www.easypolls.net/poll.html?p=59b82002e4b08651d47502d3](http://www.easypolls.net/poll.html?p=59b82002e4b08651d47502d3)
Accompanying problems associated with this Chapter

- Quiz 2
- Quiz 3