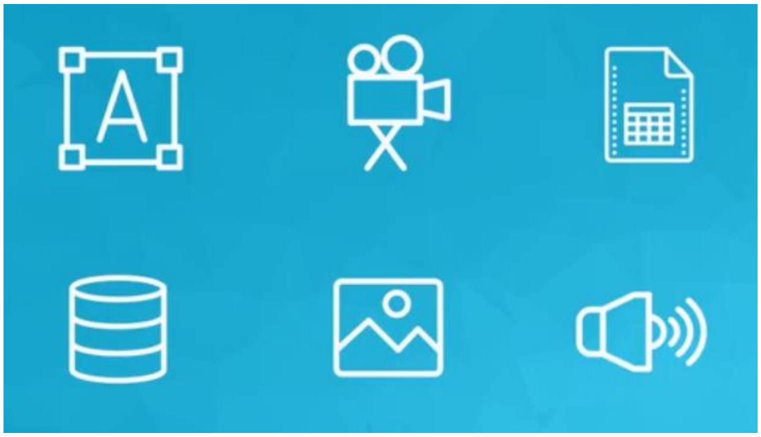
Descriptive Statistics

What is Data?

The word "data" is defined as distinct pieces of information.

You may think of data as simply numbers on a spreadsheet, but data can come in many forms. From text to video to spreadsheets and databases to images to audio, and I'm sure I'm forgetting many other forms.



Utilizing data is the new way of the world. Data is used to understand and improve nearly every facet of our lives. From early disease detection to social networks that allow us to connect and communicate with people around the world.

No matter what field you're in, from insurance and banking, to medicine, to education, to agriculture,, to automotive, to manufacturing, and so on.

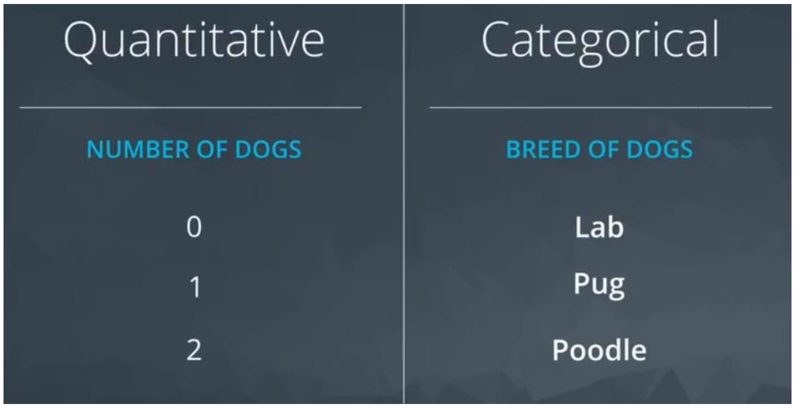
You can utilize data to make better decisions and accomplish your goals. We will be getting you started on the right foot to using your data in this course.

Data Types

We'll be taking a look at the different data types that exists in the world around us.

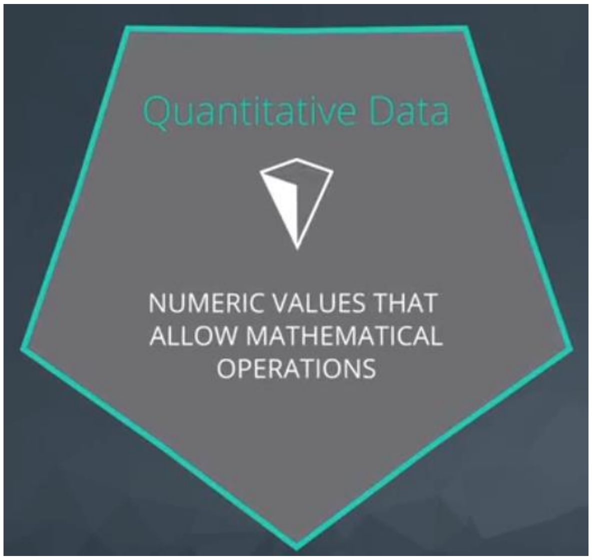
When sitting at coffee shops, I enjoy watching the dogs pass. I often wonder, how many crossed my path? I wonder if more pass on weekdays or weekends. Maybe the number differs from Mondays to Tuesdays. I also pay attention to the breeds of the dogs. I wonder if more collies stopped by on Monday than on Wednesday.

I wonder what's the most common breed. Is that breed the most common at all coffee shops? If I walked across the street to my favorite breed, would the most common breed change? This introduces two main data types: quantitative data, like the number of dogs, and categorical data, like the breed.



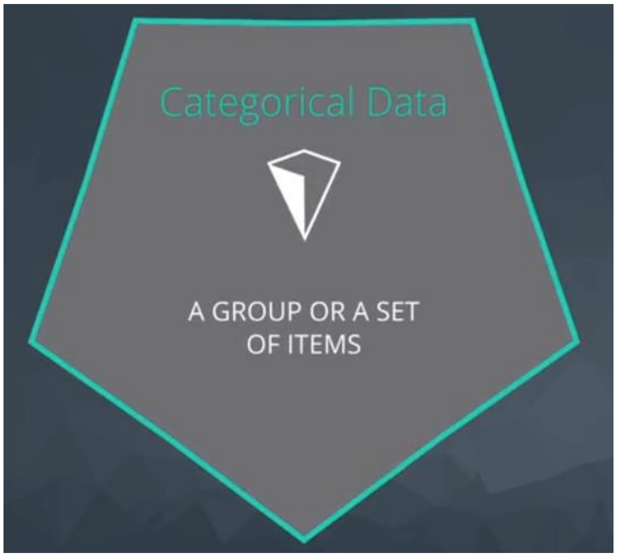
Quantitative	Categorical
NUMBER OF DOGS	BREED OF DOGS
0	Lab
1	Pug
2	Poodle

Quantitative data takes on numeric values that allow us to perform mathematical operations. In the previous example, we saw this with the number of dogs.



Quantitative Data

NUMERIC VALUES THAT ALLOW MATHEMATICAL OPERATIONS

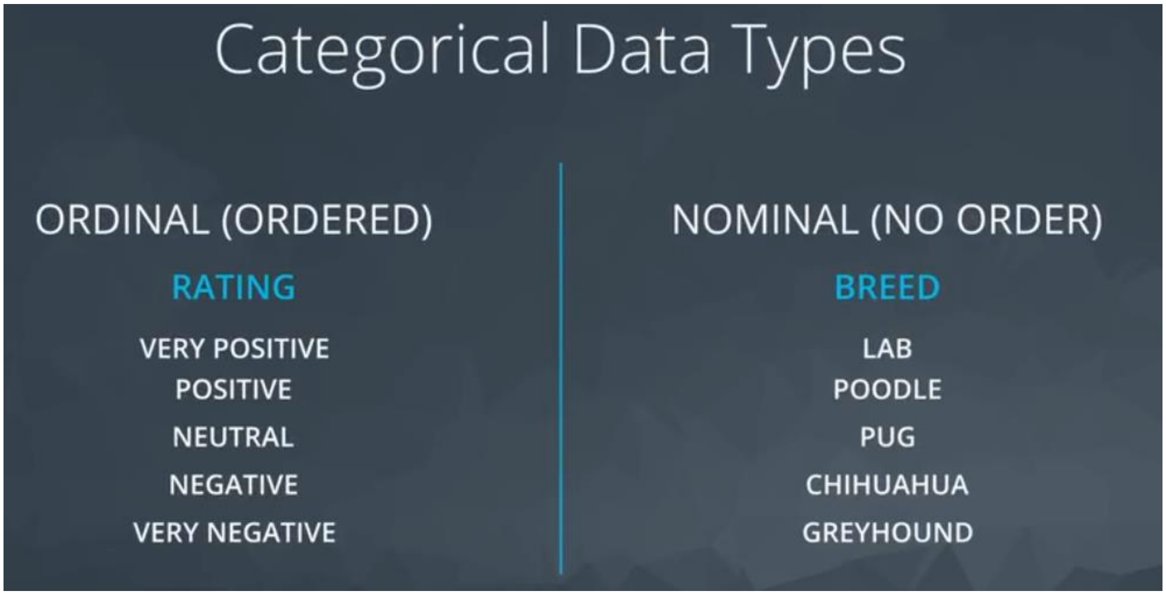


Categorical Data

A GROUP OR A SET OF ITEMS

If I see five dogs on Monday, and six dogs on Tuesday, I've seen a total of 11 dogs so far this week. Alternatively, categorical data frequently are used to label a group or set of items. We saw this with the breeds of the dogs.

Categorical Data Types



Categorical Data Types

ORDINAL (ORDERED)	NOMINAL (NO ORDER)
RATING	BREED
VERY POSITIVE	LAB
POSITIVE	POODLE
NEUTRAL	PUG
NEGATIVE	CHIHUAHUA
VERY NEGATIVE	GREYHOUND

We can divide categorical data further into two types: **Ordinal** and **Nominal**.

Categorical Ordinal data take on a ranked ordering (like a ranked interaction on a scale from Very Poor to Very Good with the dogs).

Categorical Nominal data do not have an order or ranking (like the breeds of the dog).

Quantitative Data Types

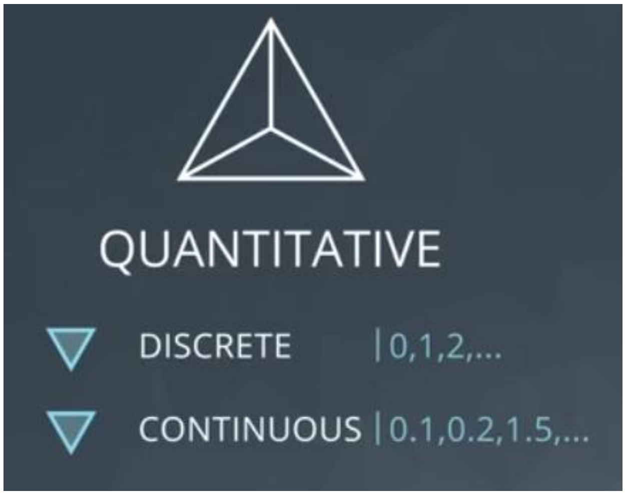
Continuous vs. Discrete

We can think of quantitative data as being either **continuous** or **discrete**.

Continuous data can be split into smaller and smaller units, and still a smaller unit exists. An example of this is the age of the dog - we can measure the units of the age in years, months, days, hours, seconds, but there are still smaller units that could be associated with the age.

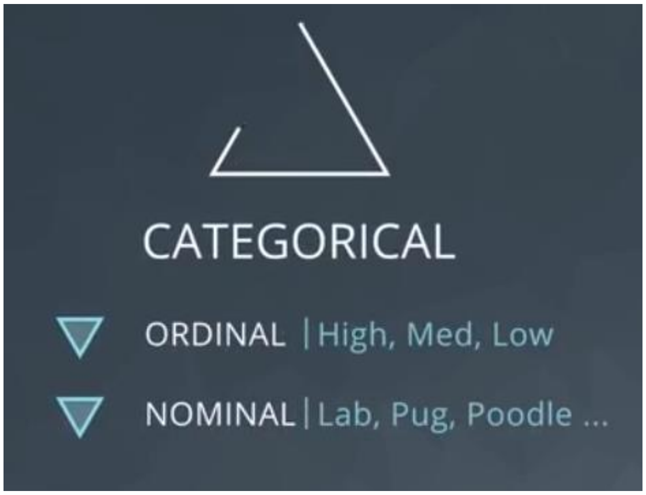
Discrete data only takes on countable values. The number of dogs we interact with is an example of a discrete data type.

Data Types Summary



QUANTITATIVE

- DISCRETE | 0,1,2,...
- CONTINUOUS | 0.1,0.2,1.5,...



CATEGORICAL

- ORDINAL | High, Med, Low
- NOMINAL | Lab, Pug, Poodle ...

The table below summarizes our data types. To expand on the information in the table, you can look through the text that follows.

Data Types		
Quantitative:	**Continuous**	**Discrete**
	Height, Age, Income	Pages in a Book, Trees in Yard, Dogs at a Coffee Shop
Categorical:	**Ordinal**	**Nominal**
	Letter Grade, Survey Rating	Gender, Marital Status, Breakfast Items

Below is a little more detail of the information shared in the above table.

Another Look

To break down our data types, there are two main blocks:

Quantitative and **Categorical**

Quantitative can be further divided into Continuous or Discrete.

Categorical data can be divided into Ordinal or Nominal.

You should have now mastered what types of data in the world around us falls into each of these four buckets: Discrete, Continuous, Nominal, and Ordinal. In the next sections, we will work through the numeric summaries that relate specifically to quantitative variables.

Quantitative vs. Categorical

Some of these can be a bit tricky - notice even though zip codes are a number, they aren’t really a quantitative variable. If we add two zip codes together, we do not obtain any useful information from this new value. Therefore, this is a categorical variable.

Height, **Age**, the **Number of Pages in a Book** and **Annual Income** all take on values that we can add, subtract and perform other operations with to gain useful insight. Hence, these are quantitative.

Gender, **Letter Grade**, **Breakfast Type**, **Marital Status**, and **Zip Code** can be thought of as labels for a group of items or individuals. Hence, these are categorical.

Continuous vs. Discrete

To consider if we have continuous or discrete data, we should see if we can split our data into smaller and smaller units. Consider time - we could measure an event in years, months, days, hours, minutes, or seconds, and even at seconds we know there are smaller units we could measure time in. Therefore, we know this data type is continuous. **Height**, **age**, and **income** are all examples of continuous data. Alternatively, the **number of pages in a book**, **dogs I count outside a coffee shop**, or **trees in a yard** are discrete data. We would not want to split our dogs in half.

Ordinal vs. Nominal

In looking at categorical variables, we found **Gender**, **Marital Status**, **Zip Code** and your **Breakfast items** are nominal variables where there is no order ranking associated with this type of data. Whether you ate cereal, toast, eggs, or only coffee for breakfast; there is no rank ordering associated with your breakfast.

Alternatively, the **Letter Grade** or **Survey Ratings** have a rank ordering associated with it, as ordinal data. If you receive an A, this is higher than an A-. An A- is ranked higher than a B+, and so on... Ordinal variables frequently occur on rating scales from very poor to very good. In many cases we turn these ordinal variables into numbers, as we can more easily analyze them, but more on this later!

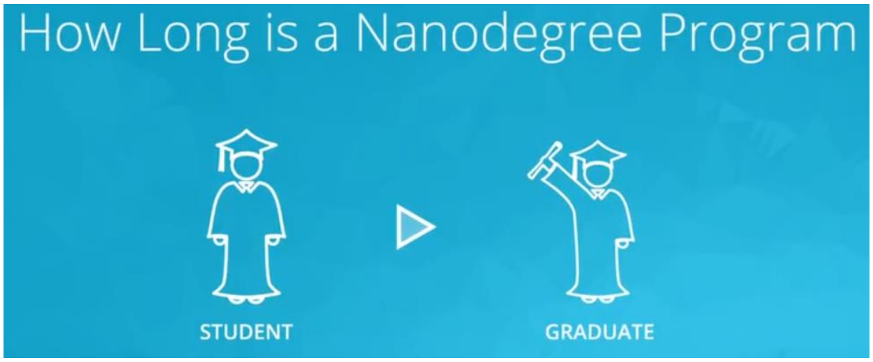
Final Words

In this section, we looked at the different data types we might work with in the world around us. When we work with data in the real world, it might not be very clean - sometimes there are typos or missing values. When this is the case, simply having some expertise regarding the data and knowing the data type can assist in our ability to ‘clean’ this data. Understanding data types can also assist in our ability to build visuals to best explain the data. But more on this very soon!

Introduction to Statistics

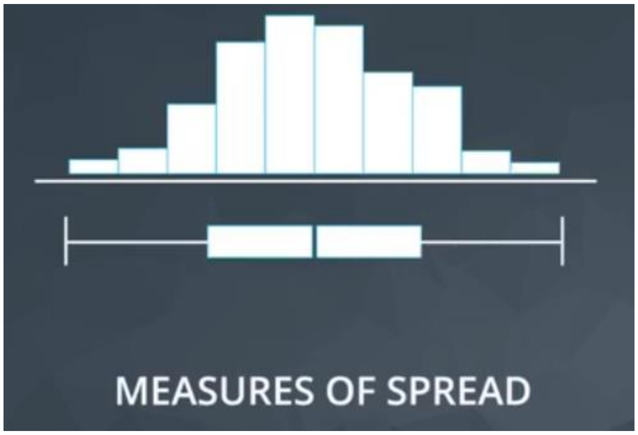
In the next lessons, we will discuss how to use statistics to describe quantitative data. You will gain insight into a process of how data is collected and how to answer questions using your data. Throughout this lesson, I hope you learn to be critical of your analysis that happened under the hood and what the numbers actually mean.

As an example of an analysis that we do here at Udacity, we look at how long a nanodegree program takes students to complete. We try to provide an estimate of the number of months or hours that students will spend.

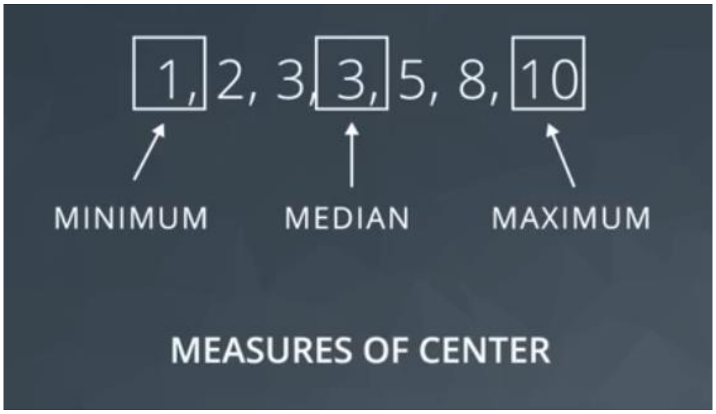


One way we might start is by reporting the average amount of time it takes to complete the nanodegree program.

But that doesn't tell the whole story. I'm sure there are differences in completion depending on what students knew before entering the program. The shortest amount of time needed to complete the nanodegree program might just be a few weeks. How did those people complete the course so fast? Well, the longest might be a couple of years. What proportion of students finish fast within two months? What proportion take longer than eight months? **Using a variety of measures, like measures of center, give you an idea of the average student.**



Measures of spread give you an idea of how students differ. Visuals can provide us a more complete picture of how long it takes any student to complete a program.



The material in the next sections will show you how to use these measures in a way that is informative and understandable to others.

Analyzing Quantitative Data

Four Aspects for Quantitative Data

There are four main aspects to analyzing **Quantitative** data.

1. Measures of Center
2. Measures of Spread
3. The Shape of the data.
4. Outliers

Analyzing Categorical Data

Though not discussed in the video, analyzing categorical data has fewer parts to consider. **Categorical** data is analyzed usually be looking at the counts or proportion of individuals that fall into each group. For example if we were looking at the breeds of the dogs, we would care about how many dogs are of each breed, or what proportion of dogs are of each breed type.

Measures of Center

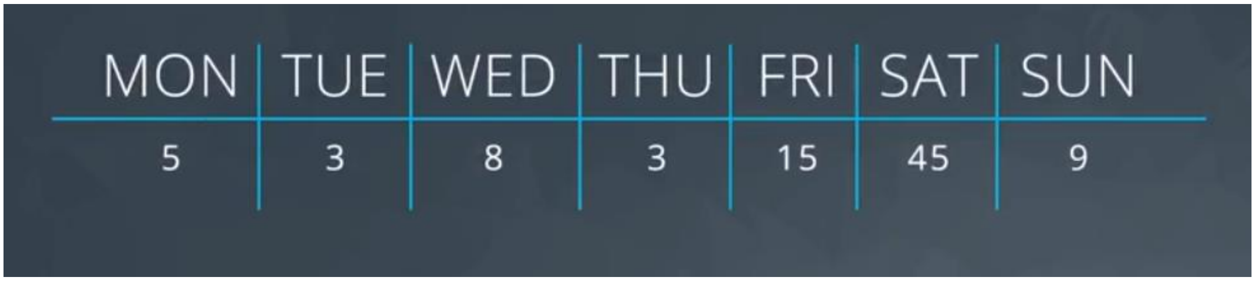
There are three measures of center:

1. **Mean**
2. **Median**
3. **Mode**

The Mean

We focused on the calculation of the mean. The mean is often called the average or the **expected value** in mathematics. We calculate the mean by adding all of our values together, and dividing by the number of values in our dataset.

The remaining measures of the median and mode will be discussed in detail in the upcoming quizzes and videos.



MON	TUE	WED	THU	FRI	SAT	SUN
5	3	8	3	15	45	9

To illustrate how each of these measures is calculated, consider this table of the number of dogs I see at the coffee shop in a week. From the table, we can see that on Monday, I saw five dogs. On Tuesday, I saw three dogs. On Wednesday, I saw eight dogs and so on.

A friend might ask you, how many dogs would you expect to see on any given day? We might choose to respond to this in a lot of different ways. Like, it depends on the day or it depends on the week.

But commonly, the word **expect** is associated with the **mean** or the average of our data set. The mean is calculated as the sum of all of the values in our data set divided by how many data points we have. As you can see calculated here, this value is 12.57 dogs.

That is, the sum of the number of dogs observed on each day divided by the number of days in a week. **The mean isn't always the best measure of center.**

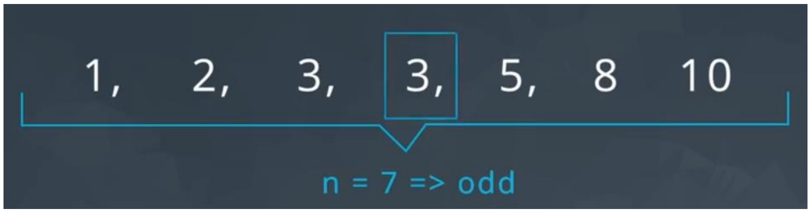
For this data set, you can see that the mean doesn't really seem like it's in the middle of the data at all. There are only two of the seven days that have recorded more dogs than the reported mean. It also is splitting our dogs into decimal values which will seem strange when we're reporting back to our friend.

The Median

The **median** splits our data so that 50% of our values are lower and 50% are higher. How we calculate the median depends on if we have an even number of observations or an odd number of observations.

Median for Odd Values

If we have an **odd** number of observations, the **median** is simply the number in the **direct middle**. For example, if we have 7 observations, the median is the fourth value when our numbers are ordered from smallest to largest. If we have 9 observations, the median is the fifth value.

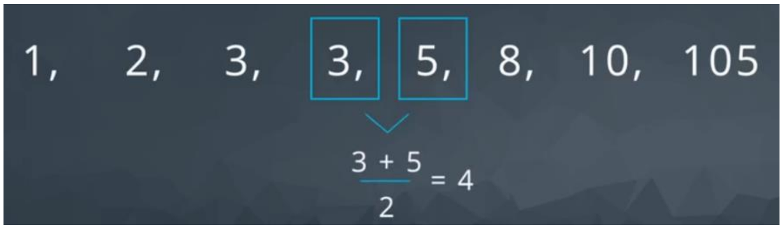


Median for Even Values

If we have an **even** number of observations, the **median** is the **average of the two values in the middle**. For example, if we have 8 observations, we average the fourth and fifth values together when our numbers are ordered from smallest to largest.

In order to compute the median we MUST sort our values first.

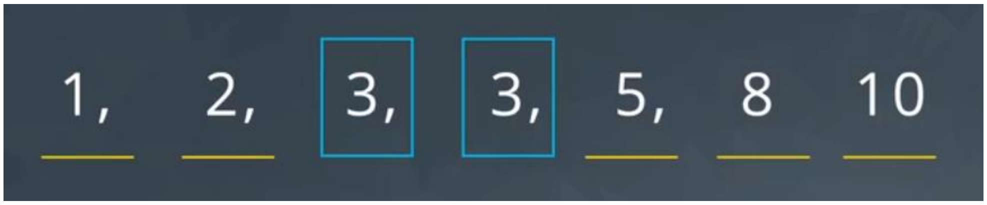
Whether we use the mean or median to describe a dataset is largely dependent on the **shape** of our dataset and if there are any **outliers**. We will talk about this in just a bit!



The Mode

The **mode** is the most frequently observed value in our dataset.

There might be multiple modes for a particular dataset, or no mode at all.



No Mode

If all observations in our dataset are observed with the same frequency, there is no mode. If we have the dataset:

1, 1, 2, 2, 3, 3, 4, 4

There is no mode, because all observations occur the same number of times.

Many Modes

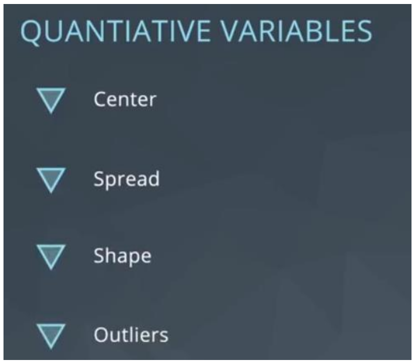
If two (or more) numbers share the maximum value, then there is more than one mode. If we have the dataset:

1, 2, 3, 3, 3, 4, 5, 6, 6, 6, 7, 8, 9

There are two modes 3 and 6, because these values share the maximum frequencies at 3 times, while all other values only appear once.

Notation

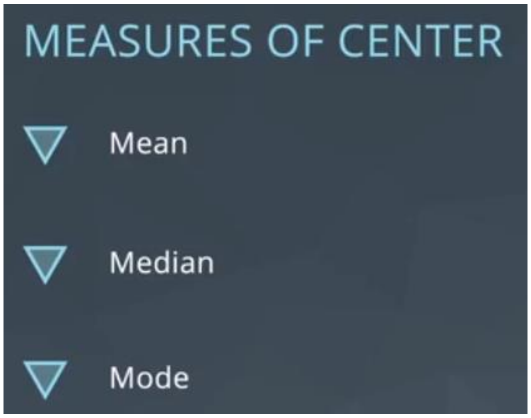
We listed the four main aspects of analyzing quantitative data; **center**, **spread**, **shape** and **outliers**.



QUANTIATIVE VARIABLES

- Center
- Spread
- Shape
- Outliers

We also looked specifically at measures of center, by introducing means, medians and modes. Before we look at measures of spread, it's important to understand notation.



MEASURES OF CENTER

▽ Mean

▽ Median

▽ Mode

Notation is a common language used to communicate mathematical ideas. **Think of notation as a universal language used by academic and industry professionals to convey mathematical ideas.** In the next videos, you might see things that seem confusing. Use the quizzes to assist with your understanding of the concepts.

You likely already know some notation. Plus, minus, multiply, division, and equal signs all have mathematical symbols that you are likely familiar with. Each of these symbols replaces an idea for how numbers interact with one another. In the coming concepts, you will be introduced to some additional ideas related to notation. Though you will not need to use notation to complete the project, it does have the following properties:

1. **Understanding how to correctly use notation makes you seem really smart.** Knowing how to read and write in notation is like learning a new language. A language that is used to convey ideas associated with mathematics.

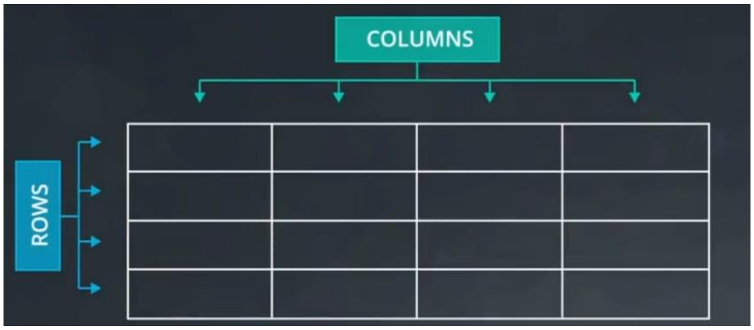
2. **It allows you to read documentation, and implement an idea to your own problem.** Notation is used to convey how problems are solved all the time. One really popular mathematical algorithm that is used to solve some of the world's most difficult problems is known as Gradient Boosting. The way that it solves problems is explained

here: **https://en.wikipedia.org/wiki/Gradient_boosting(opens in a new tab)**. If you really want to understand how this algorithm works, you need to be able to read and understand notation.

3. **It makes ideas that are hard to say in words easier to convey.** Sometimes we just don't have the right words to say. For those situations, I prefer to use notation to convey the message. Similar to the way an emoji or meme might convey a feeling better than words, notation can convey an idea better than words. Usually those ideas are related to mathematics, but I am not here to stifle your creativity.

Notation for Random Variables

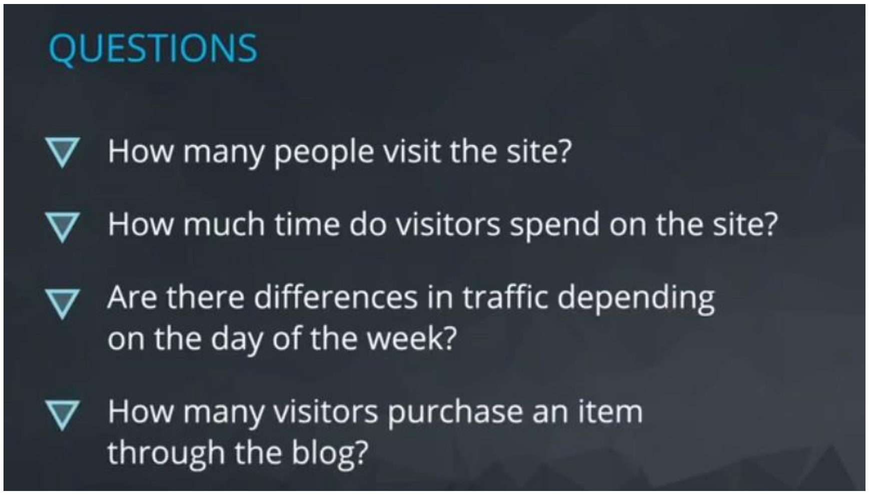
As a first example, let's apply this new idea of notation to something you've used before. Spreadsheets. Spreadsheets are a common way we hold data in the real world. In our spreadsheet, we have rows and columns.



To better understand how we use spreadsheets to hold data, let's work through an example.

Before even collecting data, we usually start with a question or many questions.

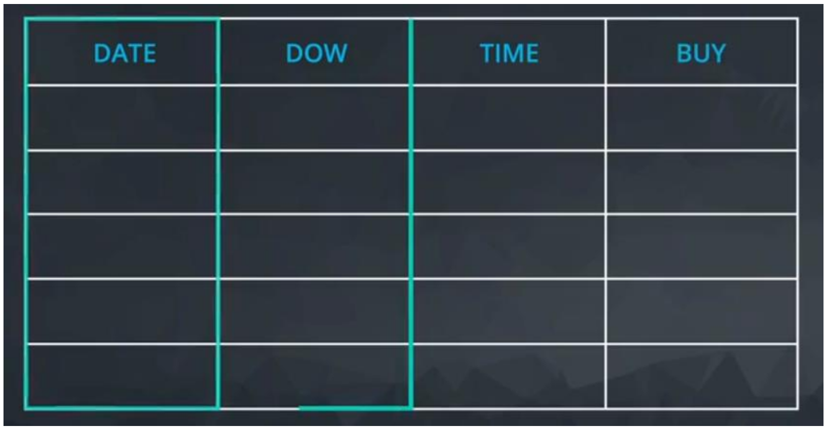
Consider I run a small blog about my best and worst adventures with the dogs at the coffee shop. Which also sells trinkets related to those adventures. Everything from fetch toys to leashes, to doggie bags, and everything in between.



QUESTIONS

- How many people visit the site?
- How much time do visitors spend on the site?
- Are there differences in traffic depending on the day of the week?
- How many visitors purchase an item through the blog?

Questions I might have are: How many people visit my site? Or how much time do visitors spend on my site? Are there differences in traffic, depending on the day of the week? How many visitors purchase an item through the blog? In order to answer these questions, say we keep track of the date of the visit, the day of the week of the visit, the amount of time spent on the site, and whether or not an individual buys an item. We can think of each of these as a column.



DATE	DOW	TIME	BUY

A column in our dataset is associated with a random variable. Explaining what a random variable is in English is complicated. But in notation, it's simple. In English, a random variable is a placeholder for the possible values of some process.

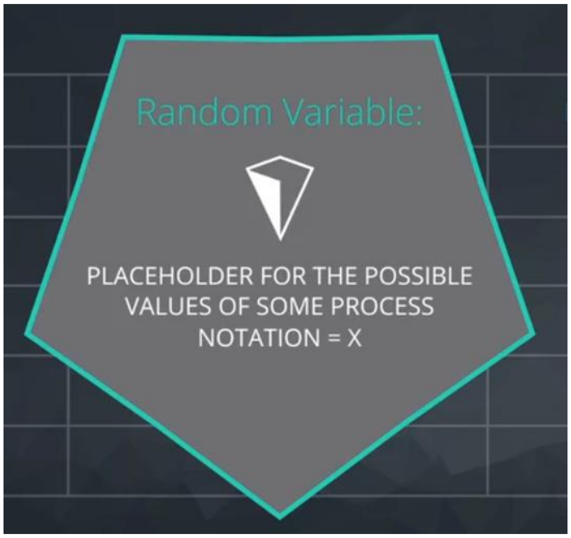
In notation, it's X. For our website, the date of the visit, the day of the week of the visit, the amount of time spent on the site, and whether or not an individual buys an item are all variables.

Let's say we have a visitor on Thursday, June 15.

The visitor stays on our site for five minutes and doesn't buy an item.

Then a second visitor, visits the site on the exact same day for 10 minutes and they do buy an item. Notice how each of these individuals has been added to our spreadsheet.

We might have many more visitors, and we could update our spreadsheet accordingly.

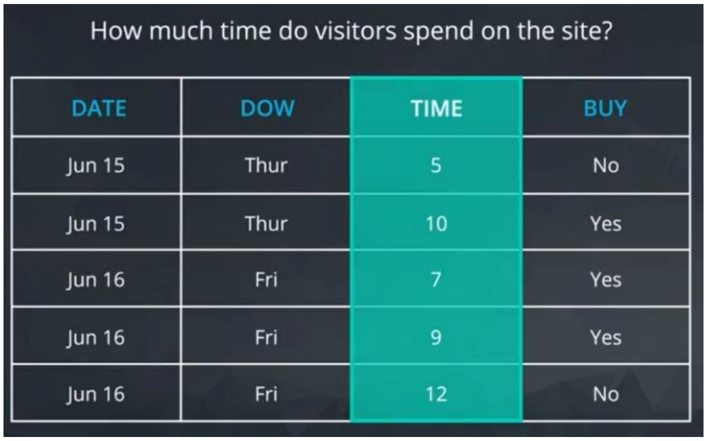


Random Variable:

PLACEHOLDER FOR THE POSSIBLE VALUES OF SOME PROCESS NOTATION = X

When using spreadsheets, we frequently analyze a full column to answer our questions of interest.

For example, to answer the question of how much time do visitors spend on our site? We need to look at this column.



How much time do visitors spend on the site?

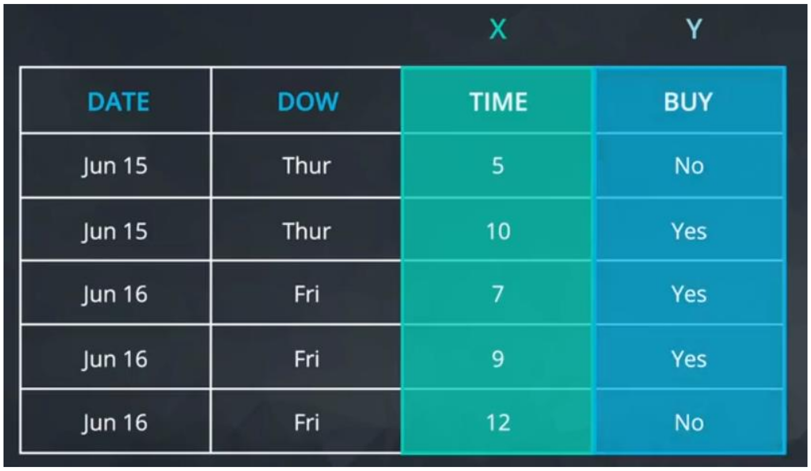
DATE	DOW	TIME	BUY
Jun 15	Thur	5	No
Jun 15	Thur	10	Yes
Jun 16	Fri	7	Yes
Jun 16	Fri	9	Yes
Jun 16	Fri	12	No

To answer the question of, Are there differences in traffic, depending on the day of the week? We need to look at this column (DOW). And to answer the question of, How many purchases occur through our blog? We need to look at this column (BUY). Mathematically, we usually consider a random variable or column using a capital letter.

Commonly, we use the capital letter X, but we can just as easily use Y, Z or any other capital letter.

We might say, consider the random variable X, which signifies the amount of time an individual spends on our website. Therefore, X relates to this entire column.

Consider we also have a random variable Y, which signifies whether or not an individual purchases an item from the website, so Y relates to this entire column.



		X	Y
DATE	DOW	TIME	BUY
Jun 15	Thur	5	No
Jun 15	Thur	10	Yes
Jun 16	Fri	7	Yes
Jun 16	Fri	9	Yes
Jun 16	Fri	12	No

Random & Observed Values

Random variables are represented by capital letters. Once we observe an outcome of these random variables, we notate it as a lower case of the same letter.

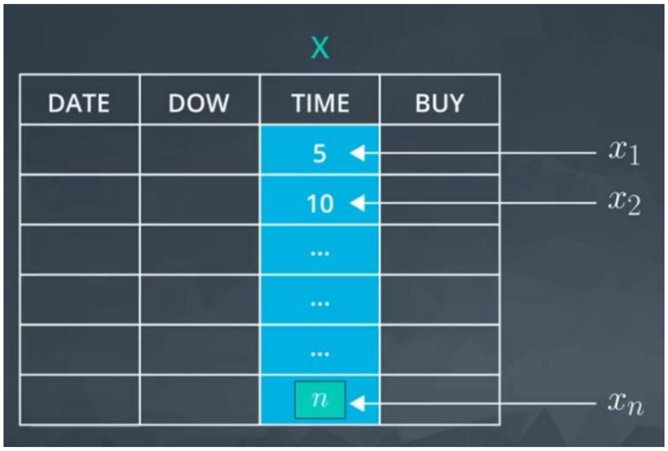
Example 1

For example, the **amount of time someone spends on our site** is a **random variable** (we are not sure what the outcome will be for any particular visitor), and we would notate this with **X**. Then when the first person visits the website, if they spend 5 minutes, we have now observed this outcome of our random variable. We would notate any outcome as a lowercase letter with a subscript associated with the order that we observed the outcome.

If 5 individuals visit our website, the first spends 10 minutes, the second spends 20 minutes, the third spends 45 mins, the fourth spends 12 minutes, and the fifth spends 8 minutes; we can notate this problem in the following way:

X is the amount of time an individual spends on the website.

$$\boldsymbol{x\_1} = 10, \quad \boldsymbol{x\_2} = 20 \quad \boldsymbol{x\_3} = 45 \quad \boldsymbol{x\_4} = 12 \quad \boldsymbol{x\_5} = 8.$$



DATE	DOW	TIME	BUY
		5	
		10	
		...	
		...	
		...	
		n	

The capital **X** is associated with this idea of a **random variable**, while the observations of the random variable take on lowercase **x** values.

Example 2

Taking this one step further, we could ask:

What is the probability someone spends more than 20 minutes in our website?

In notation, we would write:

$$\boldsymbol{P(X > 20)?}$$

Here **P** stands for **probability**, while the parentheses encompass the statement for which we would like to find the probability. Since **X** represents the amount of time spent on the website, this notation represents the probability the amount of time on the website is greater than 20. We could find this in the above example by noticing that only one of the 5 observations exceeds 20. So, we would say there is a **1** (the 45) **in 5 or 20%** chance that an individual spends more than 20 minutes on our website (based on this dataset).

Example 3

If we asked: **What is the probability of an individual spending 20 or more minutes on our website?** We could notate this as:

$$\boldsymbol{P(X \geq 20)?}$$

We could then find this by noticing there are two out of the five individuals that spent 20 or more minutes on the website. So this probability is **2 out of 5 or 40%**.

Better Way?

We know that the mean is calculated as the sum of all our values divided by the number of values in our dataset.

In our current notation, adding all of our values together can be extremely tedious. If we want to add 3 values of some random variable together, we would use the notation:

$$x_1 + x_2 + x_3$$

If we want to add 6 values together, we would use the notation:

$$x_1 + x_2 + x_3 + x_4 + x_5 + x_6$$

To extend this to add one hundred, one thousand, or one million values would be ridiculous! How can we make this easier to communicate?!

Aggregations

An **aggregation** is a way to turn multiple numbers into fewer numbers (commonly one number). **Summation** is a common aggregation. The notation used to sum our values is a greek symbol called sigma Σ.

Example 1

Imagine we are looking at the amount of time individuals spend on our website. We collect data from nine individuals:

$\mathbf{x_1}$ = 10, $\mathbf{x_2}$ = 20 $\mathbf{x_3}$ = 45 $\mathbf{x_4}$ = 12 $\mathbf{x_5}$ = 8 $\mathbf{x_6}$ = 12, $\mathbf{x_7}$ = 3 $\mathbf{x_8}$ = 68
$\mathbf{x_9}$ = 5

If we want to sum the **first three values** together in our previous notation, we write:

$$\mathbf{x_1} + \mathbf{x_2} + \mathbf{x_3}$$

In our new notation, we can write:

$$\sum_{i=1}^{3} x_i.$$

Notice, our notation starts at the first observation (�=1*i*=1) and ends at 3 (the number at the top of our summation).

So all of the following are equal to one another:

$$\sum_{i=1}^{3} x_i = \mathbf{x}_1 + \mathbf{x}_2 + \mathbf{x}_3 = 10 + 20 + 45 = 75$$

Example 2

Now, imagine we want to sum the **last three values** together.

$$\mathbf{x}_7 + \mathbf{x}_8 + \mathbf{x}_9$$

In our new notation, we can write:

$$\sum_{i=7}^{9} x_i.$$

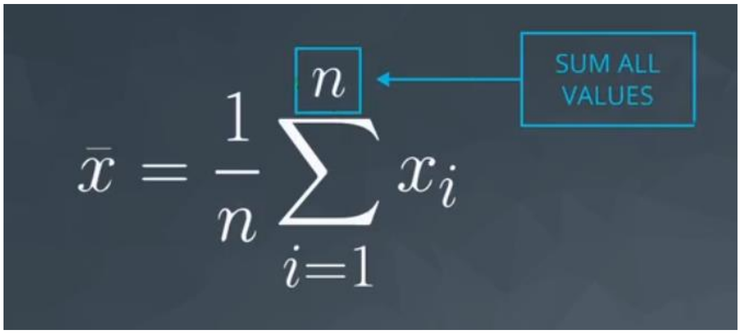
Notice, our notation starts at the seventh observation $(i = 7)$ and ends at 9 (the number at the top of our summation).

Other Aggregations

The Σ sign is used for aggregating using summation, but we might choose to aggregate in other ways. Summing is one of the most common ways to need to aggregate. However, we might need to aggregate in alternative ways. If we wanted to multiply all of our values together we would use a product sign Π , capital Greek letter pi. The way we aggregate continuous values is with something known as integration (a common technique in calculus), which uses the following symbol $\int$ which is just a long s. We will not be using integrals or products for quizzes in this class, but you may see them in the future!

Final Steps for Calculating the Mean

To finalize our calculation of the mean, we introduce **n** as the total number of values in our dataset. We can use this notation both at the top of our summation, as well as for the value that we divide by when calculating the mean.



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

Instead of writing out all of the above, we commonly write $\bar{x}$ to represent the mean of a dataset. Although, similar to the first video, we could use any variable. Therefore, we might also write $\bar{y}$, or any other letter.

We also could index using any other letter, not just i. We could just as easily use *j*, *k*, or *m* to index each of our data values. The quizzes on the next concept will help reinforce this idea.

Summary on Notation

Notation Recap

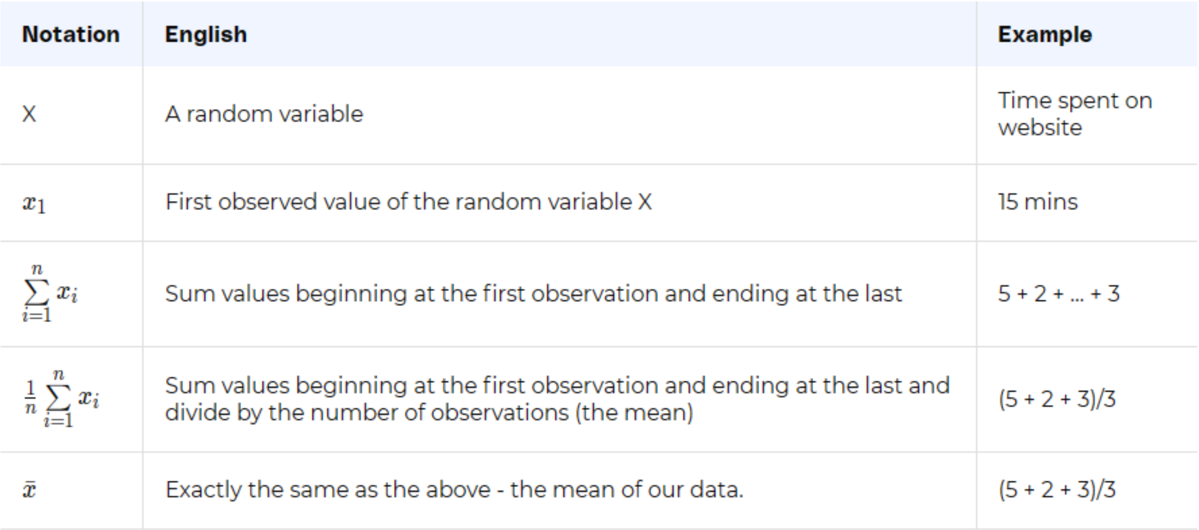
Notation is an essential tool for communicating mathematical ideas. We have introduced the fundamentals of notation in this lesson that will allow you to read, write, and communicate with others using your new skills!

Notation and Random Variables

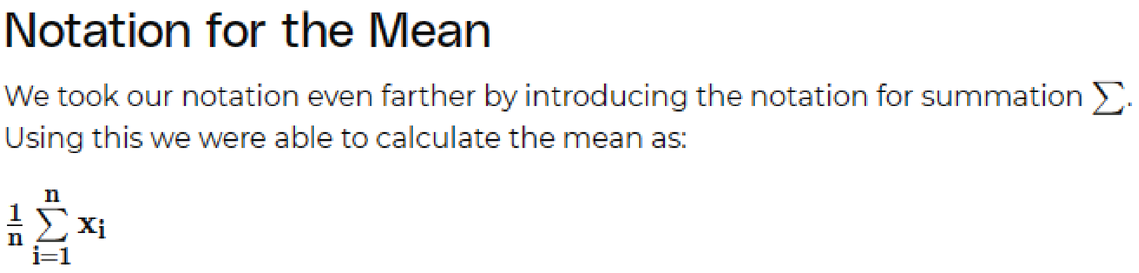
As a quick recap, **capital letters** signify **random variables**. When we look at **individual instances** of a particular random variable, we identify these as **lowercase letters** with subscripts attach themselves to each specific observation.

For example, we might have **X** be the amount of time an individual spends on our website. Our first visitor arrives and spends 10 minutes on our website, and we would say x_1 is 10 minutes.

We might imagine the random variables as columns in our dataset, while a particular value would be notated with the lower case letters.



Notation	English	Example
X	A random variable	Time spent on website
x_1	First observed value of the random variable X	15 mins
$\sum_{i=1}^{n} x_i$	Sum values beginning at the first observation and ending at the last	5 + 2 + ... + 3
$\frac{1}{n}\sum_{i=1}^{n} x_i$	Sum values beginning at the first observation and ending at the last and divide by the number of observations (the mean)	(5 + 2 + 3)/3
$\bar{x}$	Exactly the same as the above - the mean of our data.	(5 + 2 + 3)/3



Notation for the Mean

We took our notation even farther by introducing the notation for summation $\sum$. Using this we were able to calculate the mean as:

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

In the next section, you will see this notation used to assist in your understanding of calculating various measures of spread. Notation can take time to fully grasp. Understanding notation not only helps in conveying mathematical ideas, but also in writing computer programs - if you decide you want to learn that too! Soon you will analyze data using spreadsheets. When that happens, many of these operations will be hidden by the functions you will be using. But until we get to spreadsheets, it is important to understand how mathematical ideas are commonly communicated. **This isn't easy, but you can do it!**

Measures of Spread

Measures of Spread are used to provide us an idea of how spread out our data are from one another. Common measures of spread include:

1. **Range**
2. **Interquartile Range (IQR)**
3. **Standard Deviation**
4. **Variance**

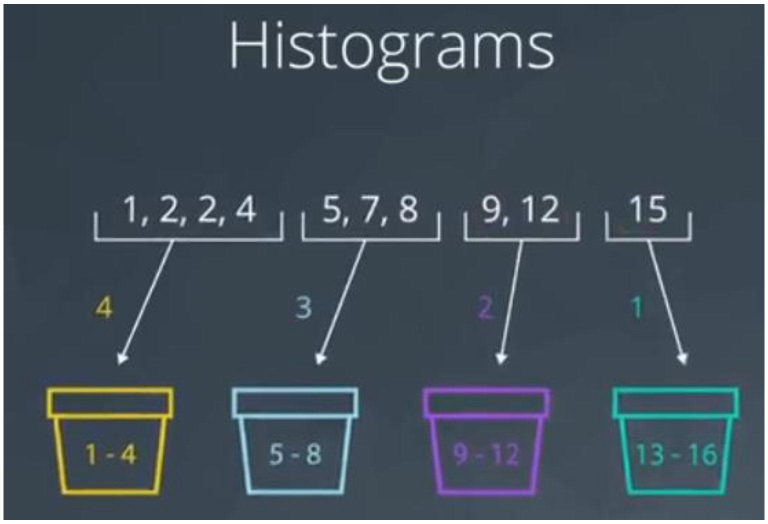
Throughout this lesson you will learn how to calculate these, as well as why we would use one measure of spread over another.

Histograms

Histograms are super useful to understanding the different aspects of quantitative data. In the upcoming concepts, you will see histograms used all the time to help you understand the four aspects we outlined earlier regarding a quantitative variable:

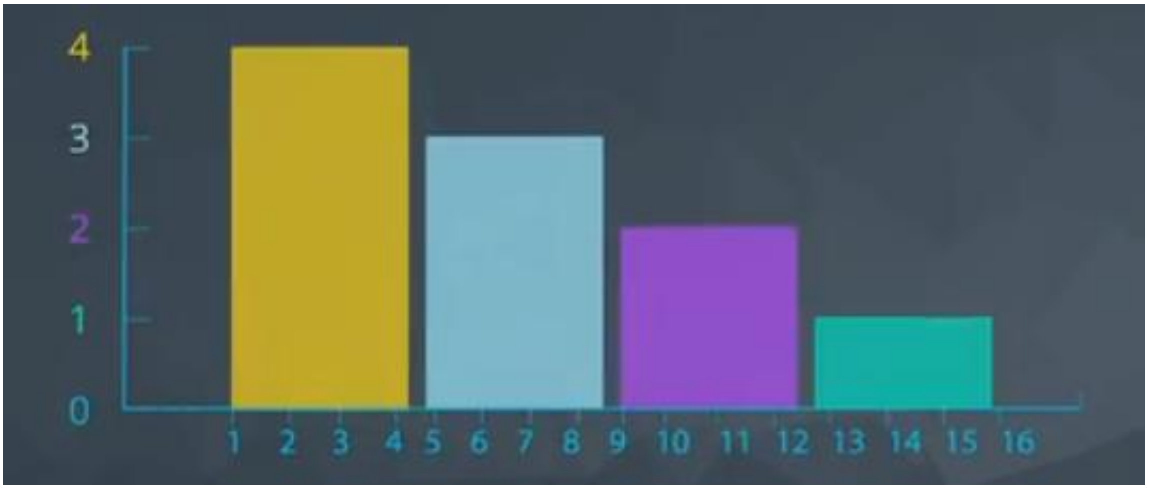
- center
- spread
- shape
- outliers

In order to understand how histograms are constructed, consider we have the following dataset. First, we need to bin our data. You as the histogram creator ultimately choose how the binning occurs. Here, I have chosen our bins as one to four, five to eight, 9-12 and 13-16. Because these first four values are between one and four, they go into the first bin. These next three values are between five and eight, so they fall in the next bin, then these two values fall in this bin and 15 falls into our last bin.



The number of values in each bin determine the height of each histogram bar.

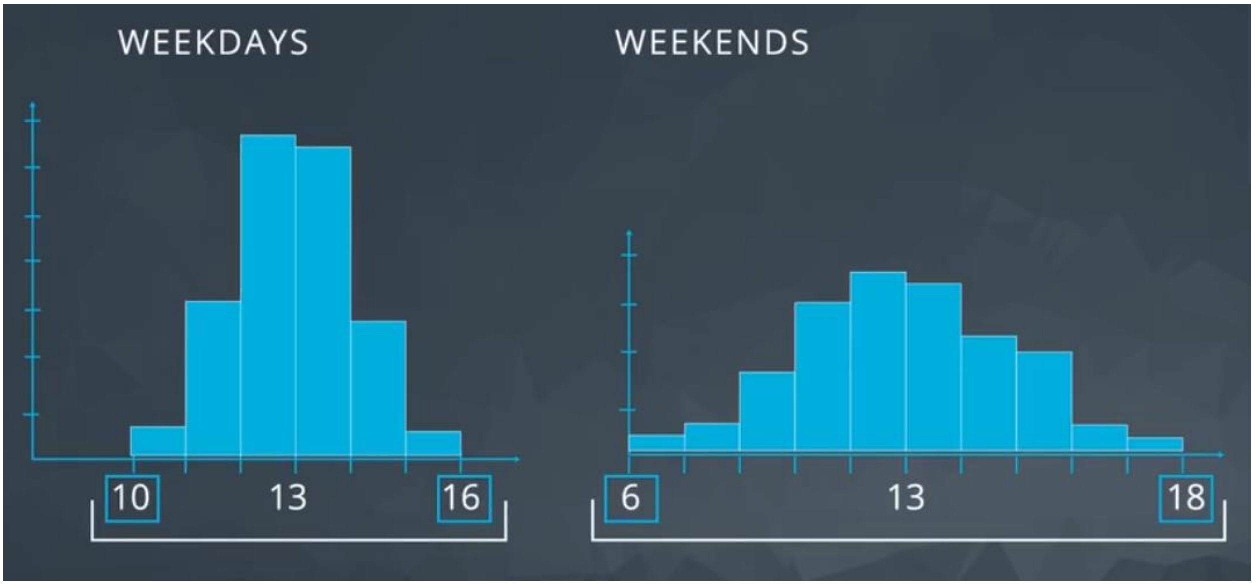
Changing the bins will result in a slightly different visual. There really isn't a right answer to choosing our bins, and in most cases software will choose the appropriate bins for us. But it is something to be aware of.



What is the difference

Here, are two histograms comparing the number of dogs that I saw on weekdays to the number of dogs I saw on weekends from last year. You will notice that the tallest bins for both weekdays and weekends are associated with 13 dogs.

So the number of dogs I expect to see are essentially the same for weekdays as on weekends.



And the measures of center would basically be the same here. Both have a mean, median, and mode that are about 13 dogs. But something is different about these two distributions.

So what's the difference? Well, the difference is **how spread out the data are for each group.** You can see that the number of dogs I see on weekdays, ranges from 10-16. While on weekends, it ranges from 6-18.

In the upcoming sections, we will look at the most common ways to measure the spread of our data.

Calculating the 5 Number Summary

The five number summary consist of 5 values:

1. **Minimum:** The smallest number in the dataset.
2. Q_1: The value such that 25% of the data fall below.
3. Q_2: The value such that 50% of the data fall below.
4. Q_3: The value such that 75% of the data fall below.
5. **Maximum:** The largest value in the dataset.

In the above video we saw that calculating each of these values was essentially just finding the median of a bunch of different dataset. Because we are essentially calculating a bunch of medians, the calculation depends on whether we have an odd or even number of values.

Range

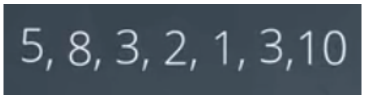
The **range** is then calculated as the difference between the **maximum** and the **minimum**.

IQR

The **interquartile range** is calculated as the difference between Q_3 and Q_1.

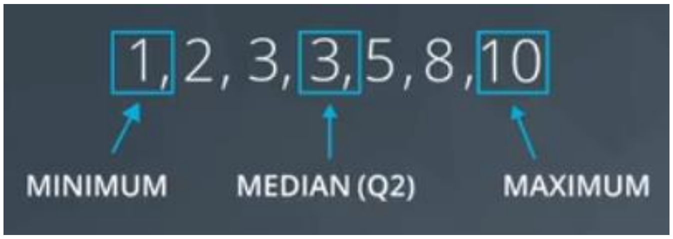
In the upcoming sections, you will practice this with Katie and on your own.

Consider we have the following dataset.



5, 8, 3, 2, 1, 3,10

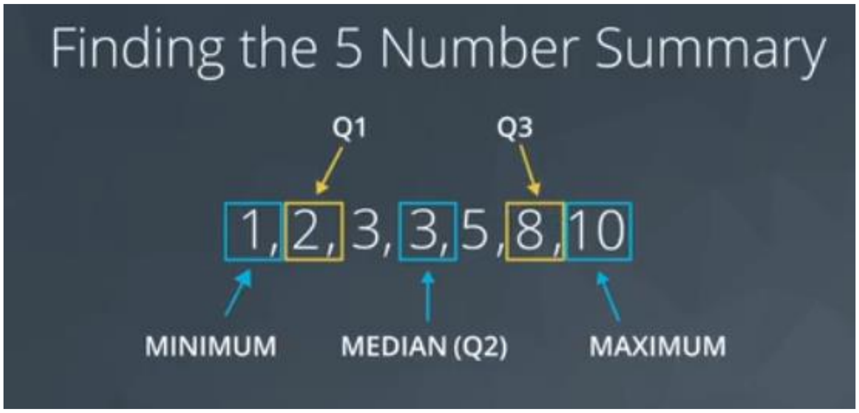
The first thing we need to do to calculate the Five Number Summary is to order our values. Once ordered, the minimum and the maximum values are easy to identify as the smallest and largest values.



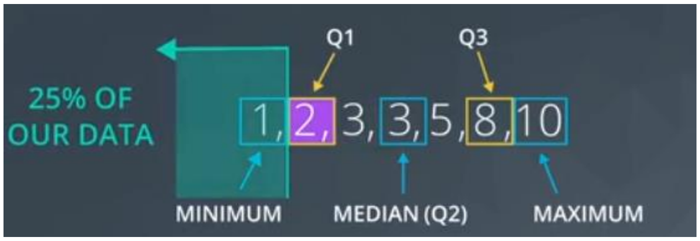
1, 2, 3, 3, 5, 8, 10

As we calculate it in the section on measures of center, the median is the middle value in our dataset. We also call this Q2 or the second quartile because The remaining two values to complete the Five Number Summary are Q1 and Q3.

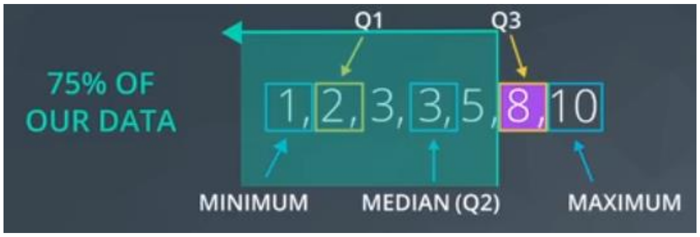
These values can be thought of as the medians of the data on either side of Q2.



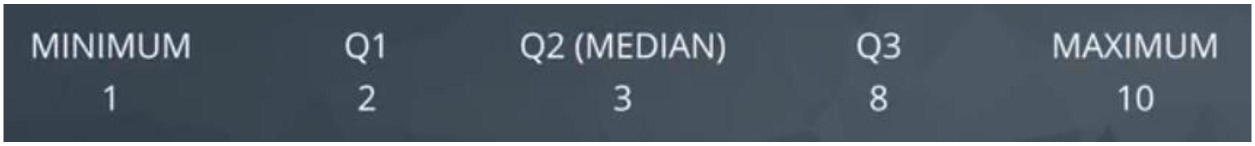
That is the median of these data points is Q1, this value is such that 25% of our data fall below it, and the median of these data points is Q3 or the third quartile.



This value is such that 75% of our data fall below this mark. Notice, Q2 was not an either a set of these points used to calculate Q1 or Q3.



This provides our Five Number Summary as the following.



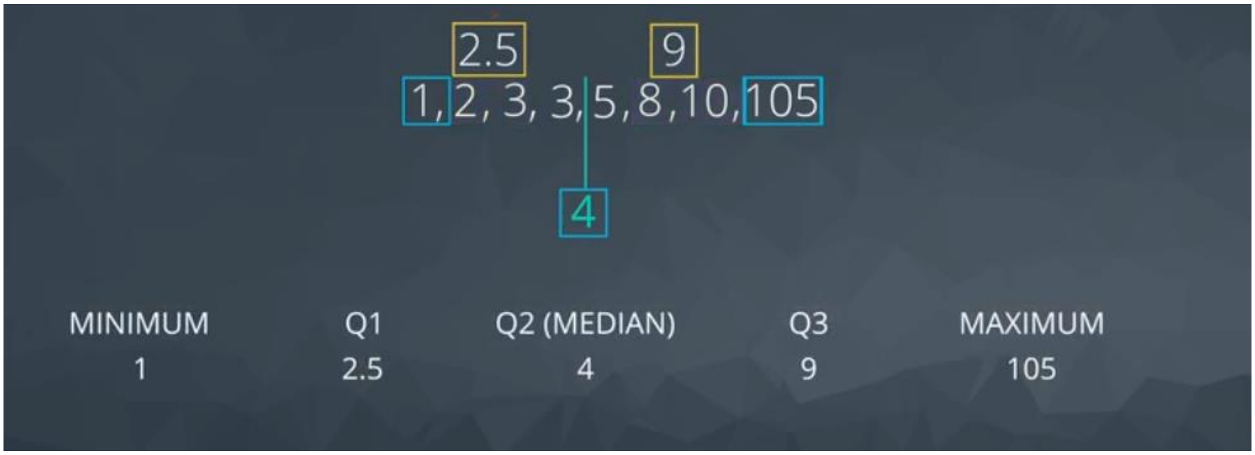
MINIMUM	Q1	Q2 (MEDIAN)	Q3	MAXIMUM
1	2	3	8	10

Let's consider another example for in our dataset has an even set of values.

Again, we first need to order the values. We can quickly identify the maximum and the minimum. Remember, with an even number of values, the median or Q2 is given as the mean of these two values here.

In order to find Q1 and Q3, we divide our dataset between the two values we use to find the median. This provides these two datasets. Finding the median of each of these will provide Q1 and Q3. For this dataset, Q1 will be the mean of these two values, and Q3, will be the mean of these two values.

This provides our Five Number Summary as the following.



2.5 9

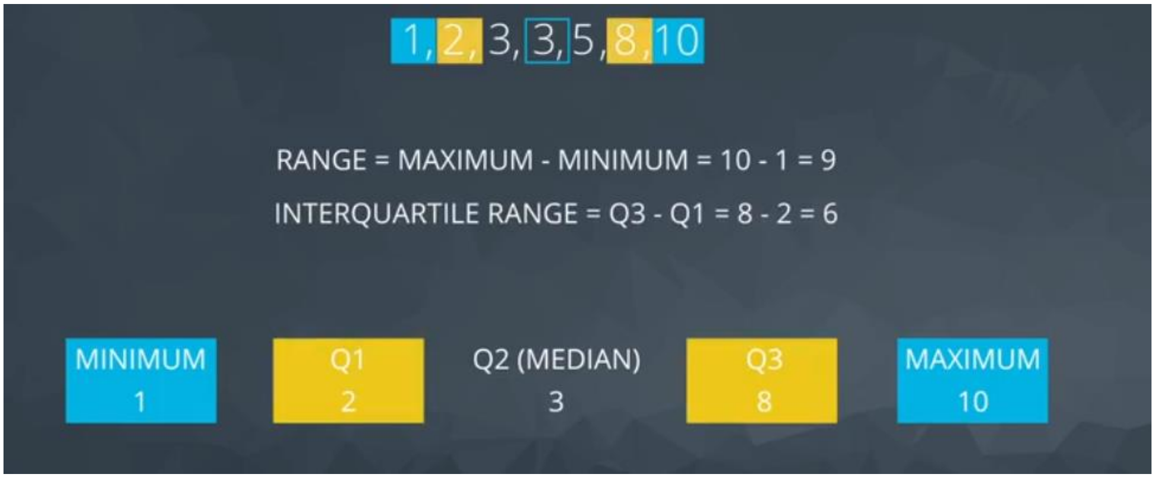
1, 2, 3, 3, 5, 8, 10, 105

4

MINIMUM	Q1	Q2 (MEDIAN)	Q3	MAXIMUM
1	2.5	4	9	105

Once we've calculated all the values for the Five Number Summary, finding the range and interquartile range is no problem. For the first dataset, the range is calculated as the maximum minus the minimum.

For the first dataset, this was 10 minus 1 or 9. And the interquartile range is calculated as Q3 minus Q1, which is 8 minus 2 or 6.



1, 2, 3, 3, 5, 8, 10

RANGE = MAXIMUM - MINIMUM = 10 - 1 = 9

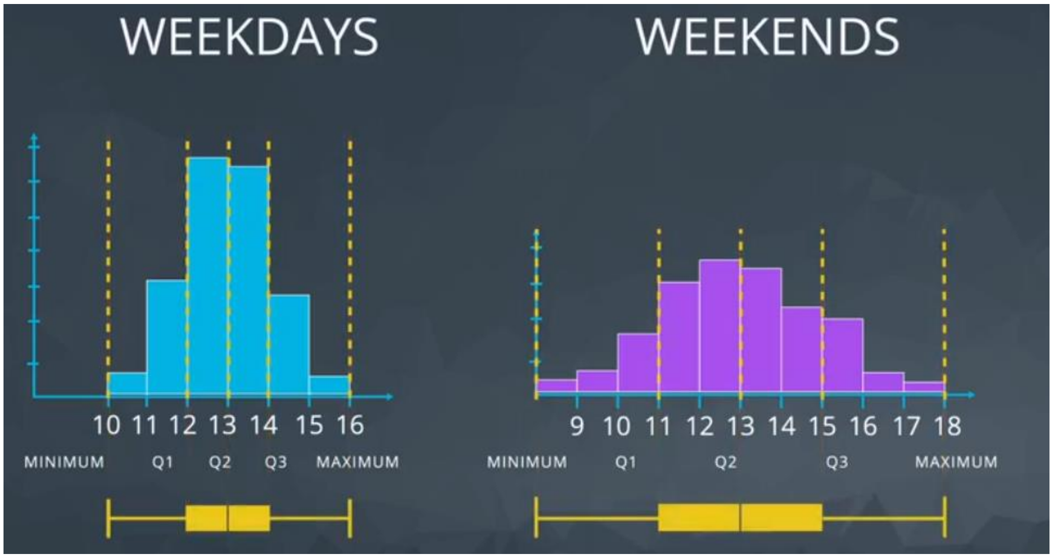
INTERQUARTILE RANGE = Q3 - Q1 = 8 - 2 = 6

MINIMUM	Q1	Q2 (MEDIAN)	Q3	MAXIMUM
1	2	3	8	10

5 Number Summary to Variance

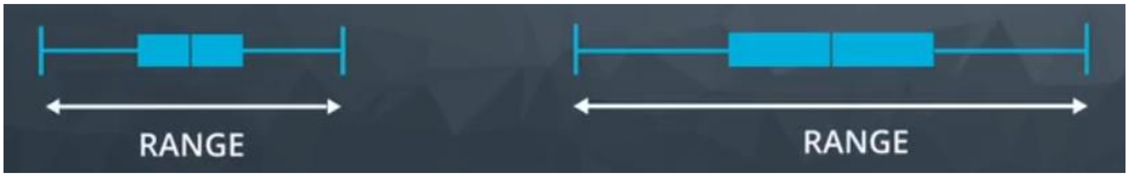
Looking back at the distributions we found for the number of dogs I see, we can mark the values of our five number summary like this.

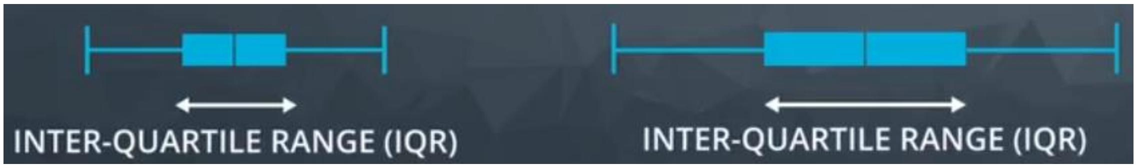
If we take just these marks, this makes a common plot for data known as a box plot.



Though I prefer a histogram in most cases, **a box plot can be useful for quickly comparing the spread of two data sets across some key metrics**, like our quartiles, and the maximum and minimum.

From both the histogram and the five-number summary, we can quickly see that the number of dogs I see on weekends varies much more than the number of dogs I see on weekdays. We can also visualize the distance from here to here as the range, while the distance from here to here is the interquartile range.





There are lots of useful metrics we can get from these box plots.

But what if we want to compare the spreads of these distributions without having to carry around all five of these values for each distribution? What if I wanted just a single value to be able to compare the two distributions spreads?

Standard Deviation and Variance

The most common way that professionals measure the spread of a data set with a single value is with the **standard deviation or variance**. Here, we will focus on the standard deviation, but we will actually learn how to calculate the variance in the process. If you have never heard of these measures before, this calculation will probably look pretty complex.

When all's said and done with this calculation, **the standard deviation will tell us on average how far every point is from the mean of the points**.

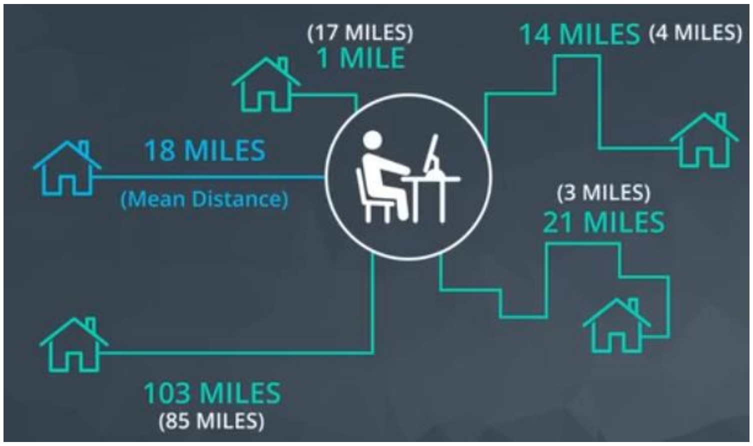
As a quick mental picture, imagine we wanted to know how far employees were located from their place of work. One person might be 15 miles, another 35, another only one mile, and another might be remote and is 103 miles. We could aggregate all of these distances together to show that the average distance employees are located from their work is 18 miles.

1 MILE
14 MILES
18 MILES
(Mean Distance)
21 MILES
103 MILES

But now, we want to know how the distance to work varies from one employee to the next. We could use the five number summary as a description.

But if we wanted just one number to talk about the spread, we'd probably choose the standard deviation.

The standard deviation is on average how much each observation varies from the mean. For this example this is, how much on average the distance each person is from work differs from the average distance all of them are from work.

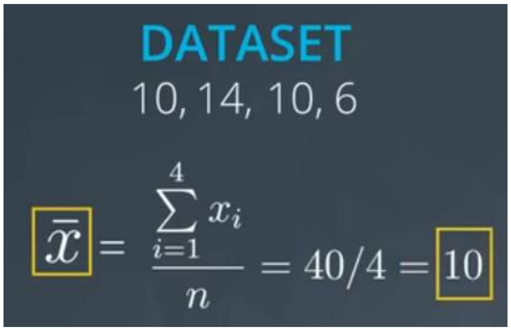


So, this one is three miles farther from work than the average while this individual is four miles closer to work than the average. The standard deviation is how far on average are individuals located from this mean distance. So, it is like the average of all of these distances. We will take a closer look at this but hopefully this gives you a strong conceptual understanding of what we'll be calculating in the next sections.

Calculation of Standard Deviation

We will work with data to calculate this measure as well as associate notation with it. It is worth noting that after this lesson, you probably won't calculate this measure by hand ever again, because you'll learn software to do it for you. The calculating it yourself will give you intuition behind what it's actually doing. And this intuition is necessary to become good at understanding data and choosing the right analysis for your situation.

Imagine we have a data set with four values, The first thing we need to do to calculate the standard deviation is to find the mean.



DATASET

10, 14, 10, 6

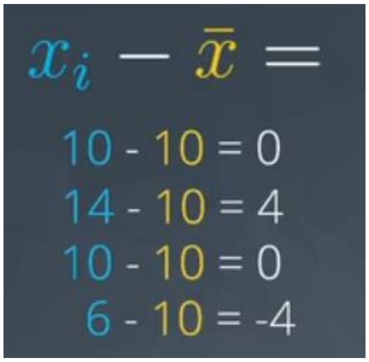
$$\boxed{\bar{x}} = \frac{\sum_{i=1}^{4} x_i}{n} = 40/4 = \boxed{10}$$

Then we want to look at the distance of each observation from this mean.

Two of these observations are exactly equal to the mean.

So the distance here is zero.

One is 4 larger the 14, while the other is 4 smaller the 6.

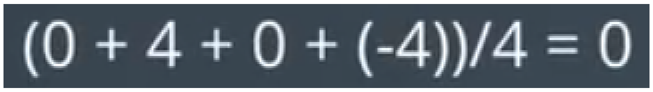


$$x_i - \bar{x} =$$

$$10 - 10 = 0$$
$$14 - 10 = 4$$
$$10 - 10 = 0$$
$$6 - 10 = -4$$

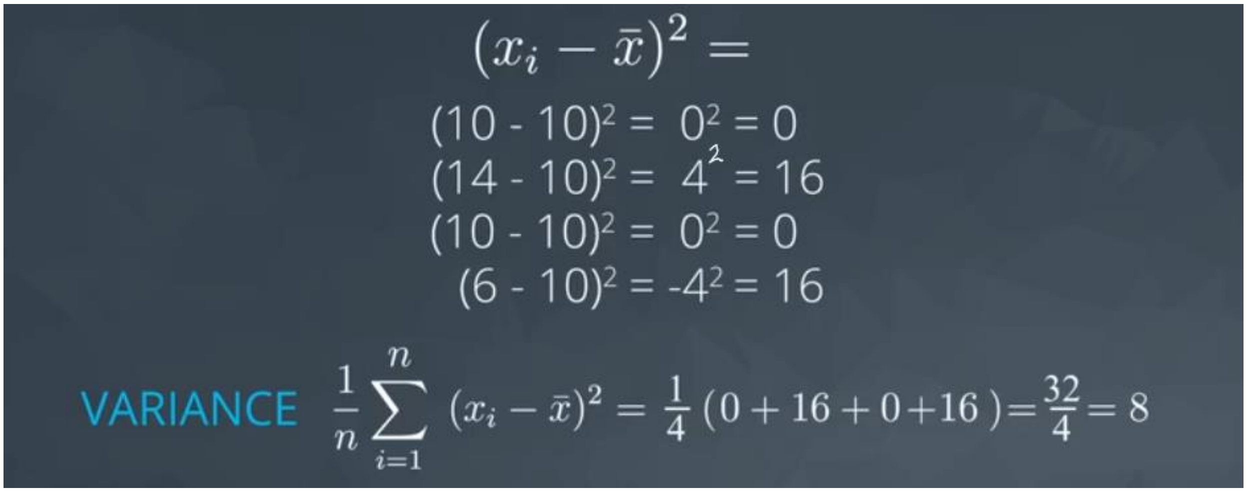
Then, if we were to average these distances, the positive would cancel with the negative value. And the value of zero isn't a great measure of the spread here.

Zero would suggest that all the values are the same or that there's no spread.



$$(0 + 4 + 0 + (-4))/4 = 0$$

So instead, we need to make all of these values positive. The way we do this when calculating the standard deviation is by squaring them all. If we do that here, our negative and positive 4 values will become 16s. Now, we could average these to find the average squared distance of each observation from the mean.



$$(x_i - \bar{x})^2 =$$

$$(10 - 10)^2 = 0^2 = 0$$
$$(14 - 10)^2 = 4^2 = 16$$
$$(10 - 10)^2 = 0^2 = 0$$
$$(6 - 10)^2 = -4^2 = 16$$

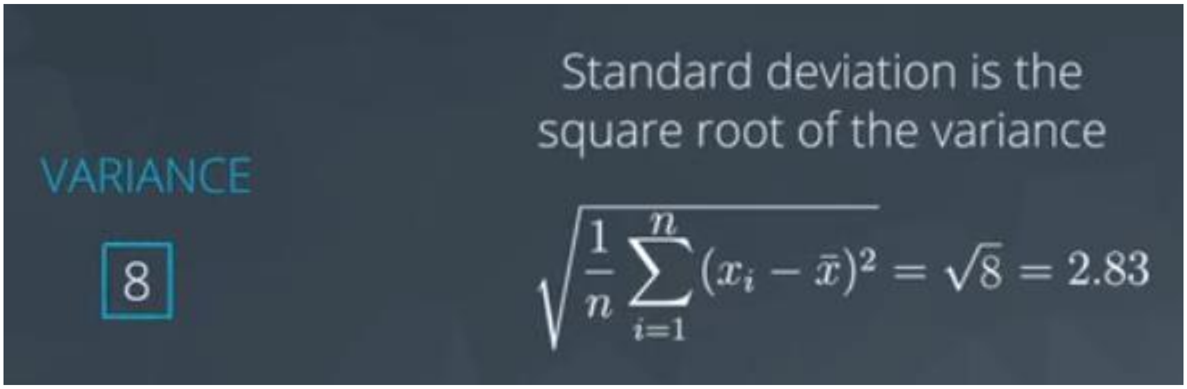
VARIANCE $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{4}(0 + 16 + 0 + 16) = \frac{32}{4} = 8$

This is called the variance. Finding the average, just as we did before, means adding all of these values and dividing by how many there are.

In our case, we had 0, 16, 0, 16 and we divided it by 4 because we have 4 observations. However, this is an average of squared values which we only did to get positive values in the first place.

So to get our standard deviation, we take the square root of this ending value.

Here, our standard deviation is 2.83.



Standard deviation is the square root of the variance

VARIANCE

$\boxed{8}$

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{8} = 2.83$$

So this is on average how far each point in our data set is from the mean, which is the definition of the standard deviation.

Other Measures of Spread

5 Number Summary

In the previous sections, we have seen how to calculate the values associated with the **five number summary** (**min**, *Q*1, *Q*2, *Q*3, **max**), as well as the measures of spread associated with these values (**range** and **IQR**).

For datasets that are **not symmetric**, the five number summary and a corresponding box plot are a great way to get started with understanding the spread of your data. **Although I still prefer a histogram in most cases, box plots can be easier to compare two or more groups.** You will see this in the quizzes towards the end of this lesson.

Variance and Standard Deviation

Two additional **measures of spread** that are used all the time are the **variance** and **standard deviation**. At first glance, the variance and standard deviation can seem overwhelming. If you do not understand the expressions below, don't panic! In this section, I just want to give you an overview of what the next sections will cover. We will walk through each of these parts thoroughly in the next few sections, but the big picture goal is to generally understand the following:

1. How the mean, variance and standard deviation are calculated.
2. Why the measures of variance and standard deviation make sense to capture the spread of our data.
3. Fields where you might see these values used.
4. Why we might use the standard deviation or variance as opposed to the values associated with the 5 number summary for a particular dataset.

The **standard deviation** is a measurement that has the **same units** as our original data, while the units of the variance are the square of the units in our original data. For example, if the units in our original data were dollars, then units of the standard deviation would also be dollars, while the units of the variance would be dollars squared.

Why the Standard Deviation?

So it might seem absurd to do this calculation of the standard deviation.

I mean, it's such a complicated way to measure the spread of our data compared to the five number summary we saw earlier. But it turns out that the standard deviation is used all the time to get a single number to compare the spread of two data sets. It is kind of nice to be able to talk about how spread out our data are from one another without having to report an entire table of values.

We can just compare the standard deviation for one group to the standard deviation of another group. And we have a way to tell which dataset is more spread out. **Having one number simplifies the amount of information that the person you're reporting to needs to consume.**

Having the single value also has other advantages with regard to what is known as inferential statistics, but that's beyond what we need to know now.

For now, we just need to know that we have a way to take all of our values and get a single number that tells us how spread out they are from one another.

Important Final Points

1. The variance is used to compare the spread of two different groups. A set of data with higher variance is more spread out than a dataset with lower variance. Be careful though, there might just be an outlier (or outliers) that is increasing the variance, when most of the data are actually very close.
2. When comparing the spread between two datasets, the units of each must be the same.
3. When data are related to money or the economy, higher variance (or standard deviation) is associated with higher risk.
4. The standard deviation is used more often in practice than the variance, because it shares the units of the original dataset.

Use in the World

The standard deviation is associated with risk in finance, assists in determining the significance of drugs in medical studies, and measures the error of our results for predicting anything from the amount of rainfall we can expect tomorrow to your predicted commute time tomorrow.

These applications are beyond the scope of this lesson as they pertain to specific fields, but know that understanding the spread of a particular set of data is extremely important to many areas. In this lesson you mastered the calculation of the most common measures of spread.

Measures of Center and Spread Summary

Recap

Variable Types

We have covered a lot up to this point! We started with identifying data types as either categorical or quantitative. We then learned, we could identify quantitative variables as either continuous or discrete. We also found we could identify categorical variables as either ordinal or nominal.

Categorical Variables

When analyzing categorical variables, we commonly just look at the count or percent of a group that falls into each **level** of a category. For example, if we had two **levels** of a dog category: lab and not lab. We might say, 32% of the dogs were lab (percent), or we might say 32 of the 100 dogs I saw were labs (count).

However, the 4 aspects associated with describing quantitative variables are not used to describe categorical variables.

Quantitative Variables

Then we learned there are four main aspects used to describe quantitative variables:

1. Measures of **Center**
2. Measures of **Spread**
3. **Shape** of the Distribution
4. **Outliers**

We looked at calculating measures of Center

1. **Means**
2. **Medians**

3. **Modes**

We also looked at calculating measures of Spread

1. **Range**
2. **Interquartile Range**
3. **Standard Deviation**
4. **Variance**

Calculating Variance

We saw that we could calculate the variance as:

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

You will also see:

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

The reason for this is beyond the scope of what we have covered thus far, but you can find an explanation **here(opens in a new tab)**.

Imagine you have a group of numbers, like [5, 6, 7, 8, 9]. You calculate their mean (average), which is 7.

Now, to understand how spread out these numbers are from the average, you calculate the standard deviation. It's like asking, "On average, how far are these numbers from 7?"

But here's the catch: if you just use "n" (the total number of items in your group) in the formula to find the standard deviation, you might get a slightly smaller number than what you'd expect if you had the data from the whole population. That's because your group (sample) might be a little closer to the mean than the whole population would be.

So, we use "n-1" instead of "n" in the formula when we're dealing with a sample. This makes the standard deviation a bit bigger and closer to what it would be for the whole population.

In simpler terms, using "n-1" instead of "n" when finding the standard deviation for a sample is like adjusting for the fact that our sample might be a bit biased towards being closer to the average than the entire population would be. It helps us get a better estimate of how spread out the numbers are in the whole population based on our smaller sample.

You can commonly find answers to your questions with a quick **Google search(opens in a new tab)**. Now is a great time to get started with this practice! This answer should make more sense at the completion of this lesson.

Standard Deviation vs. Variance

The standard deviation is the square root of the variance. In practice, you usually use the standard deviation rather than the variance. The reason for this is because the standard deviation shares the same units with our original data, while the variance has squared units.

What Next?

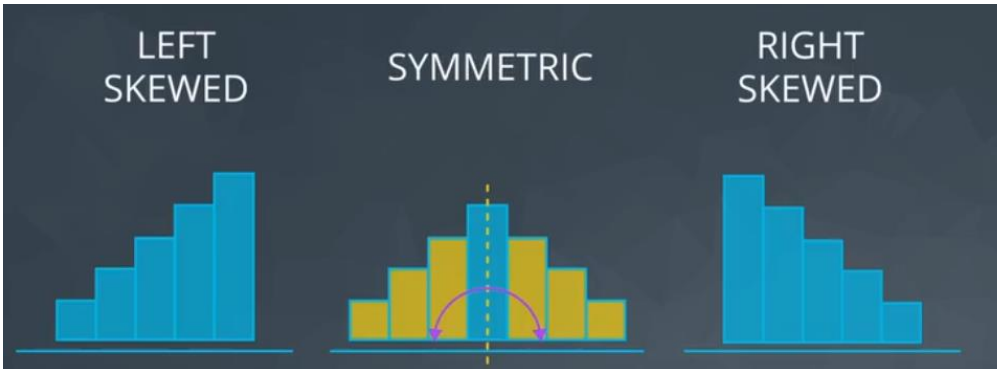
In the next sections, we will be looking at the last two aspects of quantitative variables: **shape** and **outliers**. What we know about measures of center and measures of spread will assist in your understanding of these final two aspects.

Supporting Materials

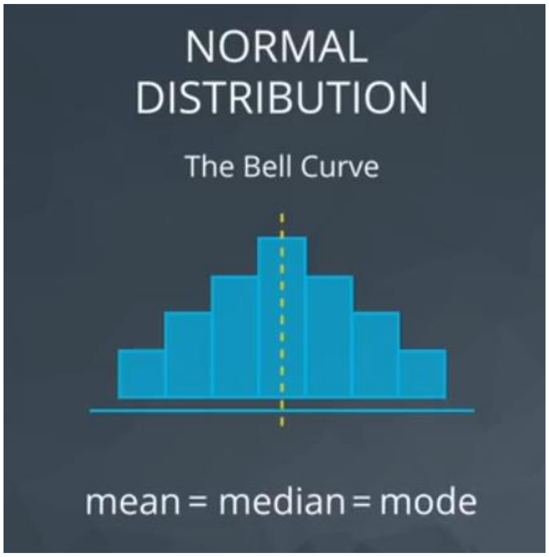
- **Calculating Variance**

Shape of Distributions

Now that we've discussed how to build a histogram, we can use this to determine the shape associated with our data. Here we have three histograms, showing the shape for three different data sets. The histogram that has shorter bins on the left and taller bins on the right, is considered a left skewed shape. This histogram that has shorter bins on the right and taller bins on the left, is considered a right skewed shape. Any distribution where you can draw a line down the middle and the right side mirrors the left side is considered symmetric.

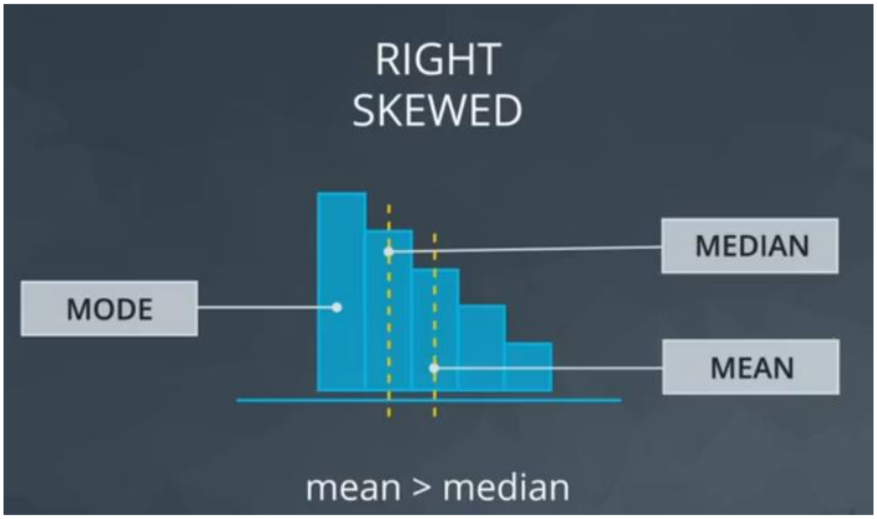


One of the most common symmetric distributions, is known as a normal distribution and it's also called the Bell Curve. The shape of the distribution can actually tell us a lot about the measures of center and spread. Symmetric distributions, like this one have a mean that's equal to the median, which also equals the mode. Each of these measures sits here in the center. The mode is essentially the tallest bar in our histogram.

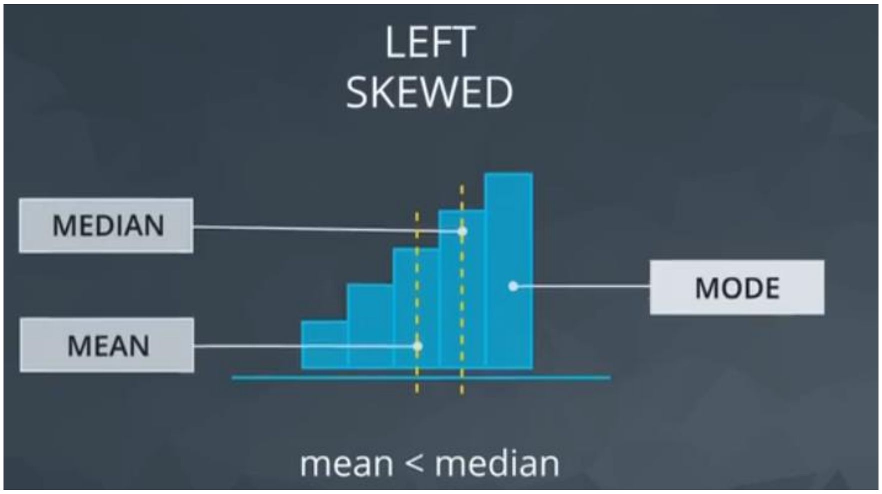


When we have skewed distributions, it's the case of the mean is pulled by the tail of the distribution, while the median stays closer to the mode.

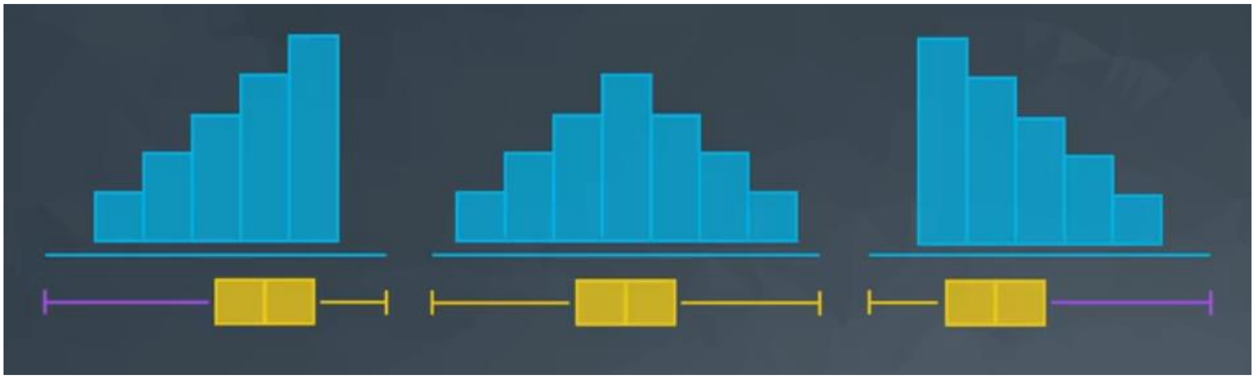
For example, in this right skewed distribution, the mean would be pulled higher, resulting in a mean that's greater than our median.



Alternatively, in a left skewed distribution, our mean is pulled down here, resulting in a mean that's less than our median.



In order to relate this to the visual of a histogram, back to the five number summary we saw on the earlier lessons, here are the corresponding box plots below each histogram.



Notice how the whiskers stretch in the direction of the skew, for each of the skewed distributions.

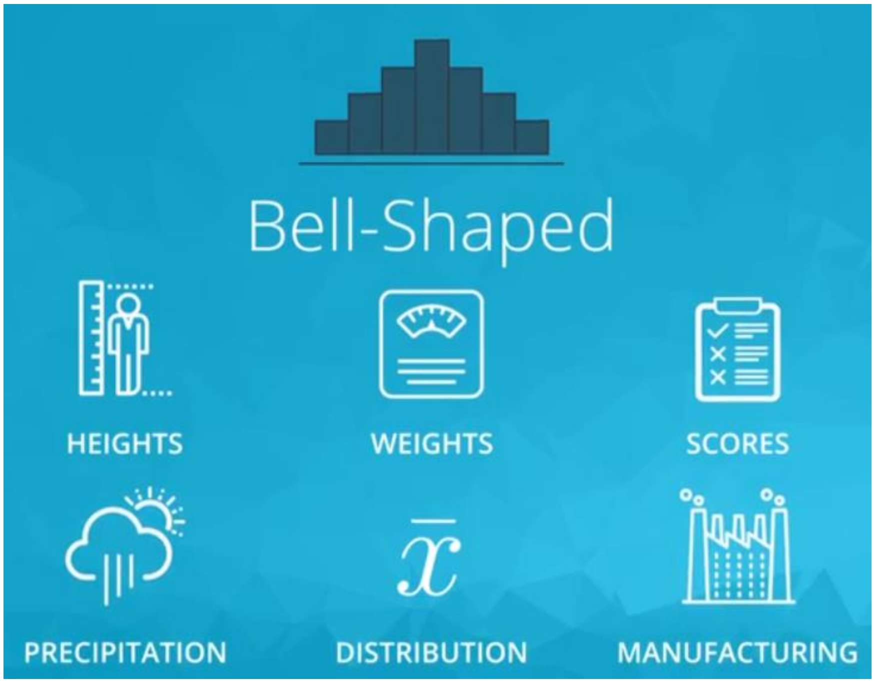
That is the longer whisker is on the left, for a left skewed distribution and it's on the right, for the right skewed distribution.

Alternatively, the symmetric histogram also has a symmetric box plot.

Data in the Real World

If you're working with data, you can always build a Quick Plot to see the shape.

Just to apply some context, some examples of approximately Bell-Shaped data include heights and weights, standardized test scores, precipitation amounts, the mean of a distribution, or errors in manufacturing processes.



Common data that follow Left Skewed Distributions include GPAs, the age of death, and asset price changes.

Left Skewed
A+
GPA
AGE OF DEATH
$
PRICE

Common data that follow approximately Right Skewed Distributions include the amount of drug left in your bloodstream over time, the distribution of wealth, and human athletic abilities.

Right Skewed
BLOOD
$
DISTRIBUTION
OF WEALTH
ATHLETIC
ABILITIES

There are links below in the instructor notes in case you want to learn more about each of these cases.

Though these three, Right Skewed, Left Skewed and Symmetric, are the most common distributions, data in the real world can be messy and it might not follow any of these distributions.

We will talk about this more in the next section.

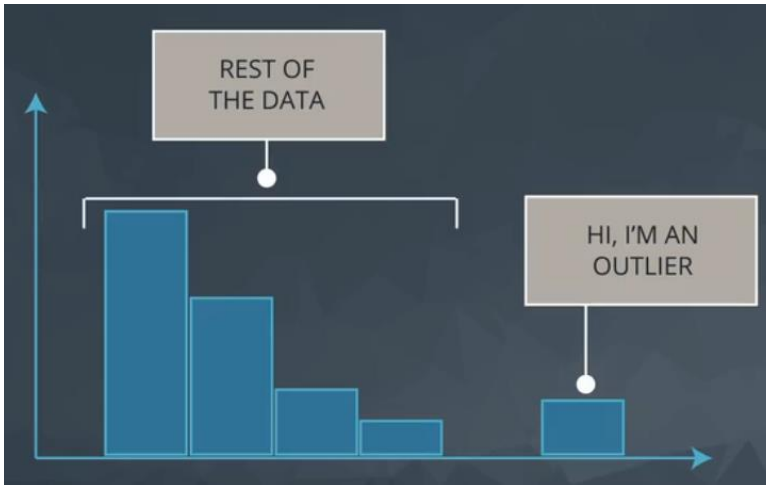
Outliers

We want to look at the final aspect used to describe quantitative variables, Outliers.

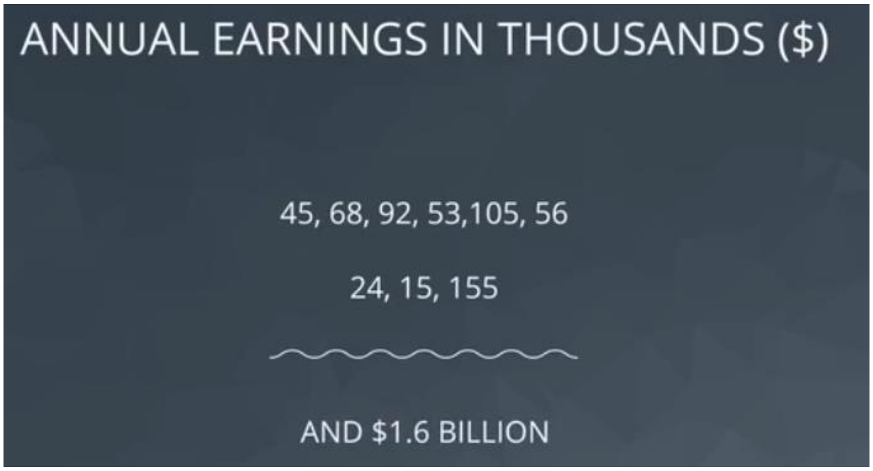
Outliers are data points that fall very far from the rest of the values in our data set.

In order to determine what is very far, there are a number of different methods.

My usual method for detecting outliers isn't very scientific. Usually, I just look at a histogram and see if the point is really far from any of the other data points. Again, a quick plot of your data can often help you understand a lot in a short amount of time.



In order to illustrate the impact that outliers can have on the way we report summary statistics, let's consider the salaries of entrepreneurs. Imagine I select ten entrepreneur earnings and I pull these nine values here as earnings in thousands of dollars, and the tenth is the CEO of Facebook.



ANNUAL EARNINGS IN THOUSANDS ($)

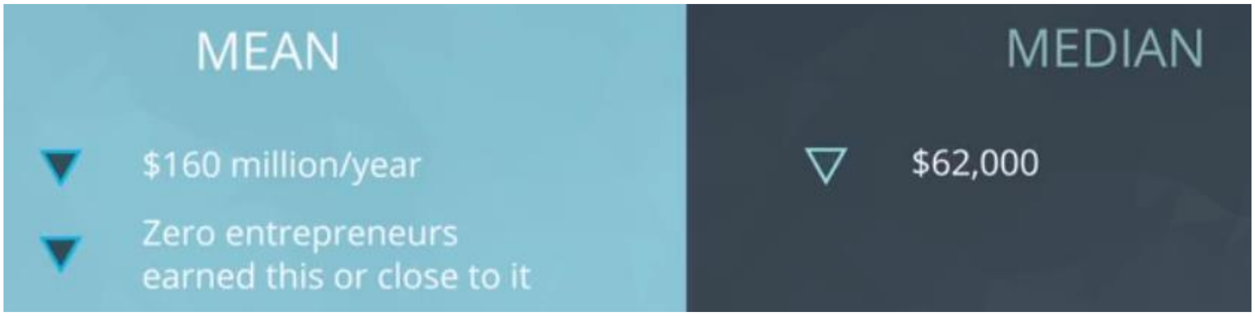
45, 68, 92, 53,105, 56

24, 15, 155

AND $1.6 BILLION

According to a post by CNN in 2016, he earned 4.4 million dollars per day.

Here, we can calculate the mean of these salaries for entrepreneurs based on this data to be approximately 160 million dollars.



MEAN

▼ $160 million/year

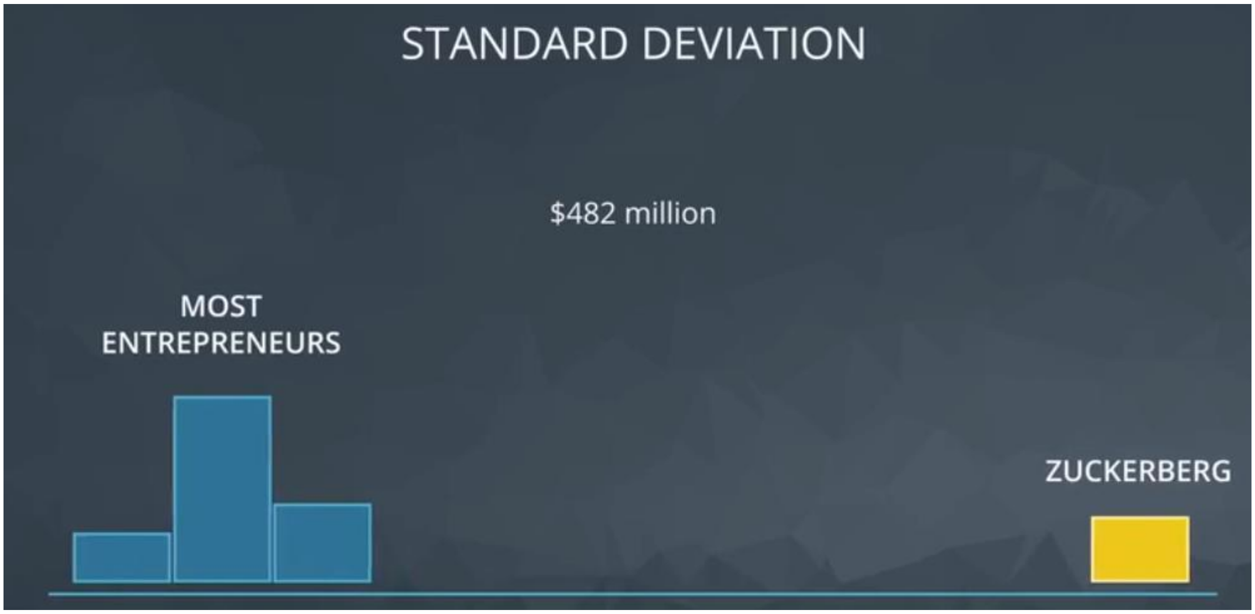
▼ Zero entrepreneurs earned this or close to it

MEDIAN

▽ $62,000

This is incredibly misleading. Literally zero of the entrepreneurs earned this salary.

None of the ten salaries are even close to this amount. A better measure of center would certainly be the median. The median here is calculated at 62,000 dollars a year and is a better indication of what an entrepreneur is likely to earn based on our data.



STANDARD DEVIATION

Our standard deviation is also not a great measure in this case. At approximately 482 million dollars, all this suggests is that our earnings for entrepreneurs are really spread out, but that really isn't fair either. Just one point is really far from the rest.

Like really really far.

Working With Outliers

Common Techniques

When outliers are present we should consider the following points.

1. Noting they exist and the impact on summary statistics.

2. If typo - remove or fix

3. Understanding why they exist, and the impact on questions we are trying to answer about our data.

In cases like the example above, we might try to understand, what was so different about the outlier when compared to the other individuals? How did this entrepreneur become so successful? And why are the earnings so large in comparison? There is an entire field aimed at this idea called the **anomaly detection**.

4. <u>Reporting the 5 number summary values is often a better indication than measures like the mean and standard deviation when we have outliers.</u>

5. Be careful in reporting. Know how to ask the right questions.

This example shows that you need to be careful about how we share our results and state our conclusions using summary statistics when we have outliers.

A single number can be very misleading about what is actually happening in our data. Some statistics are more misleading than others.

If you are the consumer of information based on data, which we all are, it's important to know how to ask the right questions regarding the statistics around you.

Outliers Advice

If you're the one doing the reporting, here are some of my personal guidelines when analyzing data.

1. First, plot your data.
2. Second, if you have outliers, determine how you should handle them.
    - This might require a domain expert of the field.
    - Should you remove them? Should you fix them? Should you keep them?
3. Third, if you're working with data that are normally distributed, the bell shape that we saw before, you can find out every little detail about the data with only the mean and the standard deviation. This may seem surprising but it's true.
4. However, if our data are skewed, the five-number summary provides much more information for these data sets than the mean and the standard deviation can provide.

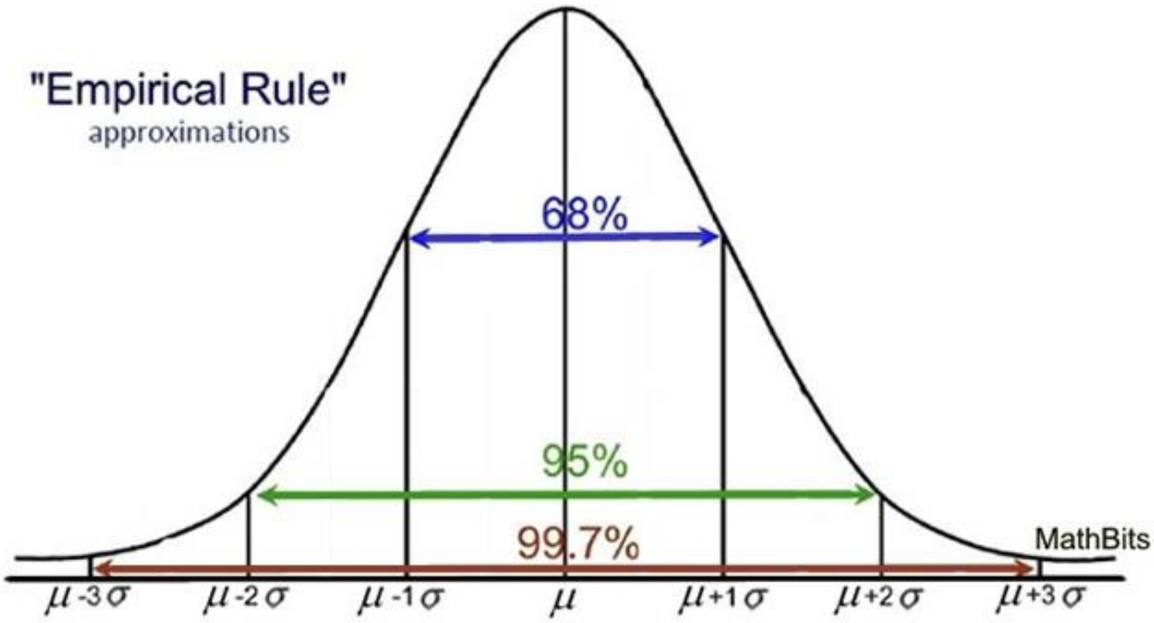
Again, the most useful and informative summary you can get is frequently a visual.

In upcoming lessons, we will focus specifically on the visuals that will best convey our message.

More On Center And Spread

When analyzing skewed data, it is common to report numeric summaries like the median and 5 number summary, as the mean and standard deviation may be misleading.

However, with symmetric data, the mean and standard deviation are commonly used, as we can understand what proportion of points might fall 1, 2, or 3 standard deviations away based on the empirical rule associated with normal distributions.



- 68% of the distribution lies within **one** standard deviation of the mean.
- 95% of the distribution lies within **two** standard deviations of the mean.
- 99.7% of the distribution lies within **three** standard deviations of the mean.

You can read more about this **here(opens in a new tab)**.

Descriptive Statistics Summary

Recap

Variable Types

We have covered a lot up to this point! We started with identifying data types as either categorical or quantitative. We then learned, we could identify quantitative variables as either continuous or discrete. We also found we could identify categorical variables as either ordinal or nominal.

Categorical Variables

When analyzing categorical variables, we commonly just look at the count or percent of a group that falls into each **level** of a category. For example, if we had two **levels** of a dog category: lab and not lab. We might say, 32% of the dogs were lab (percent), or we might say 32 of the 100 dogs I saw were labs (count).

However, the 4 aspects associated with describing quantitative variables are not used to describe categorical variables.

Quantitative Variables

Then we learned there are four main aspects used to describe quantitative variables:

1. Measures of **Center**
2. Measures of **Spread**
3. **Shape** of the Distribution
4. **Outliers**

Measures of Center

We looked at calculating measures of Center

1. **Means**
2. **Medians**
3. **Modes**

Measures of Spread

We also looked at calculating measures of Spread

1. **Range**
2. **Interquartile Range**
3. **Standard Deviation**
4. **Variance**

Shape

We learned that the distribution of our data is frequently associated with one of the three **shapes**:

1. Right-skewed

2. Left-skewed

3. Symmetric (frequently normally distributed)

Depending on the shape associated with our dataset, certain measures of center or spread may be better for summarizing our dataset.

When we have data that follows a **normal** distribution, we can completely understand our dataset using the mean and standard deviation.

However, if our dataset is **skewed**, the 5 number summary (and measures of center associated with it) might be better to summarize our dataset.

Outliers

We learned that outliers have a larger influence on measures like the mean than on measures like the median. We learned that we should work with outliers on a situation by situation basis. Common techniques include:

1. At least note they exist and the impact on summary statistics.

2. If typo - remove or fix

3. Understand why they exist, and the impact on questions we are trying to answer about our data.

4. Reporting the 5 number summary values is often a better indication than measures like the mean and standard deviation when we have outliers.

5. Be careful in reporting. Know how to ask the right questions.

Histograms and Box Plots

We also looked at histograms and box plots to visualize our quantitative data. Identifying outliers and the shape associated with the distribution of our data are easier when using a visual as opposed to using summary statistics.

What Next?

Up to this point, we have only looked at **Descriptive Statistics**, because we are describing our collected data. In the final sections of this lesson, we will be looking at the difference between **Descriptive Statistics** and **Inferential Statistics**.

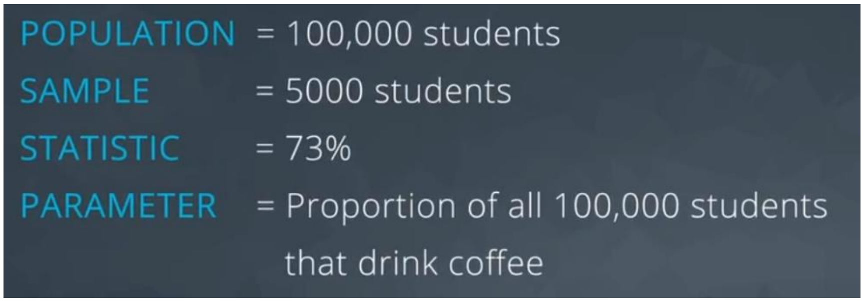
Descriptive vs. Inferential Statistics

The topics covered this far have all been aimed at descriptive statistics.

That is, describing the data we've collected. There's an entire other field of statistics known as **inferential statistics that's aimed at drawing conclusions about a population of individuals based only on a sample of individuals from that population.**

Imagine I want to understand what proportion of all Udacity students drink coffee.

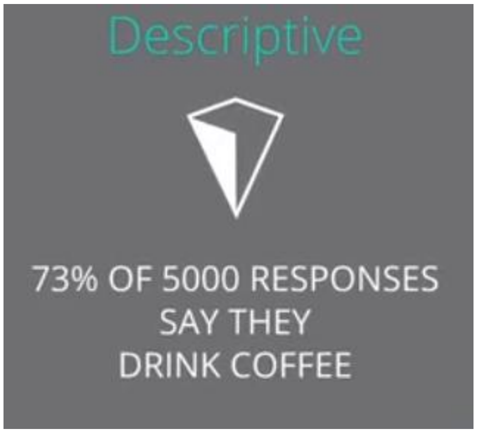
We know you're busy, and in order to get projects in on time, we assume you almost drink a ton of coffee. I send out an email to all Udacity alumni and current students asking the question, do you drink coffee? For purposes of this exercise, let's say the list contained 100,000 emails.



POPULATION	= 100,000 students
SAMPLE	= 5000 students
STATISTIC	= 73%
PARAMETER	= Proportion of all 100,000 students that drink coffee

Unfortunately, not everyone responds to my email blast. Some of the emails don't even go through. Therefore, I only receive 5,000 responses. I find that 73% of the individuals that responded to my email blast, say they do drink coffee.

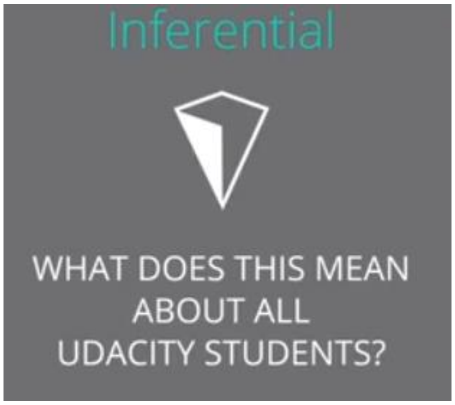
Descriptive statistics is about describing the data we have. That is, any information we have and share regarding the 5,000 responses is descriptive.



Descriptive

73% OF 5000 RESPONSES
SAY THEY
DRINK COFFEE

Inferential statistics is about drawing conclusions regarding the coffee drinking habits of all Udacity students, only using the data from the 5,000 responses.



Inferential

WHAT DOES THIS MEAN ABOUT ALL UDACITY STUDENTS?

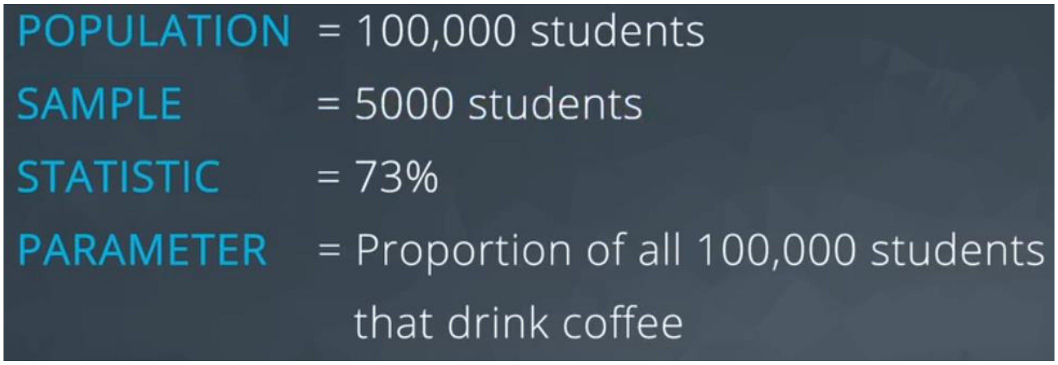
Therefore, inferential statistics in our example is all about drawing conclusions regarding all 100,000 Udacity students using only the 5,000 responses from our sample. The general language associated with this scenario is as shown here.

We have a population which is our entire group of interest. In our case, the 100,000 students. We collect a subset from this population which we call a sample.

In our case, the 5,000 students. Any numeric summary calculated from the sample is called a statistic. In our case, the 73% of the 5,000 that drink coffee. This 73% is the statistic. A numeric summary of the population is known as a parameter.

In our case, we don't know this value as it's a number that requires information from all Udacity students.

Drawing conclusions regarding a parameter based on our statistics is known as inference.



POPULATION	= 100,000 students
SAMPLE	= 5000 students
STATISTIC	= 73%
PARAMETER	= Proportion of all 100,000 students that drink coffee