Data Analysis and Statistics Presentation

RESET Lesson Plan 2022

How Statistics can be Misleading



Link to Video: https://ed.ted.com/lessons/how-statistics-can-be-misleading-mark-liddell#watch

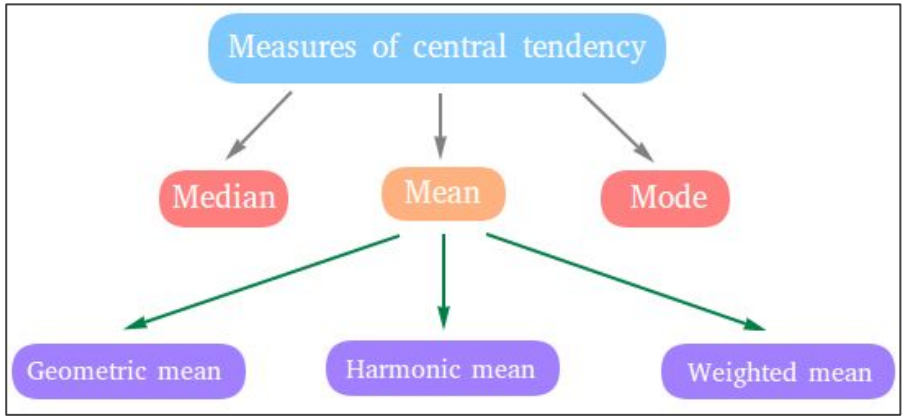
Video Reflection and Discussion
(Multiple Prompt Options)

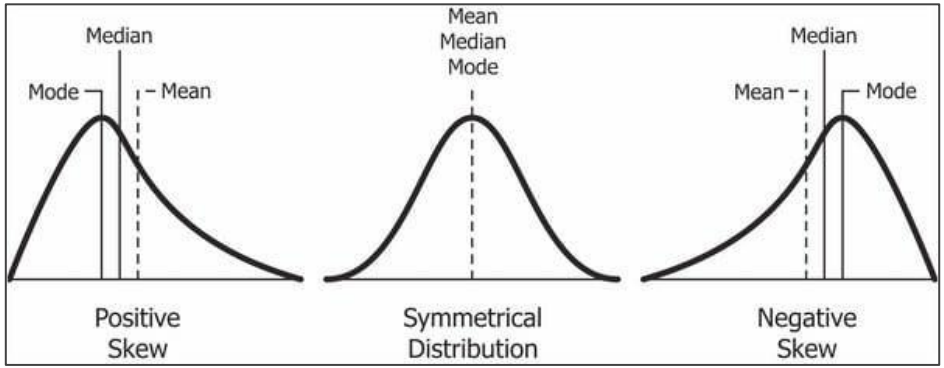
Reflect on the following prompts and discuss:

- Which types of bias should we be aware of when analyzing statistics?
- What types of motivation could be at play when people or organizations present statistics?
- Are there any ethical responsibilities that we face when working with statistics? If yes, what might they be?
- What type of variable should we be aware of when interpreting statistics?
- Why do we bother with statistics? Which tools do we need to order to analyze statistics?

Measures of Central Tendency

The measures of central tendency provide us with statistical information about a set of data. The four primary measurements that we use are the **mean, median, mode and range**. Each one of these measurements can provide us with information about our data set. This information can then be used to define how the set of data points are connected.





Mean

The first measure is the mean, which means **average**. To calculate the mean, add together all of the numbers in your data set. Then divide that sum by the number of values in the data set.

For example, let's look at the following data set: 16, 14, 12, and 18

To find the average (or the mean), we would first add all four of these values together:

16 + 14 + 12 + 18 = 60

Since there were four numbers in our data set, you would divide that sum by 4:

60 ÷ 4 = 15

So, the average number 15. The mean is used to show us the true average of a set of data.

Median - Part 1

Another measure of central tendency is the median, which is the **middle number** when listed in order from least to greatest. You may have heard the word median before, and it was likely on a highway. On a highway, you have opposite lanes of traffic. In the middle of the lanes, there is typically a grassy area or a turning area. This area in the middle of the highway is referred to as a median.

Let's find the median of this data set: 10, 6, 19, 21, and 4

The first thing you need to do with this data set is put the numbers in order from least to greatest:

4, 6, 10, 19, 21

Now find the middle number. Since there are five numbers, the middle number would be the third value.

4, 6, 10, 19, 21 Median: 10

Median - Part 2

There may be an even number of values in the data set, which would provide you with two numbers in the middle. If this occurs, you will need to average the two values.

Data Set: 3, 8, 9, 12, 12, 15

Since there are two numbers in the middle, you will average them together:

9 + 12 = 21

Then, divide by 2:

21 ÷ 2 = 10.5

The median of this set of data is 10.5.

When looking at a data set, the median is a better measure instead of the mean when there is an outlier in the data.

Mode

The mode is another measure of central tendency that tells us the **number that occurred the most often** in your data set. When looking for the mode, there can be more than one mode or no mode. The mode can tell us the most popular choice.

Data Set: 5, 6, 6, 3, and 4.

The mode of this data set would be 6.

Concept Check 1

During the first four Summer Olympic games attended by the United States, these medal counts were awarded. Calculate the **mean** number of medals won during these Olympic Games. (Round it to the nearest whole number.)

A. 47
B. 353
C. 88
D. 28.5

Olympic Site	Year	Medal Count
Athens	1896	20
Paris	1900	47
St. Louis	1904	239
London	1908	47

Concept Check 2

During the first four Summer Olympic games attended by the United States, these medal counts were awarded. Calculate the **median** number of medals won during these Olympic Games. (Round it to the nearest whole number.)

A. 219
B. 239
C. 88.25
D. 47

Olympic Site	Year	Medal Count
Athens	1896	20
Paris	1900	47
St. Louis	1904	239
London	1908	47

Concept Check 3

During the first four Summer Olympic games attended by the United States, these medal counts were awarded. Calculate the **mode** number of medals won during these Olympic Games. (Round it to the nearest whole number.)

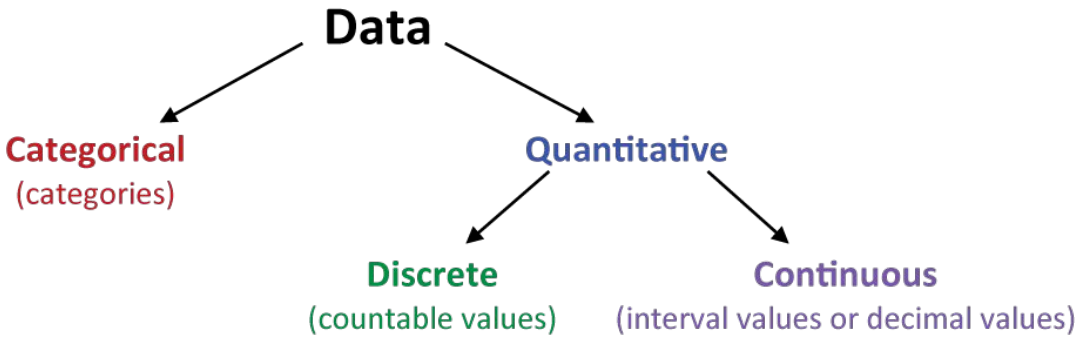
A. 47
B. 239
C. 88.25
D. No mode

Olympic Site	Year	Medal Count
Athens	1896	20
Paris	1900	47
St. Louis	1904	239
London	1908	47

Defining Data

Example Scenario: Andy is working on a class project about genetics. He needs to collect data about each of his classmates' genes and analyze this data for his class project. Where should Andy start collecting his data?

First, he will need to understand data, which is information that is collected for analysis. Data is generally used to either prove or disprove a hypothesis. Data can be grouped into two different types of information: **categorical and quantitative**.



Categorical vs. Quantitative Data

Sometimes, it can be difficult to understand the differences between categorical and quantitative data. **Categorical data** is grouped into some sort of category or multiple categories. **Quantitative data** is related to quantities. Things like height, weight, GPA, number of hours spent studying, and other types of information that are quantitative and are collected just as numbers.

A good way to remember the difference between categorical and quantitative data is to examine the answer to the question. If the answer is a **number**, then the data is **quantitative**. If the answer is a **preference**, a **characteristic**, or anything other than a number, then the data is categorical.

For example, if you ask someone, “How tall are you?” they would answer with a number: 5 foot 6 inches. If someone asked, “What's your favorite sport?” then they would answer with a preference: Basketball. Therefore, the first question gives us quantitative data and the second gives us categorical data.

Concept Check 1

What is data that is collected in groups or topics, where the number of events in each group is counted numerically?

A. Cluster sampling
B. Categorical data
C. Quantitative data
D. Numerical data
E. Stratified sampling

Concept Check 2

One trick to finding out if information is categorical or quantitative is to analyze the answer to the question. Which of the following is TRUE:

A. If the answer to the question is numerical, then the information is quantitative. If the answer to the question is a characteristic, preference, etc., then the information is categorical.
B. If the answer to the question is an interest, then the information is quantitative. If the answer to the question is a characteristic, preference, etc., then the information is categorical.
C. If the answer to the question is numerical, then the information is categorical. If the answer to the question is a characteristic, preference, etc., then the information is quantitative.
D. If the answer to the question is a characteristic, then the information is quantitative. If the answer to the question is a number, preference, etc., then the information is categorical.

Purpose of Statistics

The purpose of statistics is to describe and predict information. This can be divided into descriptive statistics and inferential statistics. Sometimes we collect data in an attempt to describe the characteristics of a population.

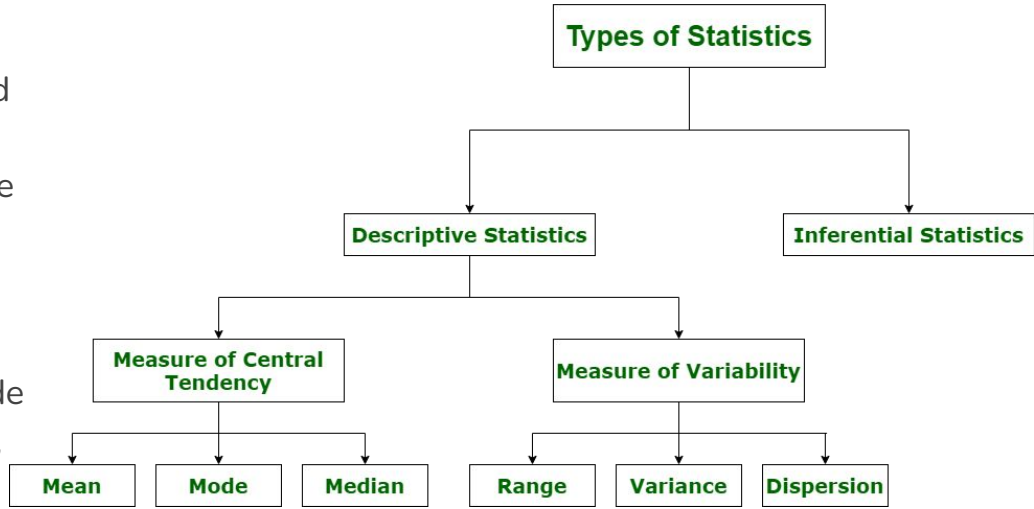
Statistics is also used to predict information. A statistical model is a combination of inferences based on collected data and population understanding used to predict information in an idealized form. This means that a statistical model can be an equation or a visual representation of information based on research that's already been collected over time.



Types of Statistical Models

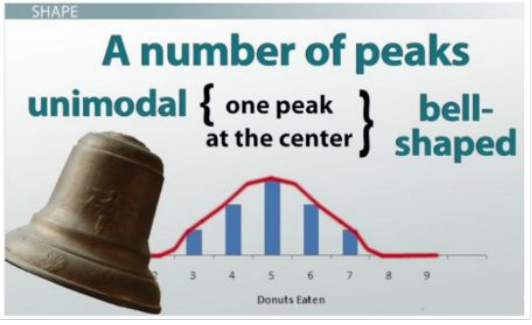
Statistical models exist because we are looking for a relationship between two, or sometimes more, variables. Essentially, all statistical models exist to find inferences between different types of variables and because there are different types of variables, there are different types of statistical models.

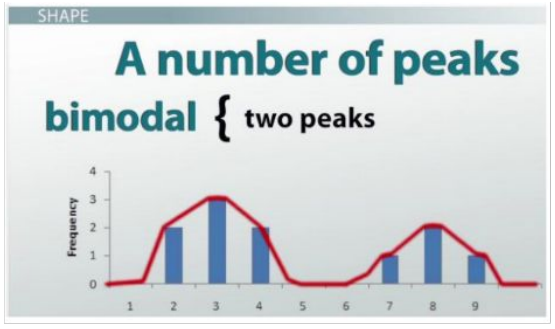
Some of the types of models, or statistical tests, include regression, analysis of variance, analysis of covariance, and chi-square. These are just a few examples of statistical models; there are many different ways we can analyze data depending on the variables.



What are Center, Shape, and Spread?

When we talk about center, shape, or spread, we are talking about the distribution of the data, or how the data is spread across the graph.



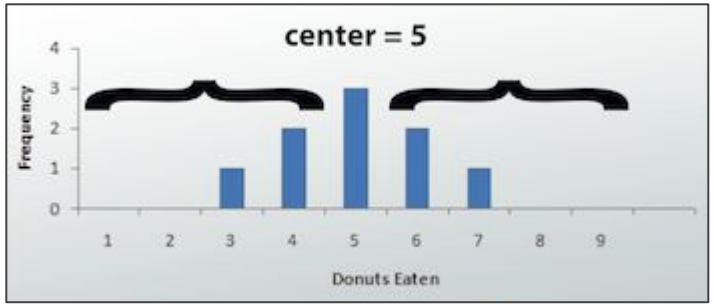






Center

The center of a distribution tells you the center or median of the data. When you look at a graph, it will be the value where approximately half of your data is on one side and the rest of your data is on the other side. The median point of your data set is the middle number if you were to put your data in ascending order. Let's say we are taking surveys of different groups of people and their donut eating habits. For the first group of people, we have this graph:



The center of this graph is 5.

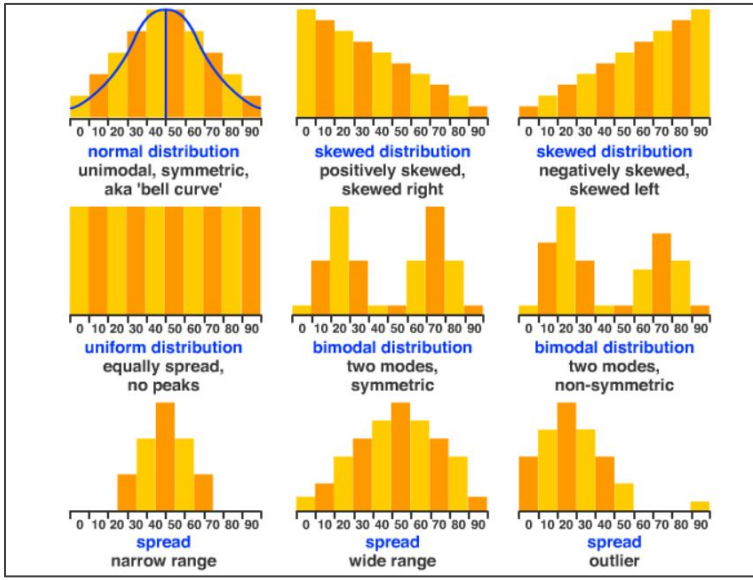
We see that our center is 5 because half of the people are to the left and the other half are to the right.

Another way to describe the center is to take the mean or average of all your data. When you describe your center in terms of mean and median, you might find that they are slightly different. Your mean might be more or less than your median.

Shape

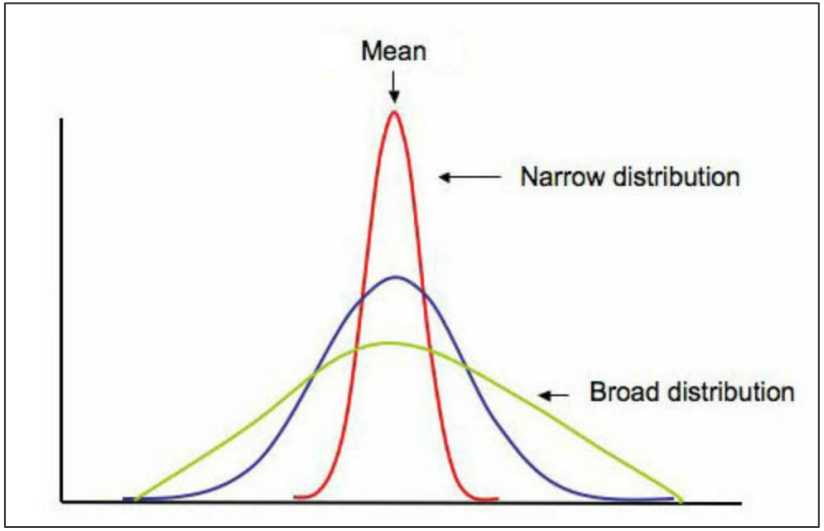
We use shape to describe the different types of graphs we will see. There are four different ways in which we can describe a graph's shape.

- We can say a graph is **symmetric** if the left and right sides of the graph are mirror images of each other.
- Sometimes, our graph will look like a rollercoaster and will have a number of peaks, or areas where the graph is higher than the surrounding areas. If there is only one peak, then we call it **unimodal**. If this one peak occurs at the center of the graph, it is also called **bell-shaped**. If it has two peaks, then we will call it **bimodal**.
- If our graph has more data on one side rather than the other, we call it **skewed**. If there are more to the right, we call it **skewed left**. If our graph has more data to the left, then we would say that our graph is **skewed right**.



Spread

We often want to know how "spread out" the distribution is to give us a measure of the variability of values taken from this distribution. The below graphic shows the general shape of three symmetric unimodal distributions with identical measures of center, but very different amounts of "spread". **The simplest measure of spread in data is the range**. It is the difference between the maximum value and the minimum value within the data set.



Concept Check 1

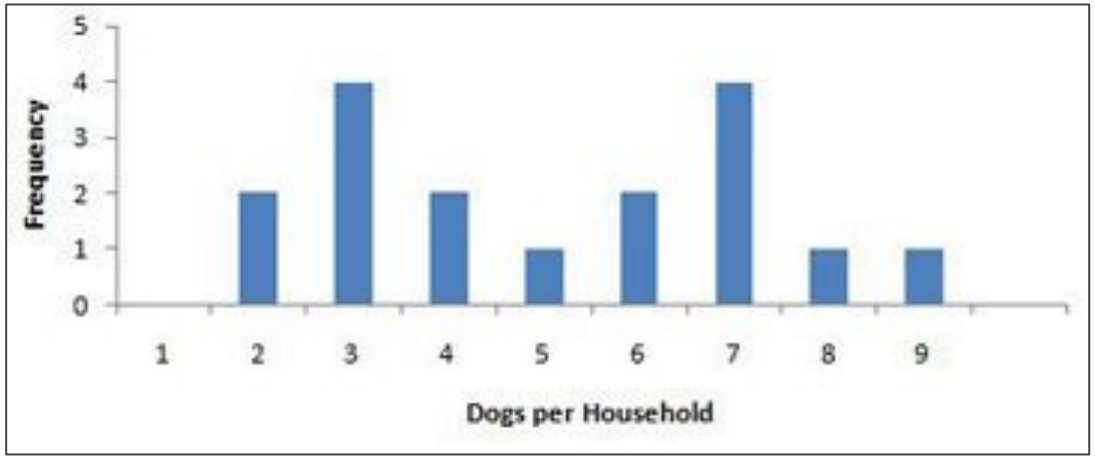
Which of the following is not a way to describe **shape**?

A. Unimodal
B. Bimodal
C. Uniform
D. Skewed
E. Range

Concept Check 2

What is the spread of this graph?

A. 2 to 9
B. 1 to 9
C. 2 to 4
D. 6 to 9
E. Large range

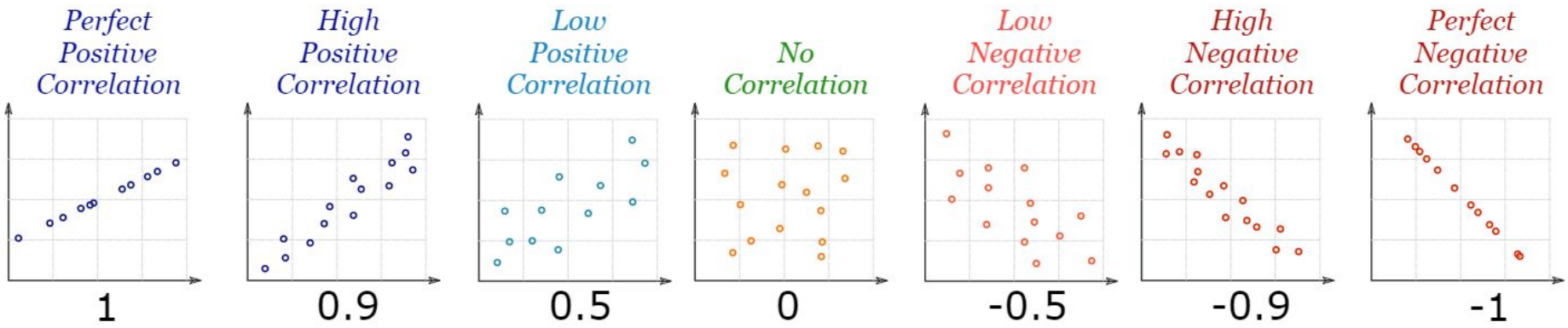


Correlation

When two sets of data are strongly linked together we say they have a High Correlation.

- Correlation is **Positive** when the values **increase** together
- Correlation is **Negative** when one value **decreases** as the other increases

A correlation is assumed to be linear (following a line).



Correlation Coefficient (R)

Correlation Coefficient can have a value:

- 1 is a perfect positive correlation
- 0 is no correlation (the values don't seem linked at all)
- -1 is a perfect negative correlation
- Any values in between -1 and 1

The value shows how good the correlation is (not how steep the line is), and if it is positive or negative.

The correlation calculation only works properly for straight line relationships. You can use the Coefficient of Determination (R^2) for nonlinear models.

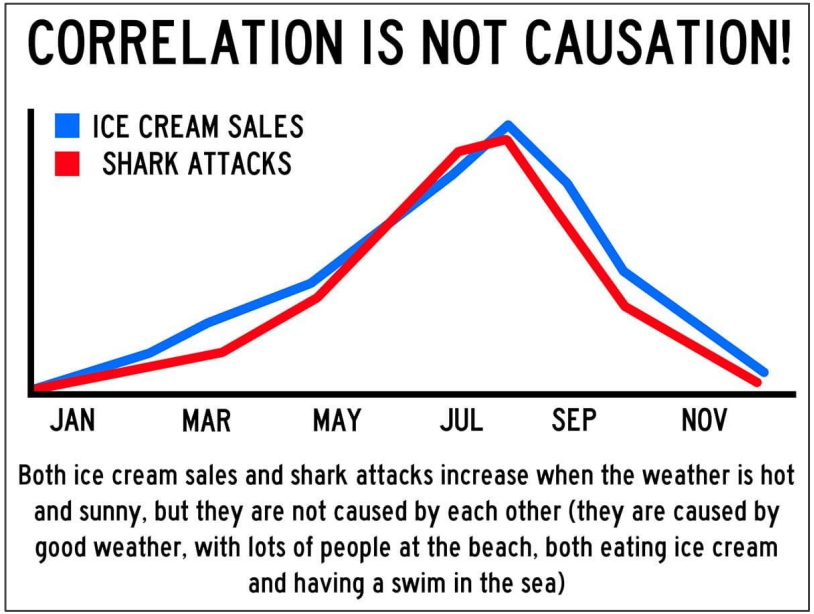
Correlation is NOT Causation

"Correlation is not causation" means that just because two things correlate does not necessarily mean that one causes the other.

What it means is that a correlation **does not prove** one thing causes the other:

- One thing **might** cause the other
- The other **might** cause the first to happen
- They **may be** linked by a different thing
- It could be random chance!

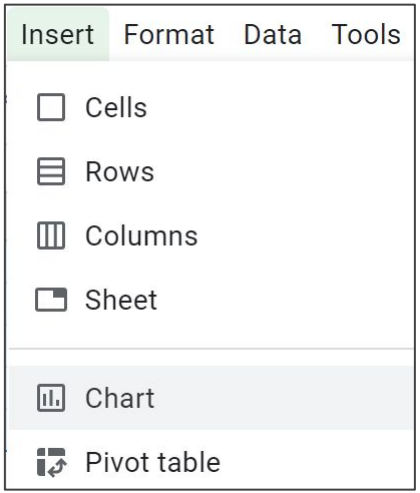
There can be many reasons the data has a good correlation.

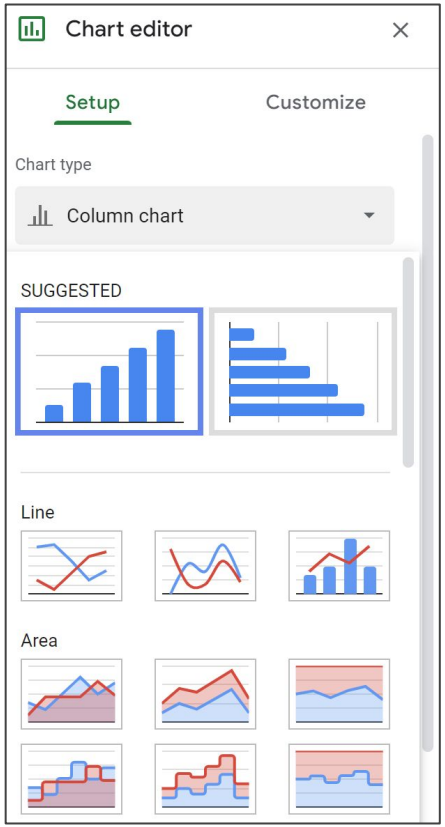


Making a Chart in Google Sheets

- Create a chart of the data
  - Insert > Chart

- Depending on your data, it will automatically create a recommended chart. The change the chart, select the approprite chart in the Chart Editor > Chart Type





Correlation and Trend Line in Google Sheets

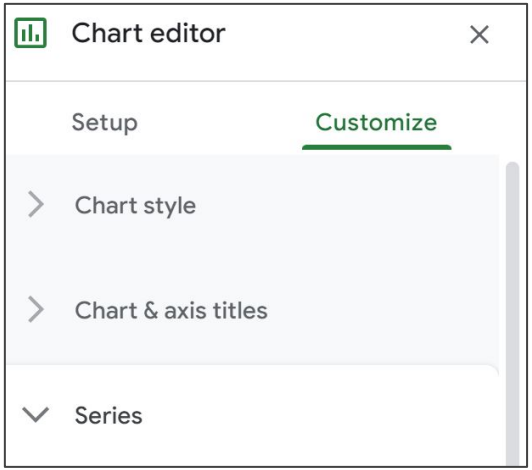
For the Chart Type, select a Line Chart.

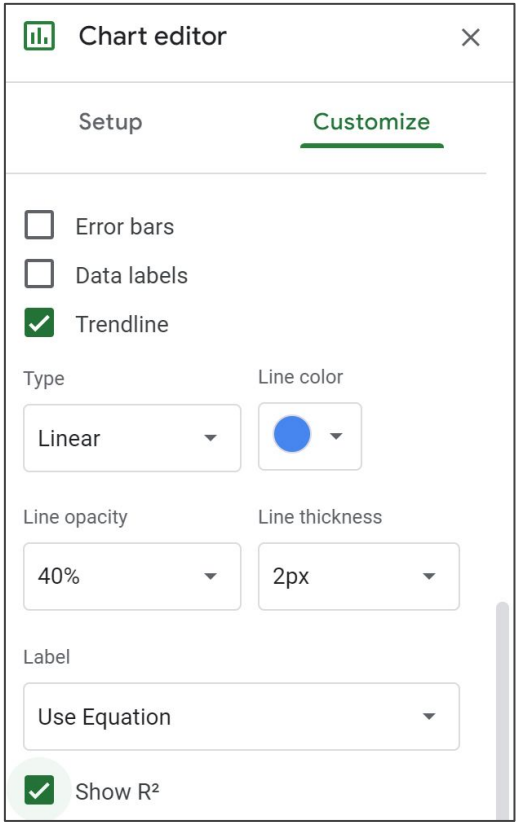
To add a Trendline and R^2:

- Customize > Series > Check the box for Trendline > Under Label select “Use Equation” > Check the box for R^2

You should now see the R^2 and the linear function trendline on your chart. The default trendline function is Linear. The trendline function can be changed in the Type dropdown menu.

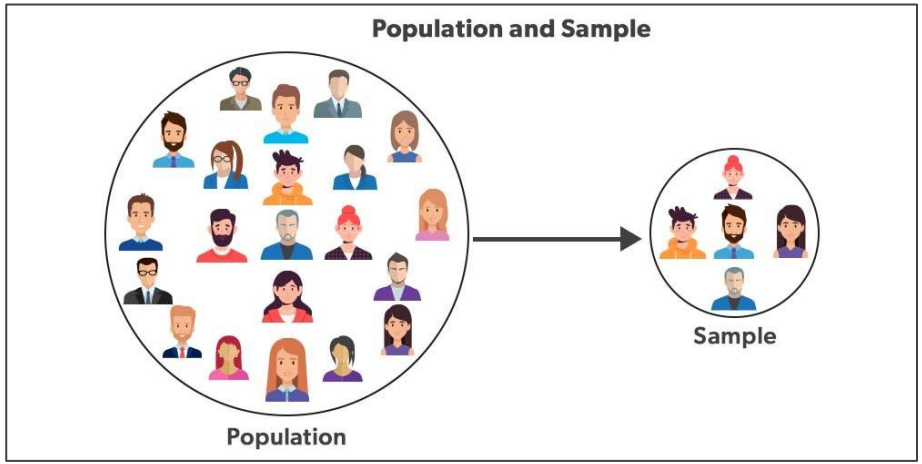
NOTE: You will need to scroll in the Chart Editor to find Trendline. The R^2 option will not appear unless Trendline is checked.

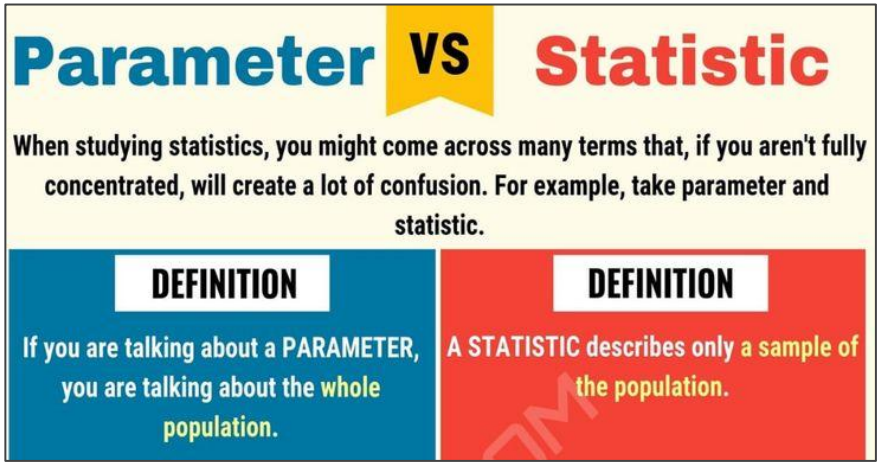




Difference between Parameters & Statistics

A parameter is a measure that describes an entire population while a statistic is a measure that describes a sample from the population. On the following slides, you will learn how to use statistics to understand populations, learn about population and samples, and differentiate parameters from statistics.





Populations

A population is **all members of a specified group**. All members of a specific group would be a population. However, most of the time it isn't practical to get information from every member of a population.

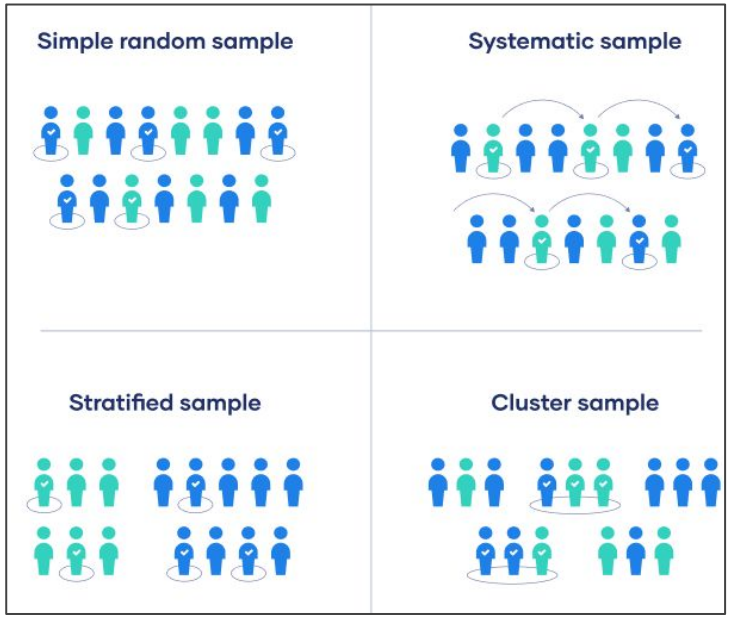
When this happens, we have to find a different way of getting information that represents the population without actually asking the whole population. Researchers may need a sample to gather information about the population.



Sample

A sample is a **part of a population** used to describe the whole group. Samples are used to help the data researcher to understand the entire population. There are many different ways you can get a sample from your population. These include:

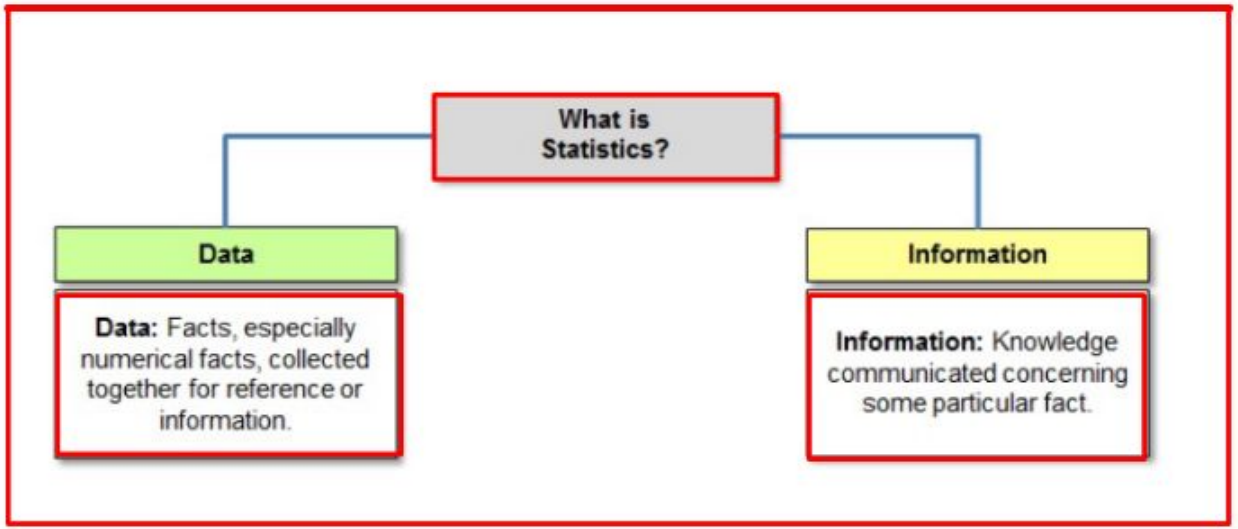
- Random sampling
- Simple random sampling
- Cluster sampling
- Stratified sampling
- Systematic sampling



Parameters & Statistics

A **parameter** is the characteristics used to describe a **population**. It is often not feasible to collect data from the entire population, so we use statistics instead.

A **statistic** is the characteristics of a **statistic** used to infer information about the population.



Concept Check 1

What are the characteristics of a sample used to infer information about the population?

A. Population
B. Sample
C. Statistic
D. Parameter

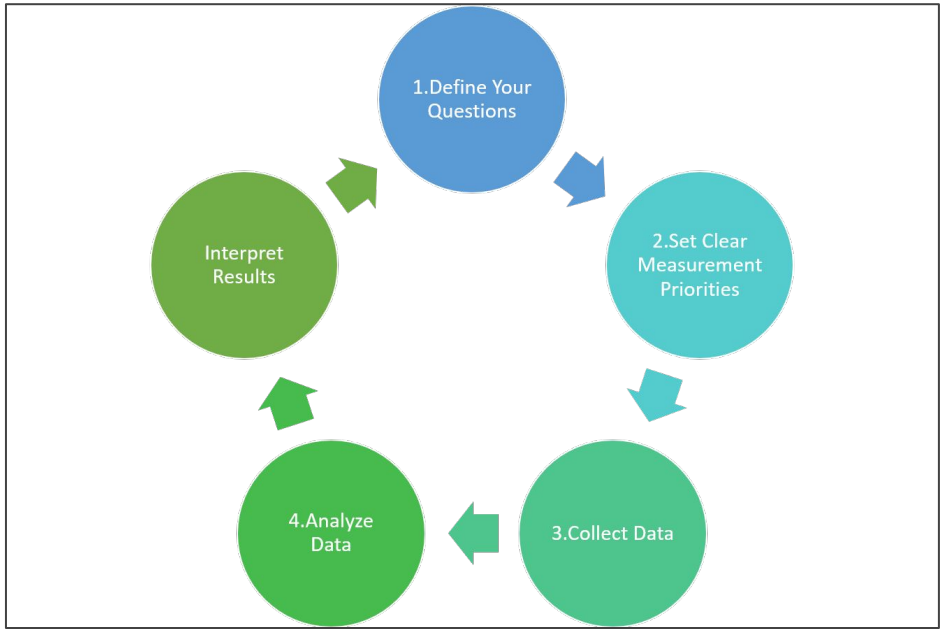
Concept Check 2

What do we call a part of a population used to describe the whole group?

A. Population
B. Sample
C. Statistic
D. Parameter
E. Variance

Data Analysis: Purpose & Steps

When performing data analysis, a process of collection, scrutinization and rationalizing has to occur during scientific statistical research in order for conclusions to be drawn. .



What Is Data Analysis?

Prompt: Have you thought about how many red fruit loops are in a box of Fruit Loop cereal or how many green peanut M&Ms are in pack of peanut M&Ms?

All of your questions can be asked by conducting a data analysis!

What is a data analysis?
Data analysis is a process of applying statistical practices to organize, represent, describe, evaluate, and interpret data.

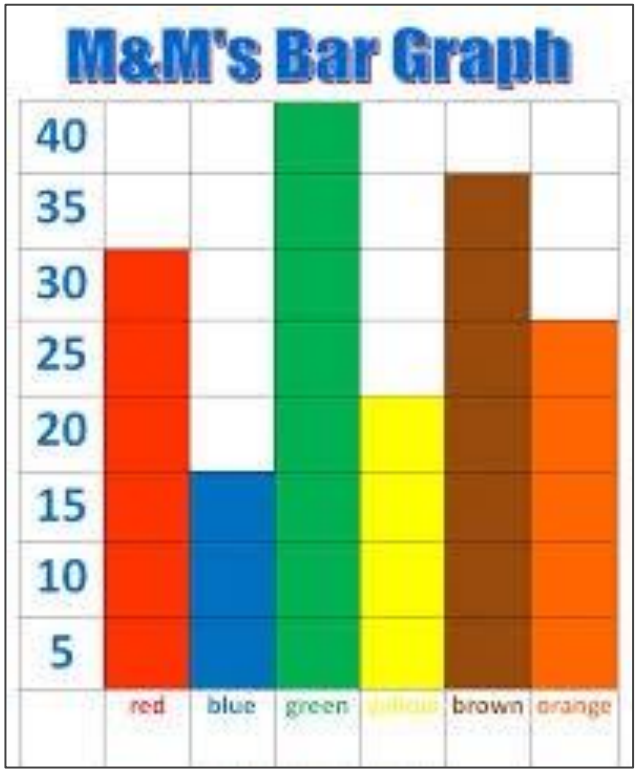
What is data?
Data is a set of qualitative and quantitative variables. For example, the number of red fruit loops and green M&Ms are examples of quantitative data because we are analyzing a number associated with the Fruit Loops and M&Ms.

Organizing and Representing Data

Organizing data is a critical component of conducting a data analysis. The organization of data takes place in the early stages of the data analysis. For example, look at the M&M's bar graph which is one way to visually represent a data set.

In this bar graph, we are able to organize the data in six different categories. The categories are the colors, which are red, blue, green, yellow, brown, and orange. The next step within organizing data is to assign a quantitative value (number) or qualitative value (words) to the data set.

A qualitative study is a method in which the researcher is able to study things within its context. An example would be a case study on the effect extracurricular activities have on high school students. A quantitative study is a method in which the researcher is able to employ mathematical practices.



Extension Activities

The following activities are available on Data Classroom. Users will need to create an account or sign in with Google.

- NBA Salary Data 2019: Mean or Median?
  - Link: https://about.dataclassroom.com/ready-to-teach/2019/10/23/this-activity-is-a-slam-dunk
- Does Thanksgiving dinner get more expensive every year? (line graph vs. linear regression)
  - Link: https://about.dataclassroom.com/ready-to-teach/Thanksgiving

Presentation Sources

- TED Ed Lesson: https://ed.ted.com/lessons/how-statistics-can-be-misleading-mark-liddell#watch
- Lesson Inspiration from Study.com: https://study.com/academy/topic/math-data-analysis-statistics-probability-help-and-review.html#overview
- Correlation Slides: https://www.mathsisfun.com/data/correlation.html