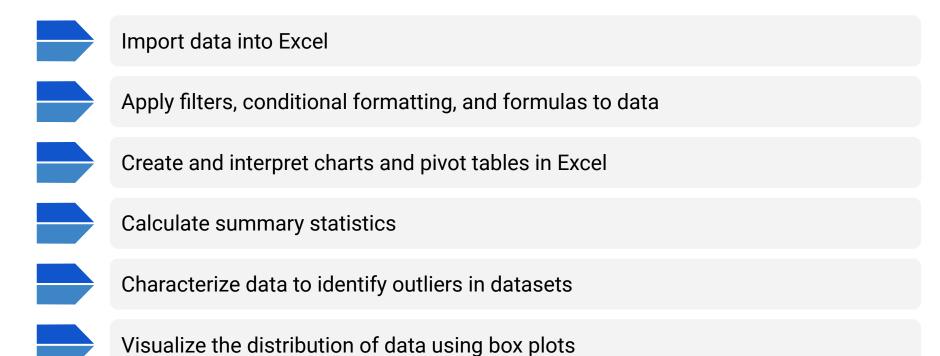
Module 1

# This Week: Excel

#### This Week: Excel

By the end of this week, you'll know how to:





# This Week's Challenge

Using pivot tables and functions to filter data, create charts that demonstrate an analysis of data sets to visualize business outcomes based on launch dates and goals.

Module 1

# Today's Agenda

# Today's Agenda

By completing today's activities, you'll learn the following skills:



**Basic Charting** 



**Summary Statistics** 



**GitHub Repositories** 



Make sure you've downloaded any relevant class files!



Instructor Demonstration
Adding Files to GitHub

## GitHub is a hosting service for source code

GitHub is a web interface for Git.

Git is version control software that can:



Track source code history



Allow for collaboration on the same code files across a team or organisation



Easily update and rollback software versions







Since 2019, GitHub is used by over 2.1 million companies.

Proficiency in Git and GitHub are highly desired skills in many industries

## We will use Git and GitHub throughout the curriculum



You will submit your homework assignments using GitHub



Your individual project work will be version controlled using Git



You will be collaborating with teammates using GitHub



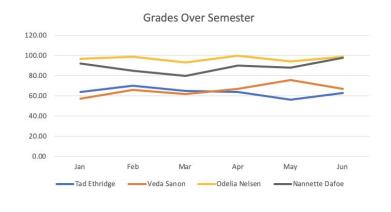
By the end of the curriculum, you should be proficient with the basic Git and GitHub functionality

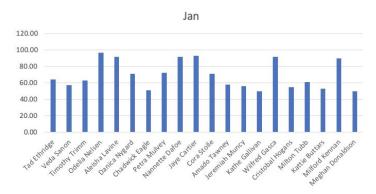


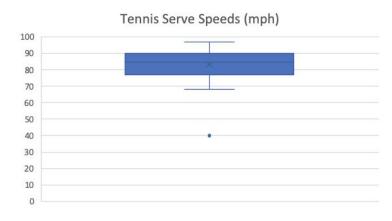
Instructor Demonstration
Basic Charting

#### It is time to learn Excel visualizations!









# We will look at a few examples and use cases

In this activity, we will:



Look at an example data set



Select data of interest



Visualise selected data



Add labels and titles to our visualization



Do not hesitate to ask questions.

Our TAs will slack out images for each operating system





# **Activity**: The Line and Bar Grades



## **Activity:** The Line and Bar Grades

You will take on the role of a teacher for this activity as you create a series of bar and line graphs that visualize the grades of your class over the course of a semester.

#### Instructions: Hint: Create a series of bar graphs that visualize the grades of all When duplicating bar graphs, it pays students in the class, with one graph for every month. to get the formatting and look of the chart where you want it for the first Create a line graph using all of the data that can be used to graph (e.g., for January), and to then compare students' grades across the semester. copy that chart and re-select the data • Use filtering in the line graph to allow you to drill down to a for the subsequent copies (keeping specific student's progress throughout the semester. the style and format, but just changing the data).

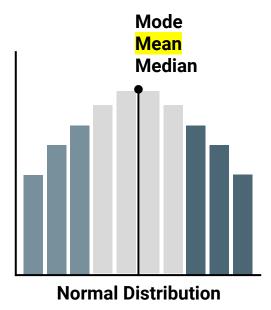


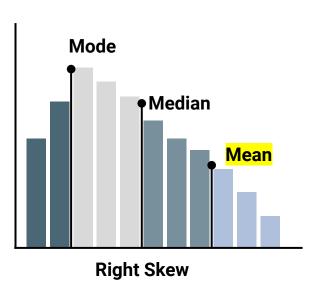
**Let's Review** 

#### Mean

Sum of all values in the sample divided by the number of values in the sample

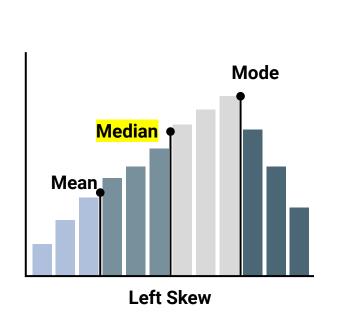


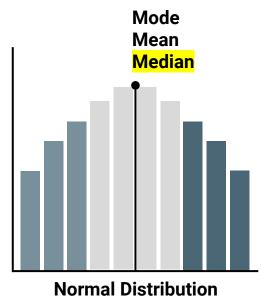


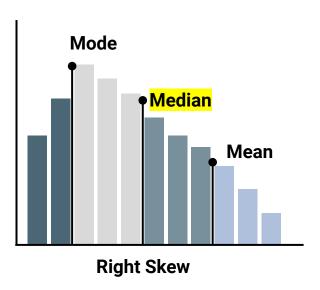


#### Median

The value at the midpoint in a set of observed values



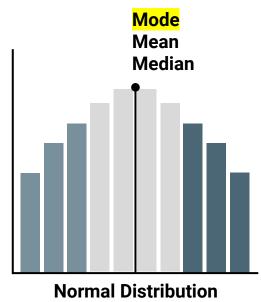


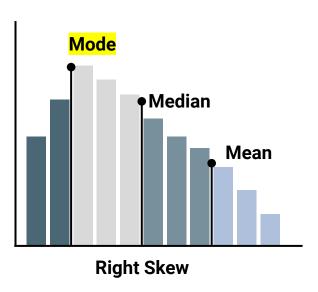


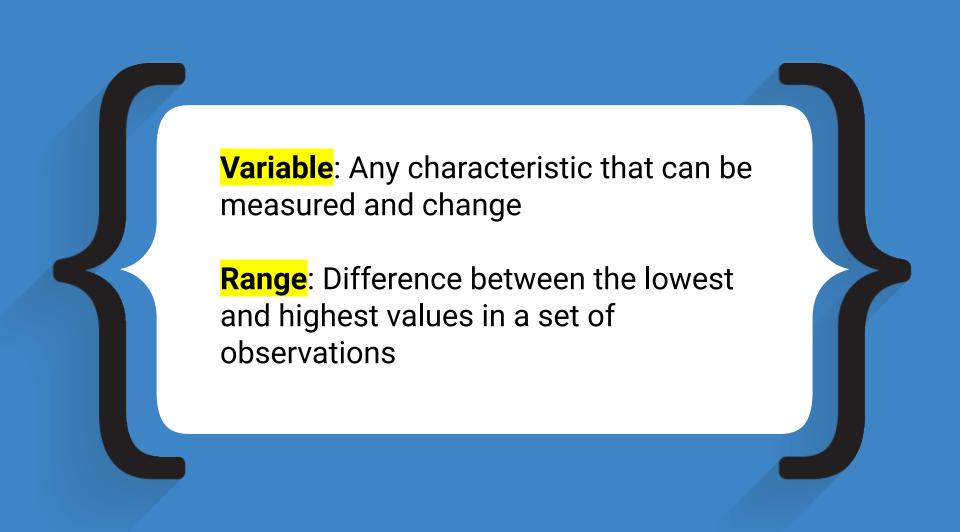
#### Mode

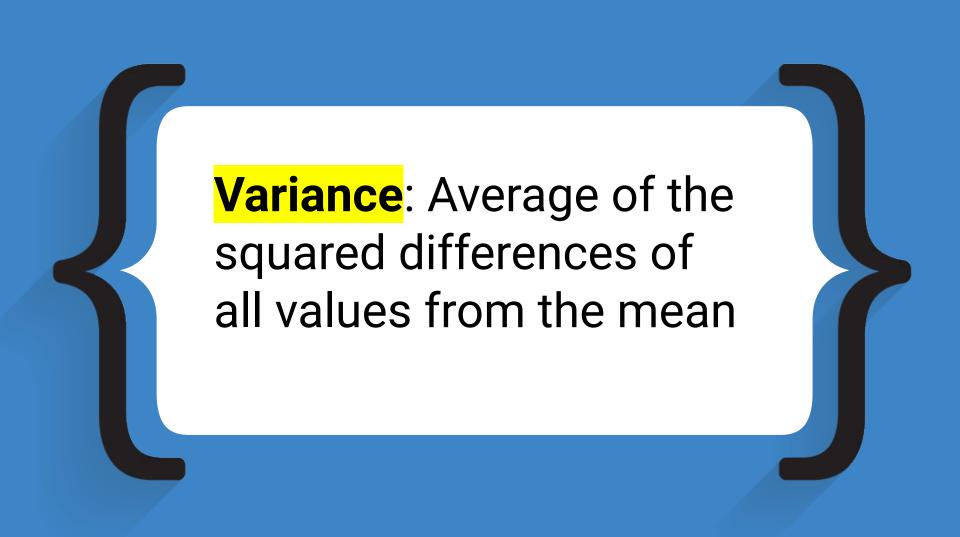
The most frequently occurring value in a set of values











#### Variance



Used to describe how far values in the data set are from the mean



Describes how much variation exists in the data



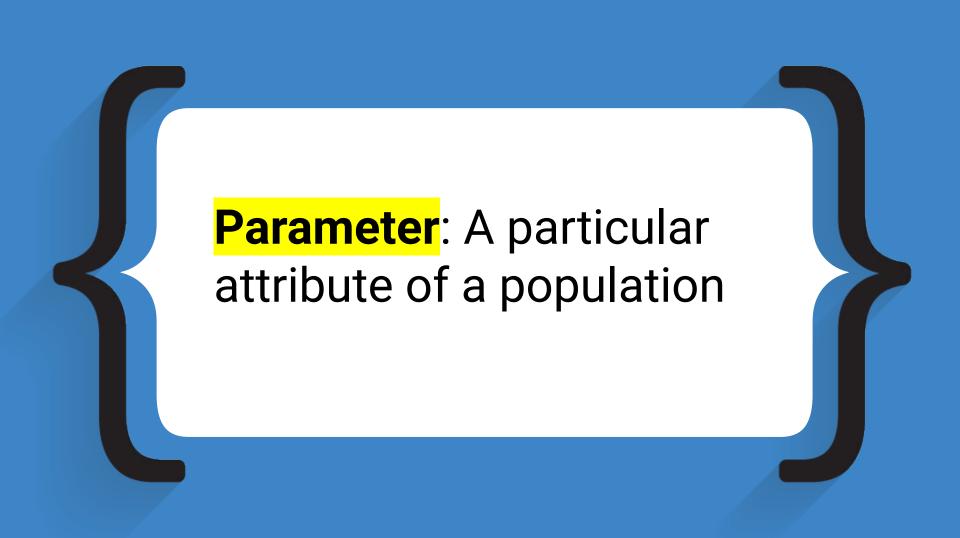
Considers the distance of each value in the data set from the centre of the data

#### The value of the one observation

The mean value of all observations

Sample variance 
$$S^2 = \frac{\sum (oldsymbol{x}_i - oldsymbol{ar{x}})^2}{ ext{The number of observations}} n - 1$$

21



#### **Parameter**

# A particular attribute of a population

Population Parameters	
$\mu$	Population mean
$\sigma$	Population standard deviation
P	Population proportion
N	Population size
X	Population data value
r	Correlation coefficient

## Extreme values may not always be reliable

In data science, extreme values are often suspicious.



Could the measurement be a mistake?



Is the data trustworthy?



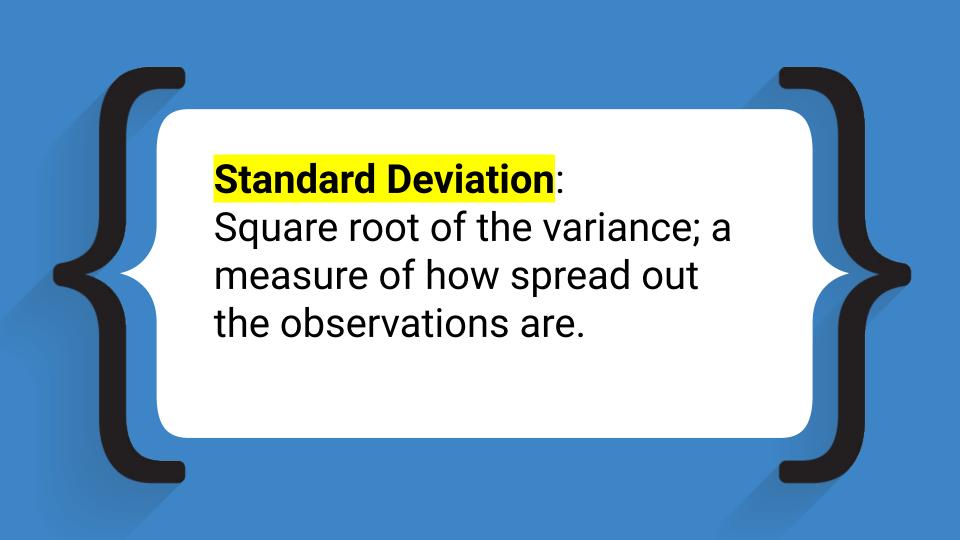
Suspicious values are called **potential outliers**.

An outlier is a data point that differs from the rest of a data set.



Outliers can inaccurately skew a data set.

They can cause us to misrepresent the actual data.



### **Standard Deviation**



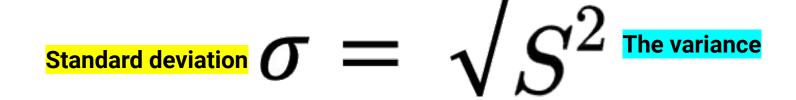
Describes how spread out the data is from the mean



Calculated from the square root of the variance

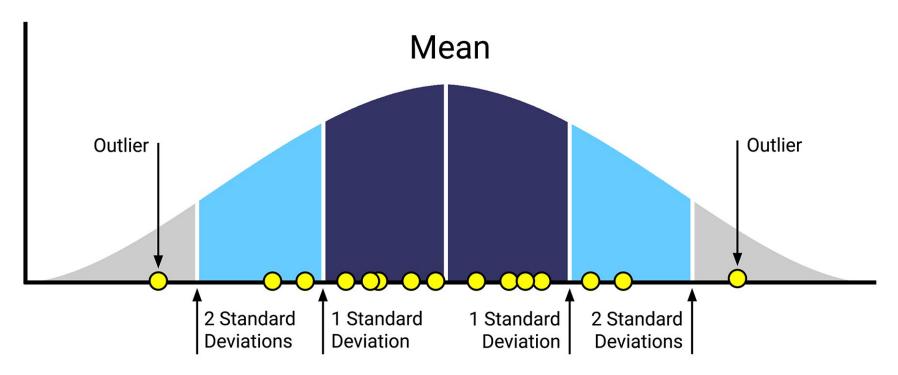


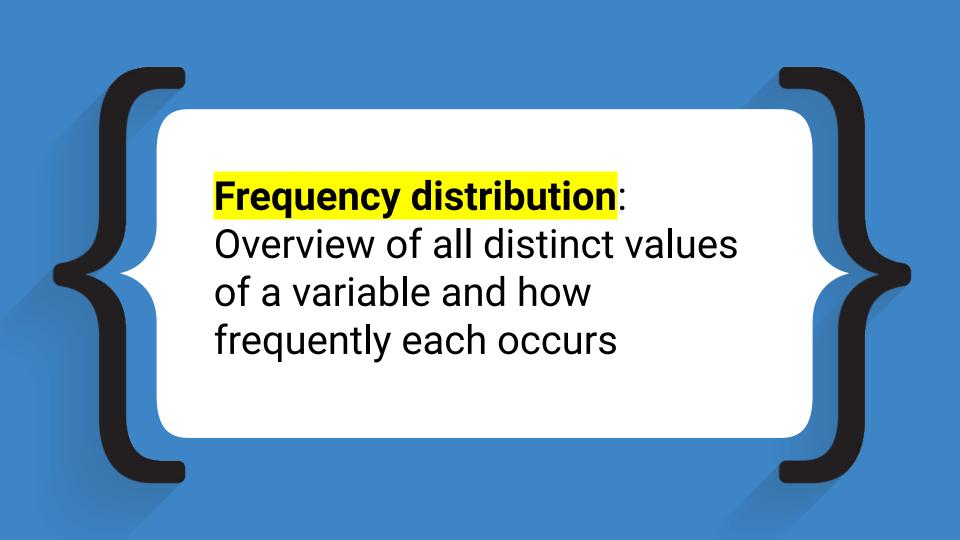
In the same units of measurement as the mean



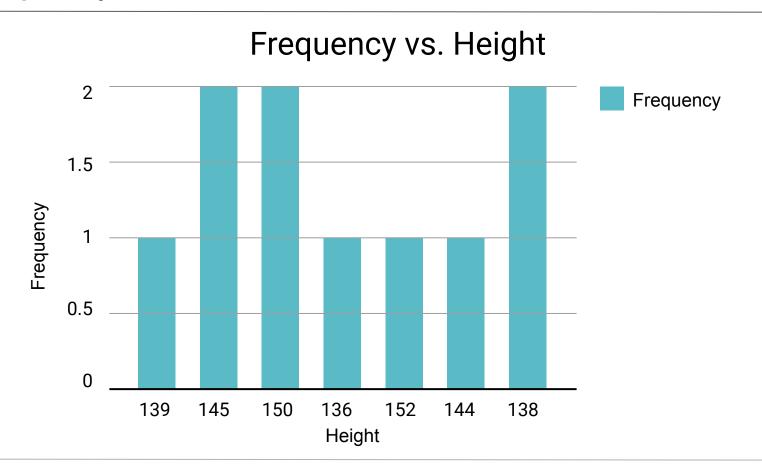
#### **Standard Deviation**

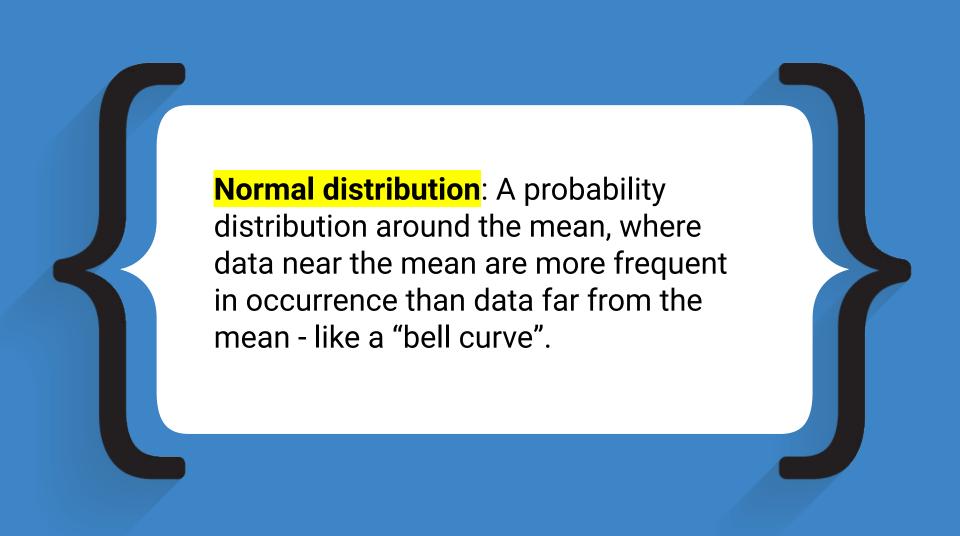
Square root of the variance; a measure used to quantify the dispersion of a set of observations.



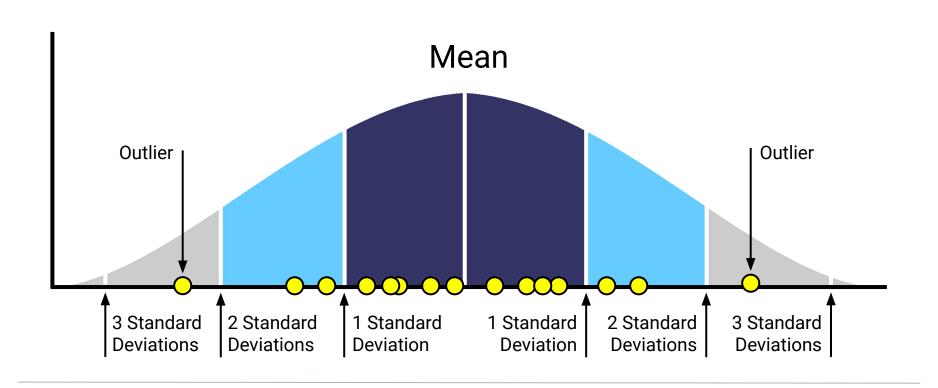


# **Frequency Distribution**

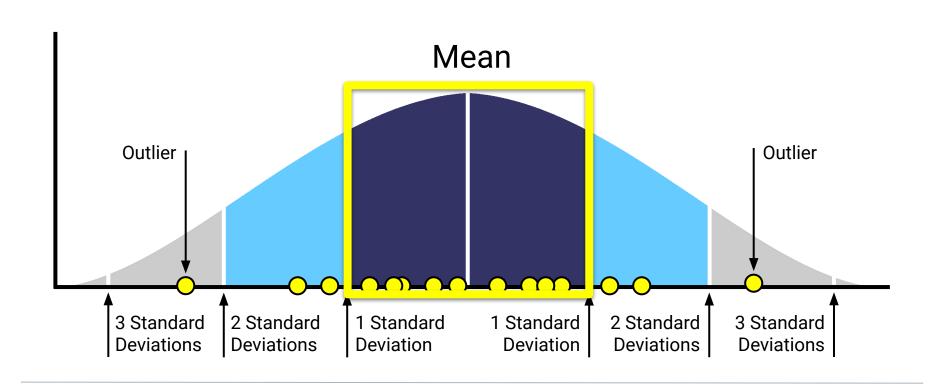




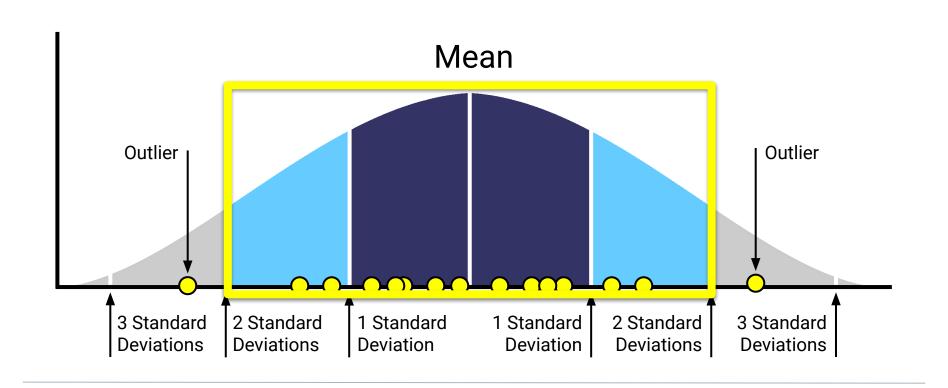
There is a symmetric bell-shaped curve of the distribution.



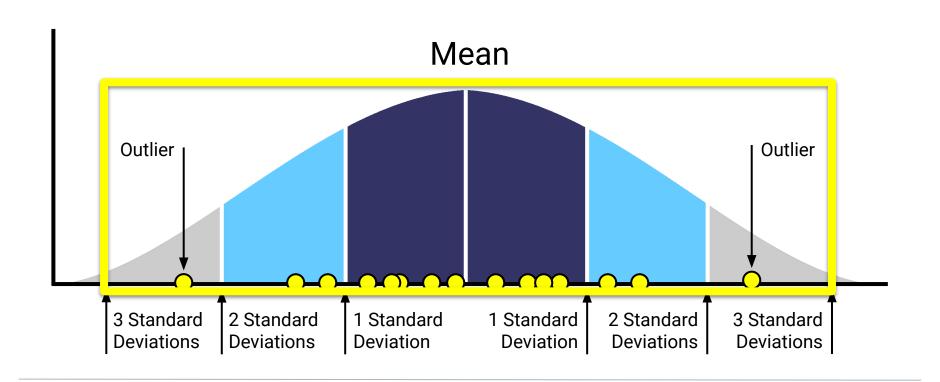
68% of the data fall within 1 standard deviation from the mean

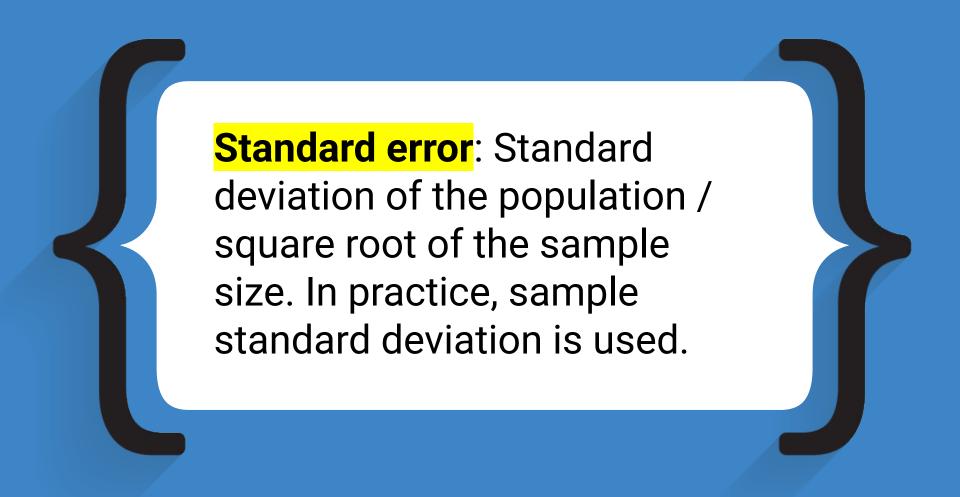


95% of the data fall within 2 standard deviations from the mean



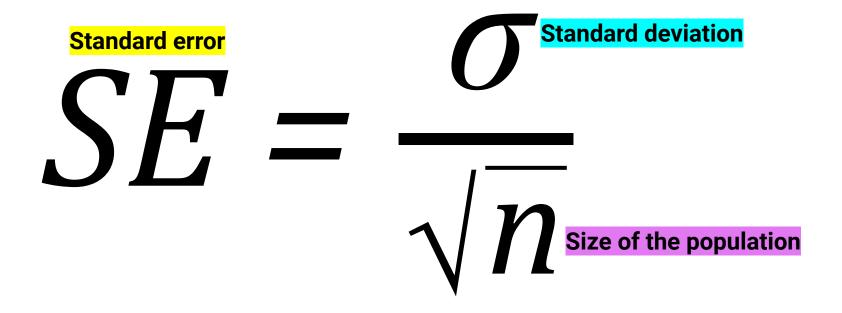
99.7% of the data fall within 3 standard deviations from the mean

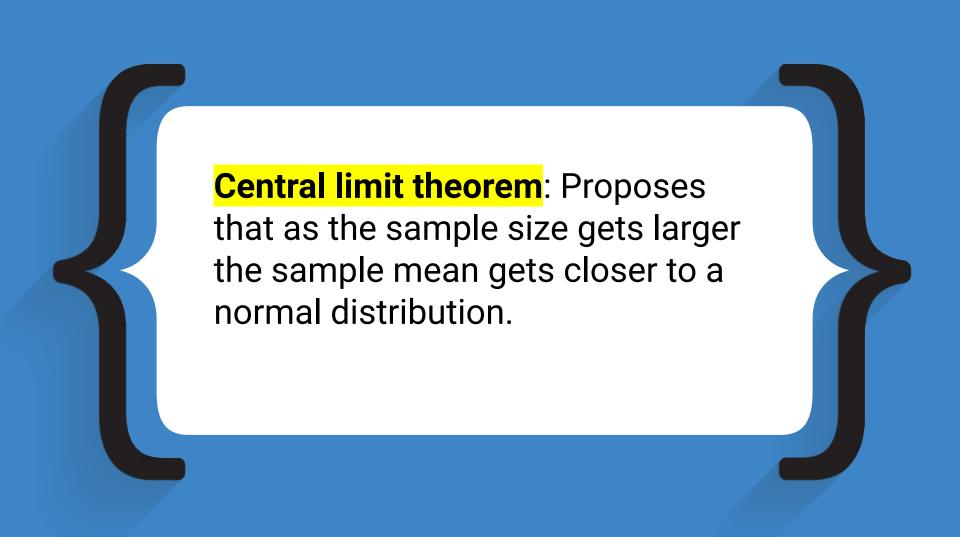




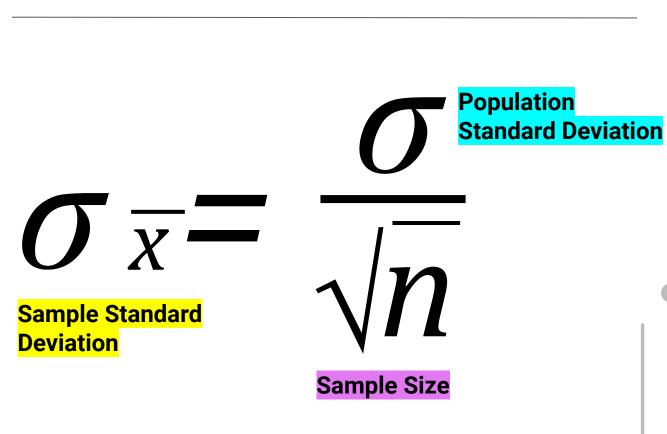
#### Standard error

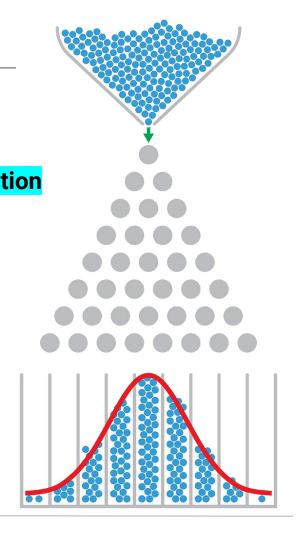
Standard deviation of the population / square root of the sample size. In practice, sample standard deviation is used.

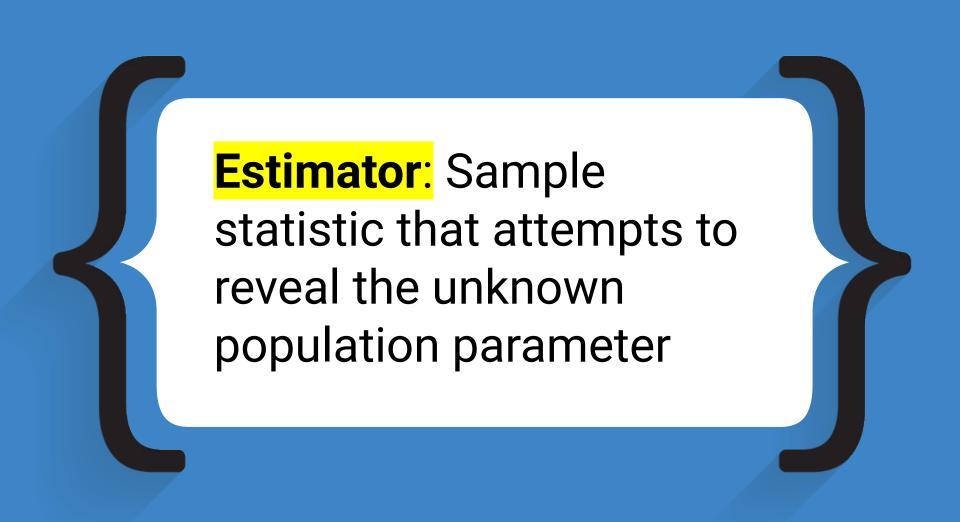




## **Central Limit Theorem**







### **Estimator**

We can estimate what the margin of error would be for a sample size based on the population.

**Population size** 

Confidence level (%)

Margin of error (%)

1000000000

95

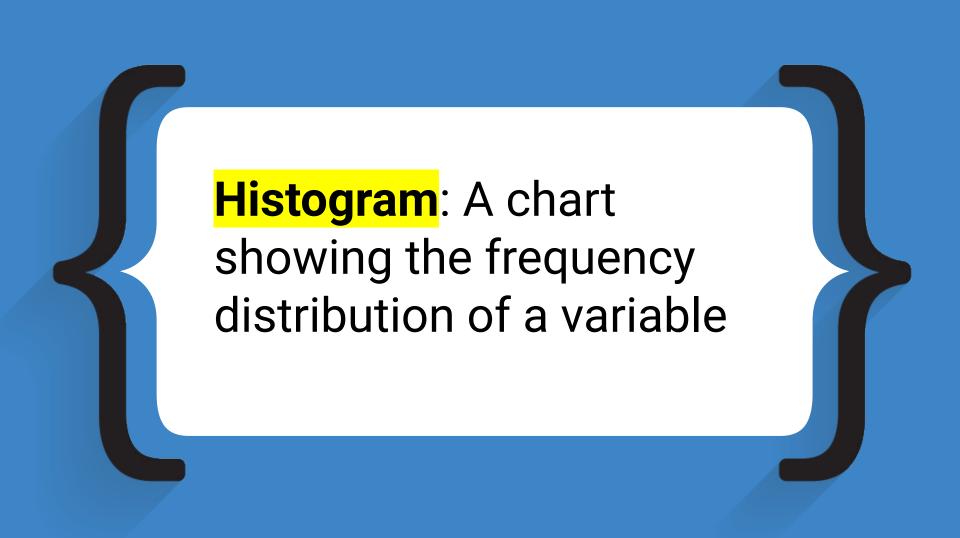
5

Sample size:

385

Margin of error:

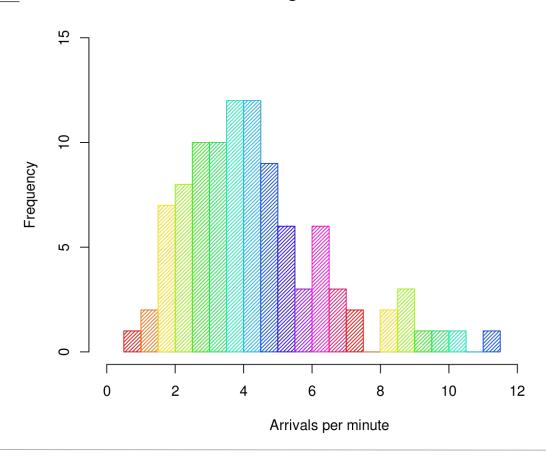
19



# Histogram

A chart showing the frequency distribution of a variable

#### Histogram of arrivals





Instructor Demonstration Quantiles, Outliers and Boxplots

## Be careful when describing real-world data



Real-world data can contain extreme values



Some summary statistics such as the mean take into account **all** values of a data set



Extreme values can **skew** these statistics!



But how can we summarize real-world data?



## We can use quantiles to describe segments of a data set!

Quantiles separate a sorted data set into equally sized fragments.

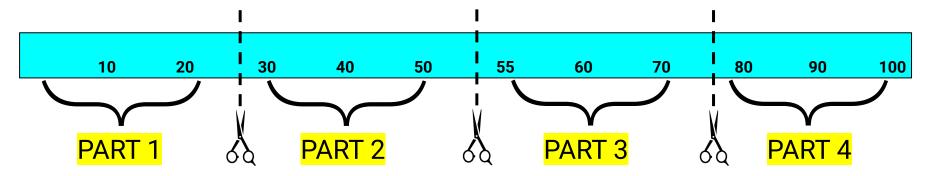
The two most popular types of quantiles are quartiles and percentiles.



Quartiles divide the data set into four equal parts



Percentiles divide the data set into 100 equal parts



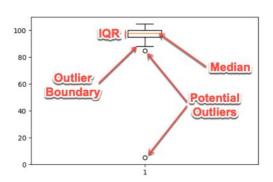


## There are two ways to identify potential outliers

01

#### Qualitatively

Use box and whisker plots to visually identify potential outlier data points





#### Quantitatively

Determine the outlier boundaries in a data set using the '1.5 IQR' rule

- IQR is the interquartile range, or the range between the 1st and 3rd quartiles
- Anything below Q1 1.5 IQR could be an outlier
- Anything above Q3 + 1.5 IQR could be an outlier





# Activity: Outliers—Drawn and Quartiled



## Variance, standard deviation, and Z-score review instructions

#### Instructions:

- Open up the activity workbook and familiarize yourself with the raw data.
  - File: Unsolved/Outliers\_Activity\_Unsolved.xlsx
- Create a new worksheet and name it 'Outlier Testing'.
- In the 'Outlier Testing' worksheet, create a summary statistics table of the Antioxidant\_content\_in\_mmol\_100g for the following statistics:
  - Mean
  - Median
  - Minimum value
  - Maximum value
  - First quartile
  - Third quartile
  - Interquartile range
- Using the calculations from the table, determine the lower and upper boundaries of the 1.5\*IQR rule.
- Determine if there are any products whose Antioxidant\_content\_in\_mmol\_100g falls outside of the 1.5\*IQR boundaries. List those products and their antioxidant content on the worksheet.
- Create a box plot of the Antioxidant\_content\_in\_mmol\_100g for all products.
  - **Note**: Be sure to add a title and label your y-axis.

