

Module 1

This Week: Excel

This Week: Excel

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



Import data into Excel



Apply filters, conditional formatting, and formulas to data



Create and interpret charts and pivot tables in Excel



Calculate summary statistics



Characterize data to identify outliers in datasets



Visualize the distribution of data using box plots



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:

01

Basic Charting

02

Summary Statistics

03

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



GitHub

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

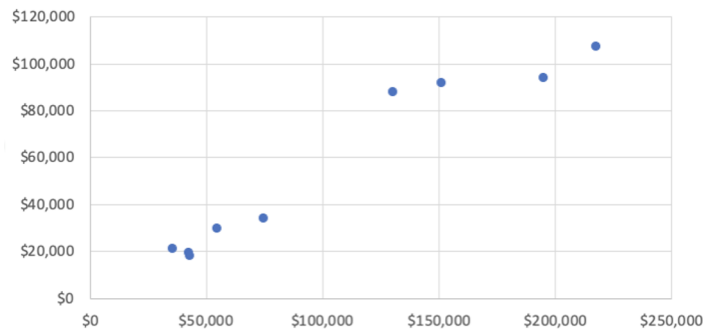


Instructor Demonstration

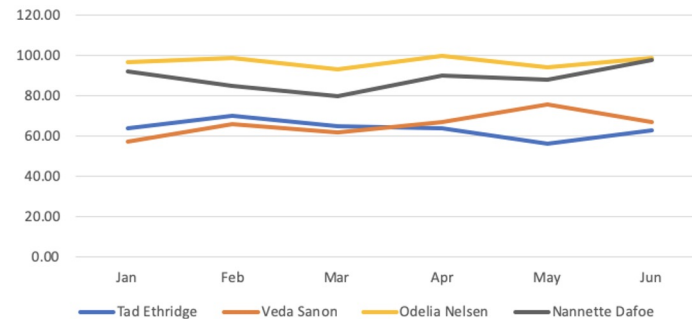
Basic Charting

It is time to learn Excel visualizations!

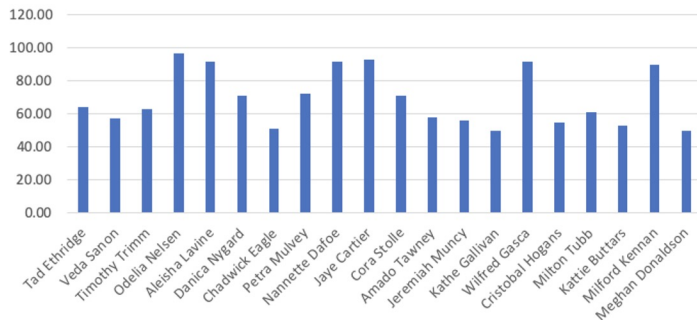
Car Price



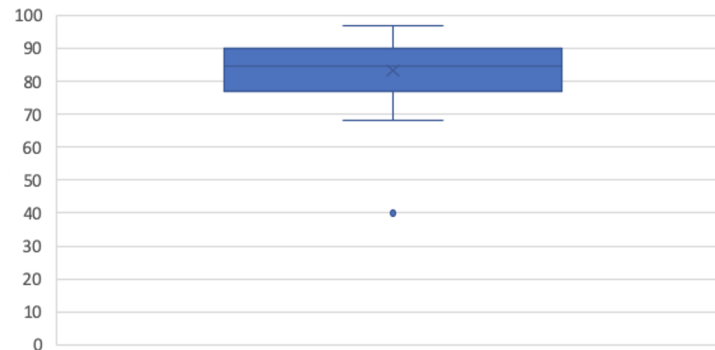
Grades Over Semester



Jan



Tennis Serve Speeds (mph)



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



Time to <code>



Activity: The Line and Bar Grades

Suggested Time:

15 minutes

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:

- Create a series of bar graphs that visualize the grades of all students in the class, with one graph for every month.
- Create a line graph using all of the data that can be used to compare students' grades across the semester.
- Use filtering in the line graph to allow you to drill down to a specific student's progress throughout the semester.

Hint:

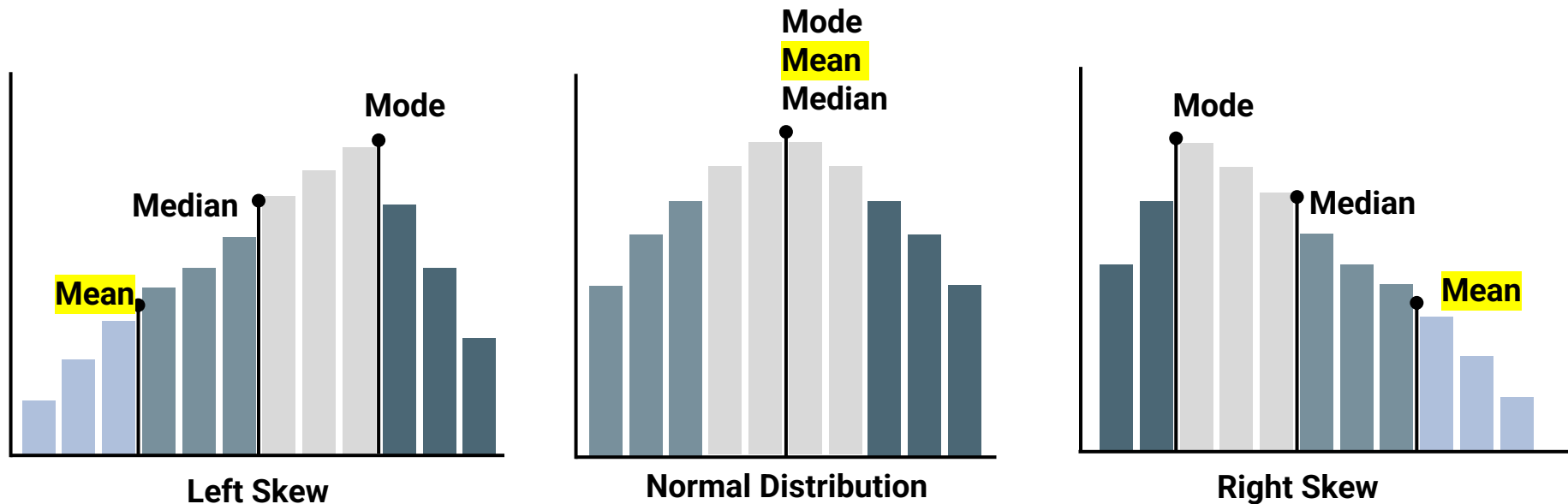
When duplicating bar graphs, it pays to get the formatting and look of the chart where you want it for the first graph (e.g., for January), and to then copy that chart and re-select the data for the subsequent copies (keeping the style and format, but just changing the data).



Time's Up! Let's Review.

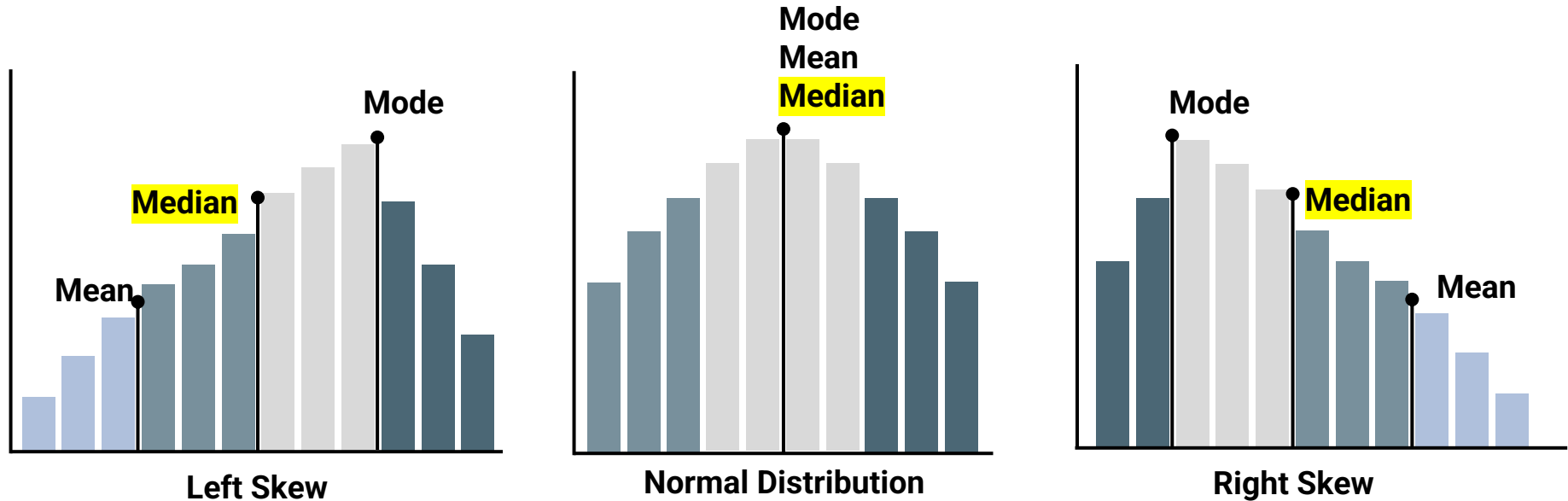
Mean

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



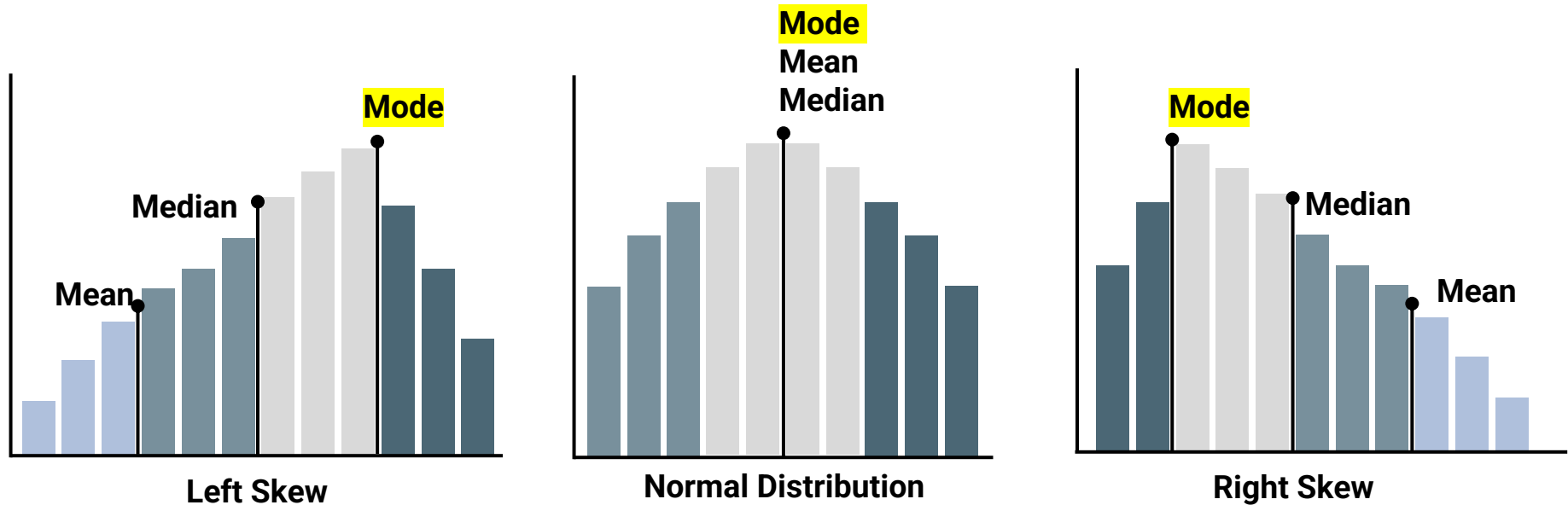
Median

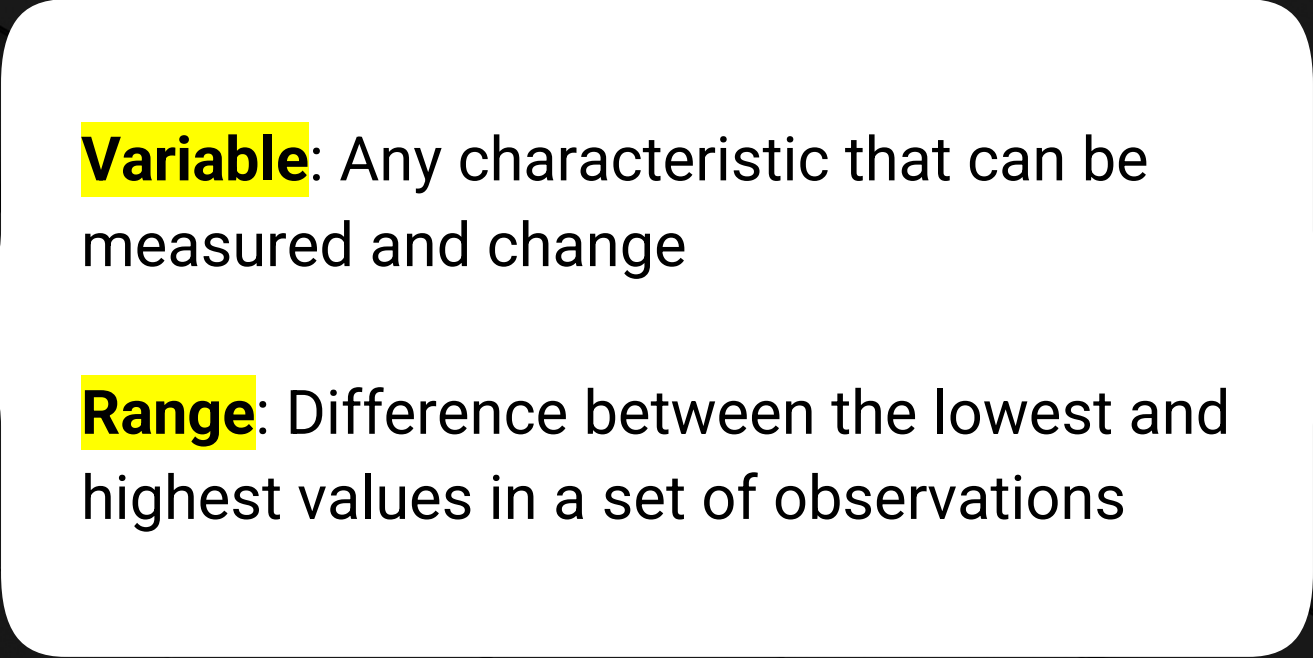
The value at the midpoint in a set of observed values



Mode

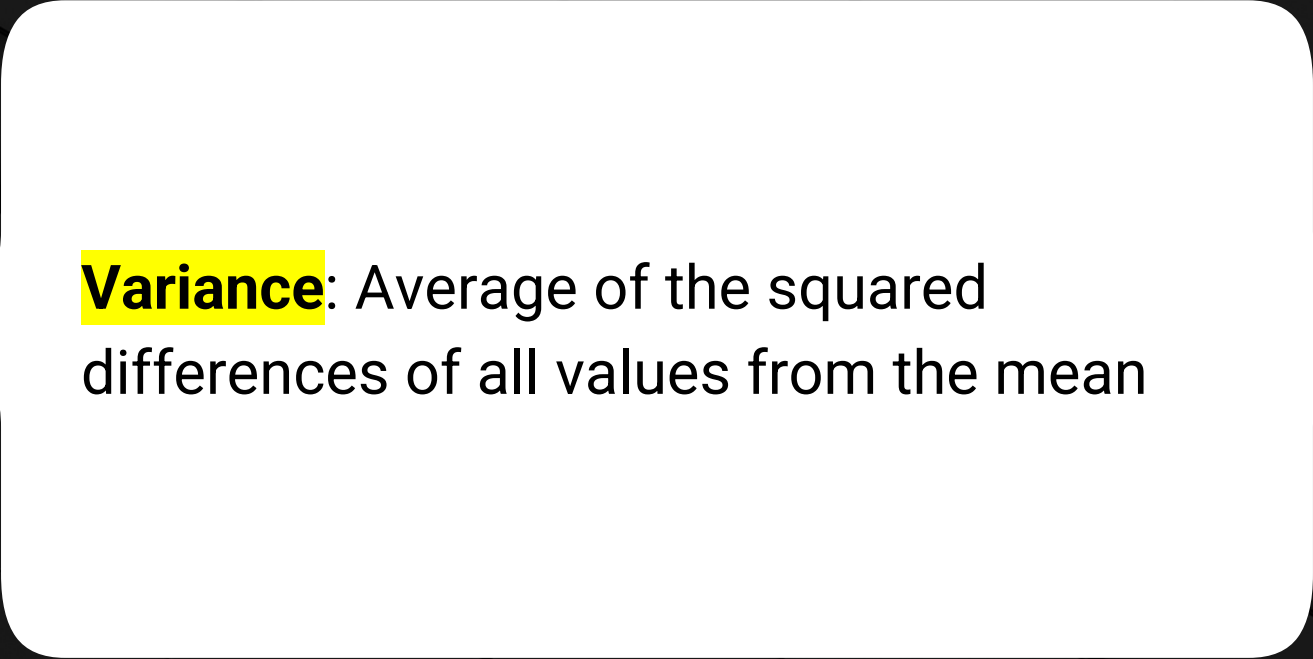
The most frequently occurring value in a set of values





Variable: Any characteristic that can be measured and change

Range: Difference between the lowest and highest values in a set of observations



Variance: Average of the squared differences of all values from the mean

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

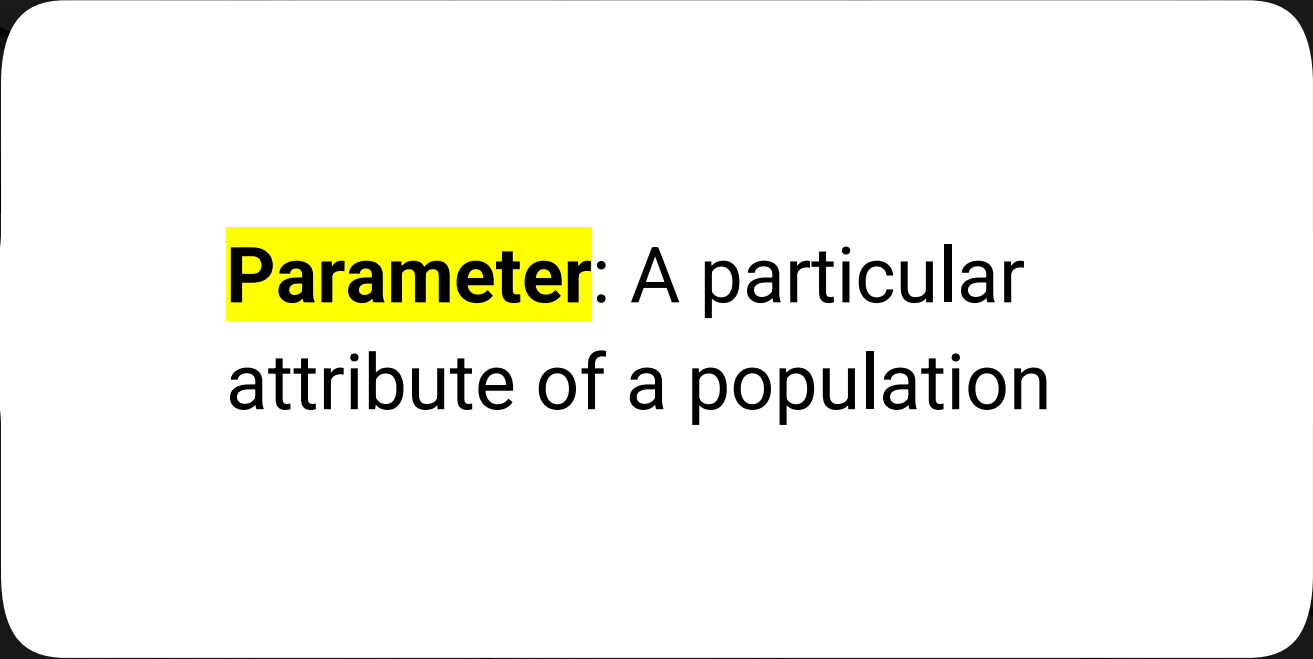
The value of the one observation

The mean value of all observations

Sample variance

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

The number of observations



Parameter: A particular attribute of a population

Parameter

A particular attribute of a population

Population Parameters

μ	Population mean
σ	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:

Square root of the variance; a measure of how spread out the observations are.

Standard Deviation



Describes how spread out the data is from the mean



Calculated from the square root of the variance

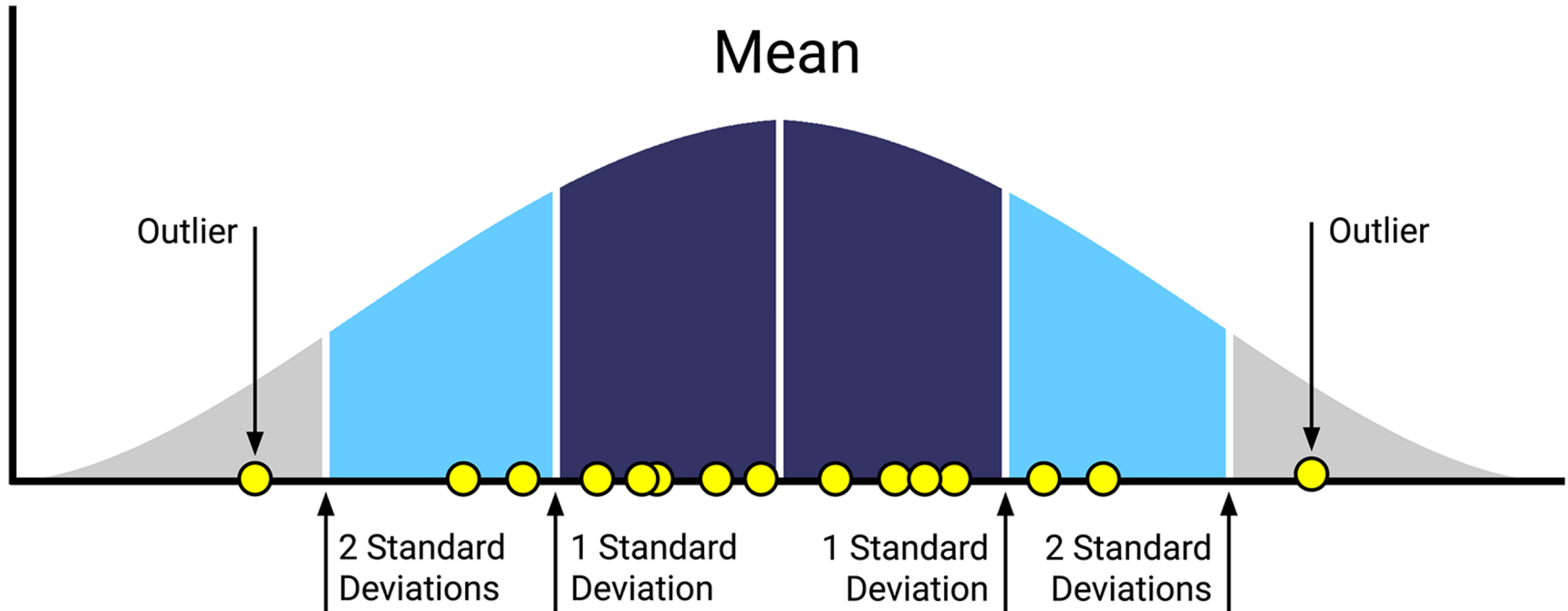


In the same units of measurement as the mean

$$\text{Standard deviation } \sigma = \sqrt{S^2} \text{ The variance}$$

Standard Deviation

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



Variance or Standard Deviation?



When two distributions are combined, the means and variance of the resulting distribution add, but the standard deviation does not.



Thus, Variance is preferred mathematically, and the standard deviation often gives nice interpretations.



Examples: Total SAT score (reading, writing and math)
Height of Americans (woman and men)

Variance or Standard Deviation?

$$\sigma_{Total} = \sqrt{Var_{women} + Var_{men}}$$

$$\sigma_{Total} = \sqrt{\sum (w - \mu_w)^2 + \sum (m - \mu_m)^2}$$

$$\sigma_{Total} \neq \sqrt{\sum (w - \mu_w)^2} + \sqrt{\sum (m - \mu_m)^2}$$

$$\begin{aligned}\sqrt{(3)^2} + \sqrt{(4)^2} &= 3 + 4 = 7 \\ \sqrt{(3)^2 + (4)^2} &= \sqrt{25} = 5\end{aligned}$$

We will use variance in the following:



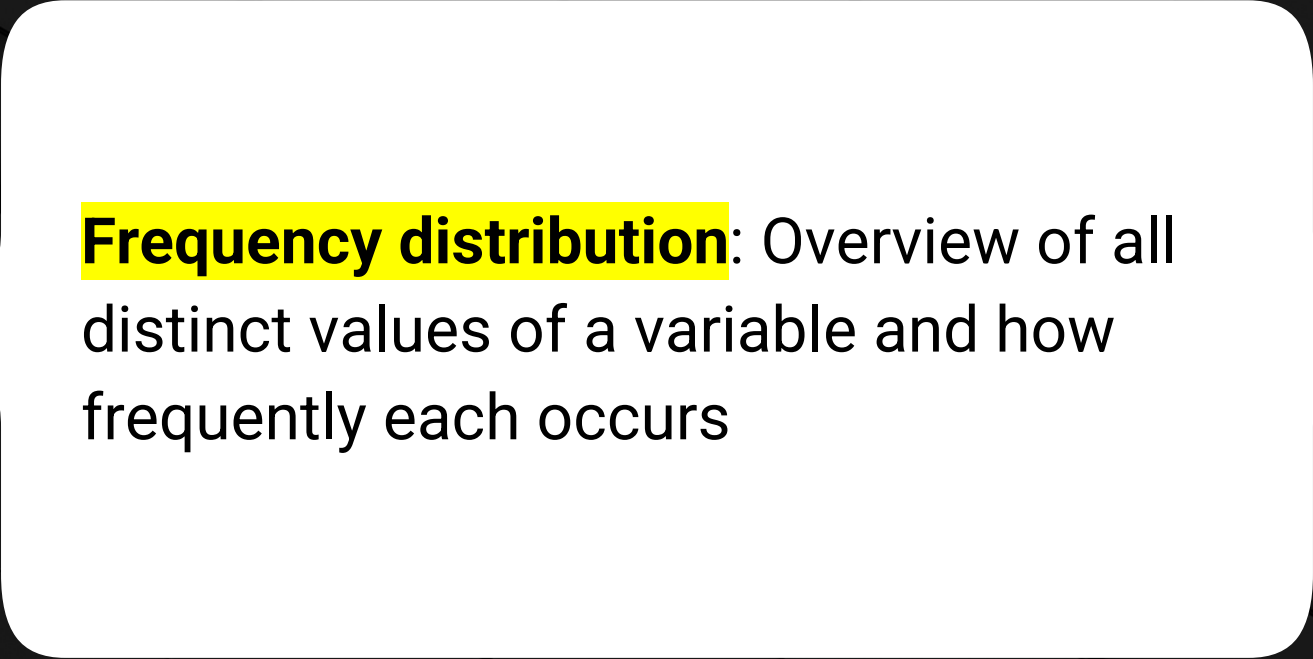
Linear Regression (least squares optimization that estimates variance)



T-test and Analysis of Variation (ANOVA)

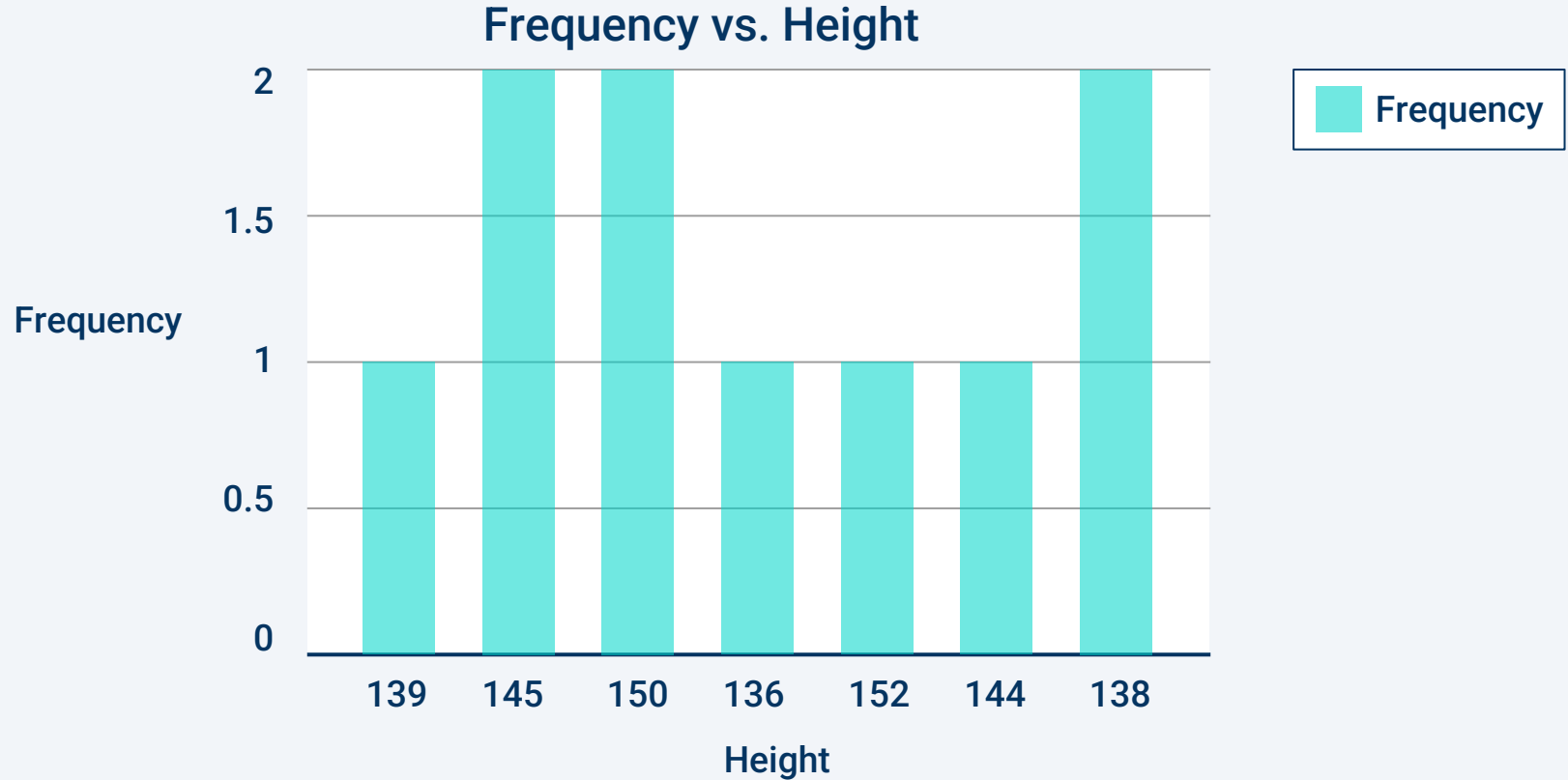


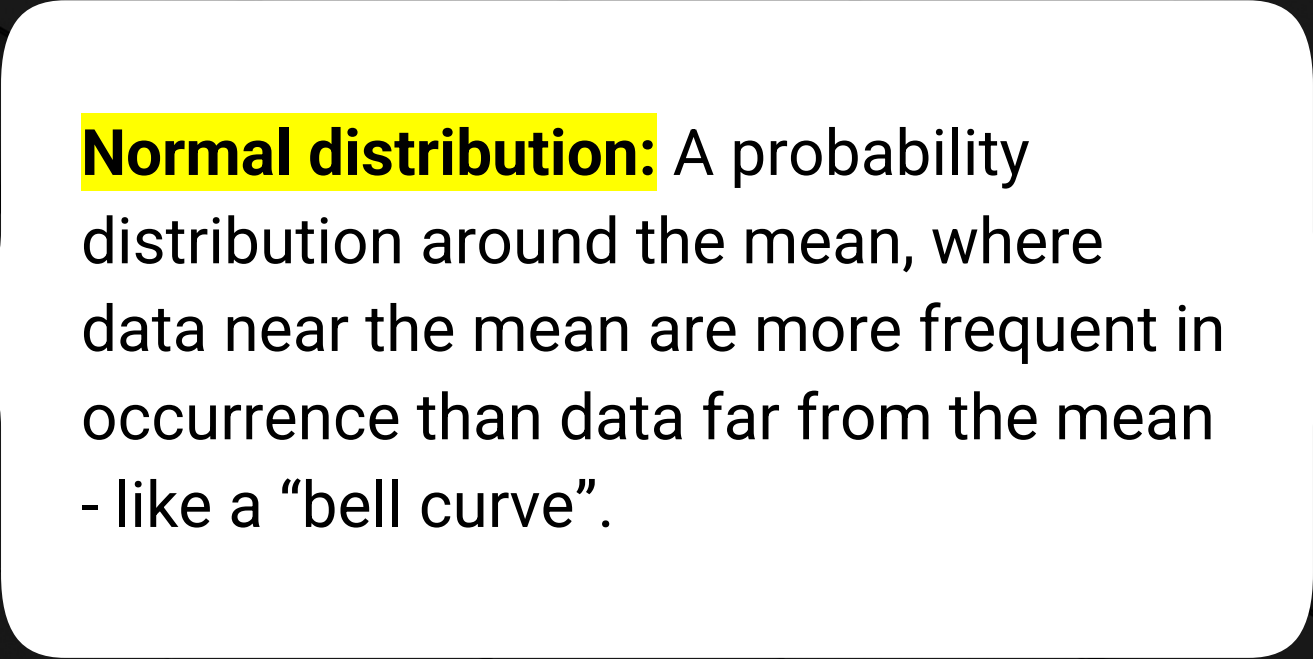
Principle Component Analysis (PCA)



Frequency distribution: Overview of all distinct values of a variable and how frequently each occurs

Frequency Distribution

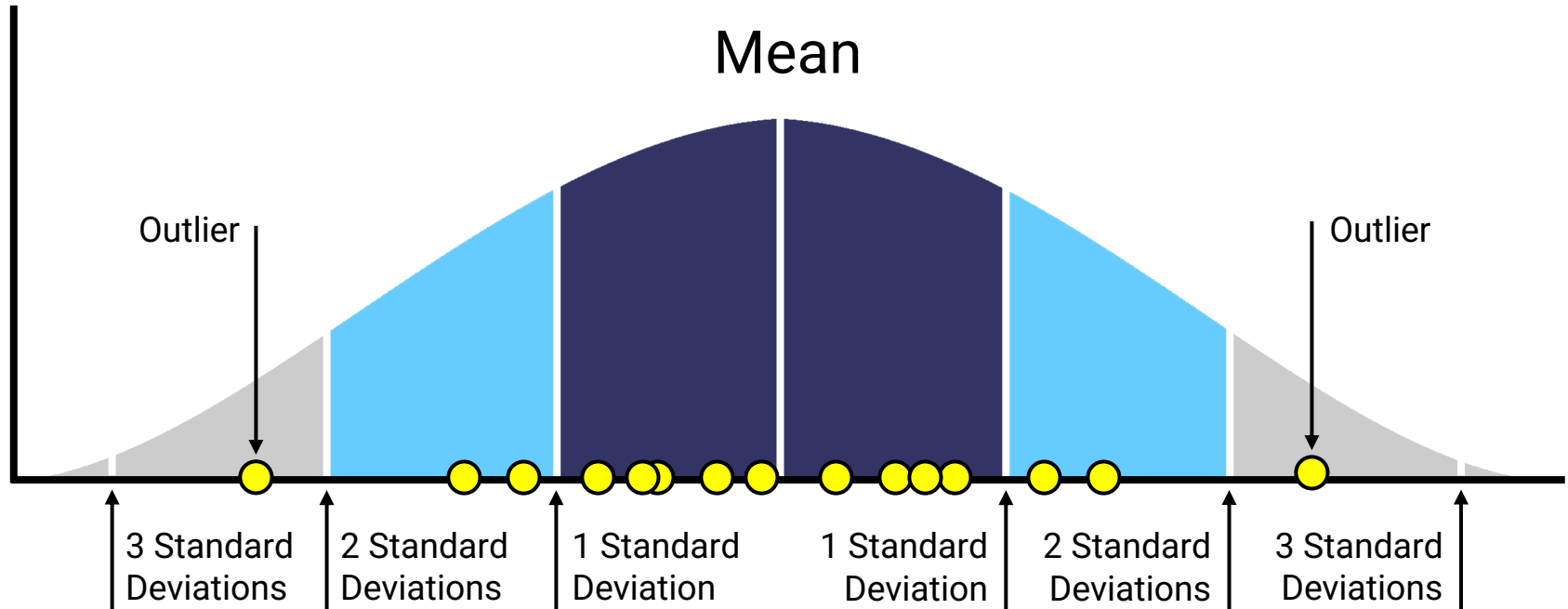




Normal distribution: A probability distribution around the mean, where data near the mean are more frequent in occurrence than data far from the mean - like a “bell curve”.

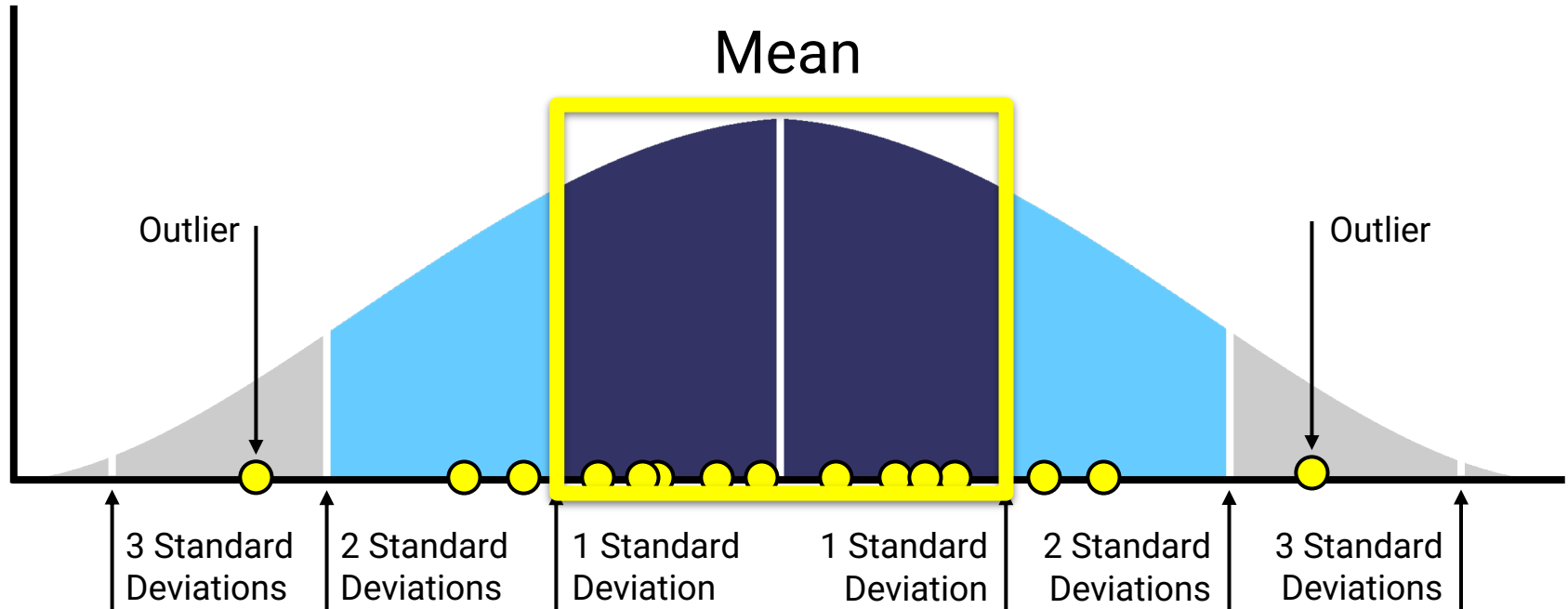
Normal Distribution Features

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



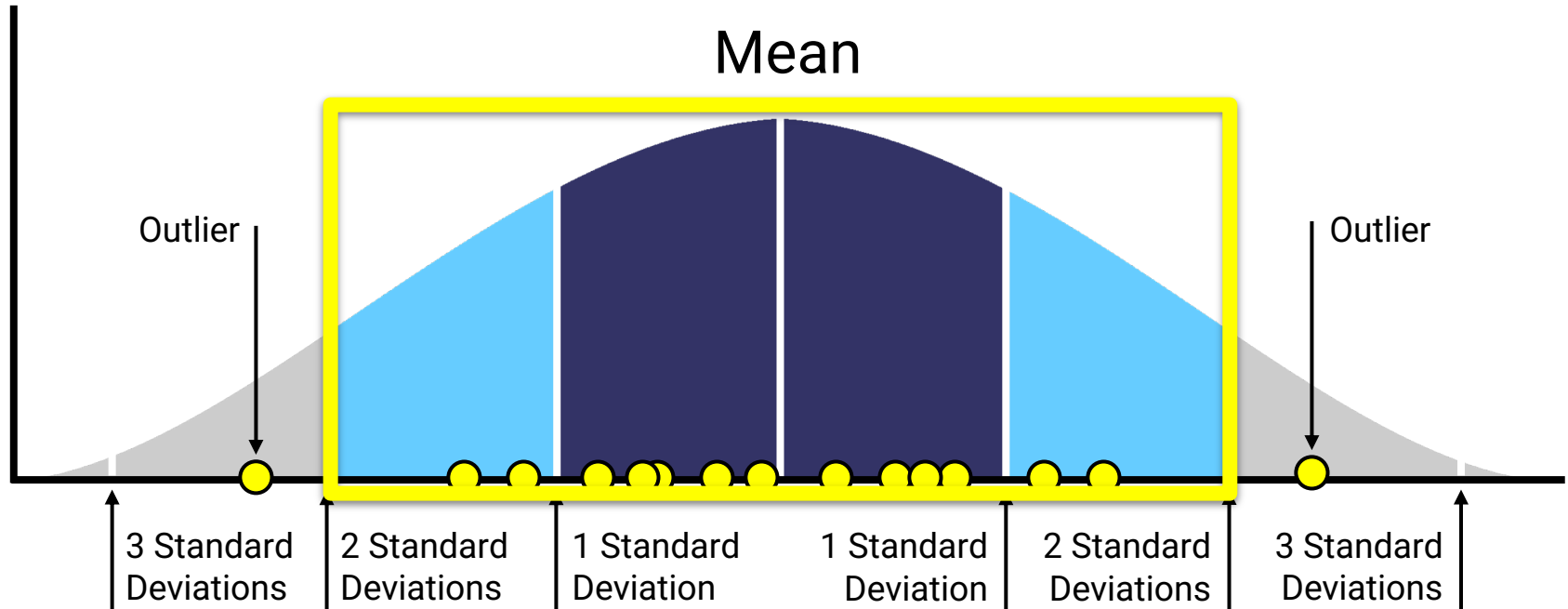
Normal Distribution Features

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



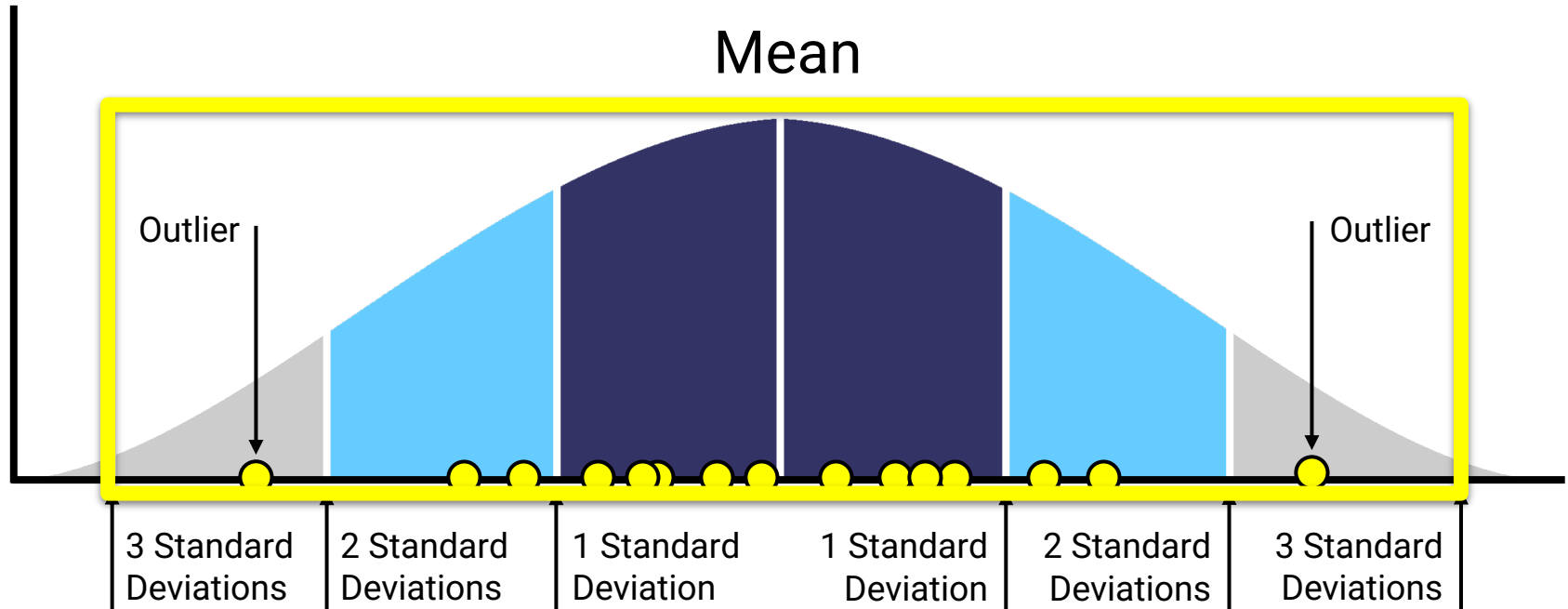
Normal Distribution Features

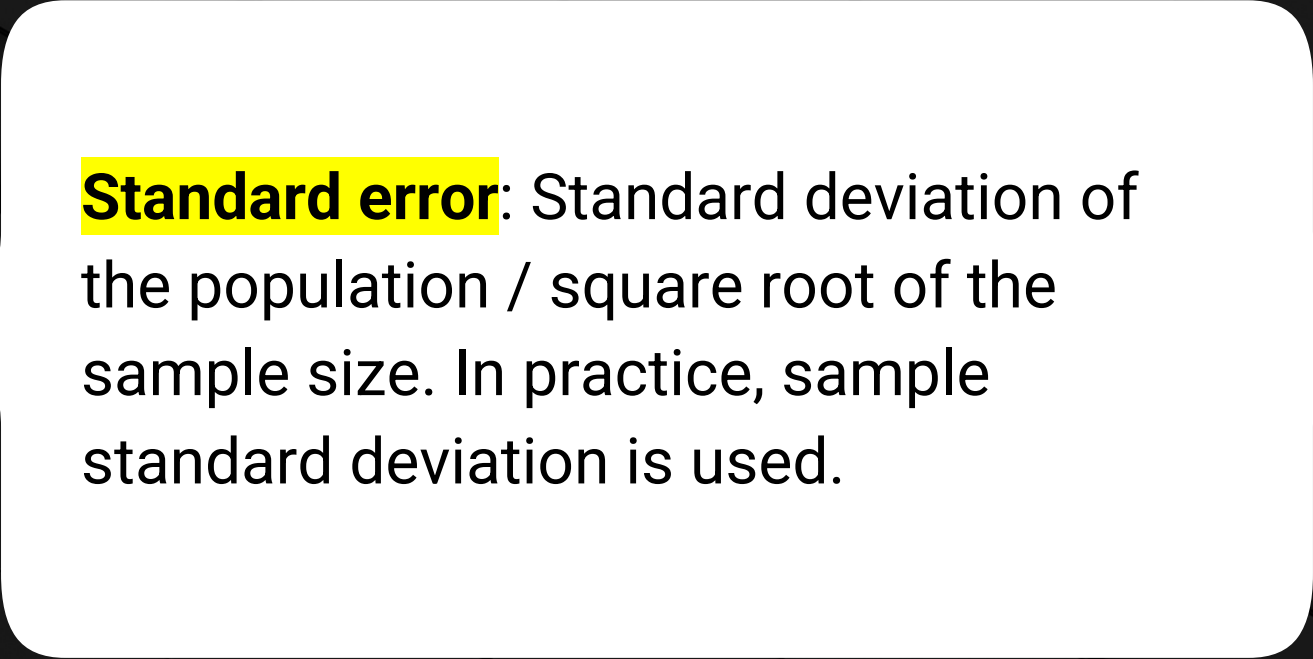
95% of the data fall within 2 standard deviations from the mean



Normal Distribution Features

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.

Standard error

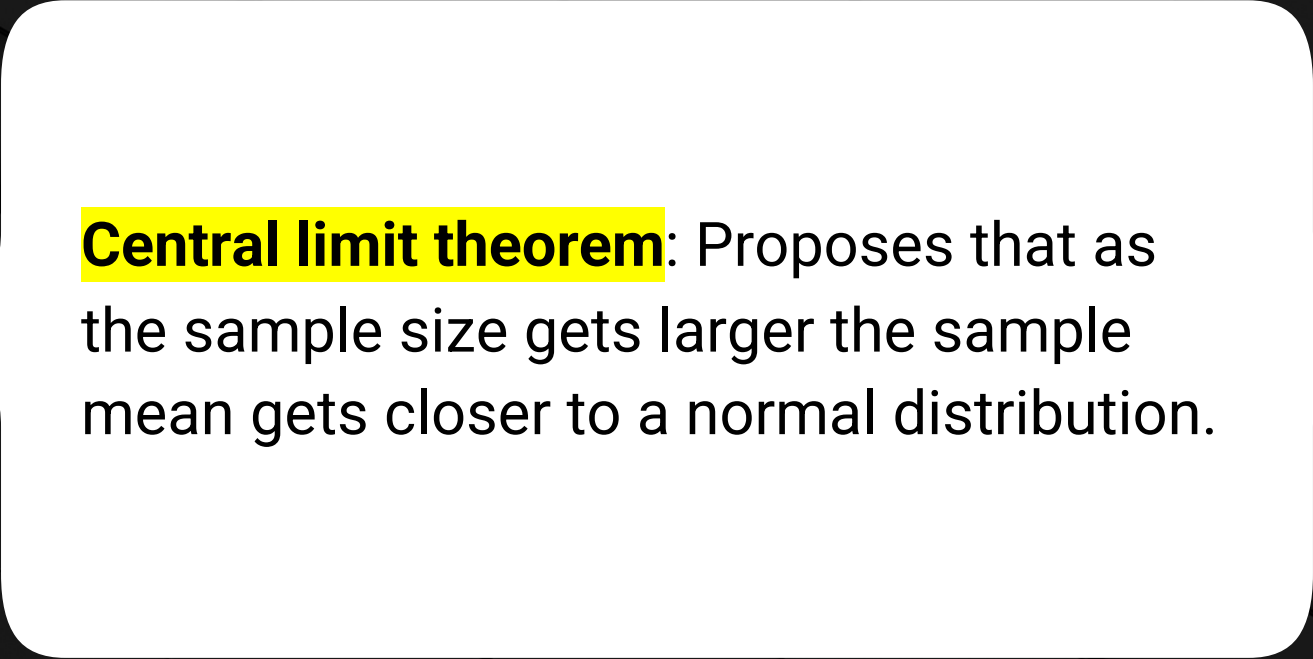
Standard deviation of the population / square root of the sample size. In practice, sample standard deviation is used.

$$SE = \frac{\sigma}{\sqrt{n}}$$

Standard error

Standard deviation

Size of the population



Central limit theorem: Proposes that as the sample size gets larger the sample mean gets closer to a normal distribution.

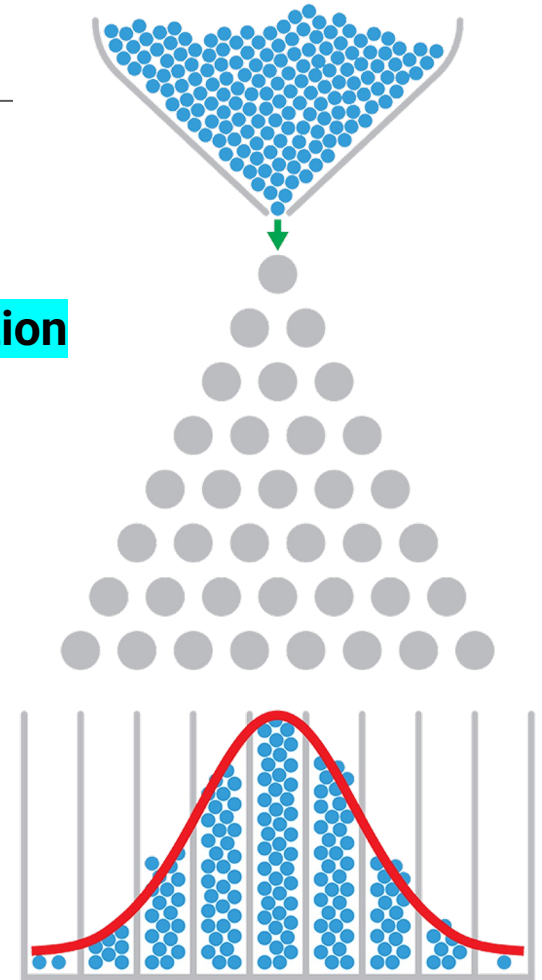
Central Limit Theorem

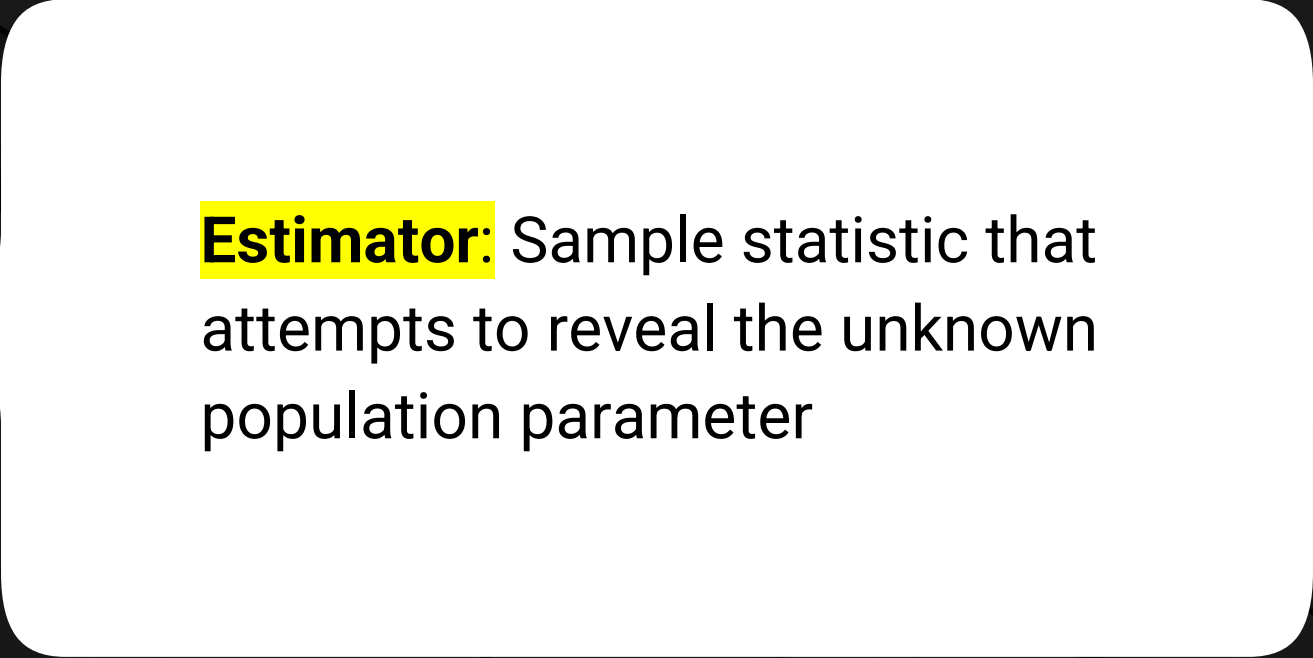
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Sample Standard Deviation

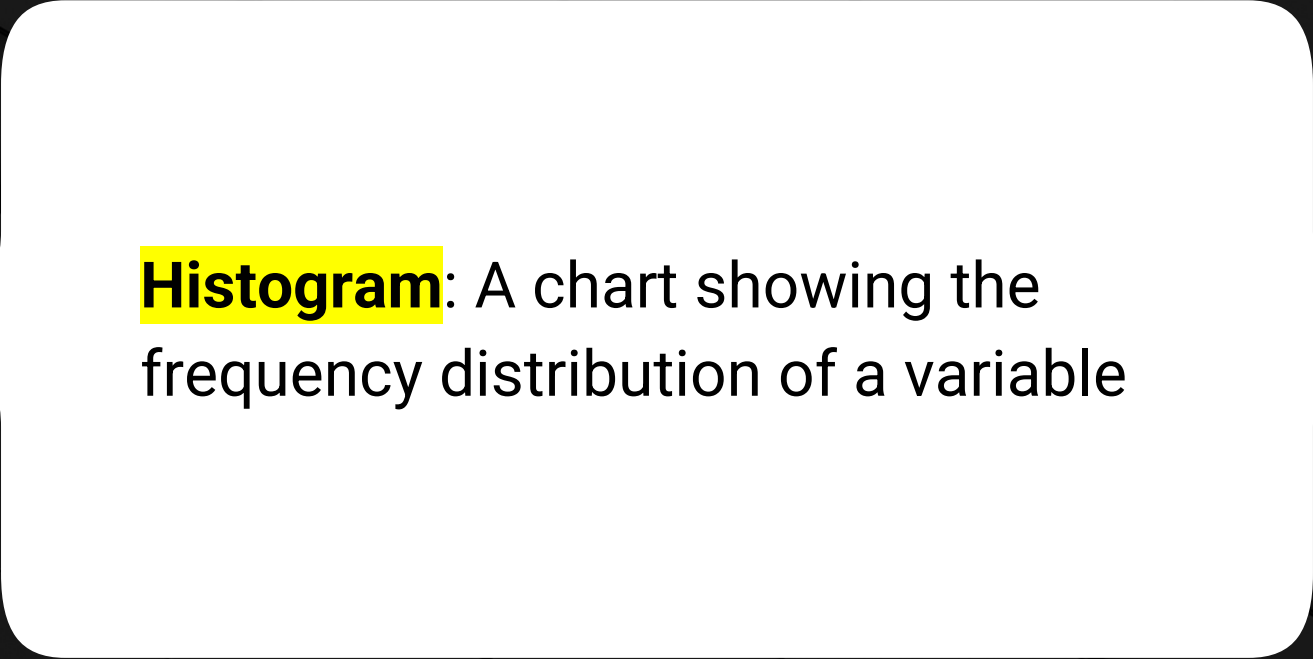
Population Standard Deviation

Sample Size





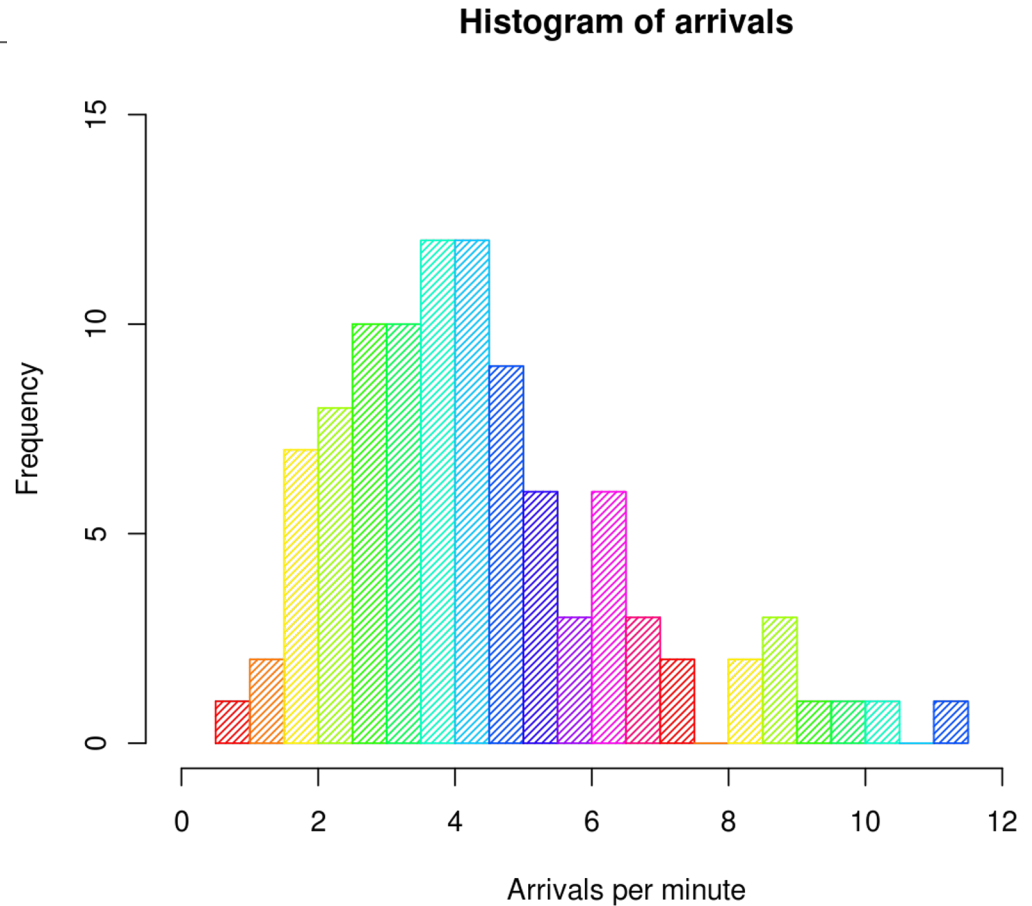
Estimator: Sample statistic that attempts to reveal the unknown population parameter



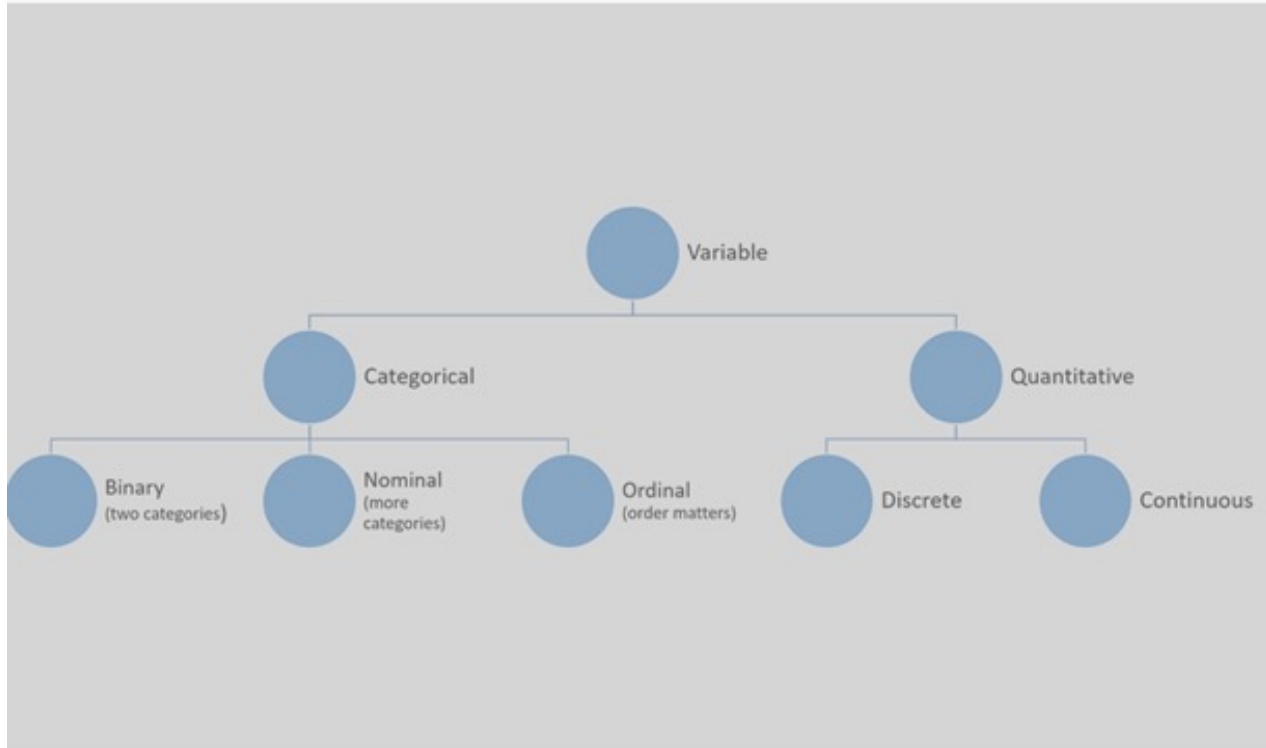
Histogram: A chart showing the frequency distribution of a variable

Histogram

A chart showing
the frequency
distribution of
a variable



Data Types





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?

Quantiles: Used to Describe Segments of a Dataset

Quantiles separate a sorted dataset into equally sized fragments.

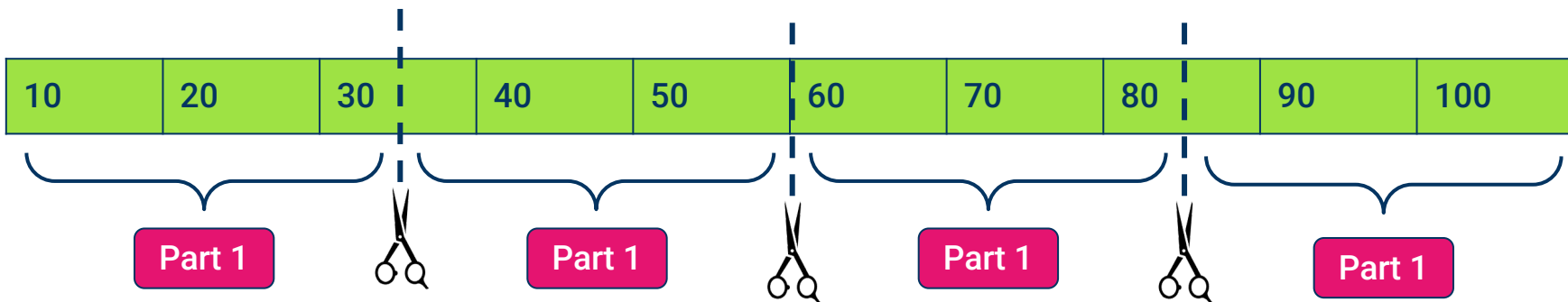
The two most popular types of quantiles are **quartiles** and **percentiles**.

01

Quartiles divide the dataset into four equally sized parts.

02

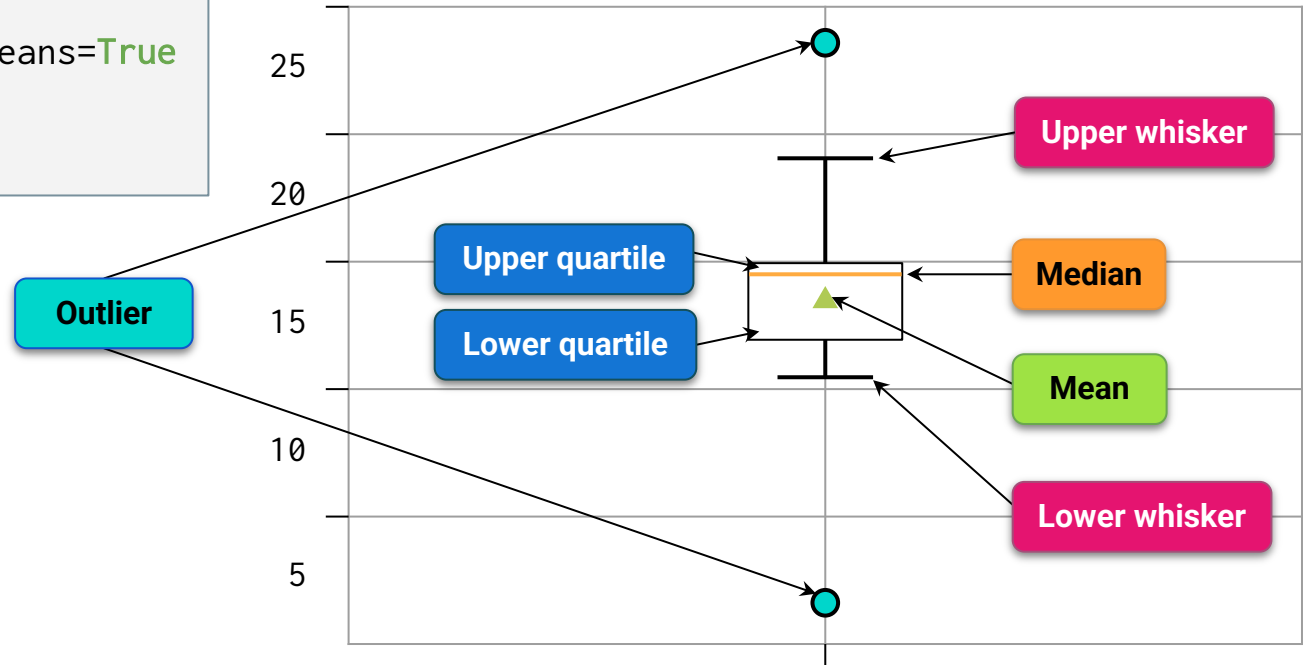
Percentiles divide the dataset into 100 equally sized parts.



Qualitatively

Use **box-and-whisker plots** to visually identify potential outliers.

```
# Create box plot  
plt.boxplot(arr, showmeans=True)  
plt.grid()  
plt.show()
```



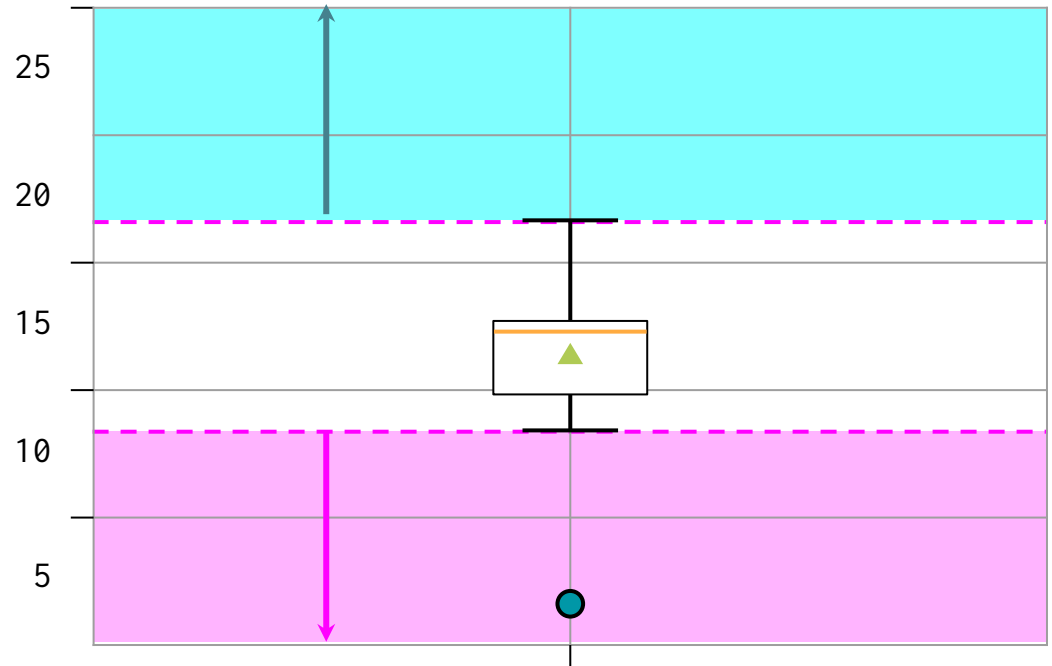
Quantitatively

Determine the outlier boundaries in a dataset by using the **$1.5 \times \text{IQR}$ rule**.

The IQR is the range between the first and the third quartile.

Anything **less than, or below,** $\text{Quartile 1} - (1.5 \times \text{IQR})$ might be an outlier.

Anything **greater than, or above,** $\text{Quartile 3} + (1.5 \times \text{IQR})$ might be an outlier.





Time to <code>



Activity: Outliers—Drawn and Quartiled

Suggested Time:

15 minutes

Activity: Outliers—Drawn and Quartiled

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 $1.5 \times \text{IQR}$ rule.
- Determine if there are any products whose Antioxidant_content_in_mmol_100g falls outside of the $1.5 \times \text{IQR}$ 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.

Class Review

In today's class, you learned the following skills:

01

Basic Charting: Line Charts - Lesson 1.3.2 Box Plots – Lesson 1.5.4

02

Summary Statistics: Lessons 1.5.1 – 1.5.3

03

GitHub Repositories: Lesson 1.6.1

By Next Class: Complete up to Lesson 2.2.4 Get DQ's Yearly Return for 2018