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

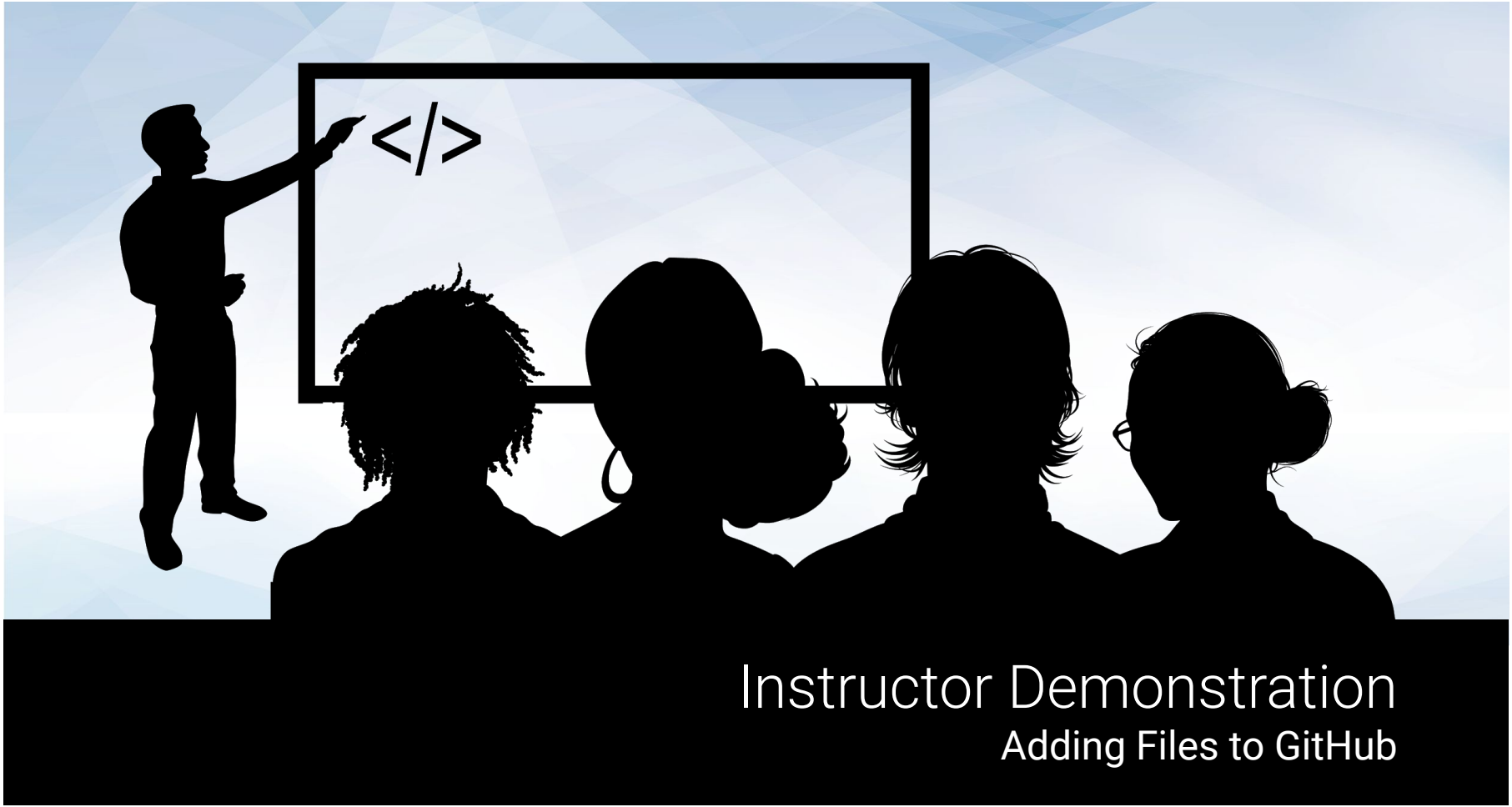
By completing today's activities, you'll learn the following skills:

01 Basic Charting

02 Summary Statistics

03 GitHub Repositories

Make sure you've downloaded any relevant class files!



Instructor Demonstration

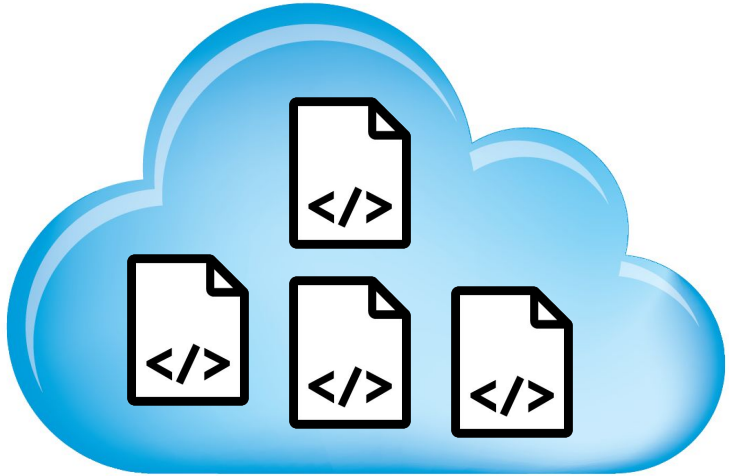
Adding Files to GitHub

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

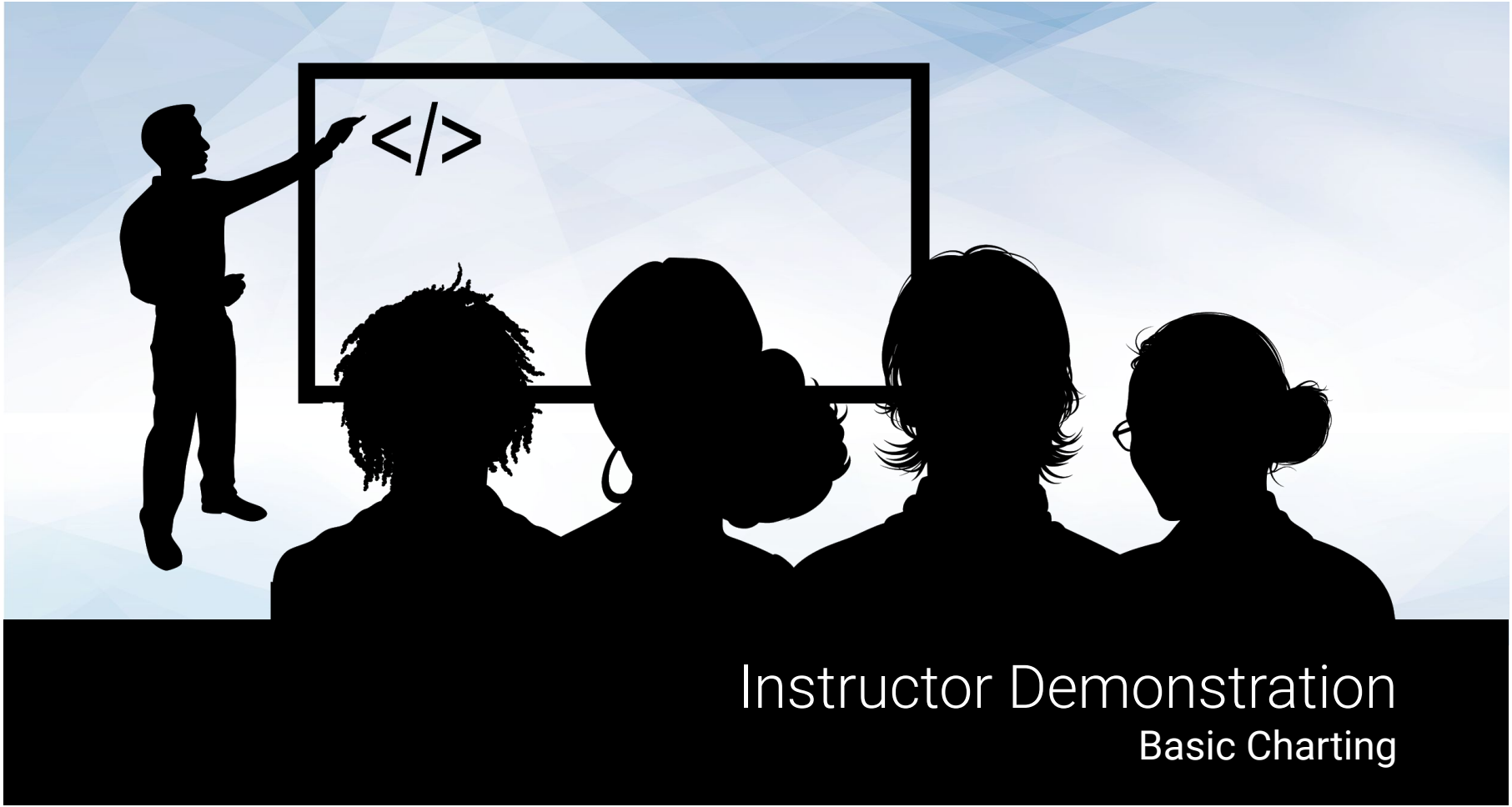
We will use Git and GitHub throughout the curriculum

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

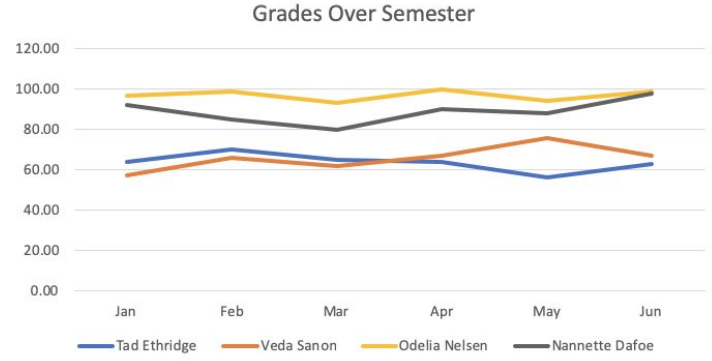
By the end of the curriculum, you should be proficient with the basic Git and GitHub functionality

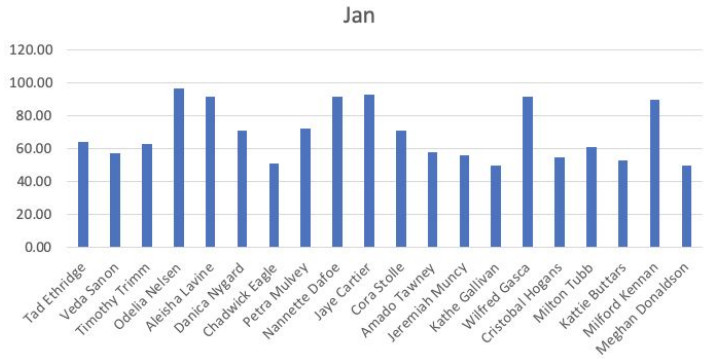


Instructor Demonstration

Basic Charting

It is time to learn Excel visualizations!





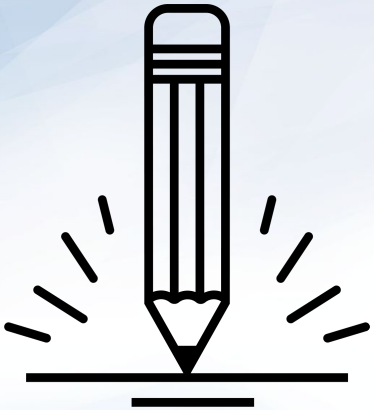
We will look at a few examples and use cases

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

< Demo Time >



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.

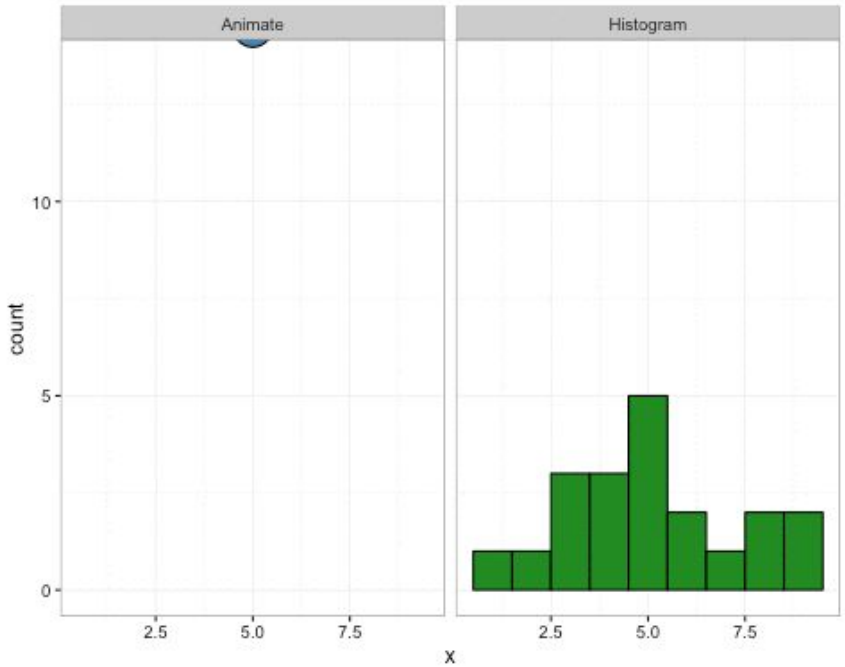
Instructions:	**Hint:**
• Create a series of bar graphs that visualize the grades of all students in the class, with one graph for every month. • Create a line graph using all of the data that can be used to compare students' grades across the semester. • Use filtering in the line graph to allow you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

Suggested Time: 15 minutes

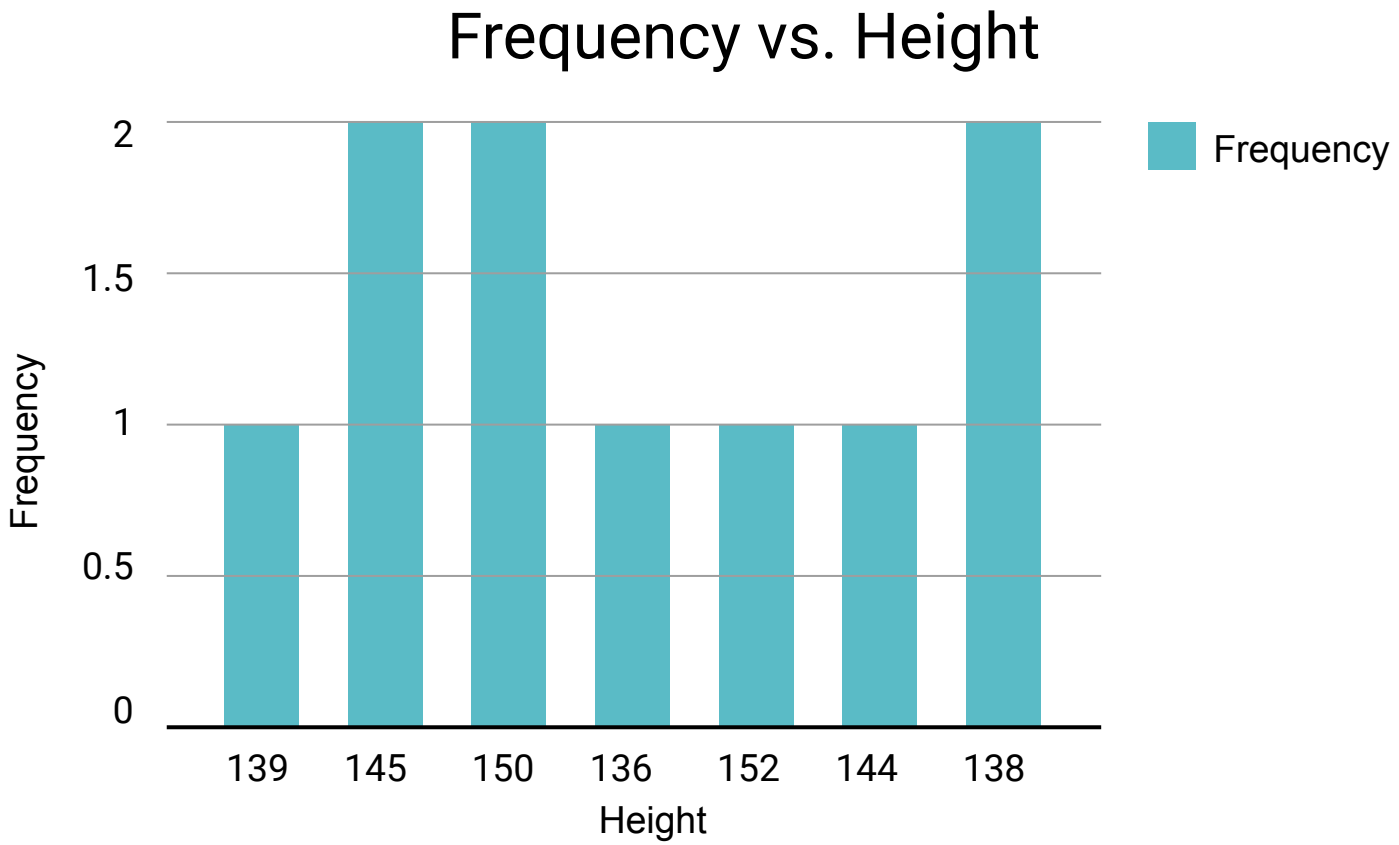
Let's Review

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

Frequency Distribution



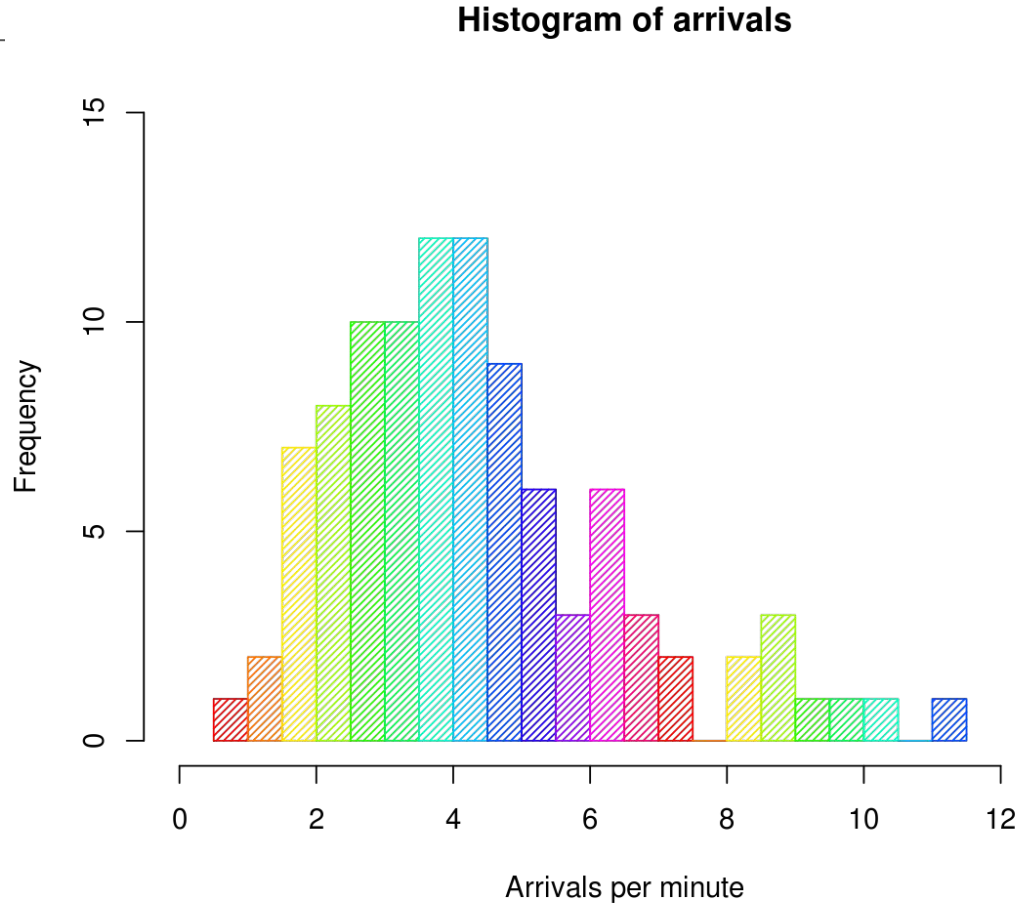
Frequency Distribution



Histogram: A chart showing the frequency distribution of a variable

Histogram

A chart showing the frequency distribution of a variable



Parameter: A particular attribute of a population

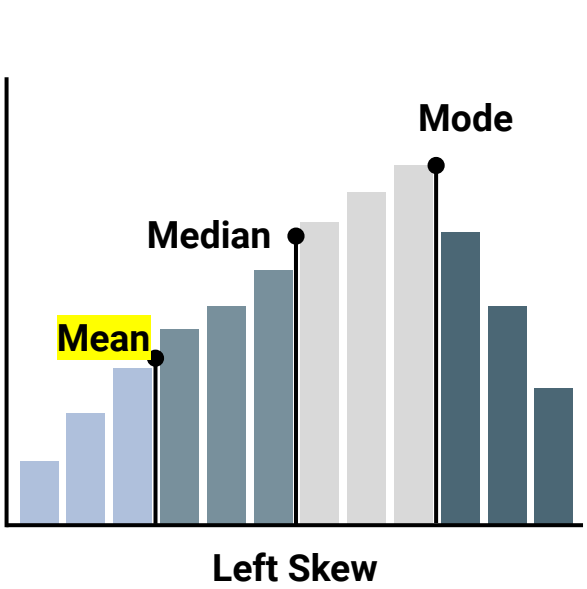
Parameter

A particular attribute of a population

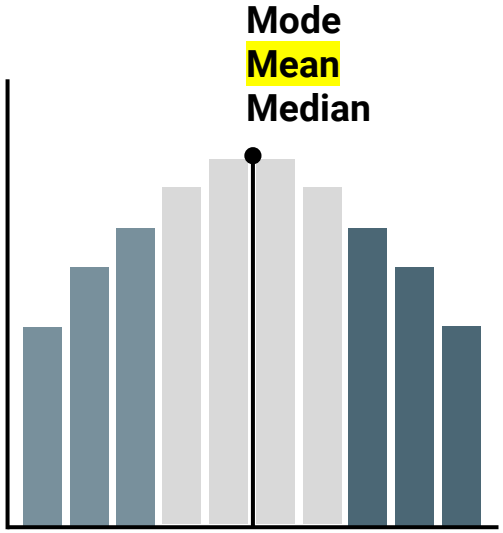
Population Parameters	
μ	Population mean
σ	Population standard deviation
P	Population proportion
N	Population size
X	Population data value
r	Correlation coefficient

Mean

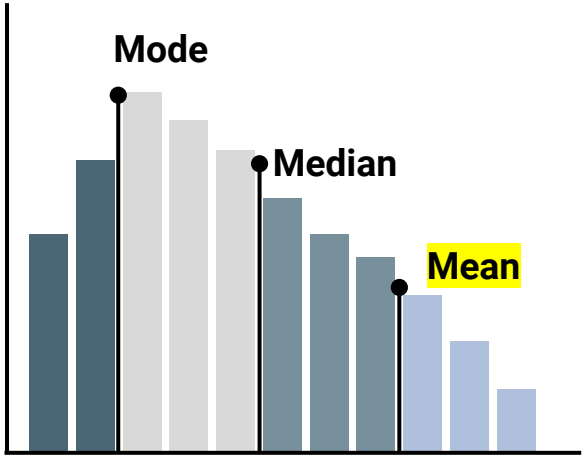
Sum of all values in the sample divided by the number of values in the sample



Left Skew



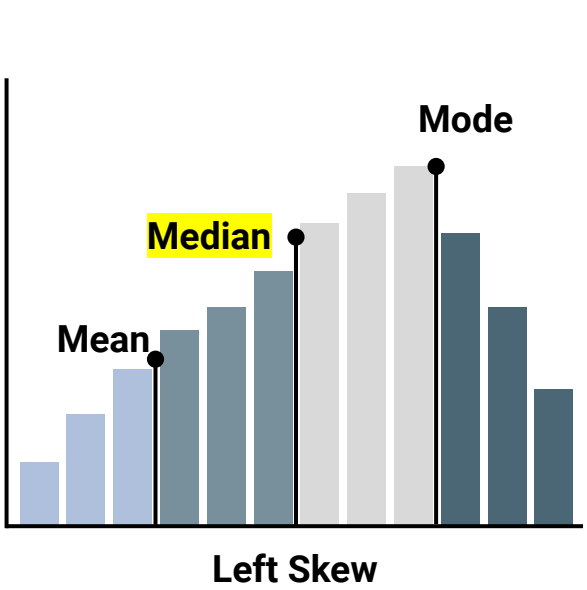
Normal Distribution



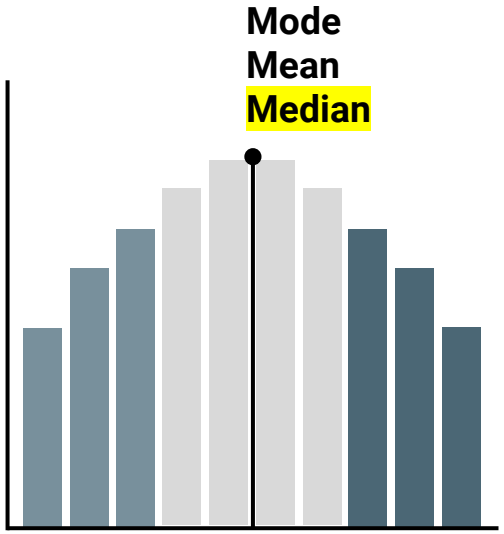
Right Skew

Median

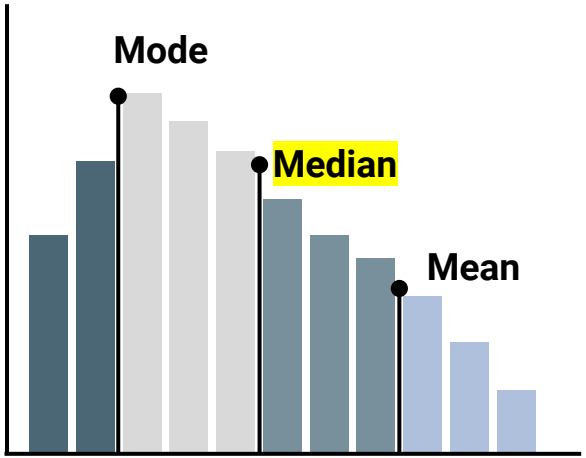
The value at the midpoint in a set of observed values



Left Skew



Normal Distribution

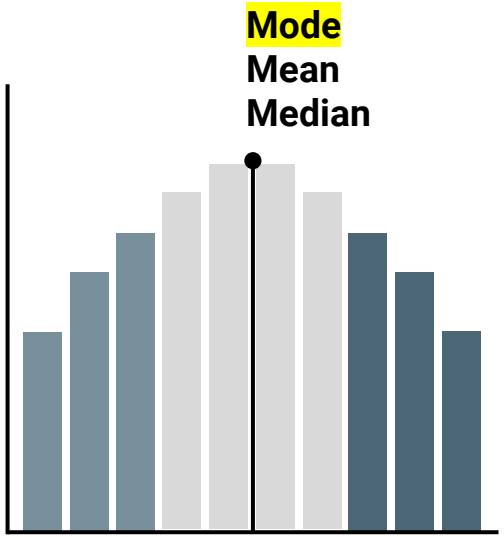


Right Skew

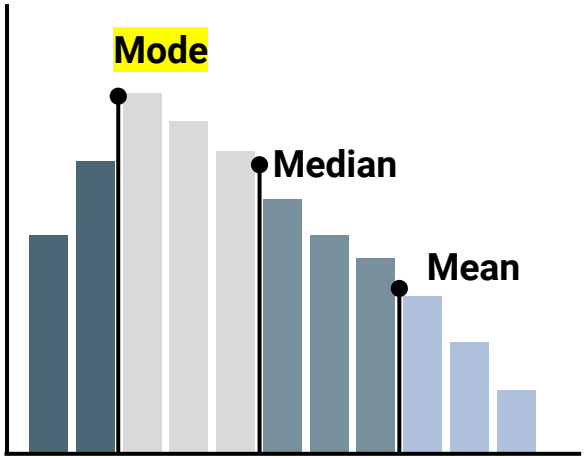
Mode

The most frequently occurring value in a set of values

Left Skew



Normal Distribution



Right Skew

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 n

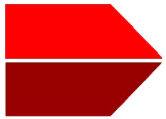
Extreme values may not always be reliable

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.

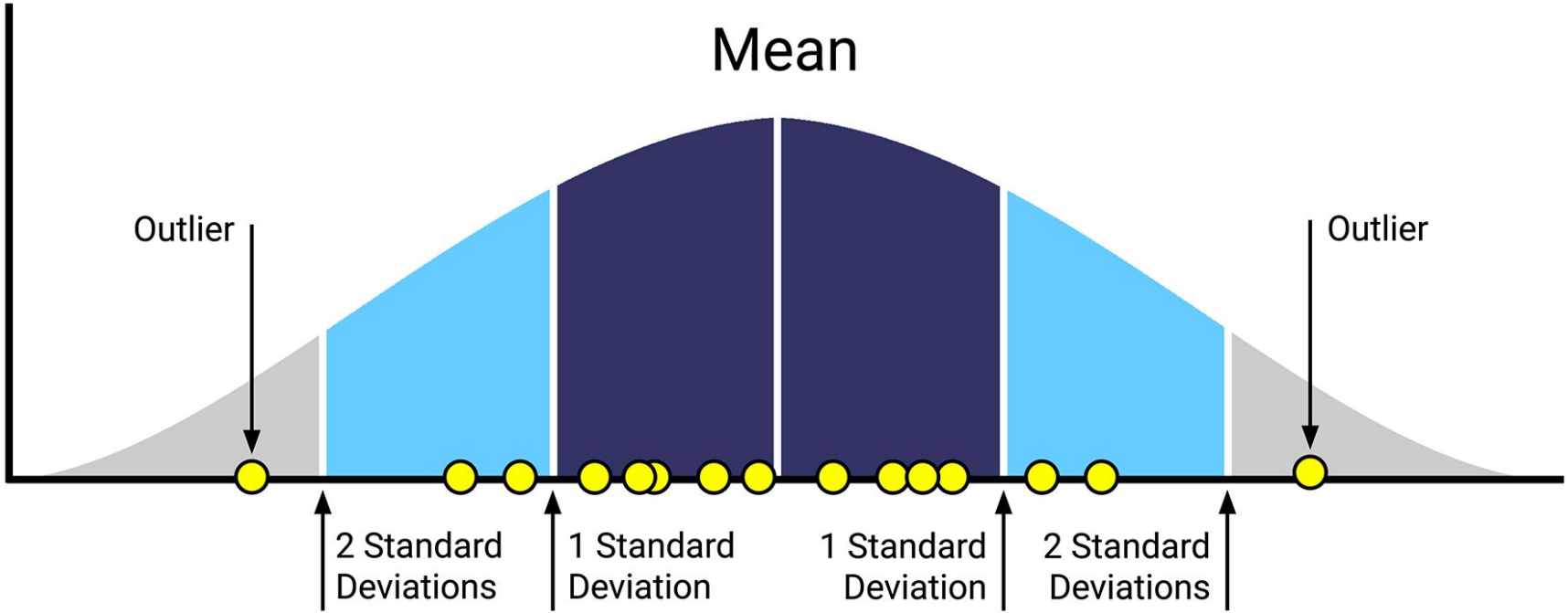
Standard Deviation

- 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

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

Standard Deviation

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

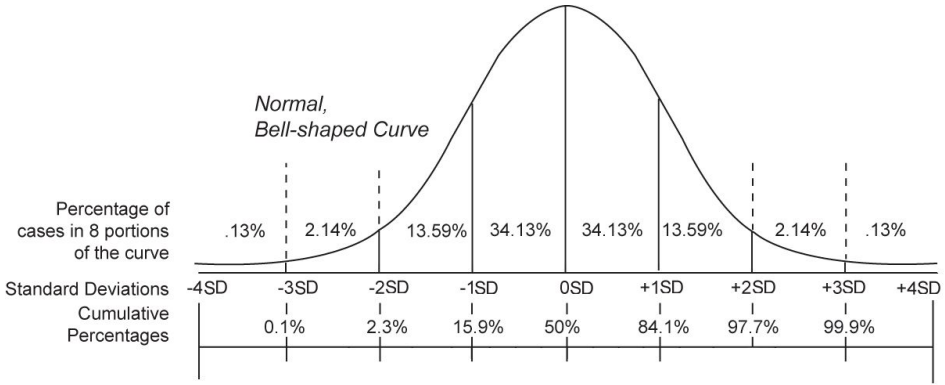


Normal distribution: A probability distribution around the mean, where data near the mean are more frequent in occurrence than data far from the mean - like a “bell curve”.

Normal Distribution Features

Normal distributions are:

- **Symmetric**
- **Unimodal** *(bell-shaped curve with one bump in the middle)*
- **Asymptotic** *(tails approach but never touch the x-axis)*
- **Mean == Median == Mode**



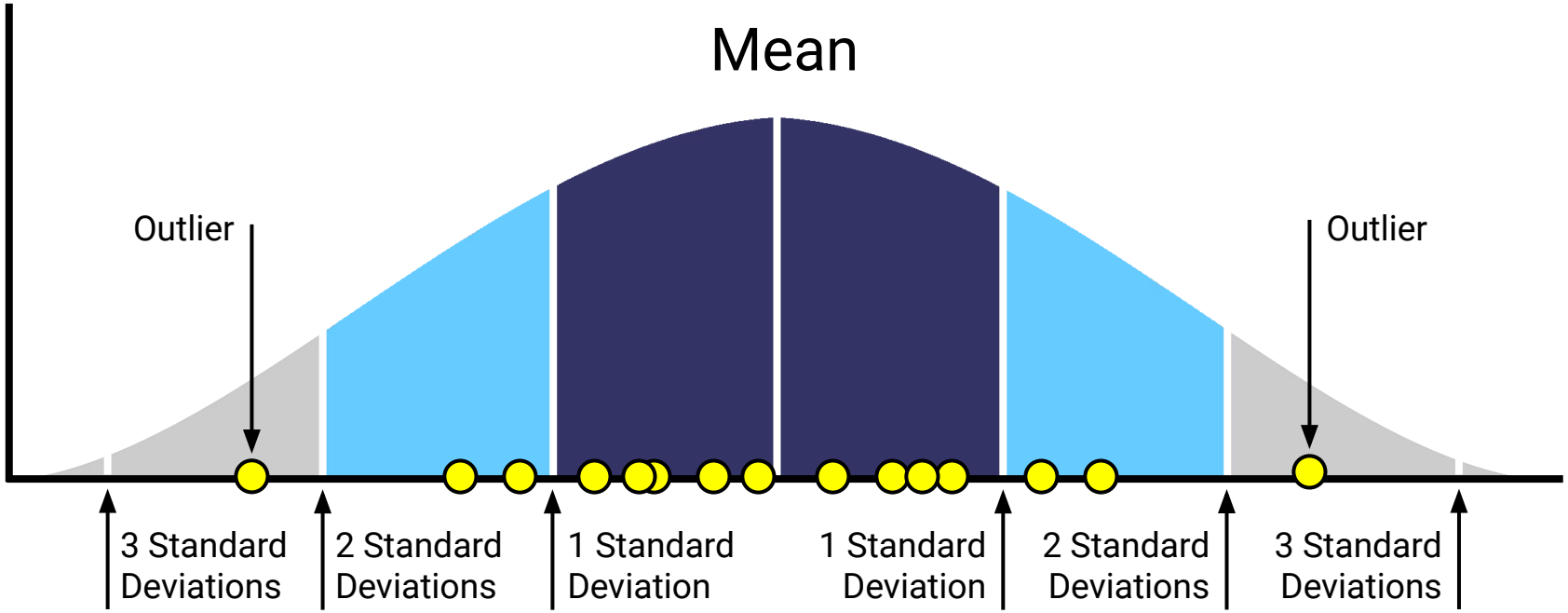
A lot of data that occur naturally can be approximated by a normal distribution, e.g.:

- Population Height
- Population Shoe Size
- Blood pressure of adult humans
- IQ Scores
- Measurement error

A normal distribution is also known as a **Gaussian Distribution** (named after Carl Friedrich Gauss)

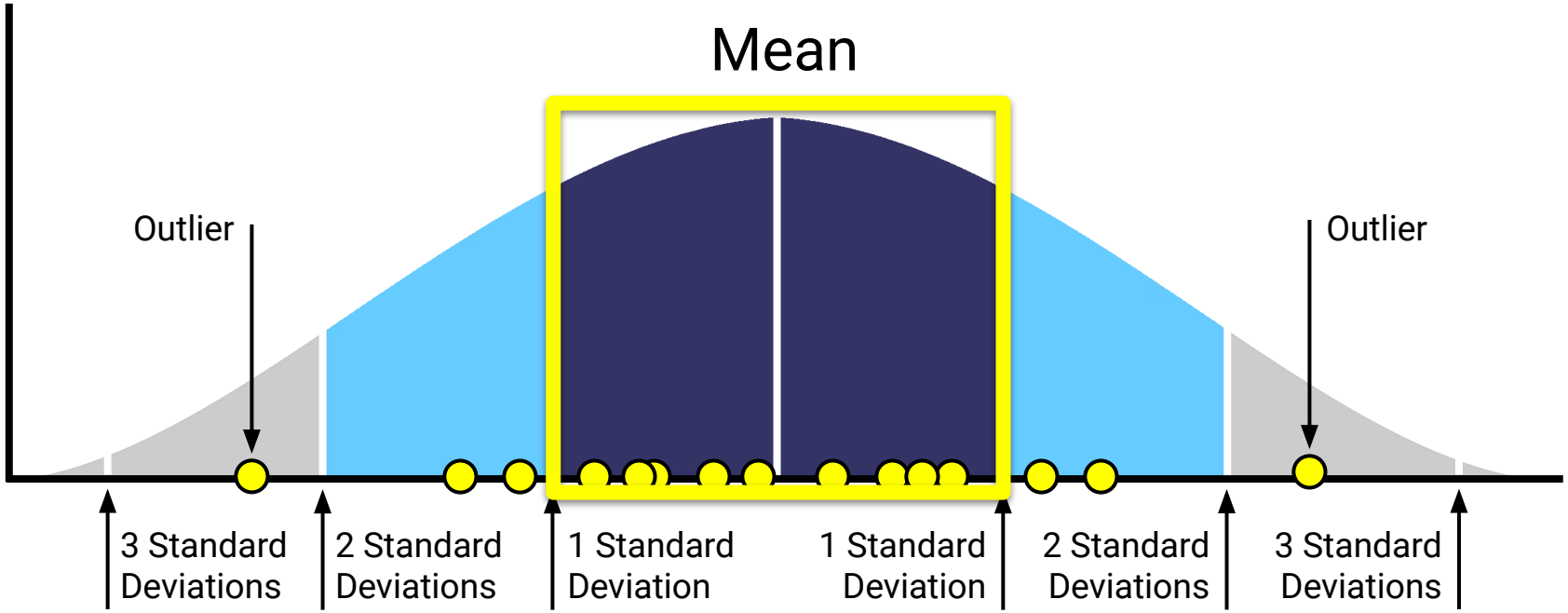
Normal Distribution Features

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



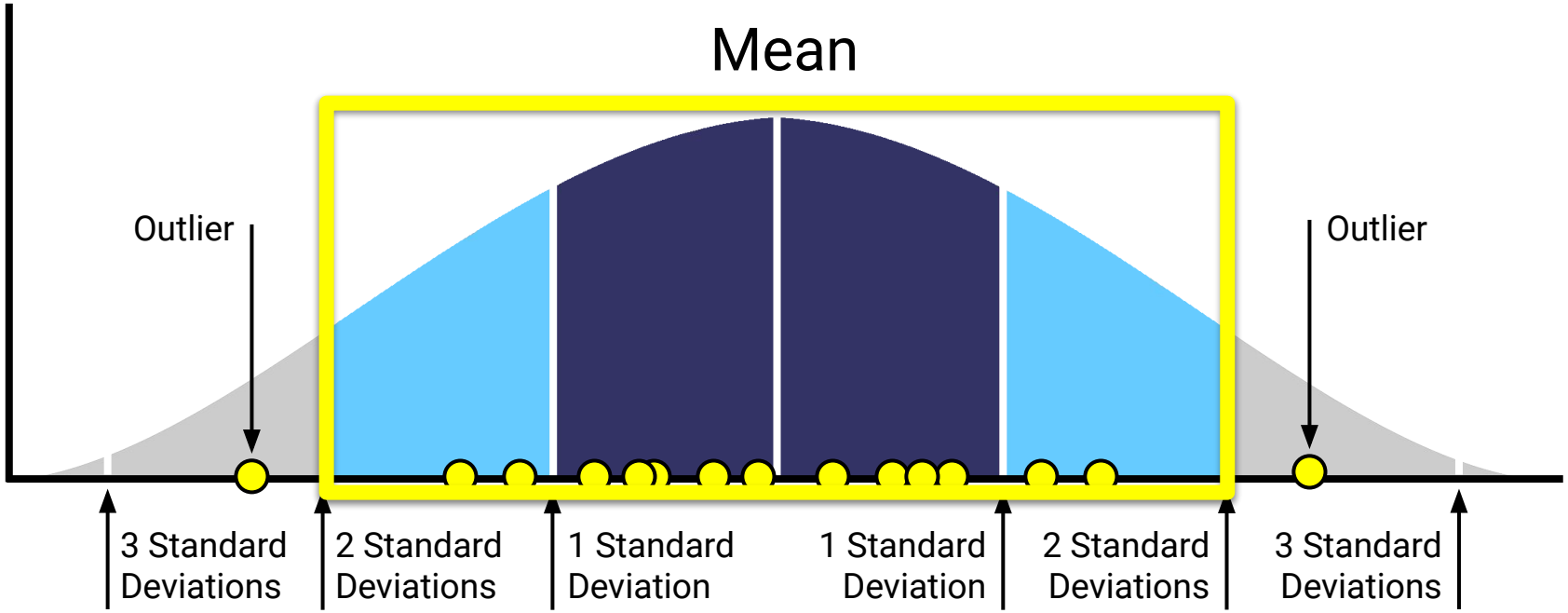
Normal Distribution Features

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



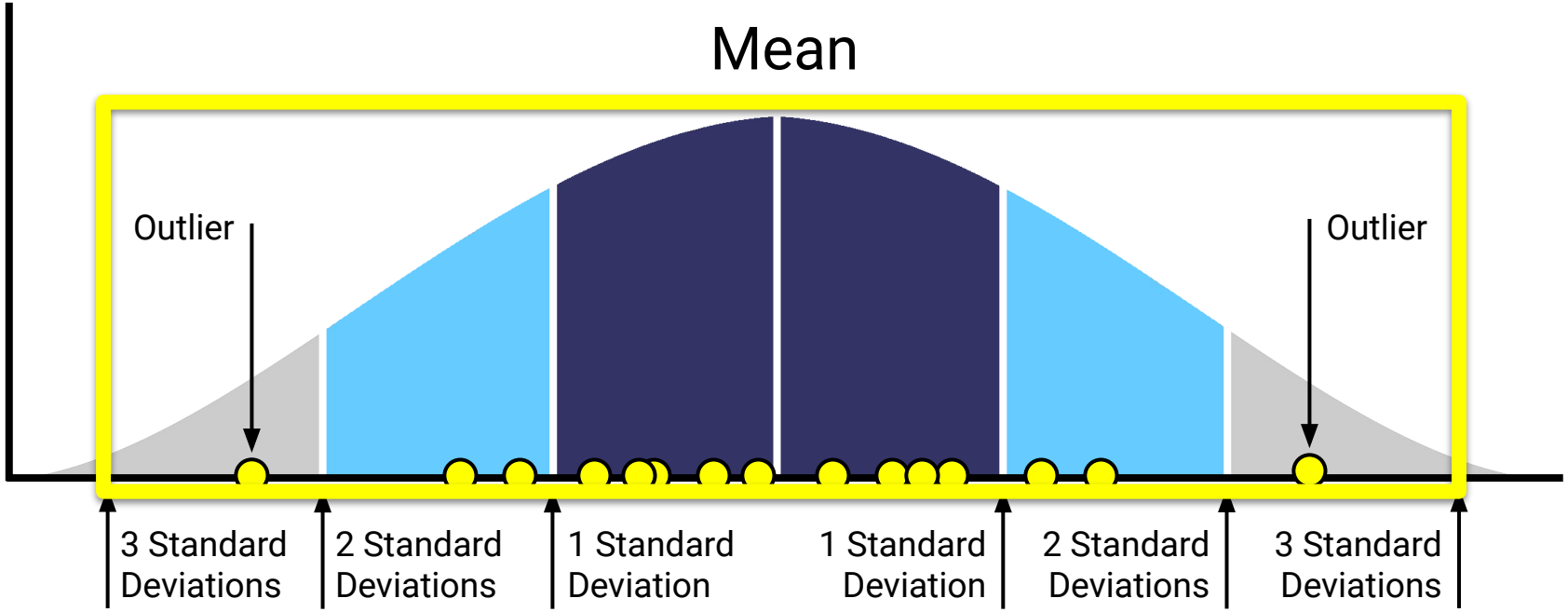
Normal Distribution Features

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



Normal Distribution Features

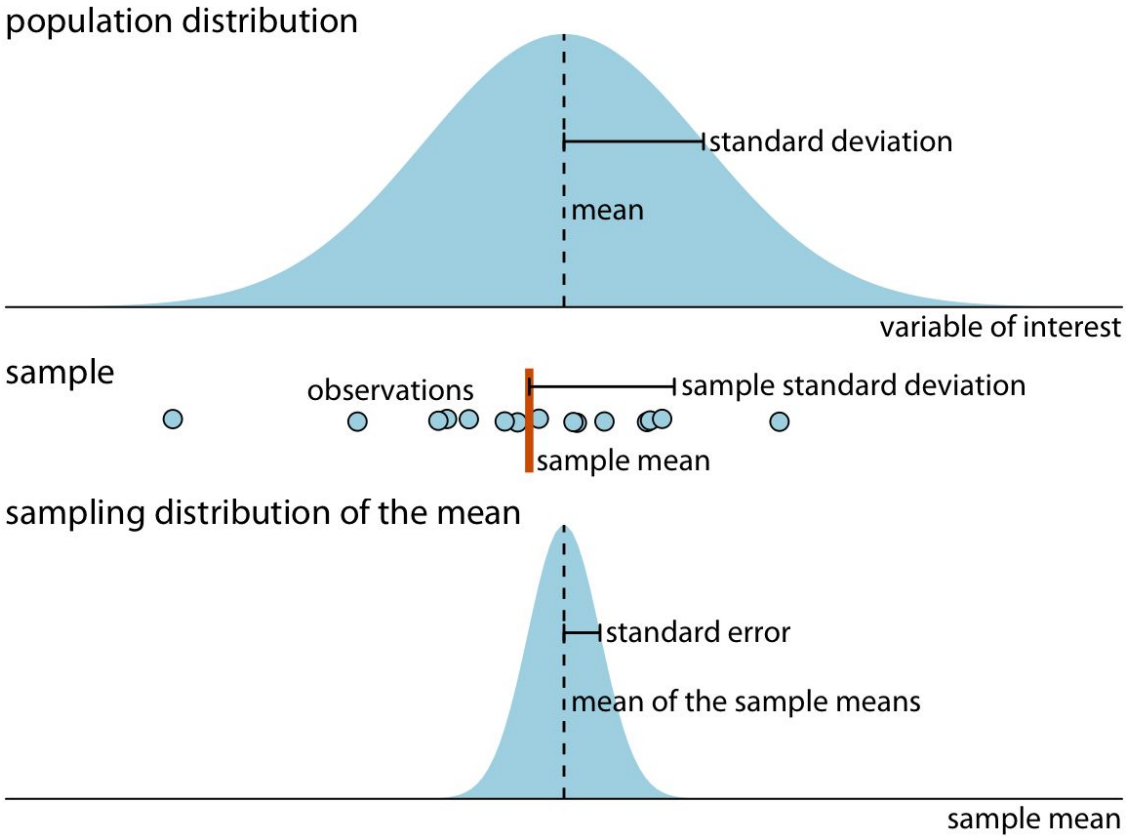
99.7% of the data fall within 3 standard deviations from the mean



Standard error: An estimate of how far the sample mean is likely to be from the population mean

Standard error

Standard error describes the error that occurs with the **random sampling** of a population. (i.e. sampling error)



Calculating Standard error

Standard deviation of the population / square root of the sample size. In practice, because we don't know the population standard deviation, sample standard deviation is used for this equation.

Standard error

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

Standard deviation

Size of the sample

Standard error decreases as the size of the sample increases

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

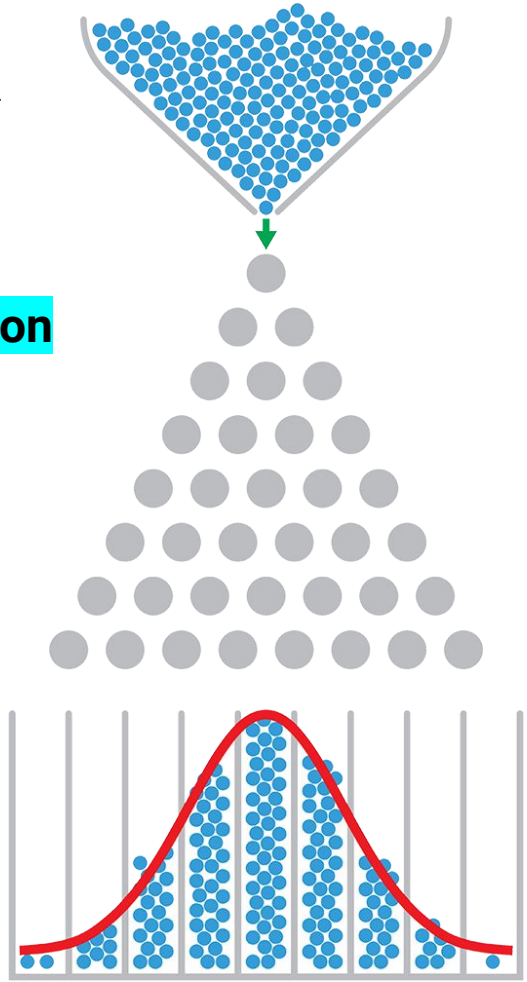
Central Limit Theorem

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

Sample Standard Deviation

Population Standard Deviation

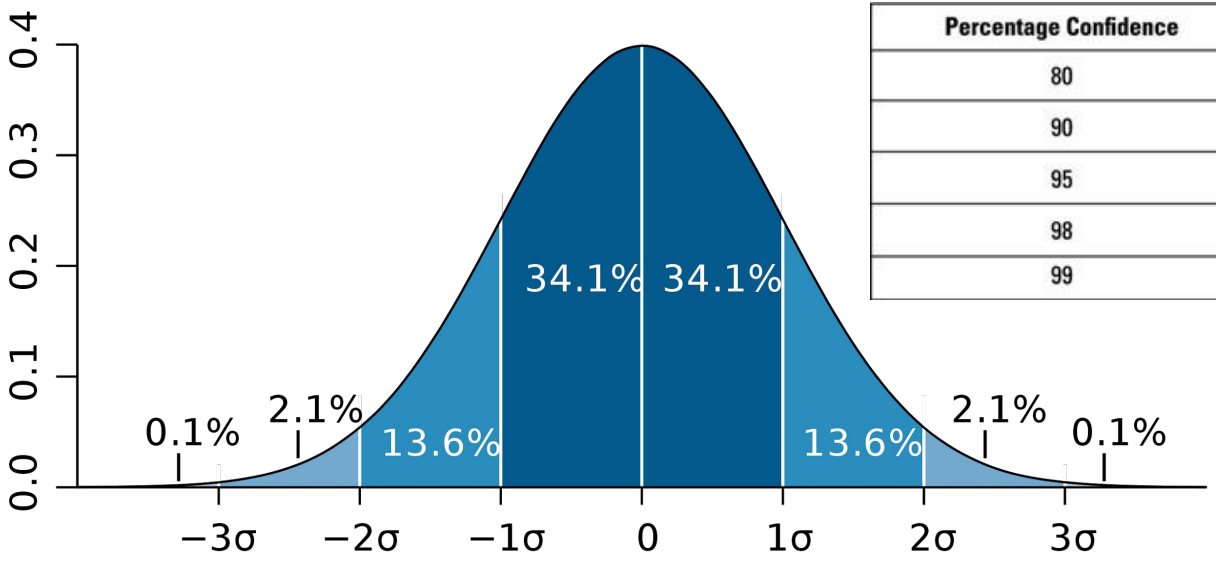
Sample Size



Confidence Interval: An estimated range of values which is likely to include an unknown population parameter

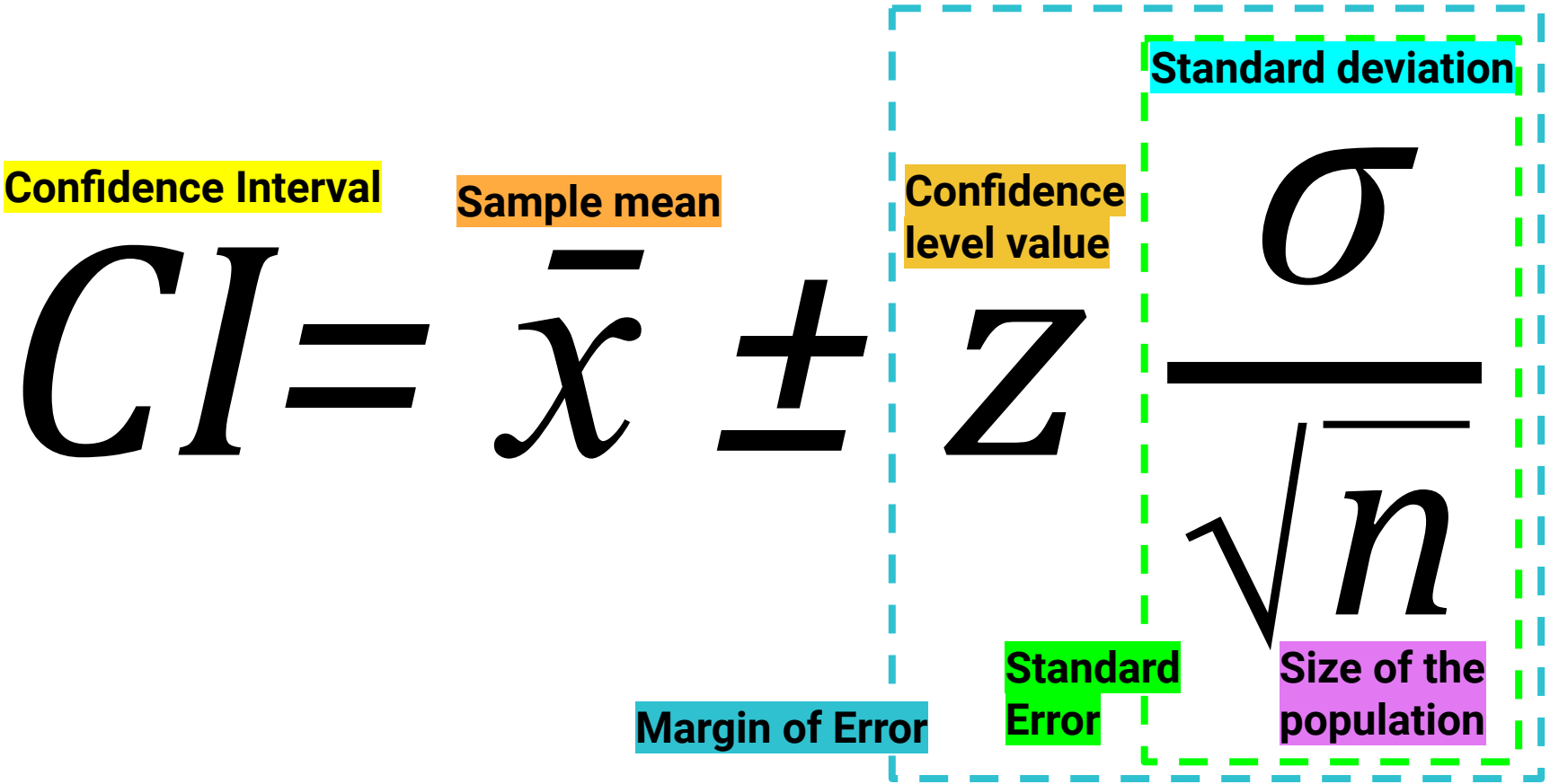
Confidence Interval

If we can assume that the sampling distribution of the mean is normally distributed, we can calculate a range of plausible values for an unknown parameter (for example, the mean). The this range is called the confidence interval, and has an associated confidence level that the true parameter is in the proposed range.



Percentage Confidence	*z**-Value
80	1.28
90	1.645
95	1.96
98	2.33
99	2.58

Calculating Confidence Interval



Estimator: An estimator is a statistic that estimates some fact about the population

Estimator / Estimand

The quantity that is being estimated is called the **estimand**, while the statistic that estimates the quantity is the **estimator**.

- E.g. the sample mean ($\bar{x}$) is an estimator for the population mean (μ)
- Estimators can be a range of values (called an **interval estimate**, e.g. confidence interval), or a single value (called a **point estimate**, e.g. standard deviation)

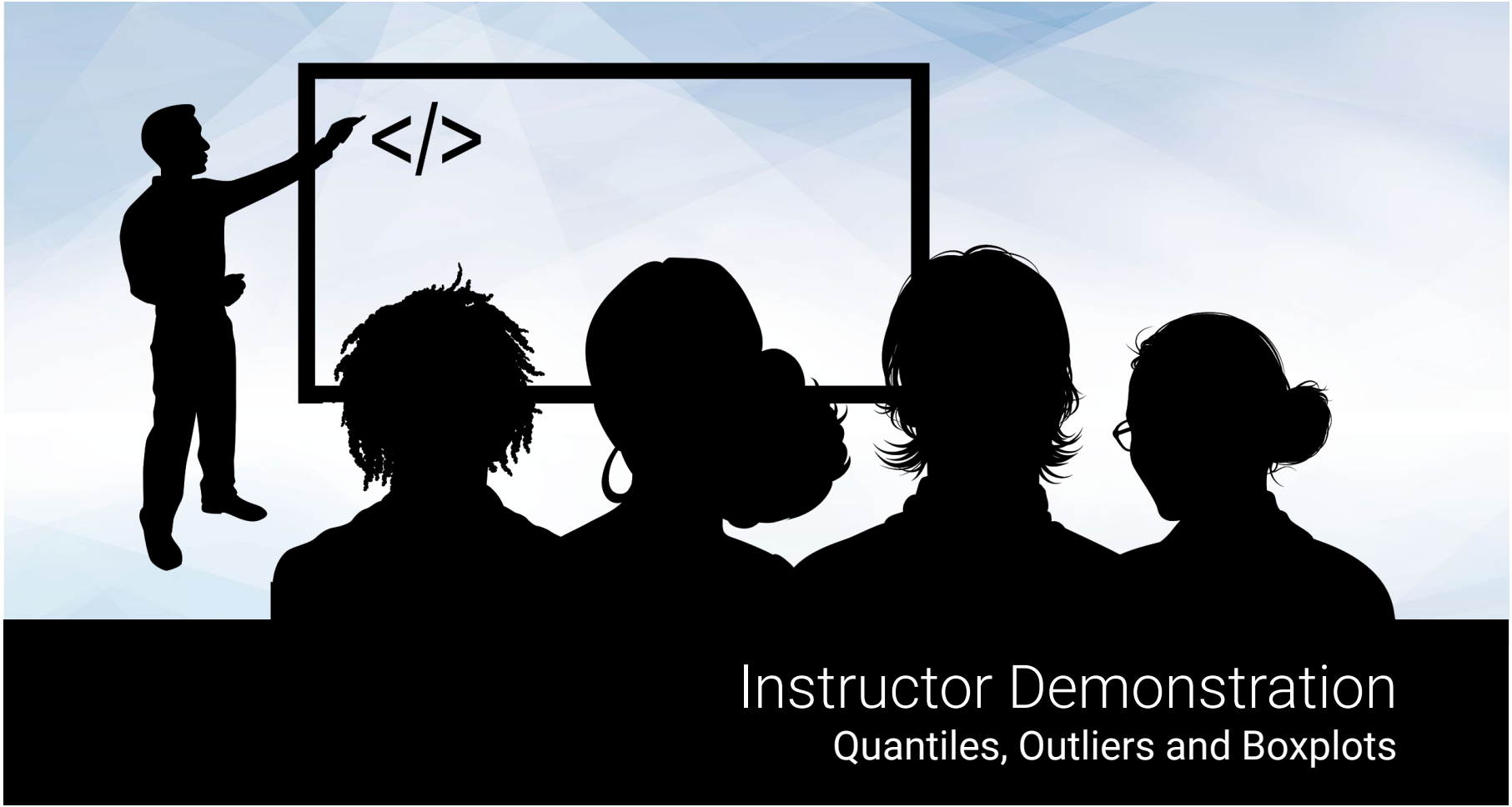
Estimator

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 (%)
10000000000	95	5

Sample size:

385



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?

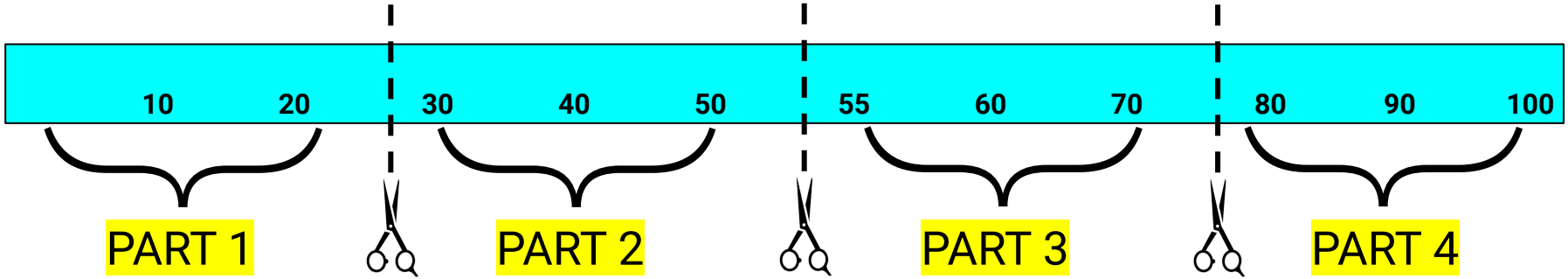
We can use quantiles to describe segments of a data set!

Quantiles separate a sorted data set into equally sized fragments.

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

01 **Quartiles** divide the data set into four equal parts

02 **Percentiles** divide the data set into 100 equal parts

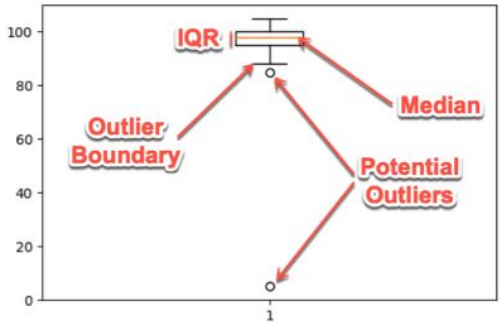


< Demo Time >

There are two ways to identify potential outliers

01 Qualitatively

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

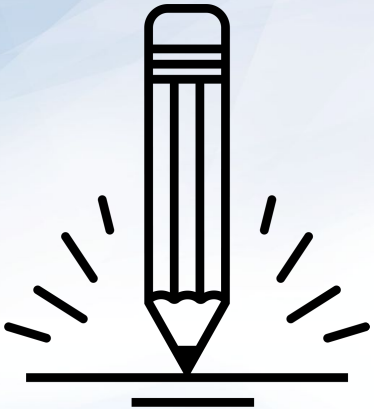


02 Quantitatively

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

< Demo Time >



Activity: Outliers—Drawn and Quartiled

Suggested Time:
15 Minutes

Variance, standard deviation, and Z-score review instructions

Instructions:

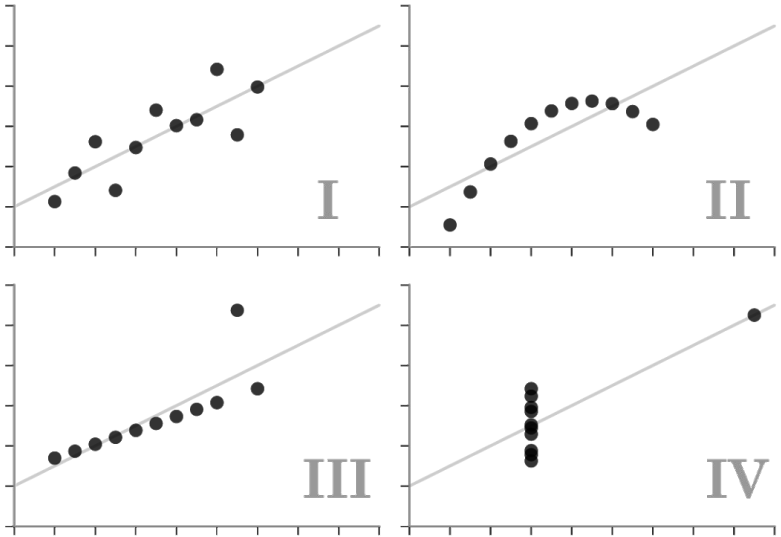
- Open up the activity workbook and familiarize yourself with the raw data.
  - File: `Unsolved/Outliers_Activity_Unsolved.xlsx`
- Create a new worksheet and name it ‘Outlier Testing’.
- In the ‘Outlier Testing’ worksheet, create a summary statistics table of the Antioxidant_content_in_mmol_100g for 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.

Suggested Time: 15 minutes

Importance of Visualizing Data / Weakness in Summary Stats.

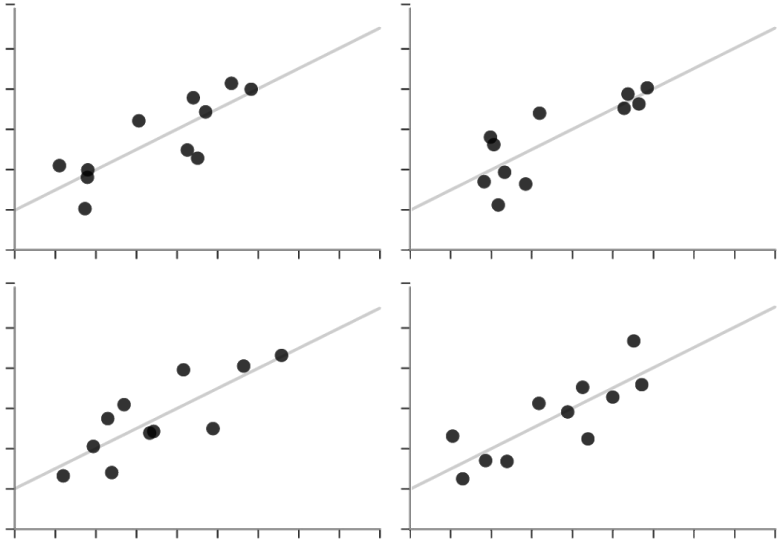
Anscombe's Quartet

Each dataset has the same summary statistics (mean, standard deviation, correlation), and the datasets are *clearly different*, and *visually distinct*.

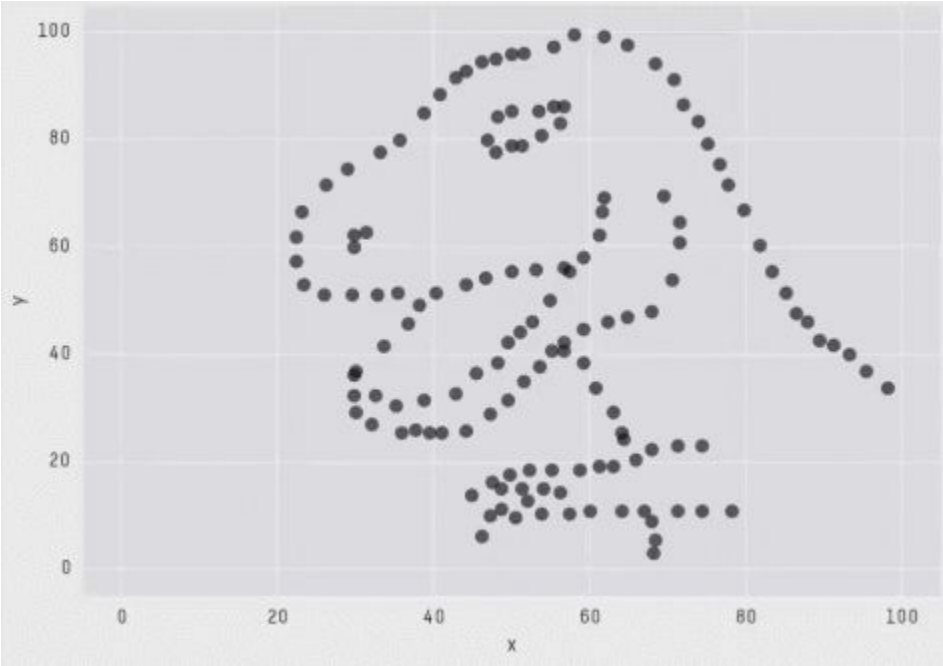


Unstructured Quartet

Each dataset here also has the same summary statistics. However, they are not *clearly different* or *visually distinct*.



Importance of Visualizing Data - Cont’



X Mean: 54.2659224
Y Mean: 47.8313999
X SD : 16.7649829
Y SD : 26.9342120
Corr. : -0.0642526