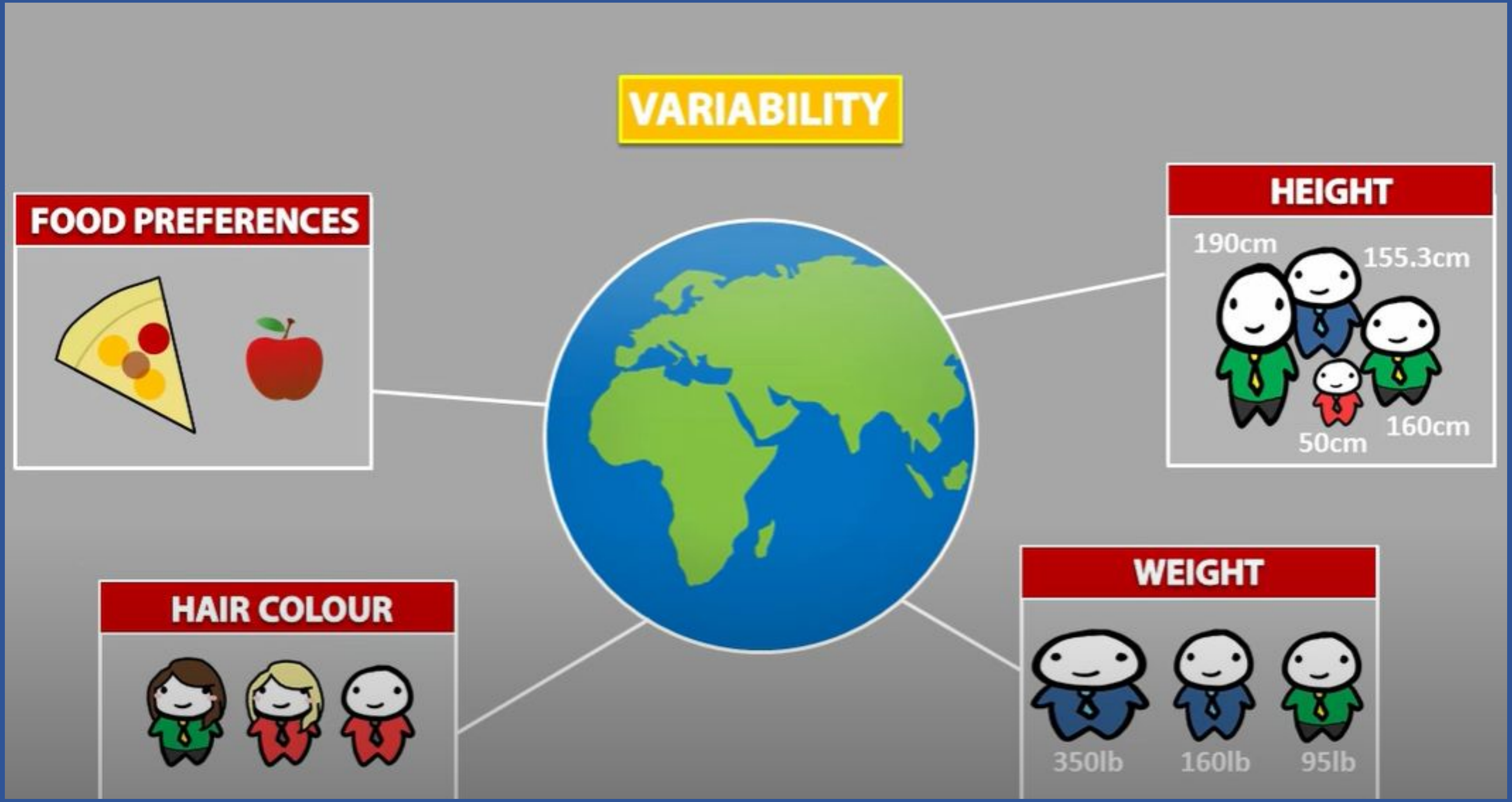
What is STATISTICS?

- Statistics is a branch of mathematics that deals with the collection, analysis, interpretation, presentation, and organization of data.
- It provides a systematic approach for making sense of data and drawing meaningful conclusions from it.

STATISTICS: MEASURE AND ANALYSE VARIABILITY



Why Statistics is Essential?

- □ Summarize Data
- □ Identify Patterns and Trends
- □ Make Inferences
- □ Test Hypotheses
- □ Predict Future Outcomes
- □ Reduce Uncertainty
- □ Scientific Research

Statistics has so many use cases!

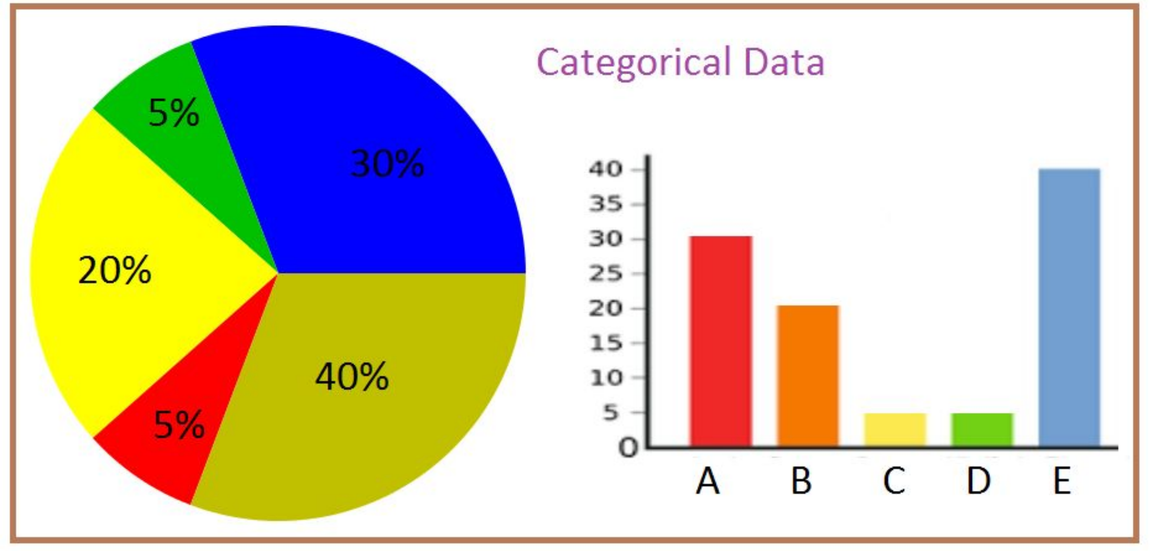
Types of Data

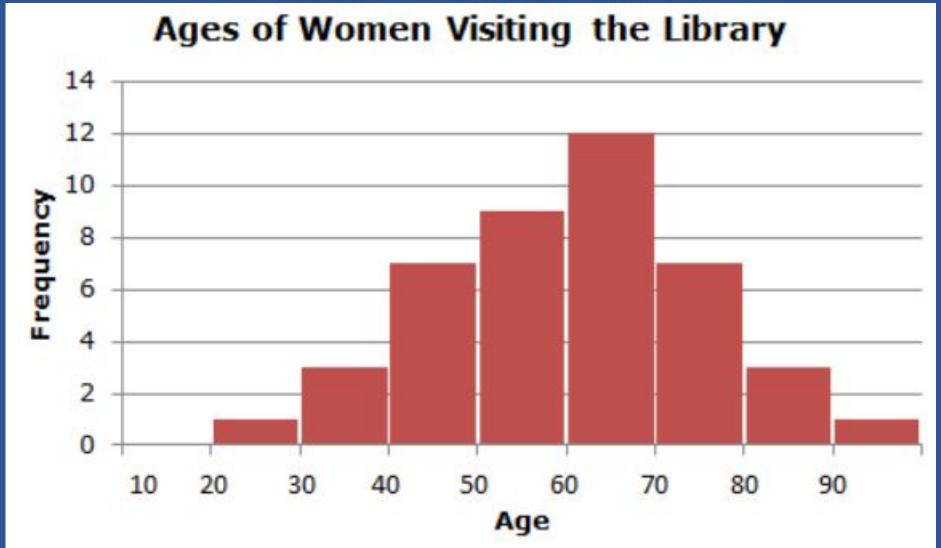
1. **Categorical Data:**

Categorical data, also known as qualitative or nominal data, represent categories, labels, or distinct groups that do not have a natural numerical order. These data points can be sorted into discrete categories or classes.

2. **Numerical Data:**

Numerical data, also known as quantitative or continuous data, represent measurements or quantities that have a natural numerical order. These data points can be expressed as numbers and can take on a wide range of values.





Types of Categorical Data

Categorical data can further be divided into two subtypes:

i. Nominal Data: These categories have no inherent order or ranking. For example, the types of cars listed above are nominal because there's no meaningful order among them.

ii. Ordinal Data: Ordinal data categories have a clear order or ranking, but the differences between them are not necessarily uniform. For instance, education levels (High School, Bachelor's, Master's, PhD) have an order, but the difference in "educational level" between High School and Bachelor's is not the same as between Bachelor's and Master's.

Types of Numerical Data

Numerical data can be further categorized into two subtypes:

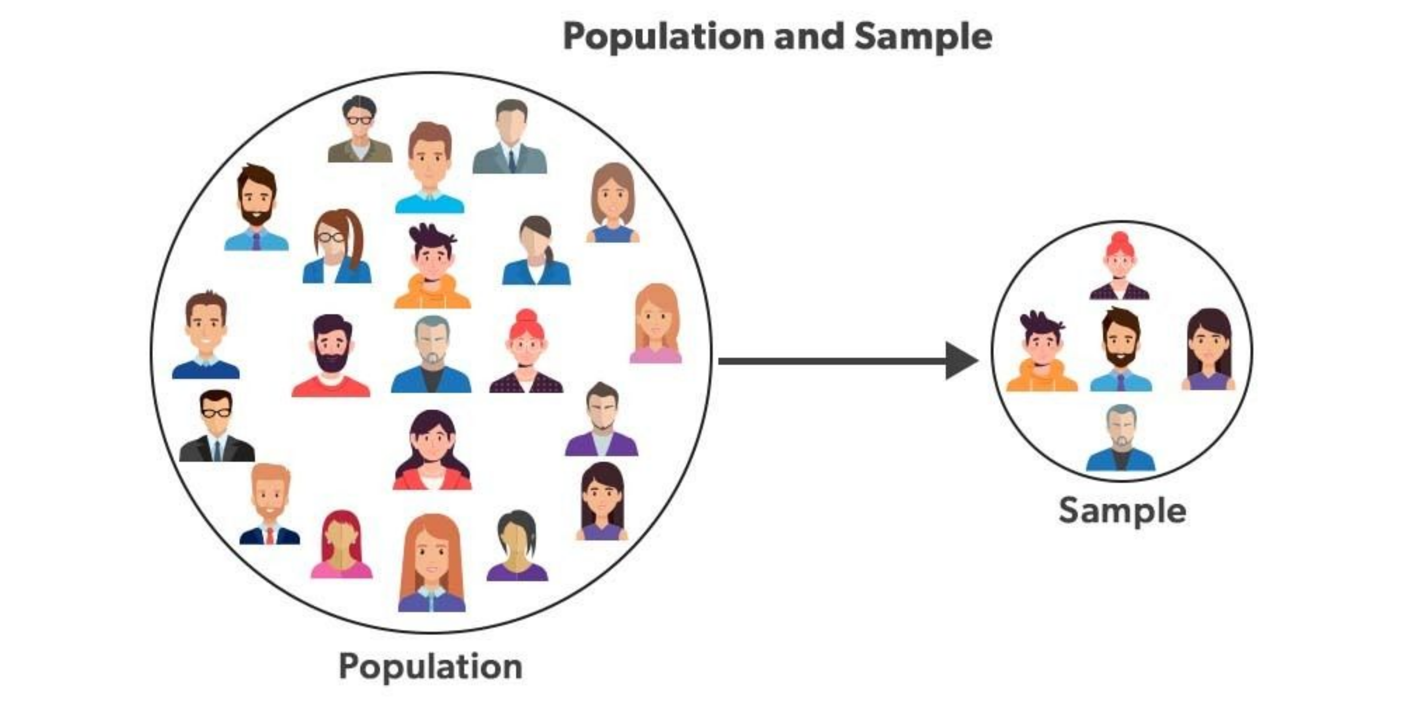
i. Discrete Data: Discrete data can only take on specific, distinct values, often counted in whole numbers. For example, the number of cars in a parking lot is discrete because you can't have a fraction of a car.

ii. Continuous Data: Continuous data can take on an infinite number of values within a given range. For example, temperature can be measured with decimal precision (e.g., 23.5°C), making it continuous.

Data Collection for Statistics

- ❑ **Population**:
  The population refers to the entire set of individuals, objects, or observations about which you want to make inferences or draw conclusions in a statistical study. It represents the complete group you are interested in studying, and it can be finite or infinite.

- ❑ **Sample**:
  A sample is a subset of the population, selected in a way that it is representative of the larger population. It is a smaller group of individuals, objects, or observations that is chosen for the purpose of collecting data and making inferences about the population as a whole.



How to pickup a sample?

- □ **Random Sampling:**

Random sampling is a method where each member of the population has an equal chance of being selected for the sample. It involves selecting individuals or items from the population in a completely random and unbiased manner.

- □ **Stratified Sampling:**

Stratified sampling is a method in which the population is divided into subgroups or strata based on certain characteristics that are important to the study. Then, a random sample is selected from each stratum. This method ensures that each subgroup is adequately represented in the final sample.

Measures of Central Tendency

Measure of Central Tendency

Measures of central tendency are statistical measures that describe the center or average value of a dataset. They provide insights into the typical or central value around which data points tend to cluster.

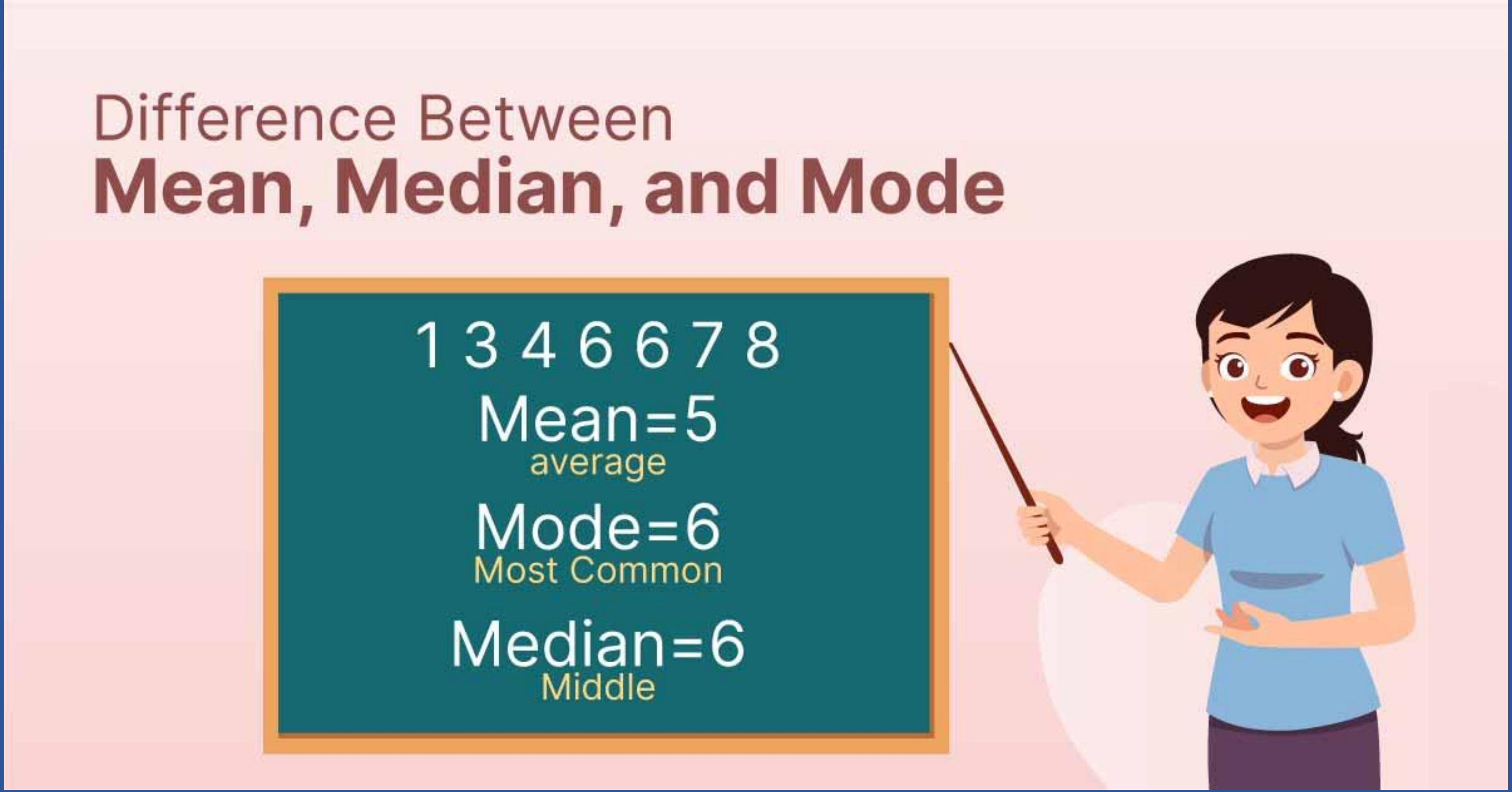
There are three main measures of central tendency:

❑ Mean: Represents the **average** value in a given dataset

❑ Median: Denoted by **middle** value present in the ordered dataset.

❑ Mode: Denotes **frequency** of occurrence of elements.

Measure of Central Tendency



Difference Between

Mean, Median, and Mode

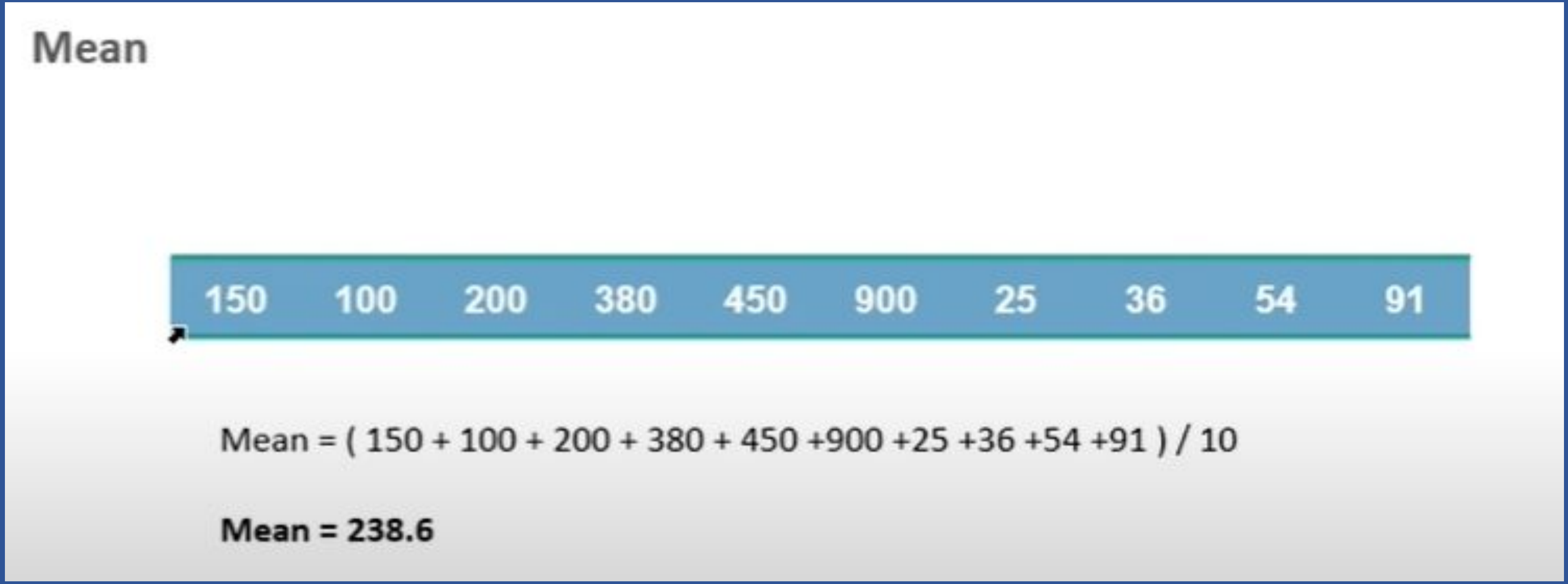
1 3 4 6 6 7 8

Mean=5
average

Mode=6
Most Common

Median=6
Middle

MEAN



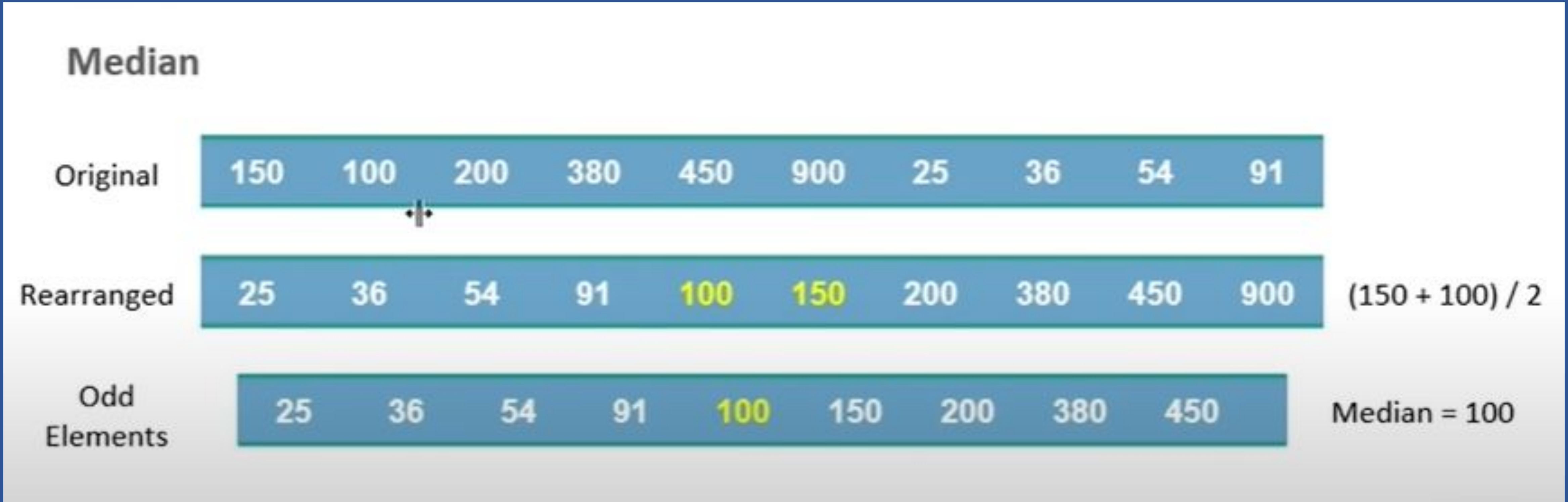
Mean

150	100	200	380	450	900	25	36	54	91

Mean = (150 + 100 + 200 + 380 + 450 +900 +25 +36 +54 +91) / 10

Mean = 238.6

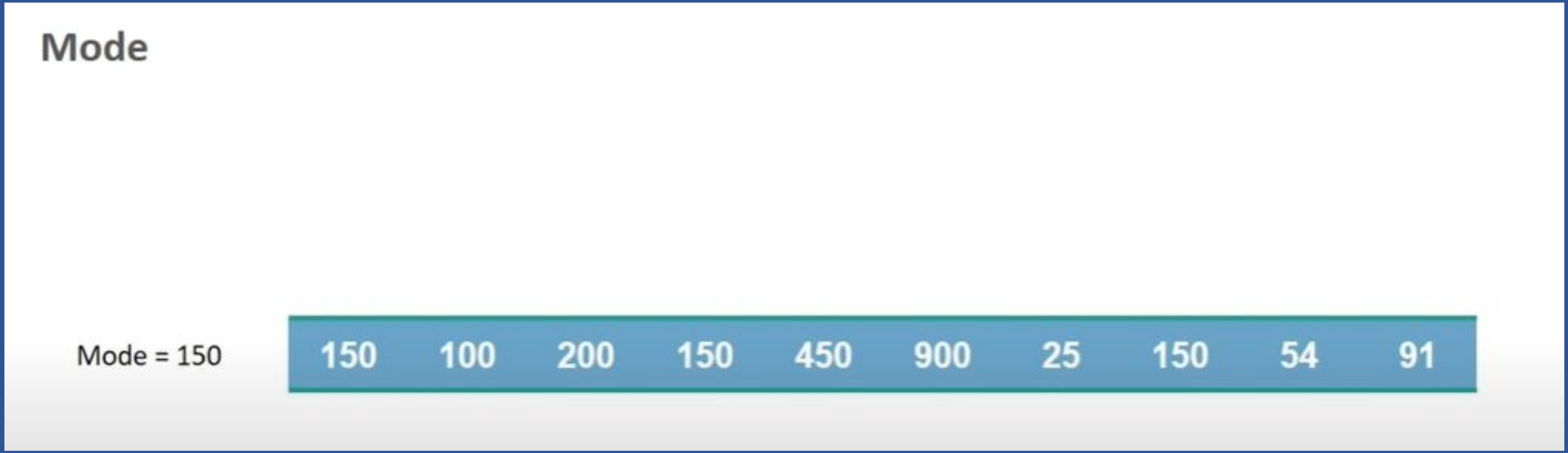
MEDIAN



Median

Original	150	100	200	380	450	900	25	36	54	91	
Rearranged	25	36	54	91	100	150	200	380	450	900	(150 + 100) / 2
Odd Elements	25	36	54	91	100	150	200	380	450		Median = 100

MODE



Mode

Mode = 150

150	100	200	150	450	900	25	150	54	91

Measures of Dispersion

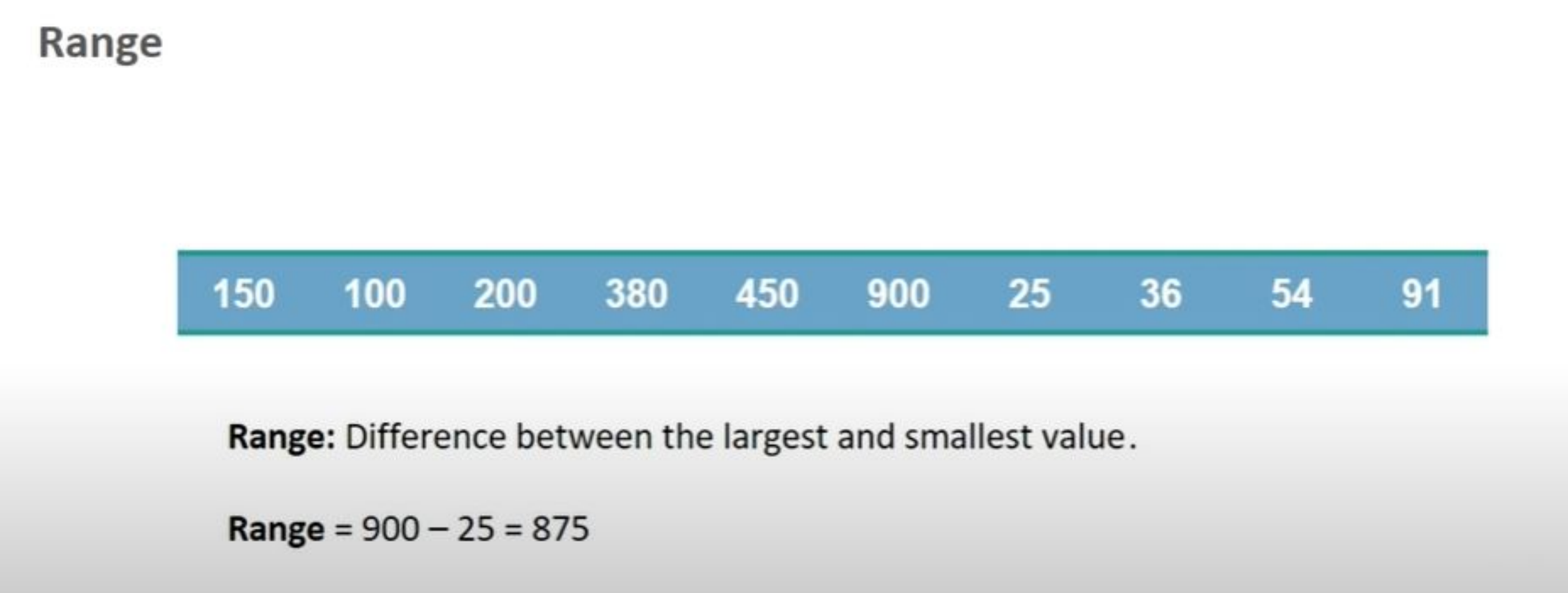
Measure of Dispersion

Measures of dispersion, also known as **measures of variability or spread**, are statistical measures that describe the extent to which data points in a dataset vary or spread out from the central value They provide important insights into the degree of variability within a dataset.

There are several common measures of dispersion including :

- ❑ **Range**: Calculated as the difference between the maximum and minimum values in a dataset. It provides a basic idea of how spread out the data is.
- ❑ **Variance**: It measures the average of the squared differences between each data point and the mean of the dataset. It quantifies how data points deviate from the mean.
- ❑ **Standard deviation**: It is a widely used measure of dispersion that is simply the square root of the variance. It provides a measure of the average amount of deviation or spread in a dataset.
- ❑ **Interquartile range(IQR):**It is a measure of variability based on the quartiles of a dataset. It represents the range between the 25th percentile (Q1) and the 75th percentile (Q3) of the data, capturing the spread of the middle 50% of the data.

RANGE



Range

150	100	200	380	450	900	25	36	54	91

Range: Difference between the largest and smallest value.

Range = 900 – 25 = 875

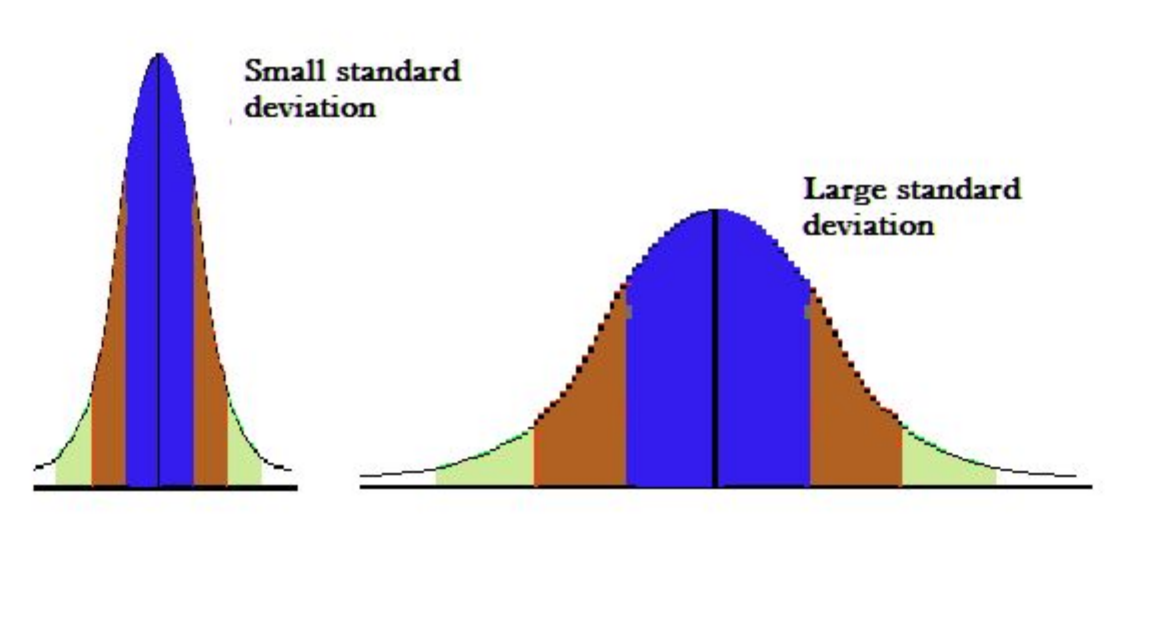
VARIANCE & STANDARD DEVIATION

Variance

Variance is defined as, “The measure of how far the set of data is dispersed from their mean value”. Variance is represented with the symbol **σ^2**. In other words, we can also say that the variance is the average of the squared difference from the mean.

Standard Deviation

How far our given set of data varies along with the mean of the data is measured in standard deviation. Thus, we define standard deviation as the “spread of the statistical data from the mean or average position”. We denote the standard deviation of the data using the symbol **σ**.
We can also define the standard deviation as the square root of the variance.



MEN
Time to measure your hair!
WOMEN

MEN
I'm no scientist but it seems most men have similar hair sizes
WOMEN

Men Group has lower standard deviation(most of them have hair length near to average hair length among men)

While

Women group has a lot of Variation or Standard Deviation

VARIANCE & STANDARD DEVIATION

All possible outcomes of rolling a die are {1; 2; 3; 4; 5; 6}, total 6 values

Before finding the variance, we need to find the mean of the data set.

Mean, $\overline{x} = (1+2+3+4+5+6)/6 = 3.5$

We can put the value of data and mean in the formula to get;

⇒ $\sigma 2 = [(1-3.5)2 + (2-3.5)2 + (3-3.5)2 + (4-3.5)2 + (5-3.5)2 + (6-3.5)2]/6$

⇒ $\sigma 2 = (6.25+2.25+0.25+0.25+2.25+6.25)/6$

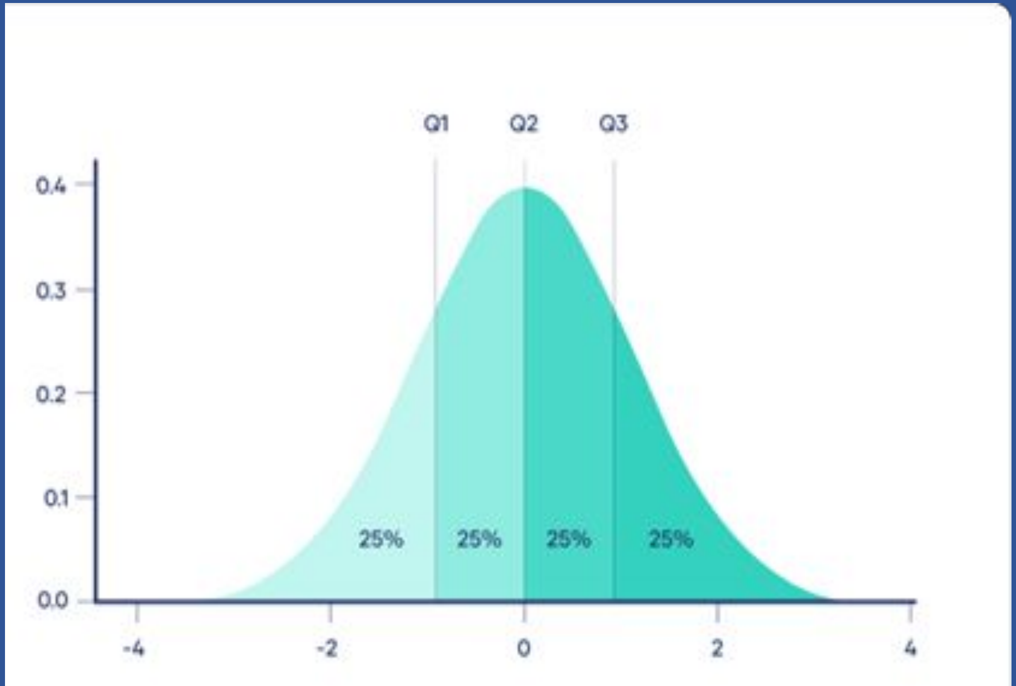
Variance **($\sigma 2$) = 2.917**

Standard Deviation $(\sigma) = \sqrt{}\,(\sigma 2)$

⇒ *Standard Deviation* $(\sigma) = \sqrt{(2.917)}$

⇒ ***Standard Deviation*** **(σ) = 1.708**

5 Number Summary

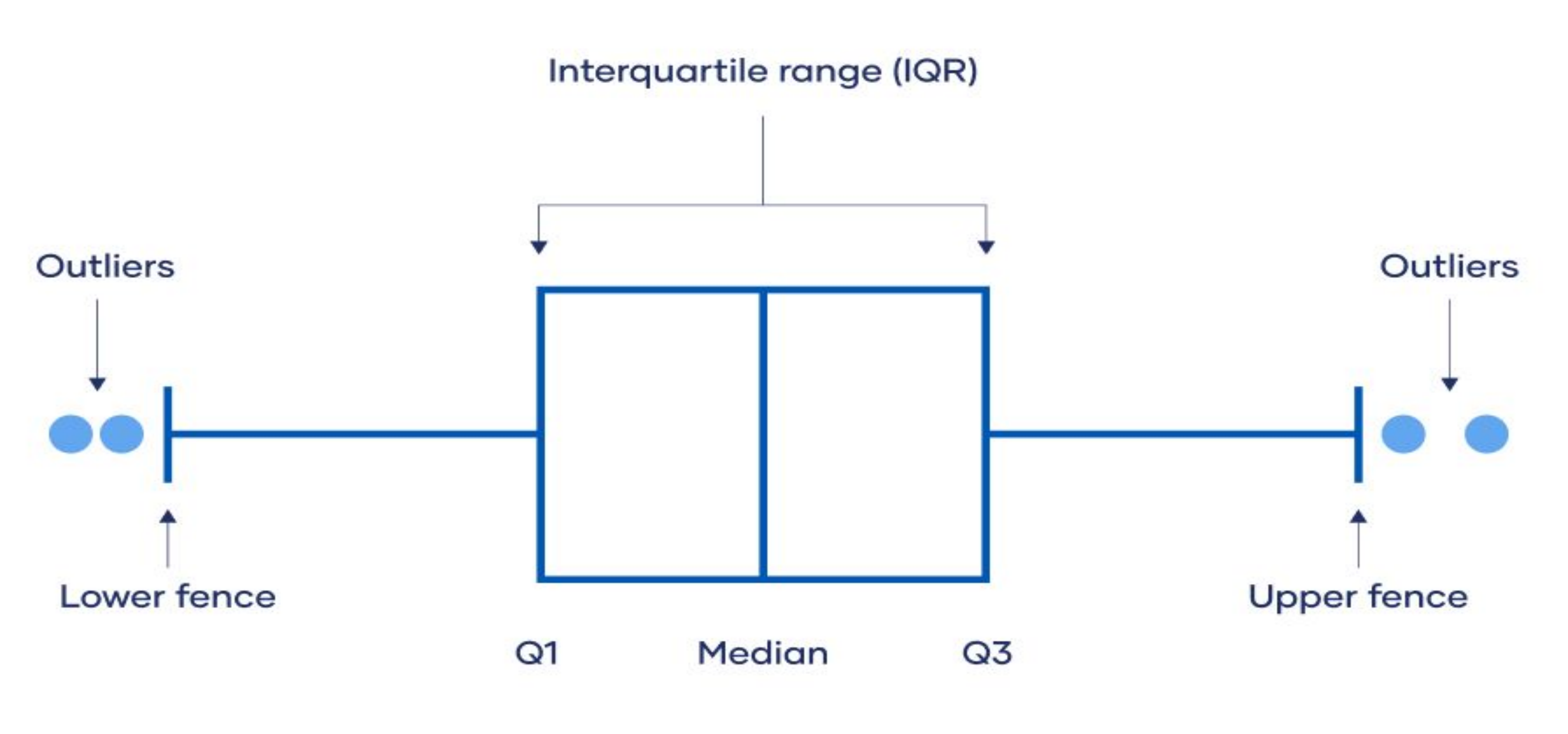


- The minimum - This is the lowest observation also called the zeroth quartile.
- The first quartile- The 25th percentile, which is the value below which 25% of the data falls.
- The second quartile(MEDIAN)- The middle value of the dataset when it's sorted, where 50% of the data falls below and 50% falls above.
- The third quartile- The 75th percentile, which is the value below which 75% of the data falls.
- The maximum - This is the highest observation also called the fourth quartile.

Inter Quartile Range(IQR)

The interquartile range is a measure of variability based on the quartiles of a dataset. It represents the range between the 25th percentile (Q1) and the 75th percentile (Q3) of the data, capturing the spread of the middle 50% of the data.

Box Plot



The interquartile range (IQR) can be used to identify outliers (outside the range of Q1 – (IQR*1.5) and Q3 + (IQR*1.5))