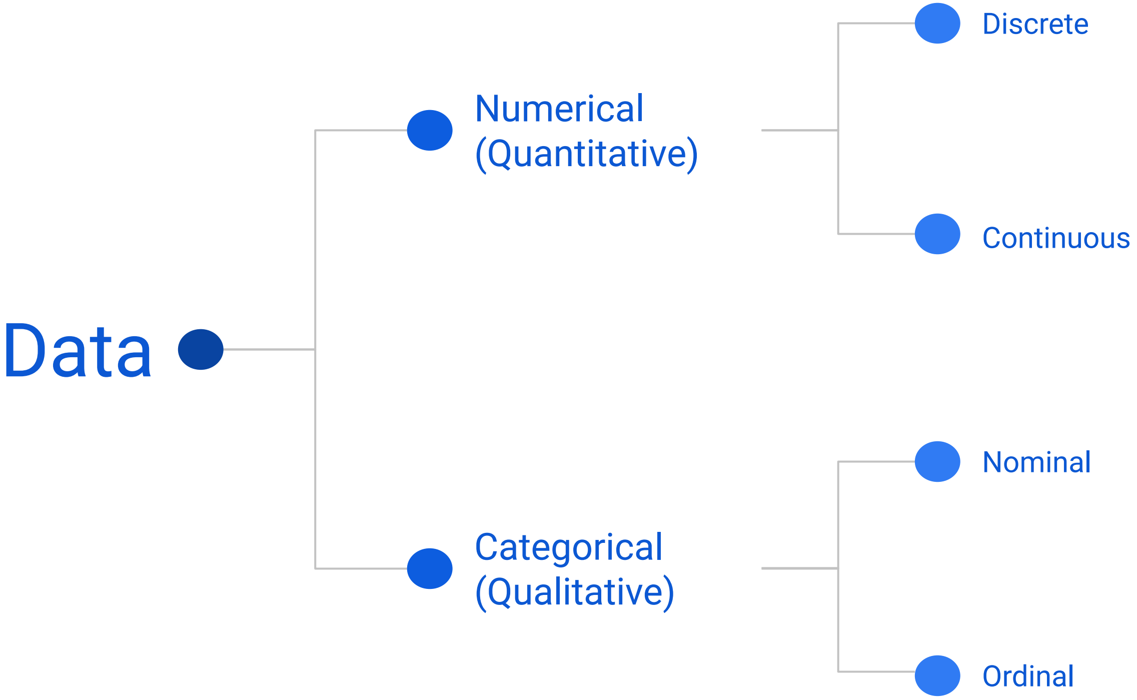
# Basic Statistics

Ramesh S

# What is Data?

- Raw observations alone are data, but they are not information or knowledge.

# Types of Data



Data

Numerical (Quantitative)

Discrete

Continuous

Categorical (Qualitative)

Nominal

Ordinal

# Numerical Data

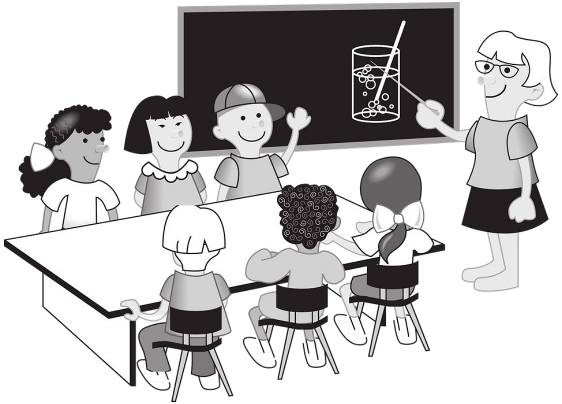
- These data have meaning as a quantity or measurement, such as
  - a person’s height
  - weight
  - IQ
  - blood pressure

***or***

- They’re a count, such as
  - the number of stock shares a person owns
  - how many teeth a dog has
  - how many pages you can read of your favorite book before you fall asleep, etc.

# Discrete Data

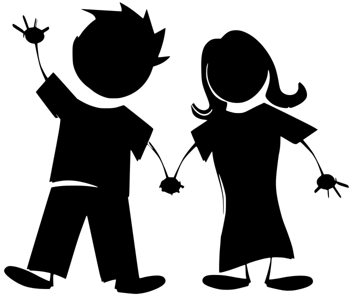
- Discrete data can take only values that can be counted
- They take on possible values that can be listed out.
  - Number of students in a class
  - Number of books in a shelf



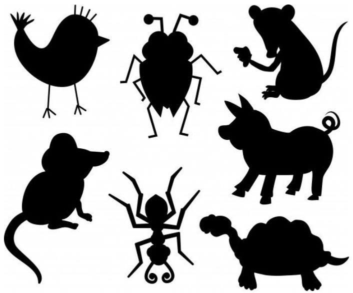
# Continuous Data

- Continuous data can take any value within a range
  - Height
  - Speed

# Categorical Data



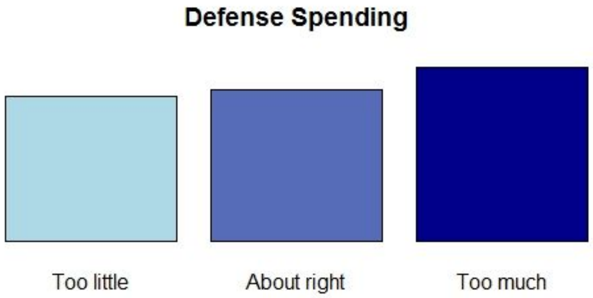
- Categorical data represent characteristics such as
    - a person's gender
    - marital status
    - hometown
    - the types of movies they like, etc.



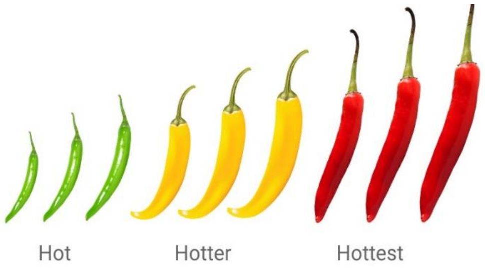
# Nominal Data

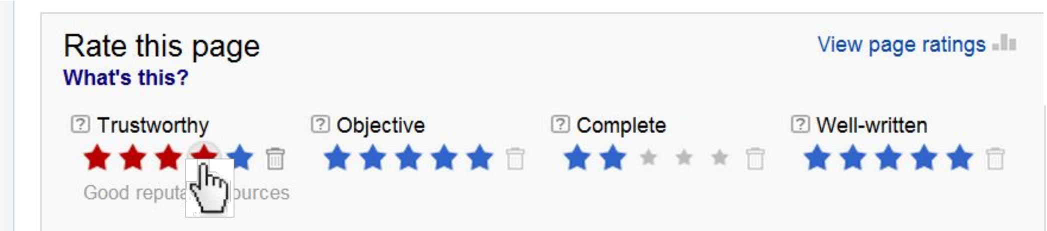
- Nominal means name and count
- Data are alphabetical or numerical in name only
- Categories without order or direction
- Restricted to keep track of people, objects and events
    - Gender
    - Marital Status
    - Any other Yes/No Data

# Ordinal Data



- Ordinal means rank or order
- They place events in order and can be sorted
- Has no absolute value (only relative position in the inequality)
  - Ranks or Grades of students
  - Intensity





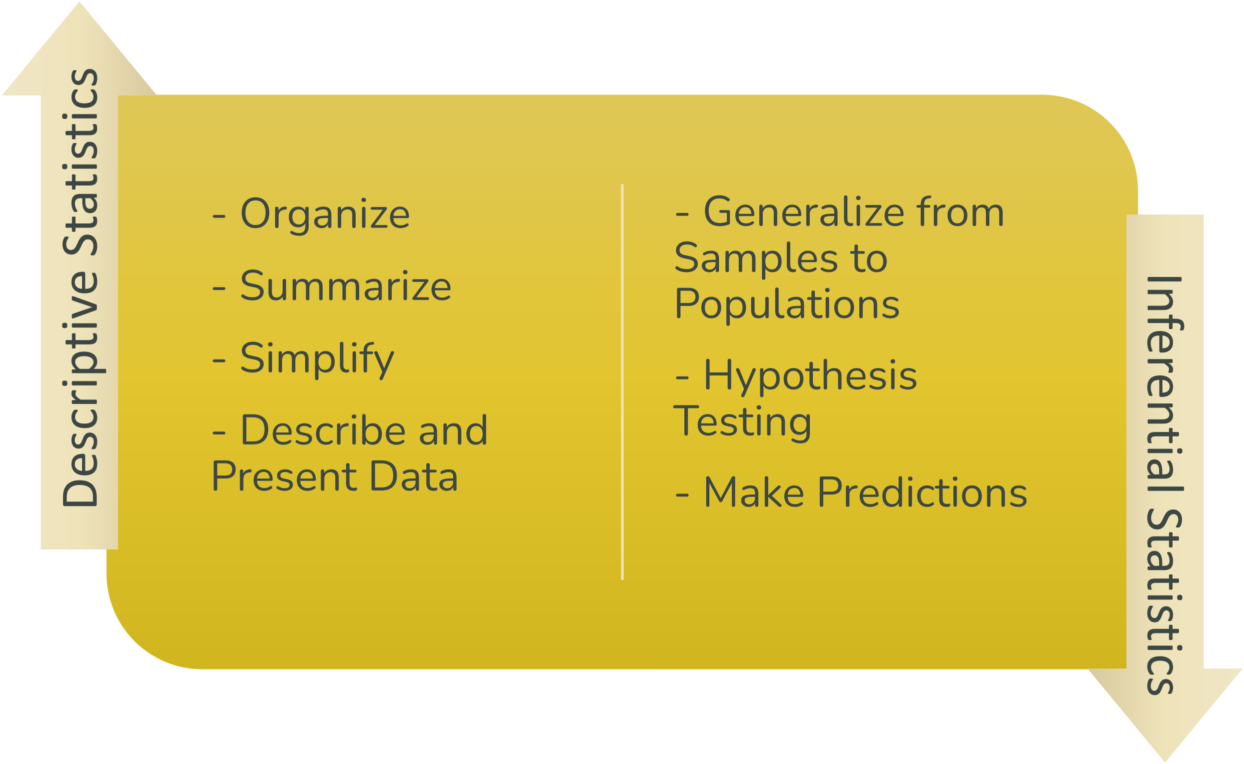
# Types of Statistics

- Descriptive statistics is for summarizing data

# Types of Statistics

- Inferential statistics for drawing conclusions from samples of data.

# Types of Statistics



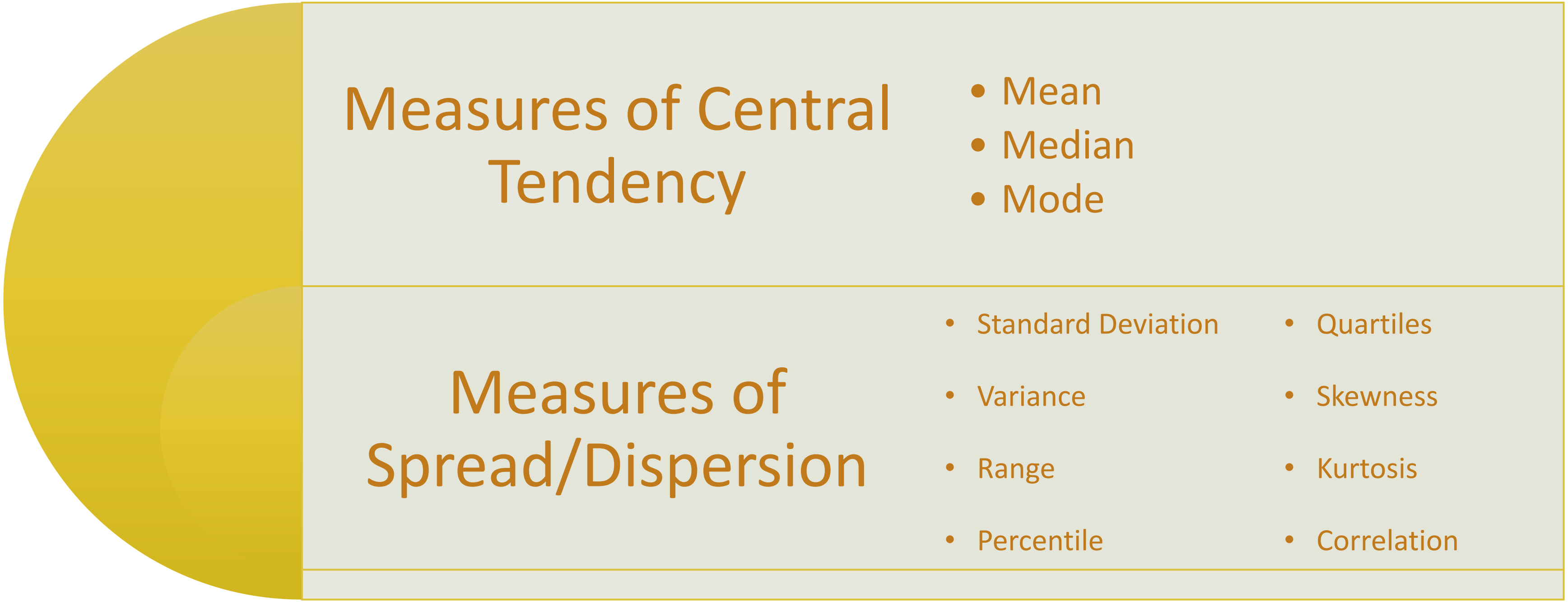
Descriptive Statistics

- Organize
- Summarize
- Simplify
- Describe and Present Data

Inferential Statistics

- Generalize from Samples to Populations
- Hypothesis Testing
- Make Predictions

# Parts of Descriptive Statistics



Measures of Central Tendency

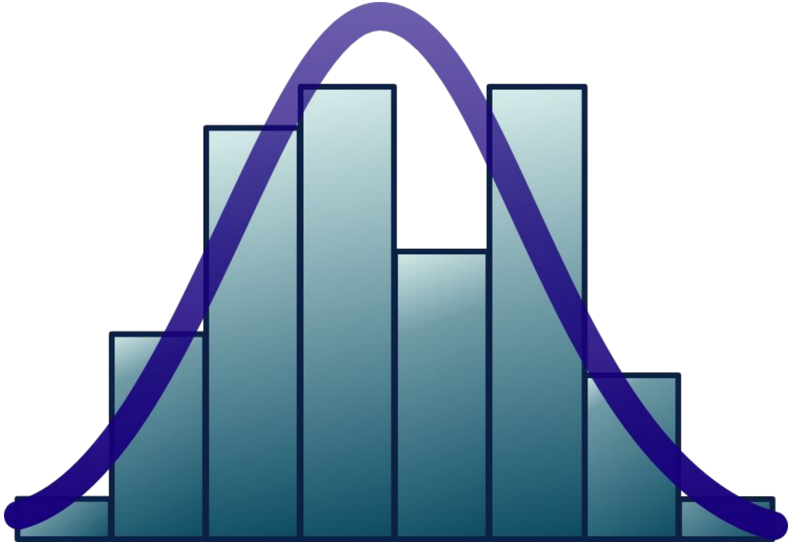
- Mean
- Median
- Mode

Measures of Spread/Dispersion

- Standard Deviation
- Variance
- Range
- Percentile
- Quartiles
- Skewness
- Kurtosis
- Correlation

# Measures of Central Tendency

- Central tendency refers to the idea that there is one number that best summarizes the entire set of measurements.

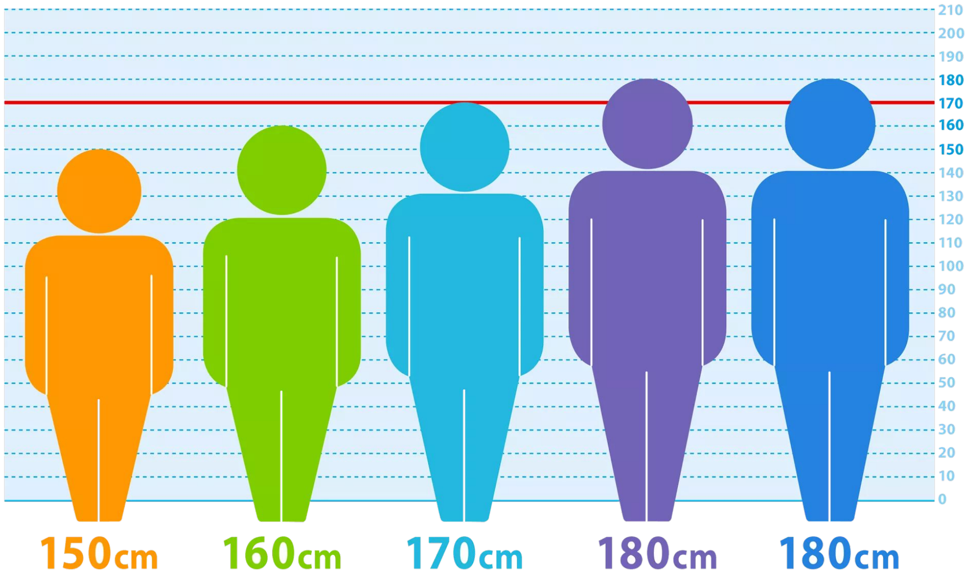


# Mean/Average

- Mean or Average is a central tendency of the data i.e. a number around which a whole data is spread out.
- In a way, it is a single number which can estimate the value of whole data set.
- The mean has one main disadvantage: it is particularly susceptible to the influence of *outliers*.

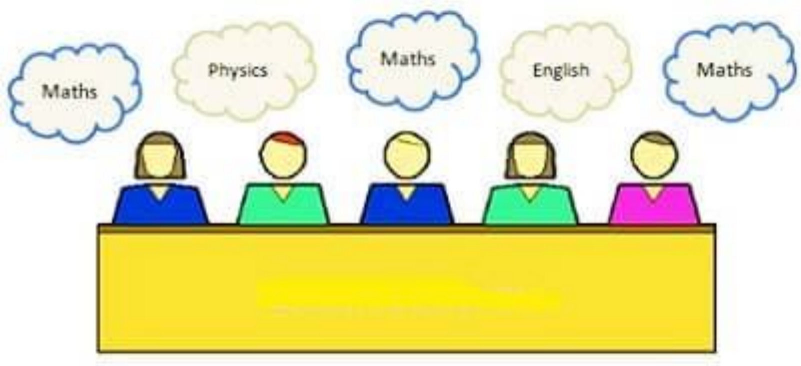
# Median

- Median is the value which divides the data in two equal parts.
- i.e. number of terms on right side of it is same as number of terms on left side of it.
- Data should be arranged in either ascending or descending order.



# Mode

- Mode is the term appearing maximum time in data set
- i.e. term that has highest frequency.



# Mean

**Advantages:**

- Takes into account every number in the data set.
- Easy and quick way to represent the entire data values by a single or unique number due to its straightforward method of calculation.
- Each set has a unique mean value.

**Disadvantages:**

- Its value is easily affected by extreme values known as the outliers.

# Median

**Advantages:**

- Takes into account every number in the data set. That means all numbers are included in calculating the mean.
- Easy and quick way to represent the entire data values by a single or unique number due to its straightforward method of calculation.
- Each set has a unique mean value.

**Disadvantages:**

- Its value is **not** easily affected by extreme values known as the outliers.

# Mode

**Advantages:**

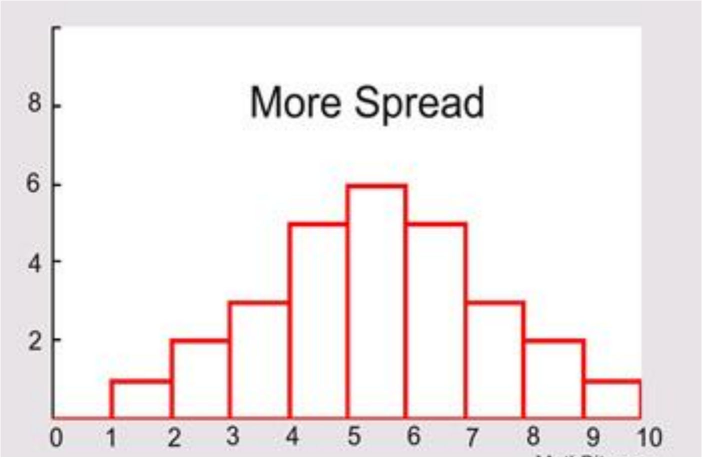
- Just like the median, the mode is not affected by outliers.
- Useful to find the most “popular” or common item. This includes data sets that do not involve numbers.

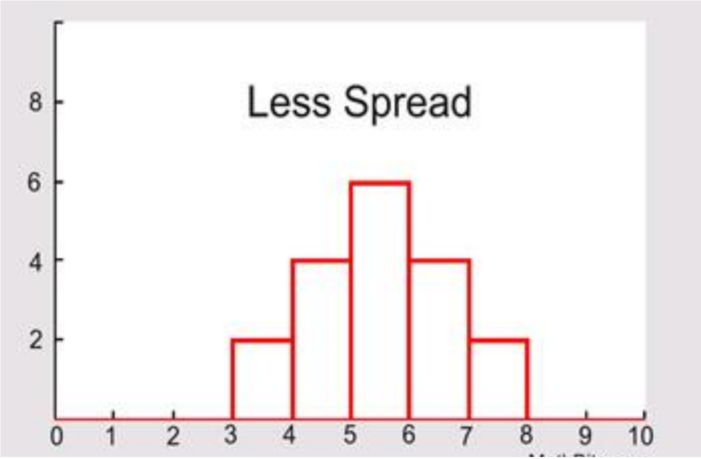
**Disadvantages:**

- If the set contains no repeating values, the mode is irrelevant.
- In contrast, if there are many values that have the same count, then mode can be meaningless.

# Measures of Spread/Dispersion

- Measure of Spread refers to the idea of variability within the data

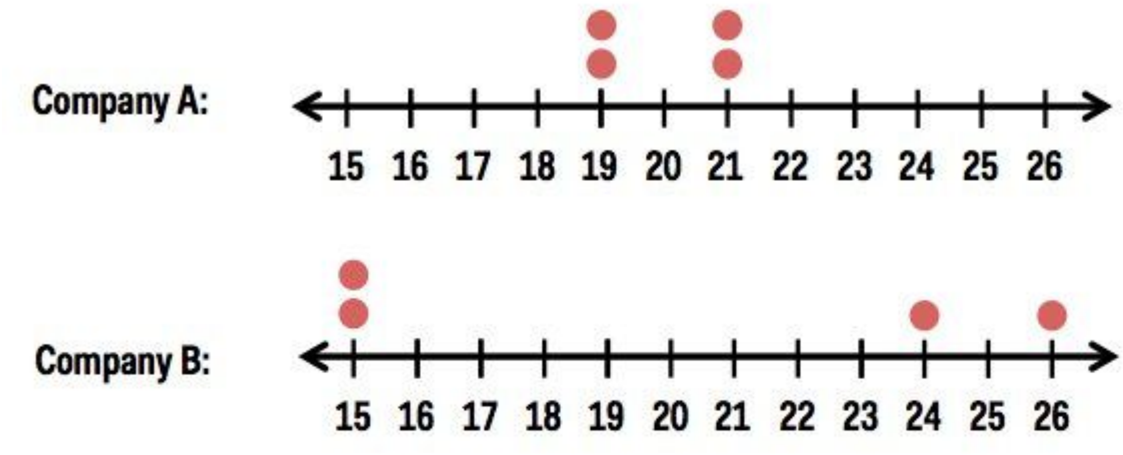




# Standard Deviation

- Standard deviation is the measurement of average distance between each quantity and mean.
- Consider two small businesses with four employees each.
- In one business, two employees make ₹19 per hour and the other two make ₹21 per hour.
- In the second business, two employees make ₹15 per hour, one makes ₹24, and the last makes ₹26

# Standard Deviation



# Standard Deviation

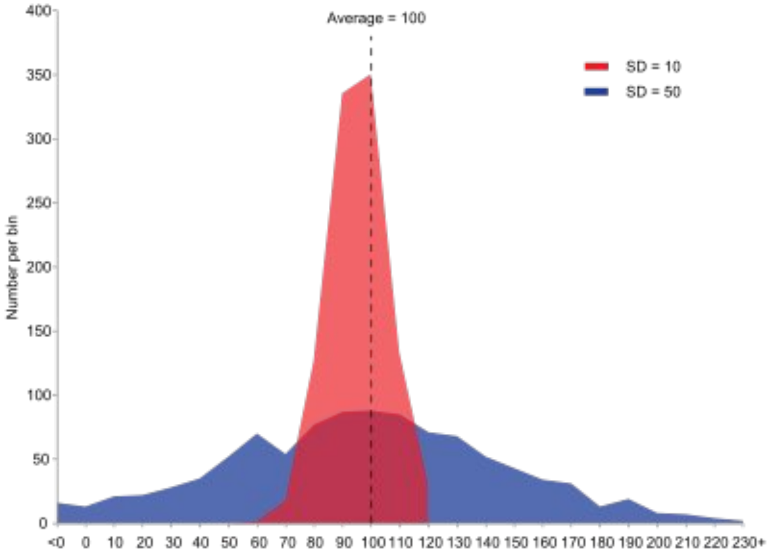
- In both companies, the average wage is ₹20 per hour, but the distribution of hourly wages is clearly different.
- In company A, all four employees' wages are tightly bunched around that average,
- At company B, there's a big spread between the two employees making ₹15 and the other two employees.
- The standard deviation of company A's employees is 1, while the standard deviation of company B's wages is about 5.
- In general, the larger the standard deviation of a data set, the more spread out the individual points are in that set.

# Standard Deviation

- Standard deviation tells about the concentration of the data around the mean of the data set.
- Standard deviation is inversely proportional to the concentration of the data around the mean i.e with high concentration, the standard deviation will be low, and vice versa.
- It cannot be negative.
- The value of standard deviation can be easily impacted by outliers as a single outlier (abnormal value) distorts the overall mean, and thereby, deviation from the mean of all elements.

# Variance

- Variance measures how far each number in the set is from the mean
- It is the squared value of the Standard Deviation



# Variance

- Variance is the measure of dispersion in a data set.
- In other words, it measures how spread out a data set is.
- It is calculated by first finding the deviation of each element in the data set from the mean, and then by squaring it.
- *Variance is the average of all squared deviations.*

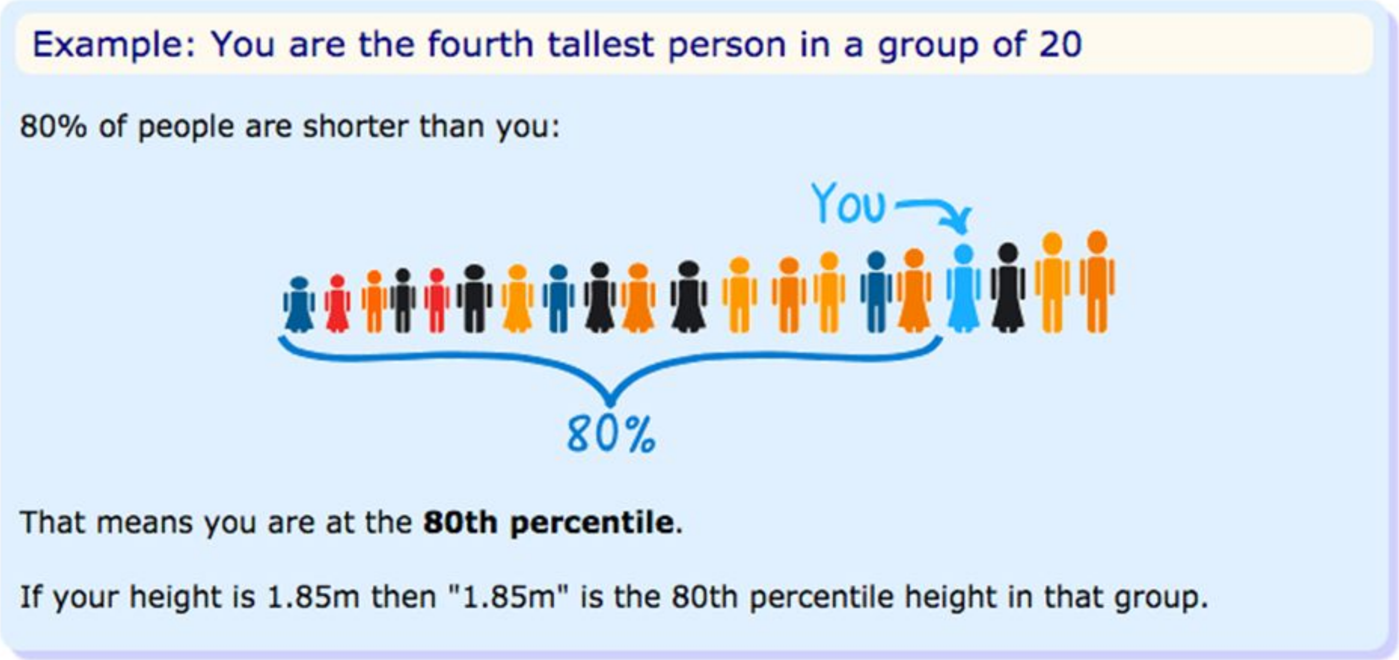
# Variance

- A weather reporter is analyzing the high temperature forecasted for a series of dates versus the actual high temperature recorded on each date.
- A low variance would show a reliable weather forecast.

# Percentile

- Percentile is a way to represent position of a values in data set.
- If *k* is *n*th percentile, it implies that *n%* of the total terms are less than *k*.

# Percentile



Example: You are the fourth tallest person in a group of 20

80% of people are shorter than you:

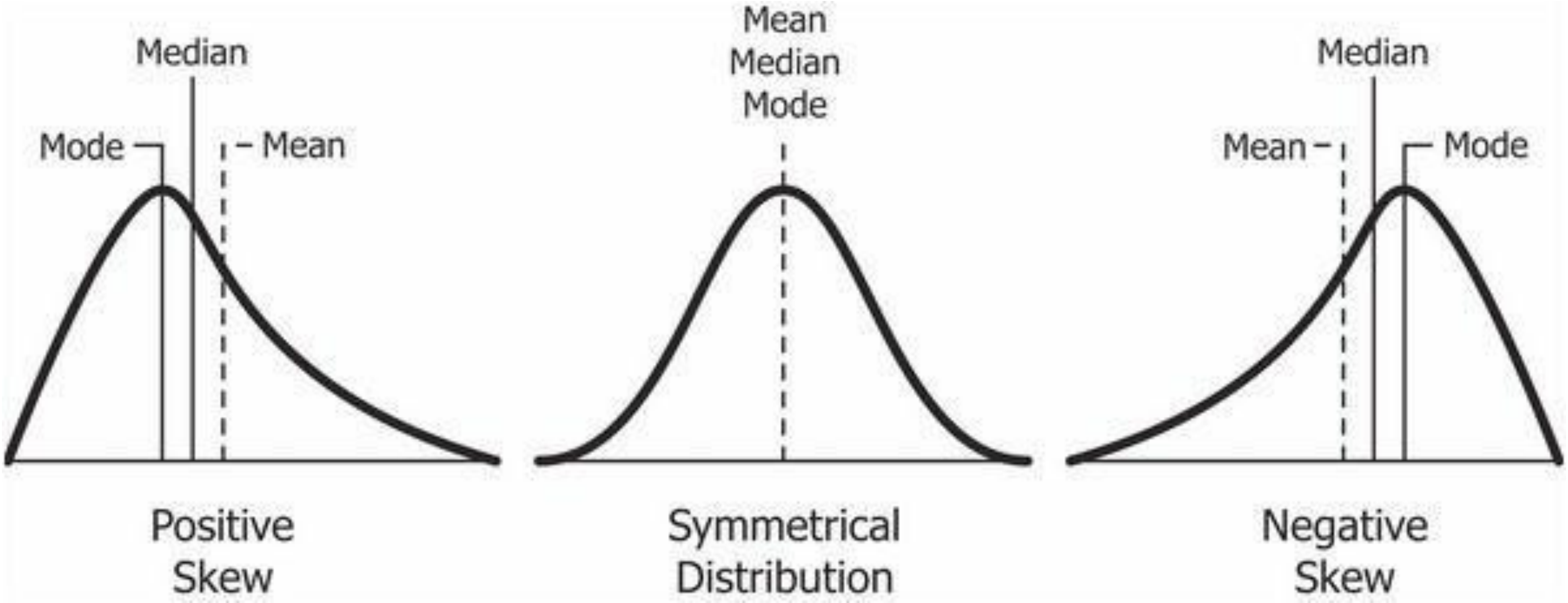
That means you are at the **80th percentile**.

If your height is 1.85m then "1.85m" is the 80th percentile height in that group.

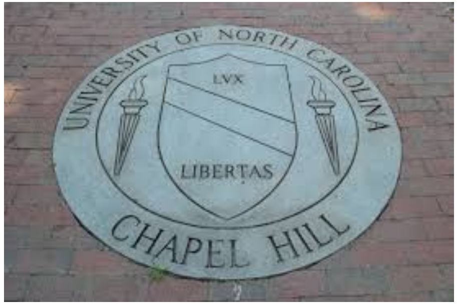
# Skewness

- Skewness is a measure of the asymmetry of the distribution of a random variable about its mean.
- The curve appears distorted or skewed either to the left or to the right.

# Types of Skewness



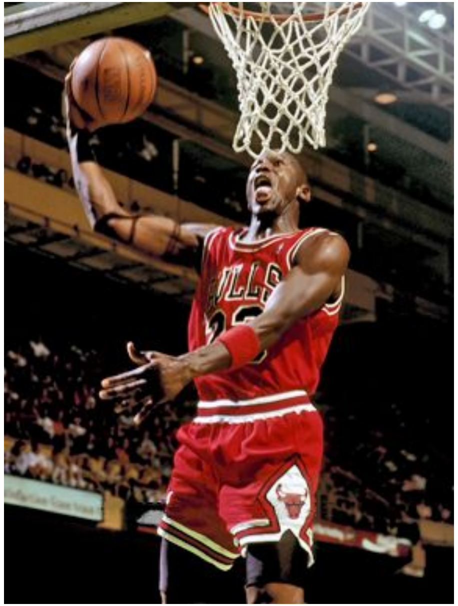
# Skewness



- For example, if you looked at the 10 people who graduated with cultural geography degrees from UNC in 1984
- You'd find the mean amount that people made in the next year is around 3.5 million dollars.
- Say what? Obviously someone who studies cultural geography isn't a millionaire at 22, so what's happening here?
- Michael Jordan was one of those 10 people who graduated, and he made 33 million dollars.

# Skewness

- However, upon realising that this distribution is more skewed, you realize that the mean is not a very good estimate of the amount of money someone would make graduating with a cultural geography degree from UNC.
- Instead, you take the median, which is around 50,000 dollars, not 33 million dollars.



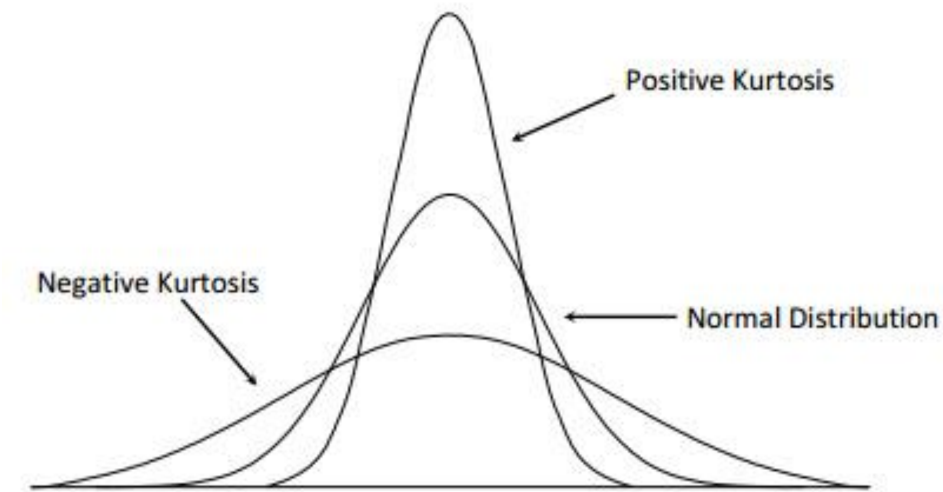
# Kurtosis

- Kurtosis is about existence of outliers.
- An outlier is an observation point that is distant from other observations.
- Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution.
- It is more related to the shape of distribution

| 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| 10 | 12 | 15 | 123 | 18 | 11 |

Outlier

# Types of Kurtosis



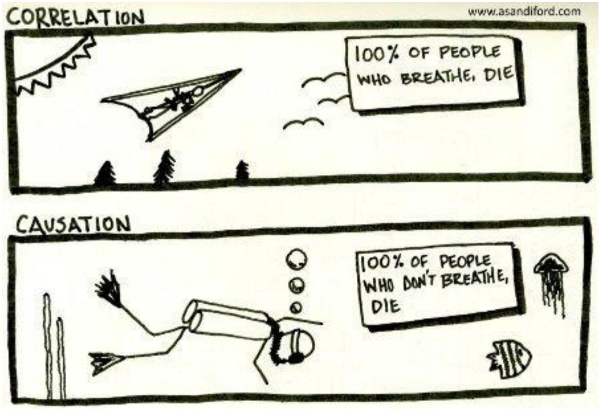
# Correlation

- Correlation is a statistical technique that can show whether and how strongly pairs of variables are related.
- It does not tell us why and how behind the relationship but it just says the relationship exists.
- A Kid **prays** that it should **rain** today so that she can bunk school.
- Fortunately, it rained 9 out of 10 times when she prayed.
- Kid now **strongly** believes that her prayers does all the raining

# Causation

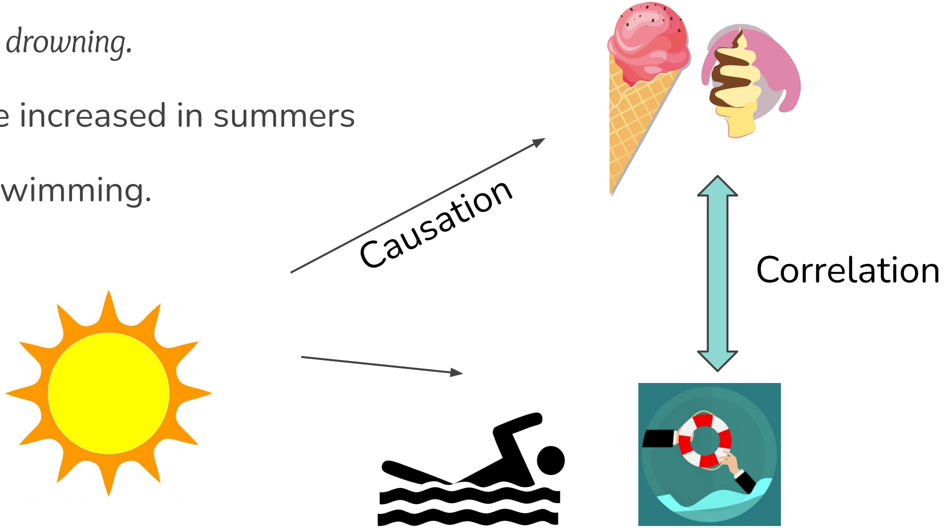
## Ban ice cream - It is causing deaths by drowning

- Causation takes a step further than correlation.
- It says any change in the value of one variable will cause a change in the value of another variable, which means one variable makes other to happen.
- It is also referred to as cause and effect.


CORRELATION
www.asandiford.com
100% OF PEOPLE WHO BREATHE, DIE
CAUSATION
100% OF PEOPLE WHO DON'T BREATHE, DIE

# Correlation and Causation

- *As ice cream sales increase, the rate of drowning deaths increases sharply.*
- *Therefore, ice cream consumption causes drowning.*
- The fact is that, ice cream sales are increased in summers
- People engage in watersports or swimming.


Causation
Correlation