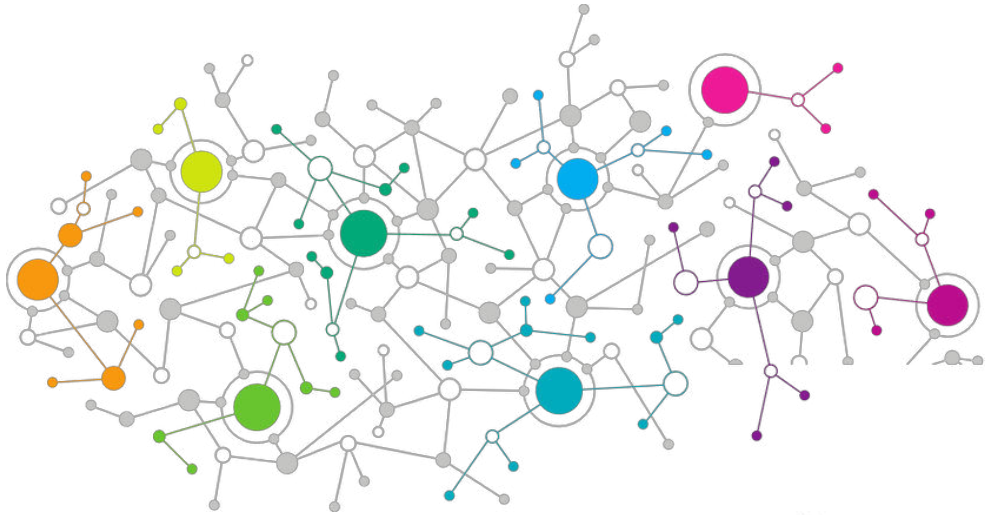
IntelliPaat

Data Science

Statistics



Agenda

01 Introduction to Statistics

02 Statistical Terminology

03 Introduction to Sampling

04 Sampling Technique

05 Categories of Statistics

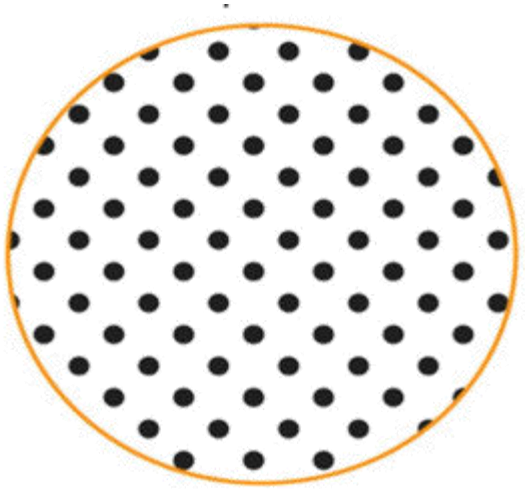
Introduction to Statistics

Statistics is a very broad subject, with applications in a vast number of different fields. In generally one can say that statistics is the methodology for collecting, analyzing, interpreting and drawing conclusions from information

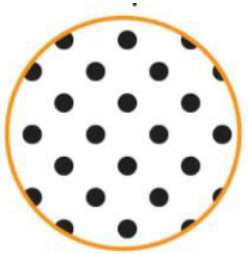
Statistical Terminology

Statistical Terminology

A **population** includes all of the elements from a set of data

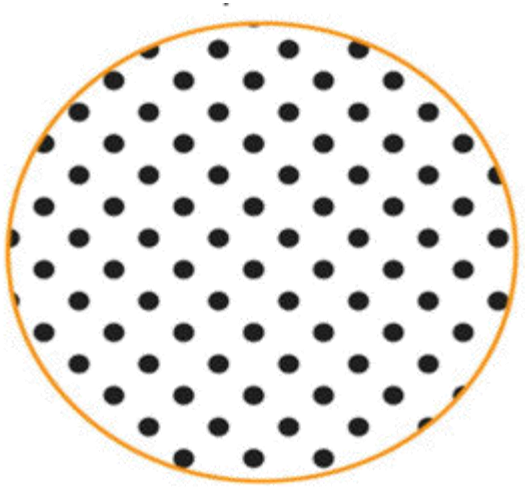


A **sample** consists one or more observations drawn from the population



Statistical Terminology

Population

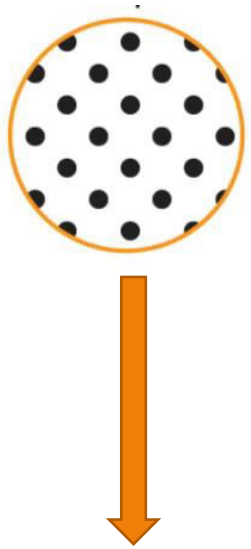


Collection of all items of interest to our study

N

Parameters

Sample



A Subset of Population

n

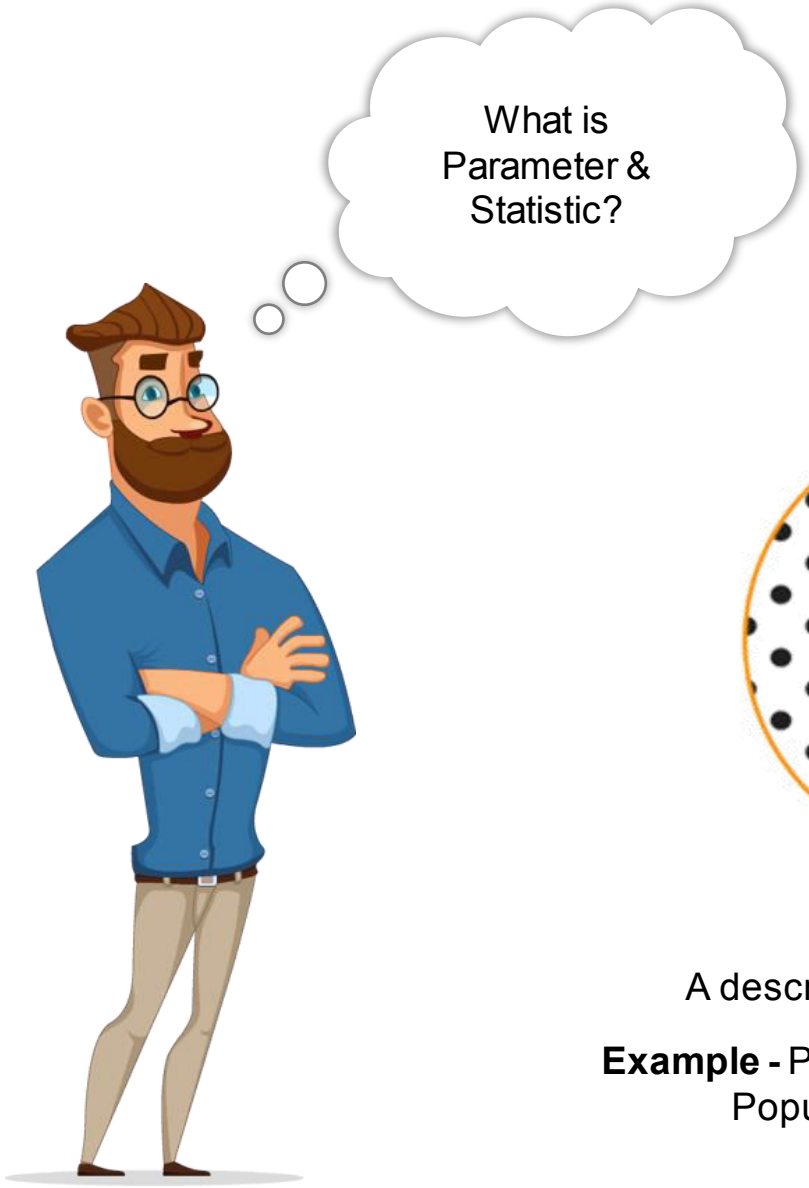
Statistics

Statistical Terminology

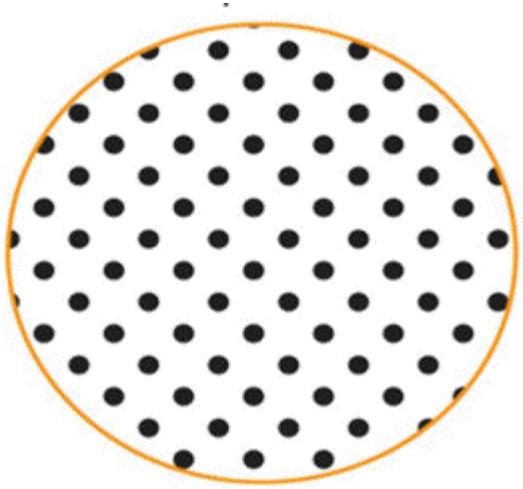
Census
Gathering data from the whole population of interest

Survey
Gathering data from the sample in order to make conclusions about the population.

Statistical Terminology



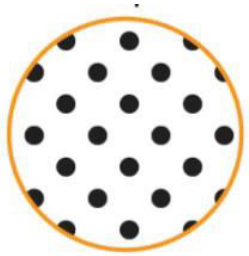
Population



Parameter
A descriptive measure of the population

Example - Population mean, Population variance, Population standard deviation etc.

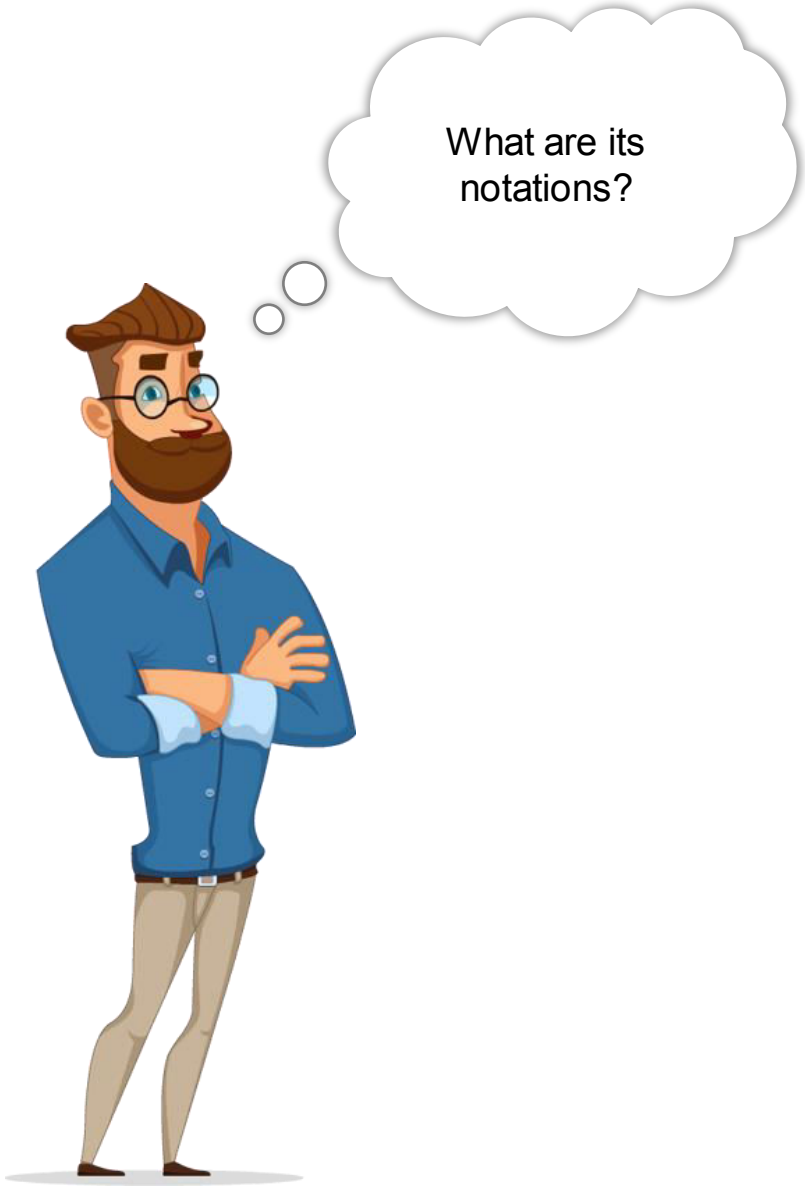
Sample



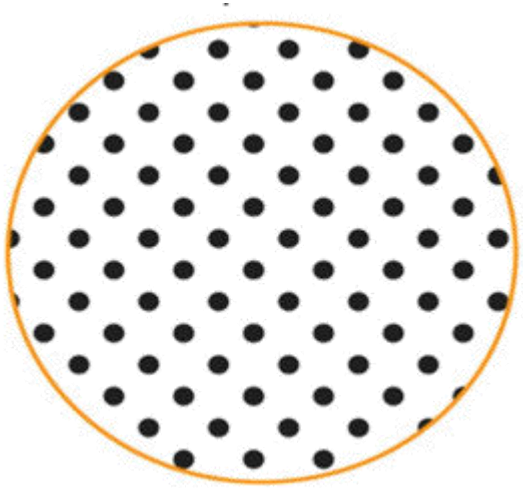
Statistic
A descriptive measure of the sample

Example - Sample mean, Sample variance, Sample standard deviation etc.

Statistical Terminology



Population



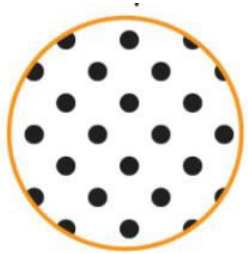
Greek – Population Parameter

Mean - μ

Variance - σ^2

Standard Deviation - σ

Sample



Roman – Sample Statistic

Mean - $\bar{x}$

Variance - s^2

Standard Deviation - s

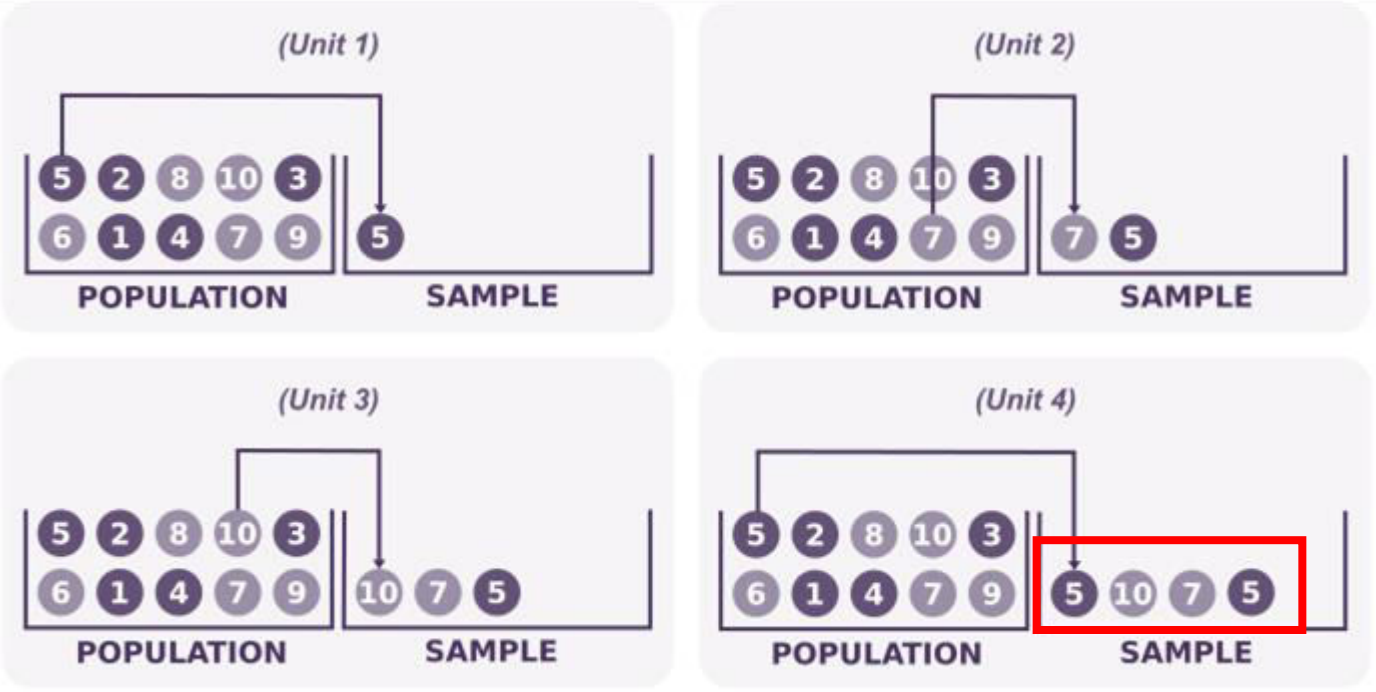
Sampling

Sampling

Sampling Techniques

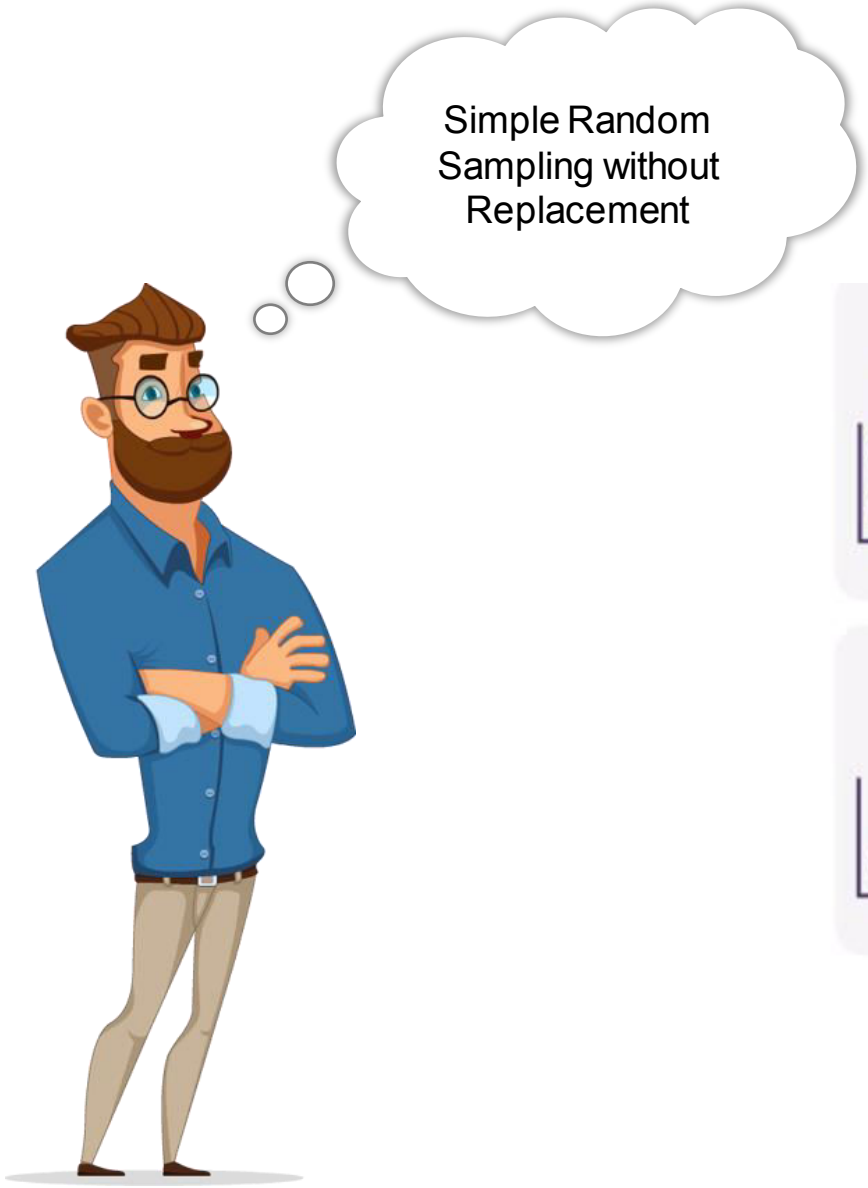
Simple Random Sampling

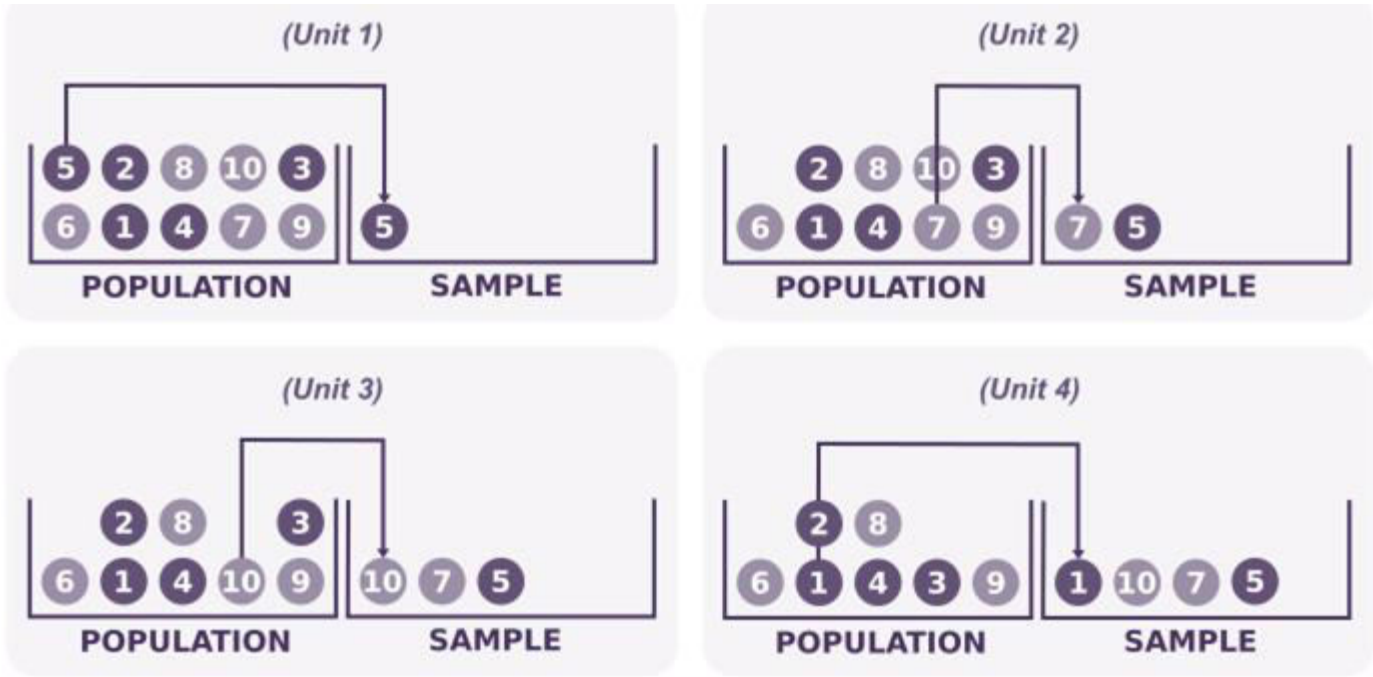
Every subject in the population has an equal chance of being selected.
Randomly pick subjects from the whole population.



SIMPLE RANDOM SAMPLING WITH REPLACEMENT

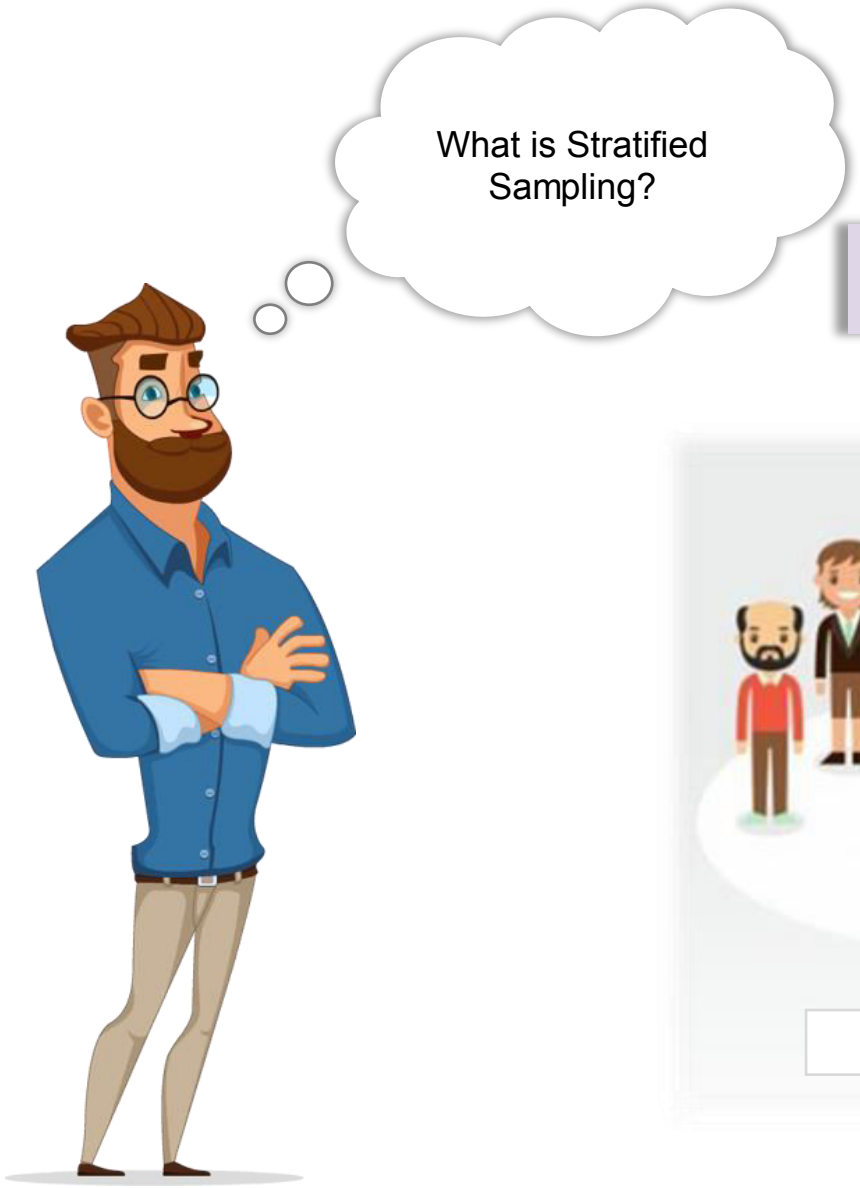
Simple Random Sampling





SIMPLE RANDOM SAMPLING WITHOUT REPLACEMENT

Stratified Sampling

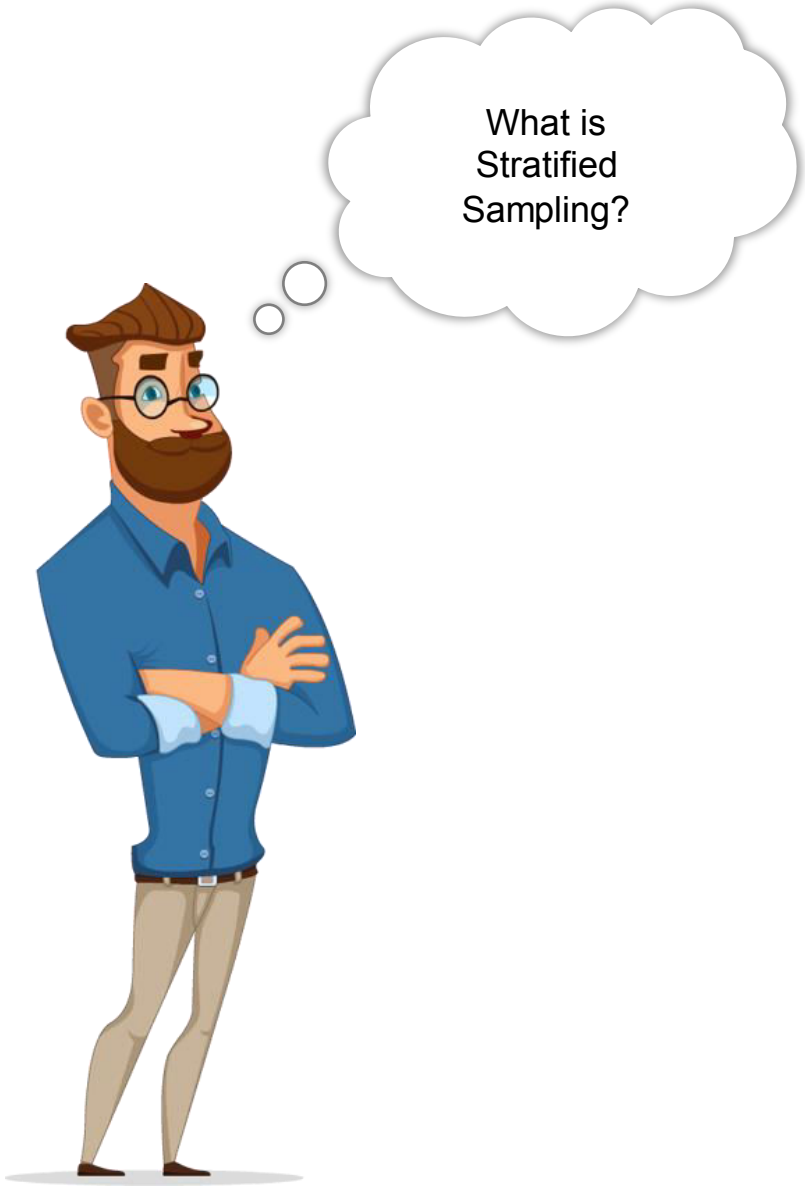


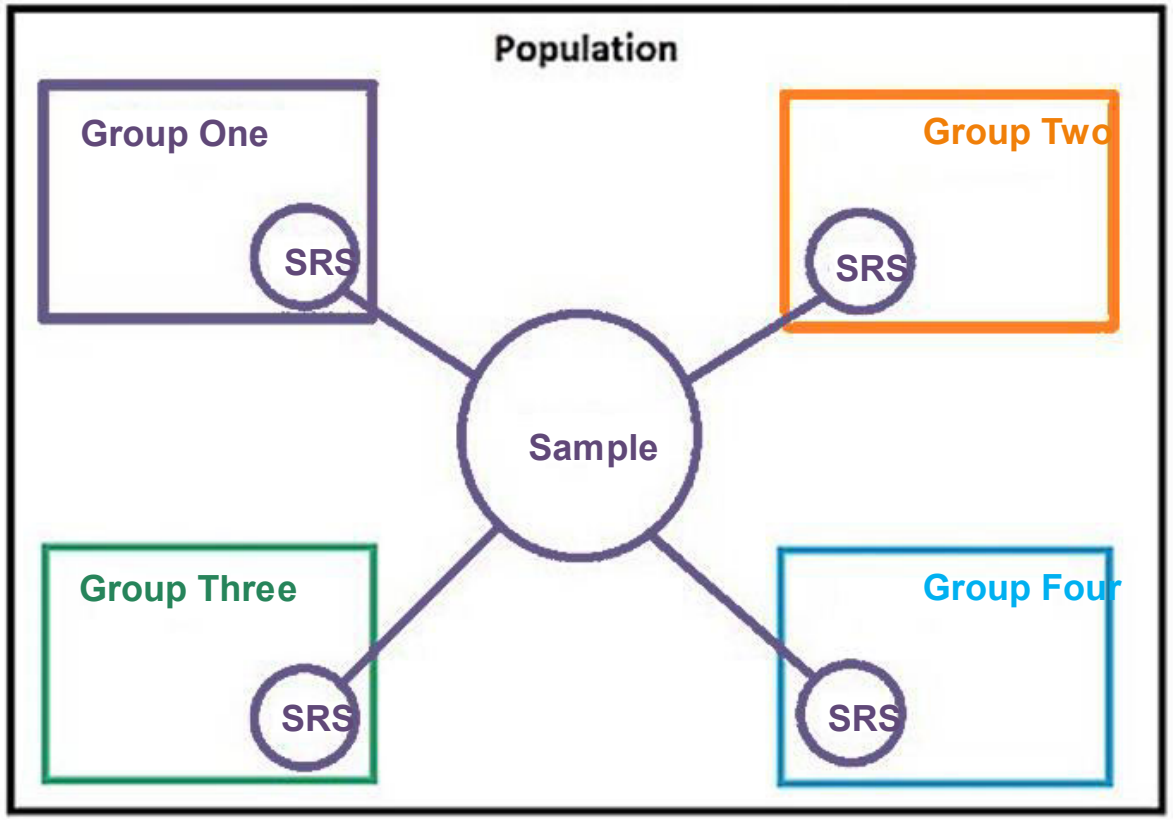
The population is partitioned into non-overlapping groups, called strata and a sample is selected by some design within each stratum

Population

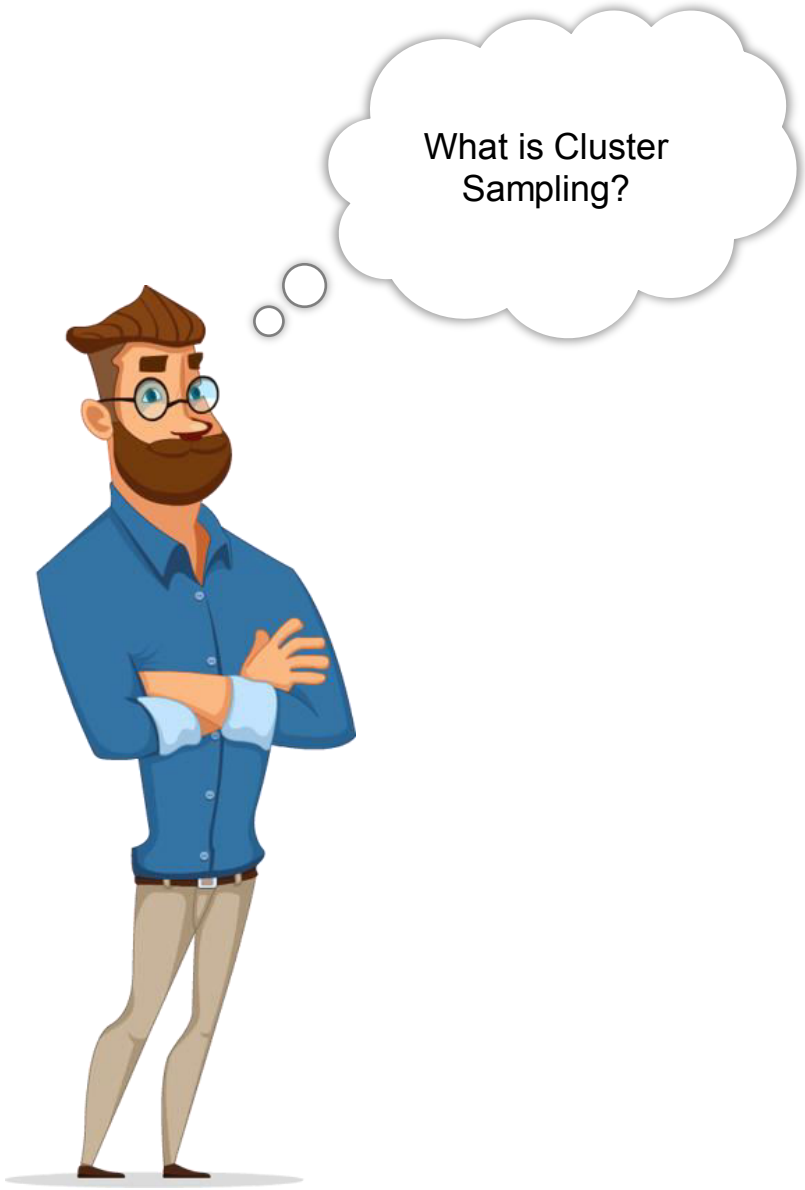
Strata

Stratified Sampling

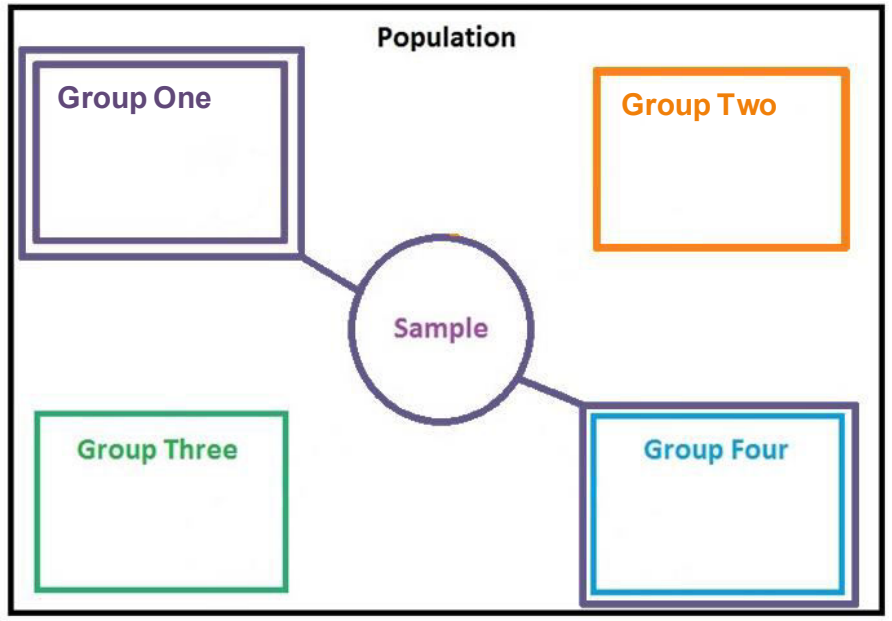




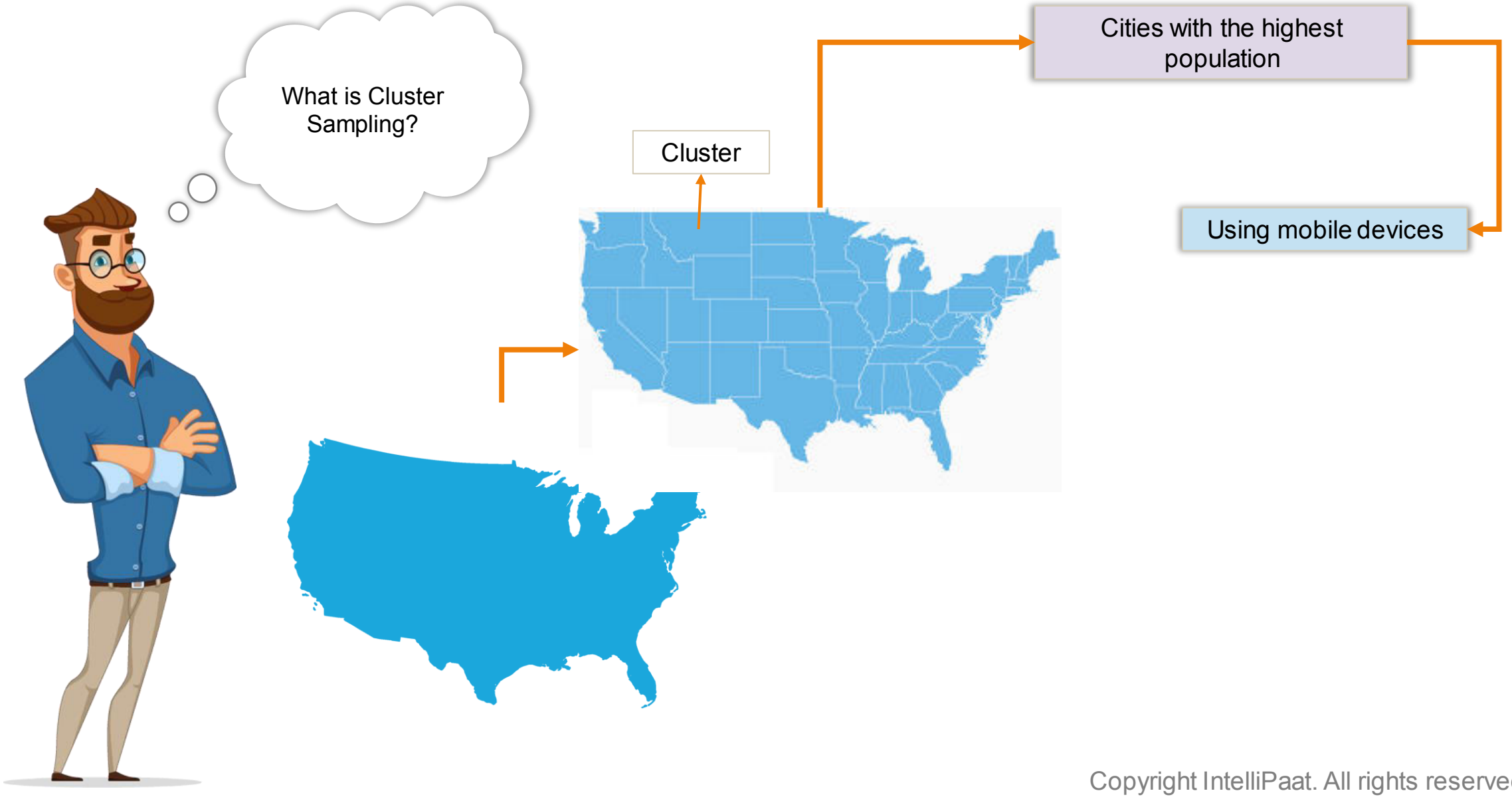
Cluster Sampling



- Divide the population into groups (clusters)
- Obtain a simple random sample of so many clusters from all possible clusters
- Obtain data on every sampling unit in each of the randomly selected clusters

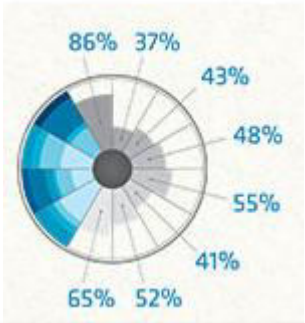


Cluster Sampling

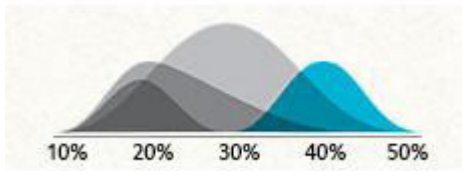


Categories of Statistics

Categories of Statistics

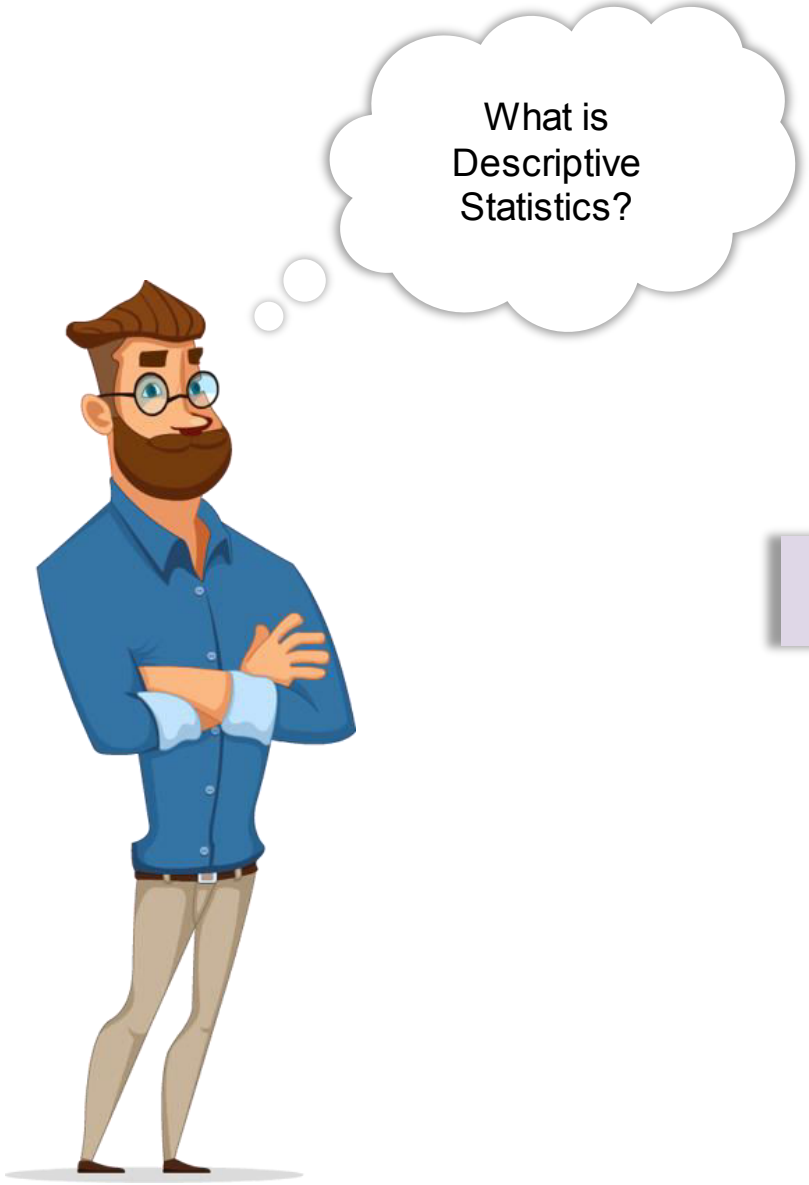


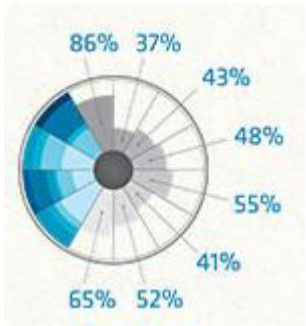
Descriptive Statistics



Inferential Statistics

Categories of Statistics

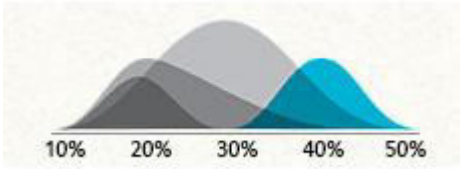




The analysis of data that helps describe, show or summarize data in a meaningful way such that, for example, patterns might emerge from the data

Just describes and summarizes data

Categories of Statistics

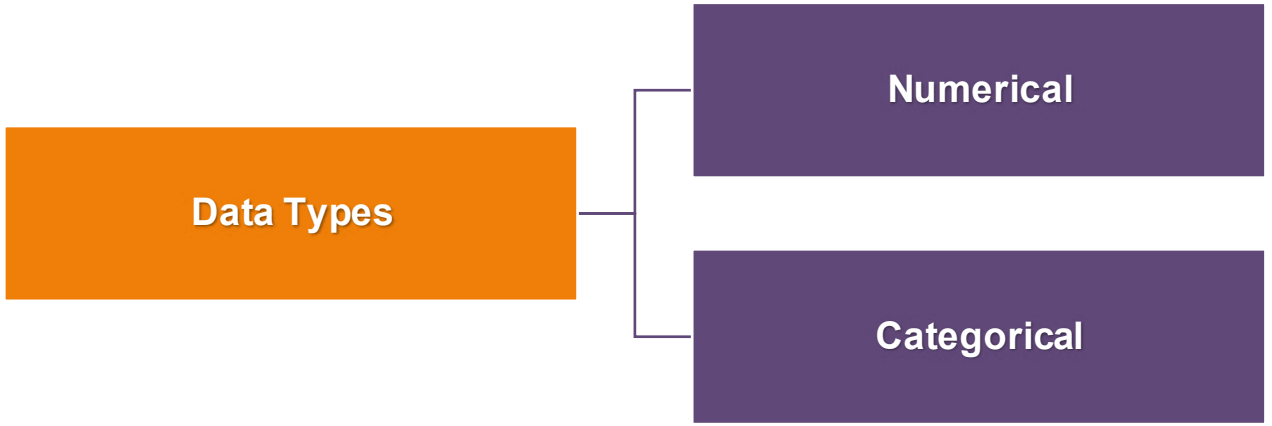


Inferential Statistics takes data from a sample and makes inferences about the larger population from which the sample was drawn

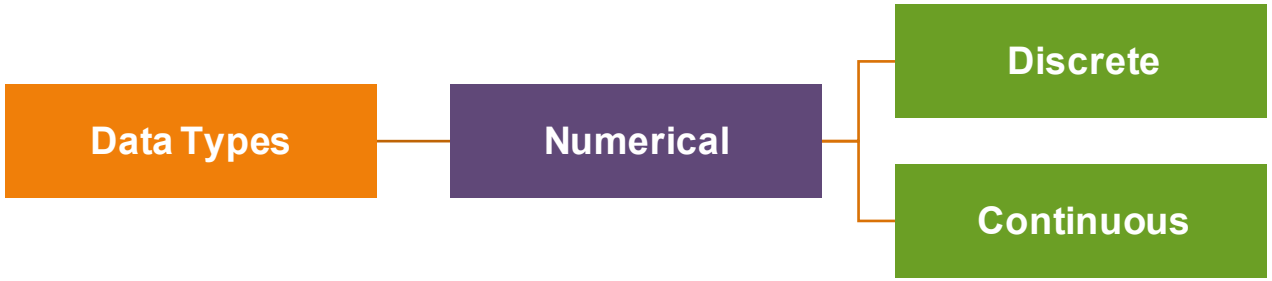
Studies a sample of same data

Types of Data

Types of Data



Numerical Data



- Deals with Number
- Data which can be measured
- Length, height, area, volume, weight, speed, time, temperature, humidity, sound

Numerical Data

Values or observations that is counted as distinct and separate and can only take particular values.

Counted

Types of Data - Numerical Data

You can measure continuous data. Values or observations may take on any value within a finite or infinite interval.

Measured

Temperature?

- ○ - 10
- ○ -5
- ○ 0
- ○ + 5
- ○ + 10
- ○ + 15

Categorical Data

You can measure continuous data. Values or observations may take on any value within a finite or infinite interval.

Measured

Categorical Data

Values or observations can be ranked (put in order) or have a rating scale attached. You can count and order, but not measure, ordinal data.

Logical Ordering

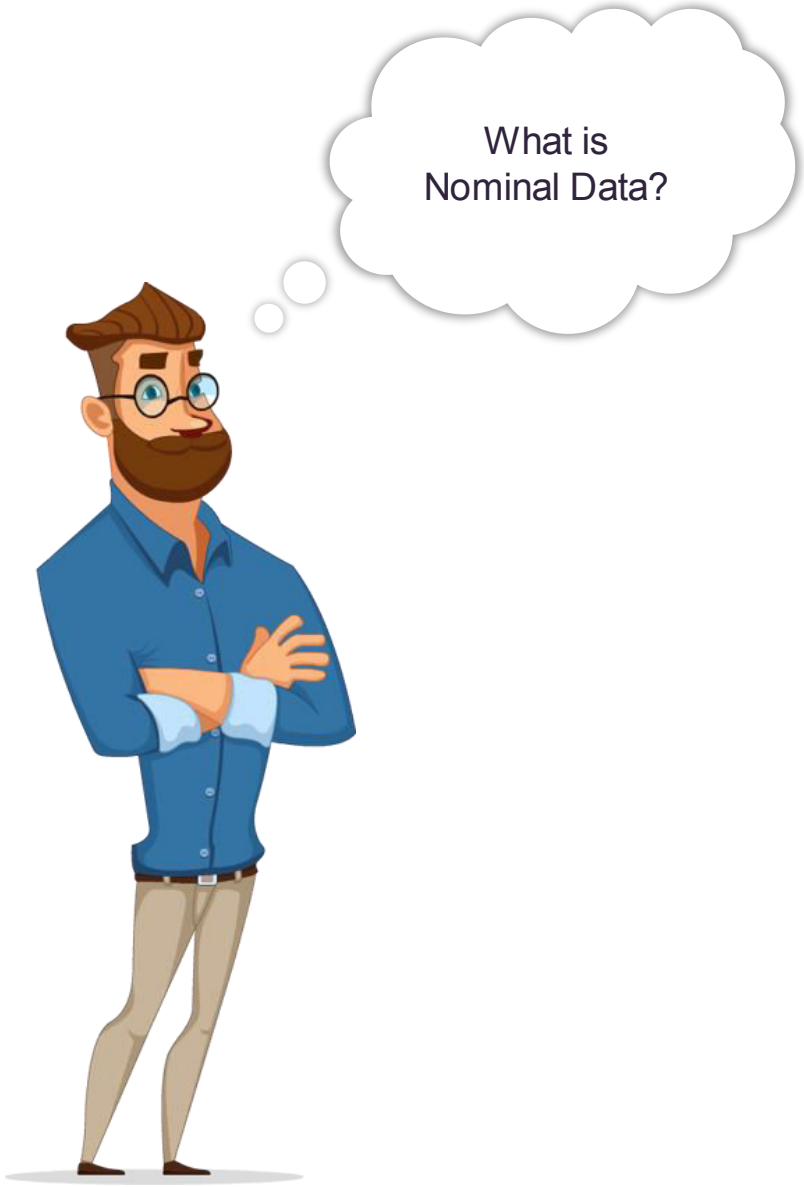
How do you feel today?

- 1 – Very Unhappy
- 2 – Unhappy
- 3 – OK
- 4 – Happy
- 5 – Very Happy

How satisfied are you with our service?

- 1 – Very Unsatisfied
- 2 – Somewhat Unsatisfied
- 3 – Neutral
- 4 – Somewhat Satisfied
- 5 – Very Satisfied

Categorical Data



Values or observations can be assigned a code in the form of a number where the numbers are simply labels. You can count but not order or measure nominal data.

Name or Label

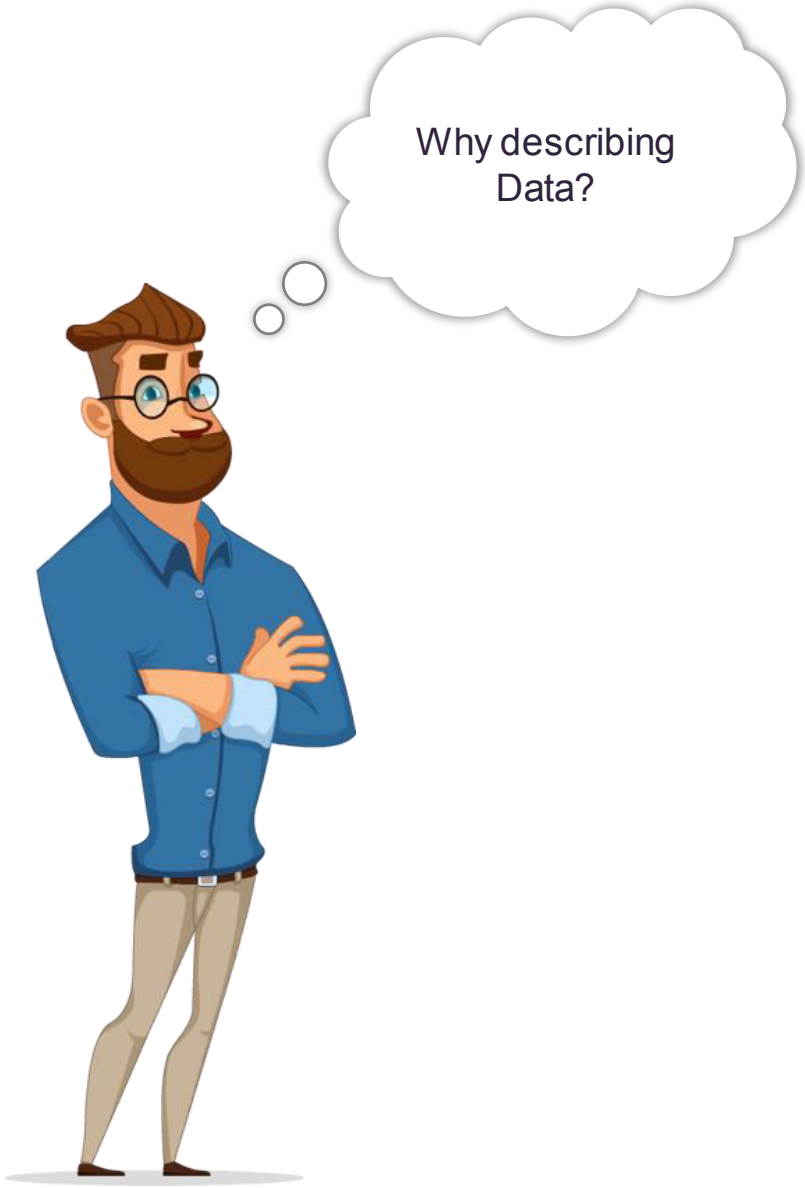
What is your gender?
- M – Male
- F – Female

What is your hair color?
- 1 – Brown
- 2 – Black
- 3 – Blonde
- 4 – Gray
- 5 – Other

Where do you live?
- A – North of the equator
- B – South of the equator
- C – Neither: In the international space station

Describing data through Statistics

Describing data through Statistics



- Describing data through Statistics
  - Measures of Central Tendency
  - Measures of Spread

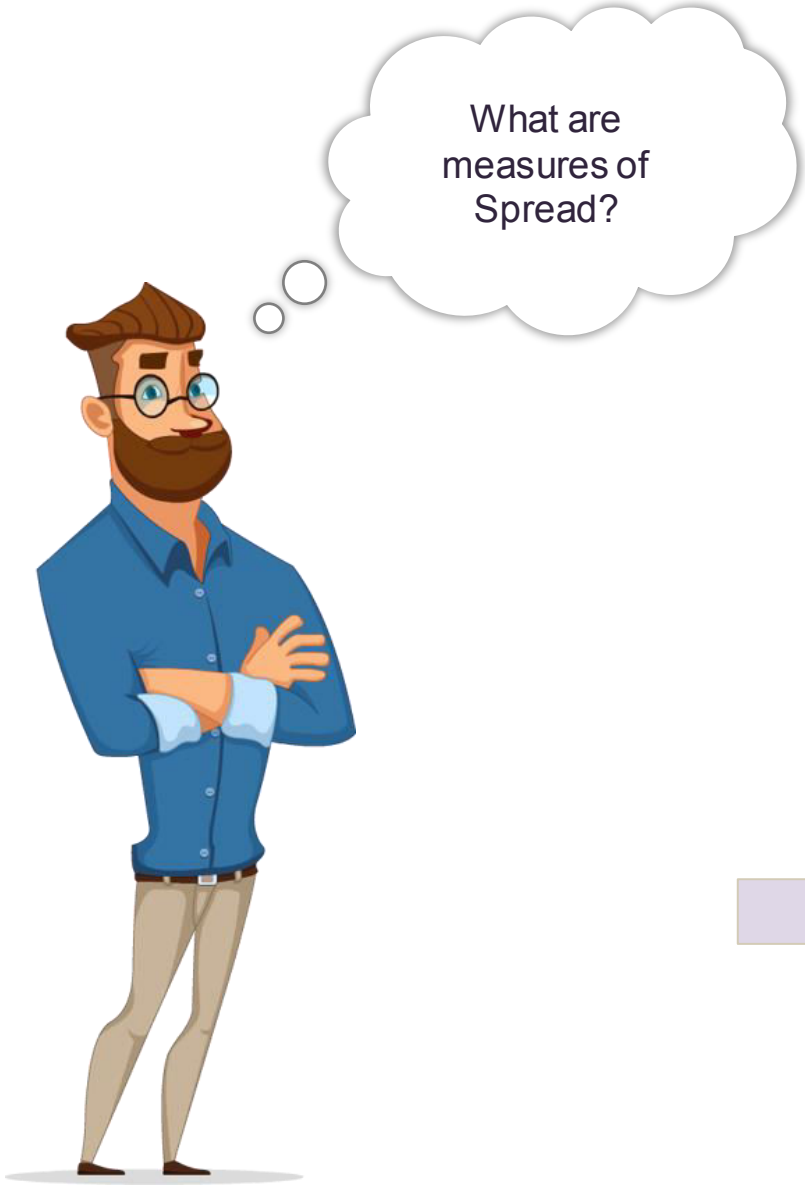
Describing data through Statistics

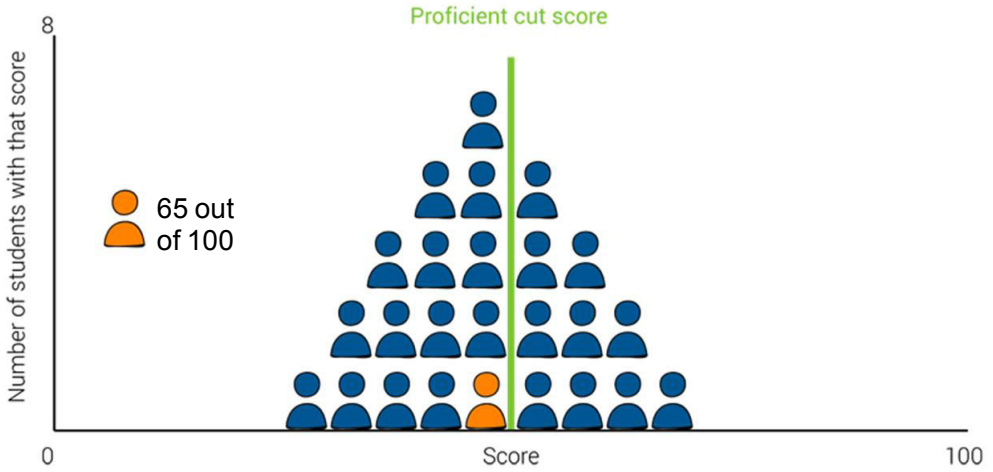
These are ways of describing the central position of a frequency distribution for a group of data

Mean | Median | Mode

Summarize a bunch of scores with a single number

Describing data through Statistics





These are ways of summarizing a group of data by describing how spread out the scores are

Range | Quartiles | Variance | Standard Deviation

Know the spread of scores within a bunch of scores

The Central Tendencies

Mean

The mean is the most common measure of center. It is what most people think of when they hear the word "average". However, the mean is affected by extreme values so it may not be the best measure of center to use in a skewed distribution

Used with both discrete and continuous data

$$Mean = \bar{x} = \frac{\sum_i^n x_i}{n} = \frac{x_1, x_2, x_3, \ldots x_i}{i}$$

$$Mean = \frac{(1+2+3+4+5)}{5} \rightarrow \frac{15}{5} \rightarrow 3$$

Trimmed Mean

A variation of the mean, which you calculate by dropping a fixed number of sorted values at each end and then taking average of the remaining values.
Trimmed mean with P smallest and largest values omitted

Marks of 5 student = 67, 15, 75, 72, 85

$$Trimmed\ Mean = \bar{x} = \frac{\sum_{i=p+1}^{n-p} x_i}{n-2p}$$

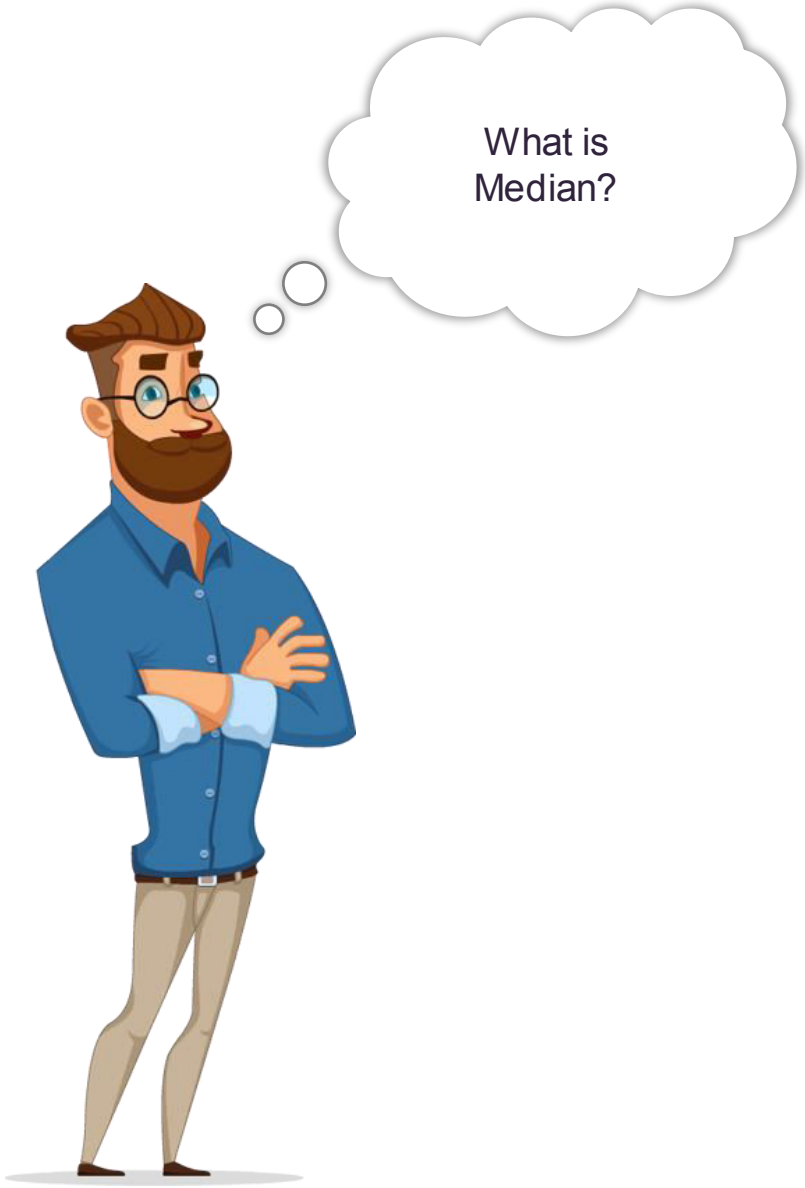
$$Trimmed\ Mean = \frac{(67+75+72)}{3} \rightarrow \frac{214}{3} \rightarrow 71.3$$

Implementation in R

customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService
7590-VHVEG	Female	0	Yes	No	1	No
5575-GNVDE	Male	0	No	No	34	Yes
3668-QPYBK	Male	0	No	No	2	Yes
7795-CFOCW	Male	0	No	No	45	No
9237-HQITU	Female	0	No	No	2	Yes
9305-CDSKC	Female	0	No	No	8	Yes
1452-KIOVK	Male	0	No	Yes	22	Yes
6713-OKOMC	Female	0	No	No	10	No
7892-POOKP	Female	0	Yes	No	28	Yes
6388-TABGU	Male	0	No	Yes	62	Yes
9763-GRSKD	Male	0	Yes	Yes	13	Yes

```
mean(customer_churn$MonthlyCharges)
```

Median



The median is the middle number on a sorted list of the data

$$\boldsymbol{Median} = \frac{n+1}{2}$$

If n =7, $\boldsymbol{Median} = \frac{7+1}{2}$ **=** 4^{th} value

$$\boldsymbol{Median} = \frac{n}{2}, \left(\frac{n}{2}\right) + 1$$

If n =6, $\boldsymbol{Median} = \frac{6}{2}, \left(\frac{6}{2}\right) + 1$ = 3^{rd} and 4^{th} value

Median

If we have more outlier in our datasets then it is best to use median then mean because mean will give poor central tendency

An Outlier is a rare chance of occurrence within a given data set

Implementation in R

customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService
7590-VHVEG	Female	0	Yes	No	1	No
5575-GNVDE	Male	0	No	No	34	Yes
3668-QPYBK	Male	0	No	No	2	Yes
7795-CFOCW	Male	0	No	No	45	No
9237-HQITU	Female	0	No	No	2	Yes
9305-CDSKC	Female	0	No	No	8	Yes
1452-KIOVK	Male	0	No	Yes	22	Yes
6713-OKOMC	Female	0	No	No	10	No
7892-POOKP	Female	0	Yes	No	28	Yes
6388-TABGU	Male	0	No	Yes	62	Yes
9763-GRSKD	Male	0	Yes	Yes	13	Yes

```
median(customer_churn$MonthlyCharges)
```

Mode

The most frequently occurring data point

Mean and median need not to be in the dataset but the Mode has to be in it.

Variability and Spread

Range

Range = Max - Min

Player 1 = 13-7

Points scored per game	7	8	9	10	11	12	13
Frequency, f	1	1	2	2	2	1	1

Player 2 = 13-7

Points scored per game	7	9	10	11	13
Frequency, f	1	2	4	2	1

Player 3 = 30-3

Points scored per game	3	6	7	10	11	13	30
Frequency, f	2	1	2	3	1	1	1

Implementation in R

customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService
7590-VHVEG	Female	0	Yes	No	1	No
5575-GNVDE	Male	0	No	No	34	Yes
3668-QPYBK	Male	0	No	No	2	Yes
7795-CFOCW	Male	0	No	No	45	No
9237-HQITU	Female	0	No	No	2	Yes
9305-CDSKC	Female	0	No	No	8	Yes
1452-KIOVK	Male	0	No	Yes	22	Yes
6713-OKOMC	Female	0	No	No	10	No
7892-POOKP	Female	0	Yes	No	28	Yes
6388-TABGU	Male	0	No	Yes	62	Yes
9763-GRSKD	Male	0	Yes	Yes	13	Yes

```
range(customer_churn$MonthlyCharges)
```

Quartiles

Quartiles tell us about the spread of a data set by breaking the data set into quarters, just like the median breaks it in half.

Lower quartile (25^{th} percentile, Q1) = $\frac{(n+1)}{4}$th

Middle quartile = Median = $\frac{2*(n+1)}{4}$th

Upper quartile (75^{th} percentile, Q3) = $\frac{3*(n+1)}{4}$th

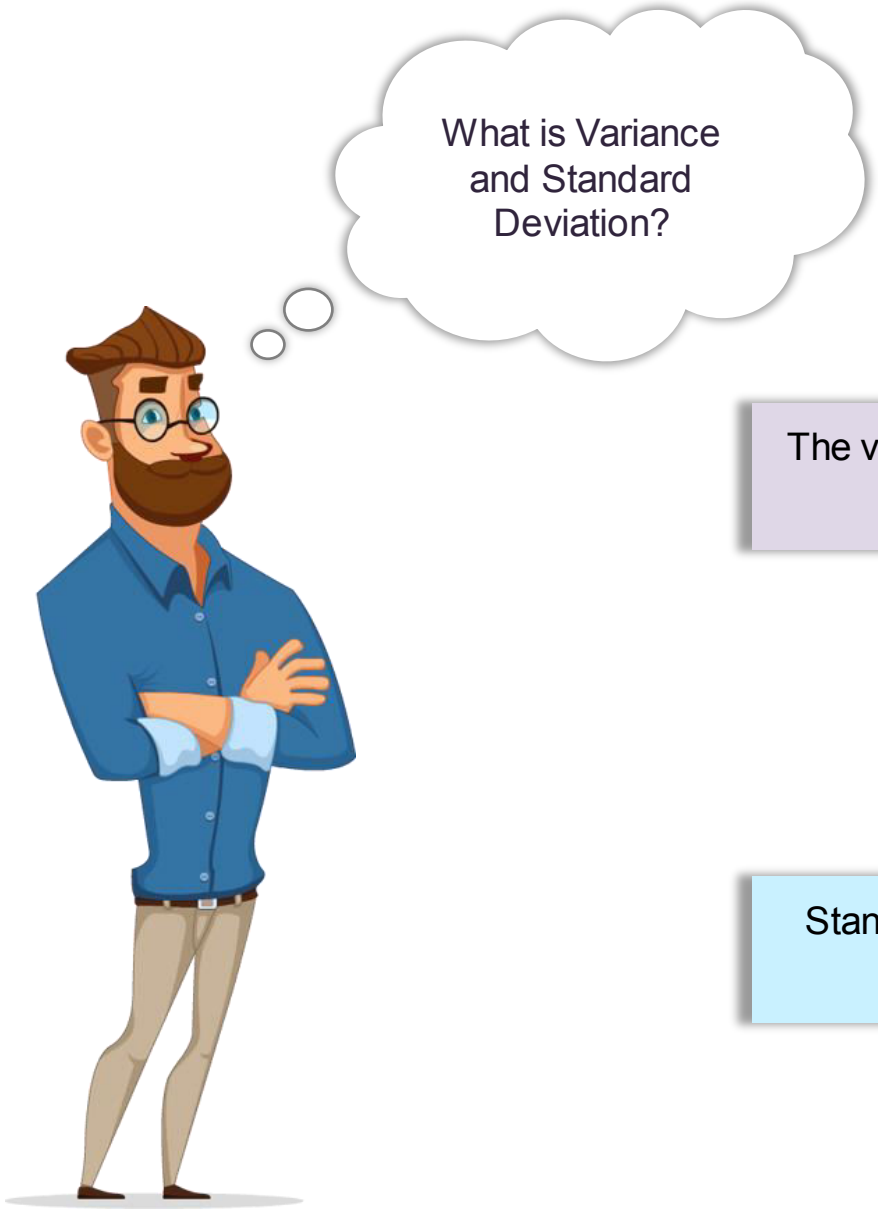
Interquartile range, IQR = Q3-Q1 (central 50% of data)

Implementation in R

customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService
7590-VHVEG	Female	0	Yes	No	1	No
5575-GNVDE	Male	0	No	No	34	Yes
3668-QPYBK	Male	0	No	No	2	Yes
7795-CFOCW	Male	0	No	No	45	No
9237-HQITU	Female	0	No	No	2	Yes
9305-CDSKC	Female	0	No	No	8	Yes
1452-KIOVK	Male	0	No	Yes	22	Yes
6713-OKOMC	Female	0	No	No	10	No
7892-POOKP	Female	0	Yes	No	28	Yes
6388-TABGU	Male	0	No	Yes	62	Yes
9763-GRSKD	Male	0	Yes	Yes	13	Yes

```
IQR(customer_churn$MonthlyCharges)
```

Variance & Standard Deviation



The variance of the data is the average squared distance between the mean and each data value.

Variance = $(Unit)^2$

Variance of length = $(m)^2$

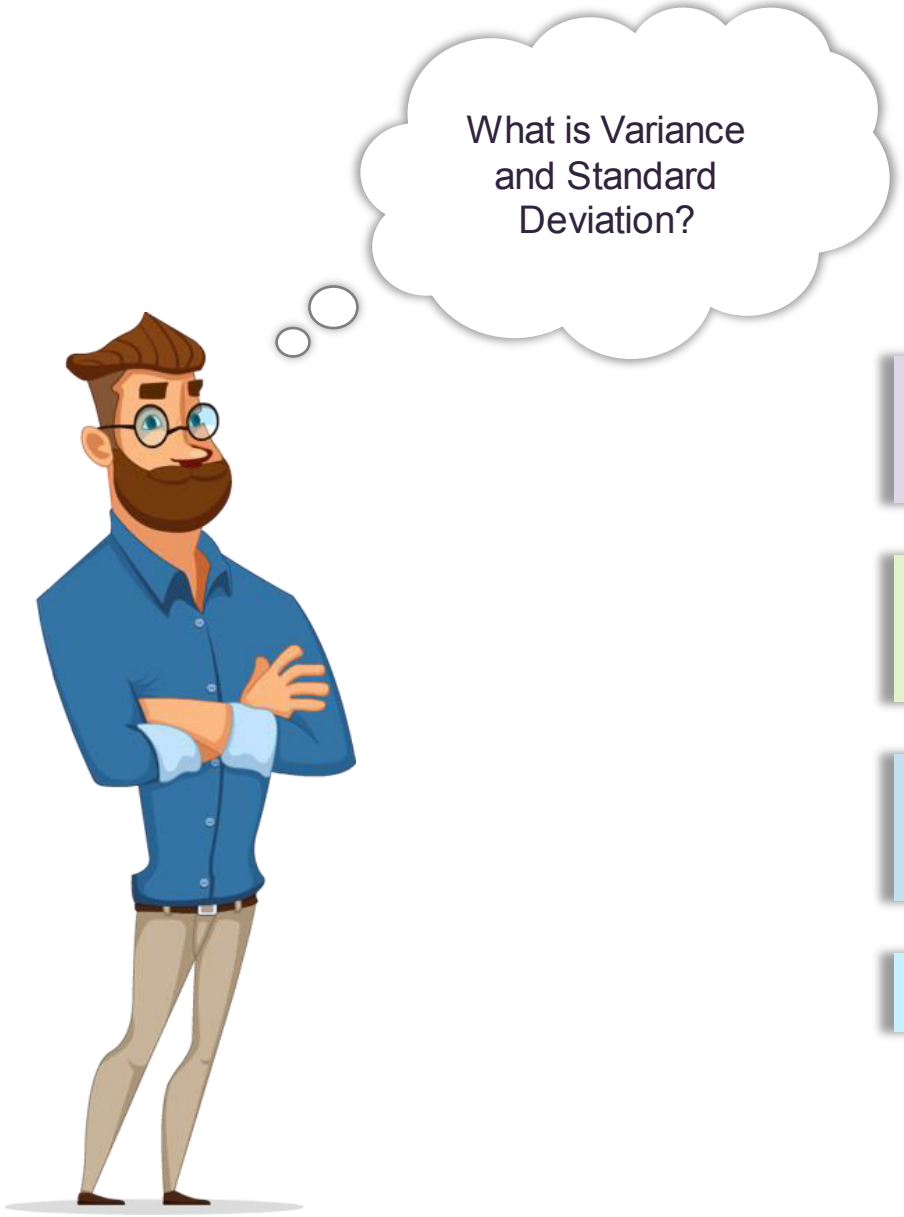
Standard deviation measures how spread out the values in a data set are around the mean. Square root of the variance

Variance & Standard Deviation

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

$$\text{Stantard Deviation} = S = \sqrt{\text{variance}}$$

Variance & Standard Deviation



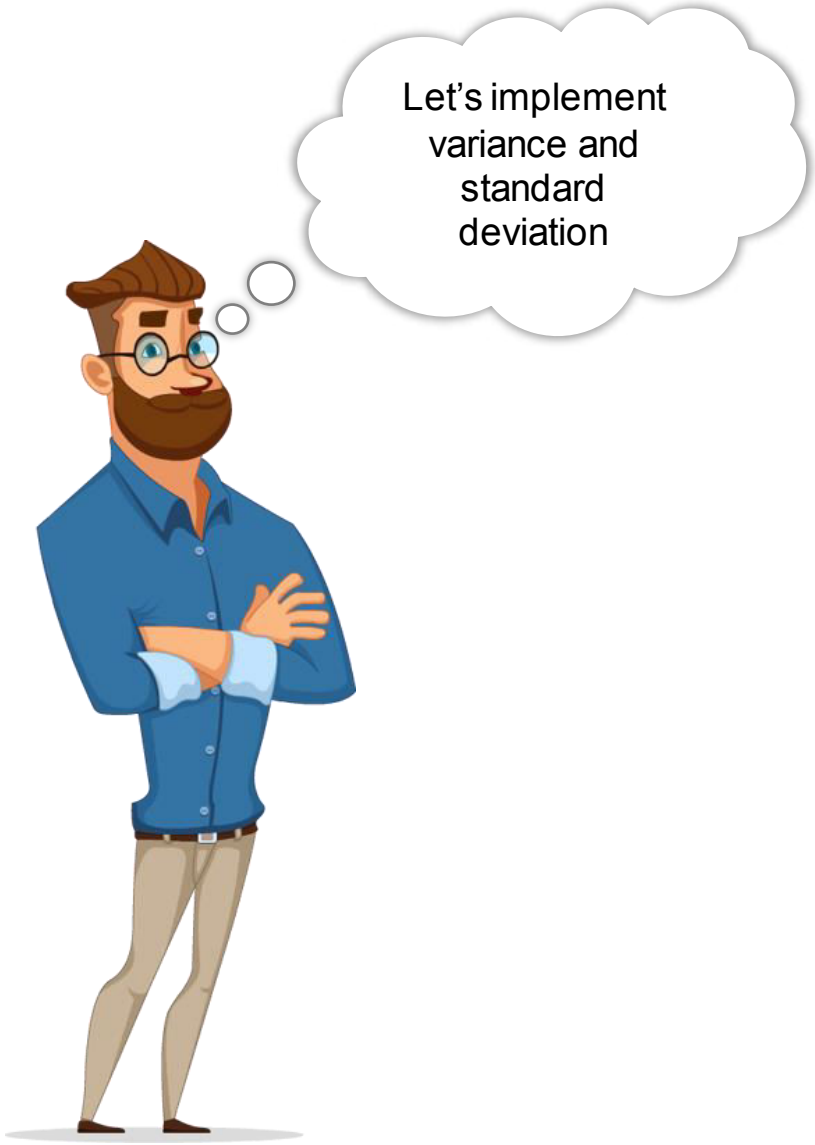
Standard deviation is the measure of spread most commonly used in statistical practice when the mean is used to calculate central tendency.

Standard deviation is also influenced by outliers one value could contribute largely to the results of the standard deviation.

Standard deviation is also useful when comparing the spread of two separate data sets that have approximately the same mean.

The more widely spread the values are, the larger the standard deviation is

Implementation in R



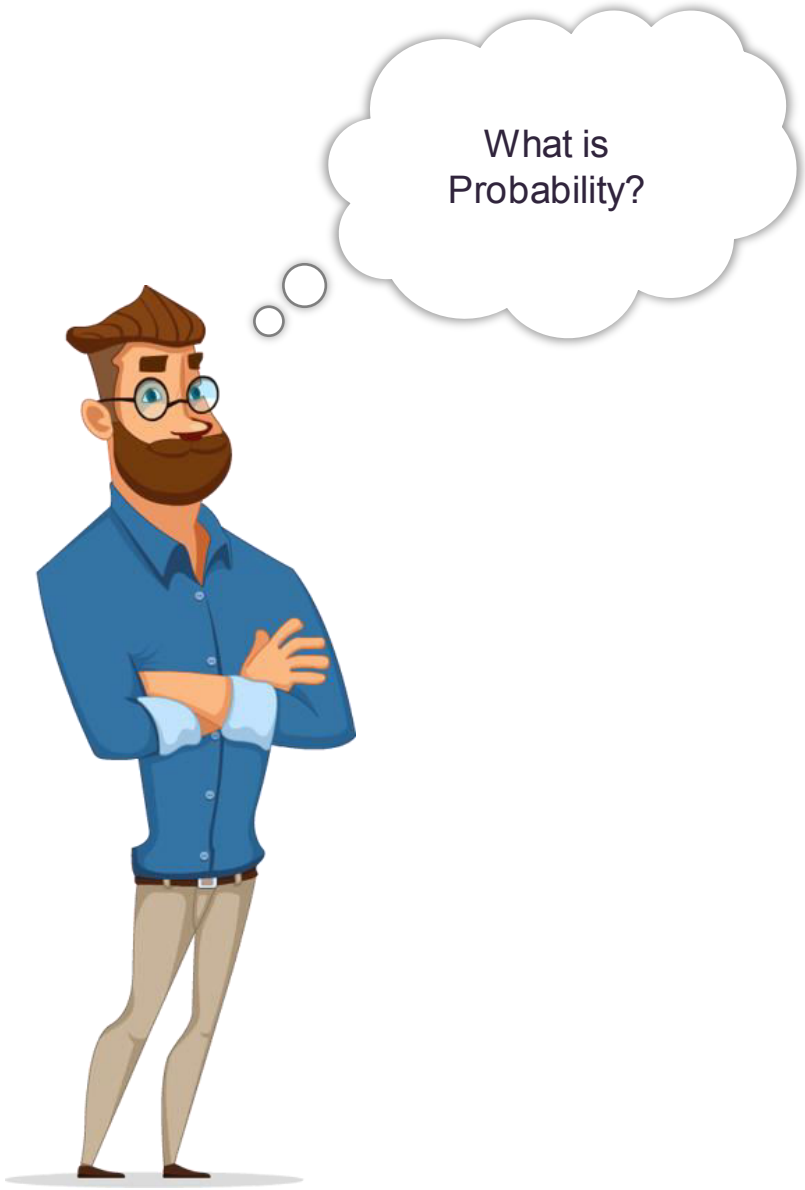
customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService
7590-VHVEG	Female	0	Yes	No	1	No
5575-GNVDE	Male	0	No	No	34	Yes
3668-QPYBK	Male	0	No	No	2	Yes
7795-CFOCW	Male	0	No	No	45	No
9237-HQITU	Female	0	No	No	2	Yes
9305-CDSKC	Female	0	No	No	8	Yes
1452-KIOVK	Male	0	No	Yes	22	Yes
6713-OKOMC	Female	0	No	No	10	No
7892-POOKP	Female	0	Yes	No	28	Yes
6388-TABGU	Male	0	No	Yes	62	Yes
9763-GRSKD	Male	0	Yes	Yes	13	Yes

```
var(customer_churn$MonthlyCharges)

sd(customer_churn$MonthlyCharges)
```

Probability

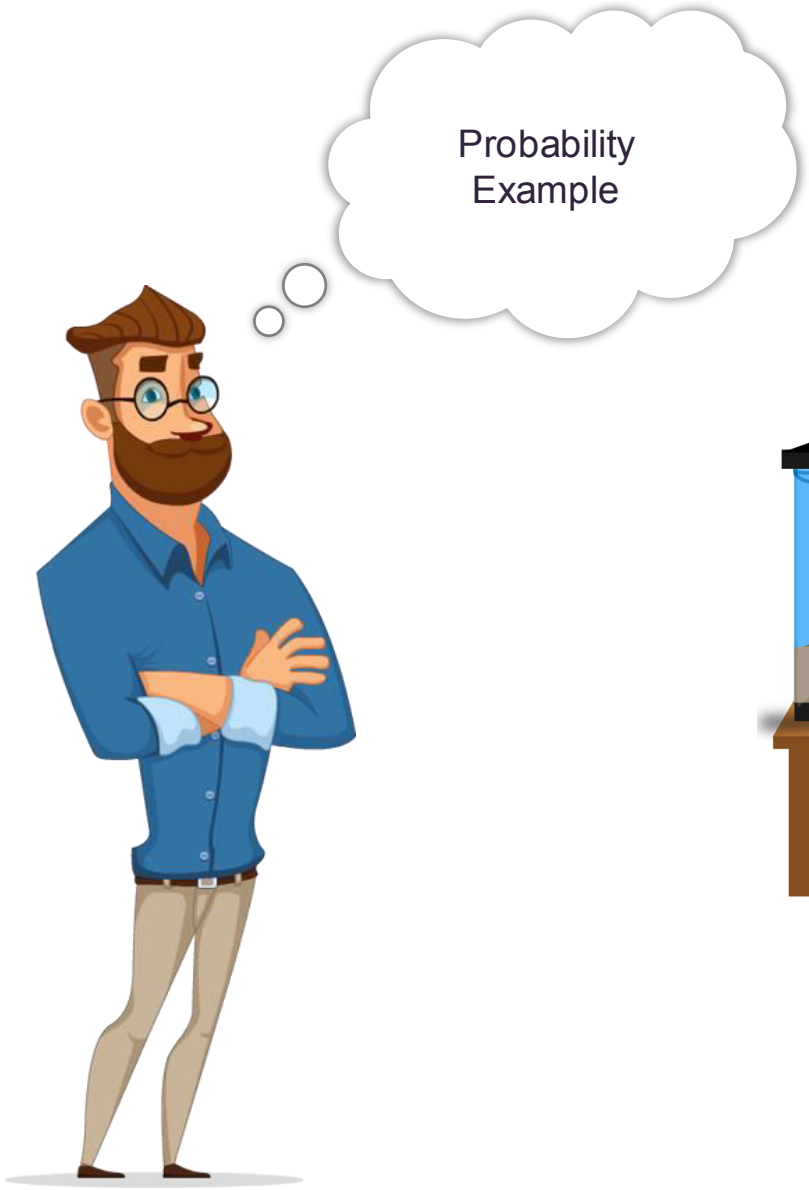
Probability

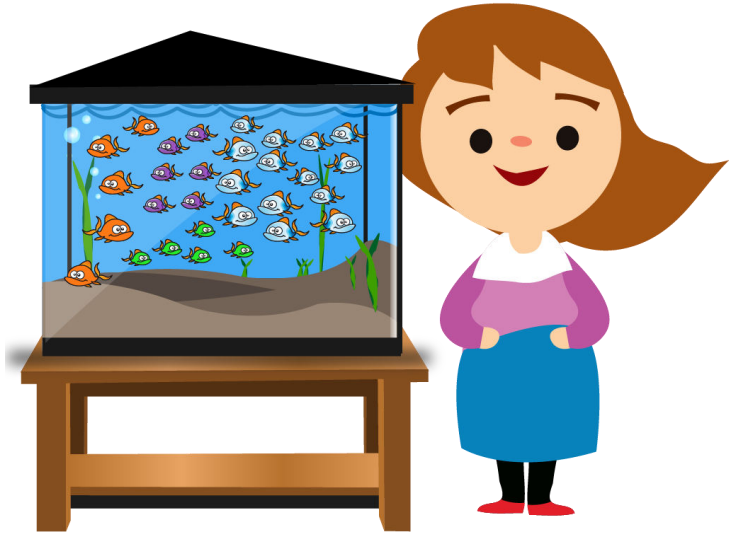


Probability can be determined post conducting a thought experiment

$$\mathrm{P}(E) = \frac{\#\ of\ times\ an\ event\ occured}{total\ \#\ of\ opportunities\ for\ the\ event\ to\ have\ occurred}$$

Probability



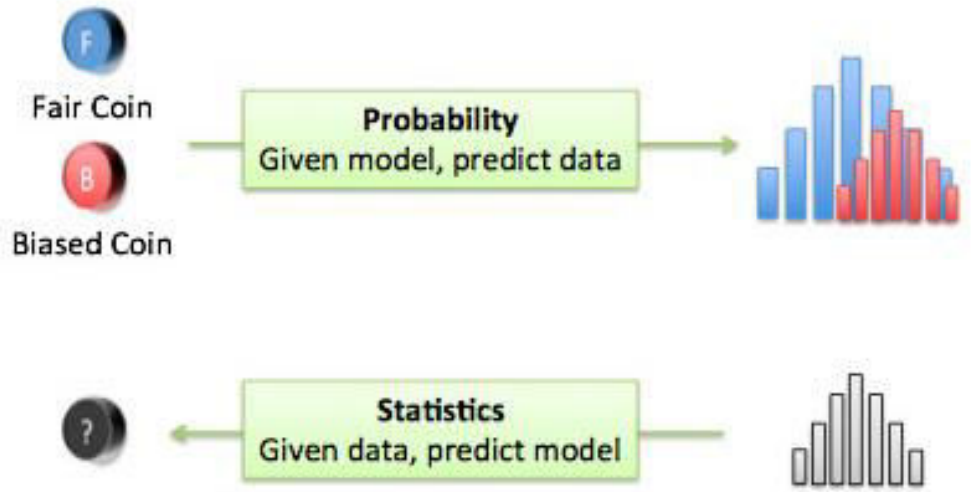


	=	14
	=	5
	=	5
	=	6
Total	=	30

Probability  = $\frac{6}{30}$ = $\frac{1}{5}$

Probability Vs Statistics

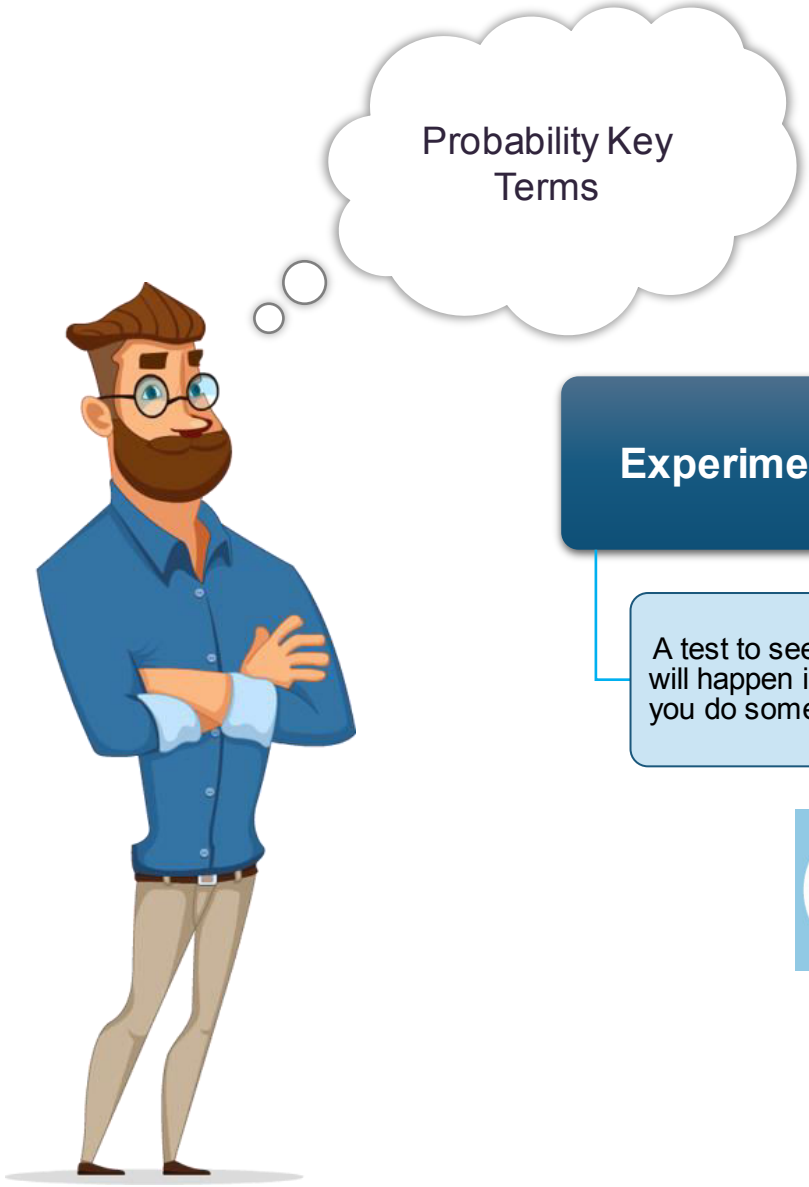
Probability Vs Statistics



Probability	• Predict the likelihood of future event • What will happen in a given ideal world
Statistics	• Analyze the past events • How ideal is the world?

Probability Terminology

Probability Terminology



Experiment

A test to see what will happen incase you do something.

Outcome

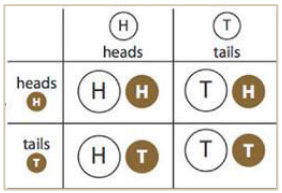
Refers to a single (one) result of an experiment.

Event

The set of a group of different outcomes of an experiment.

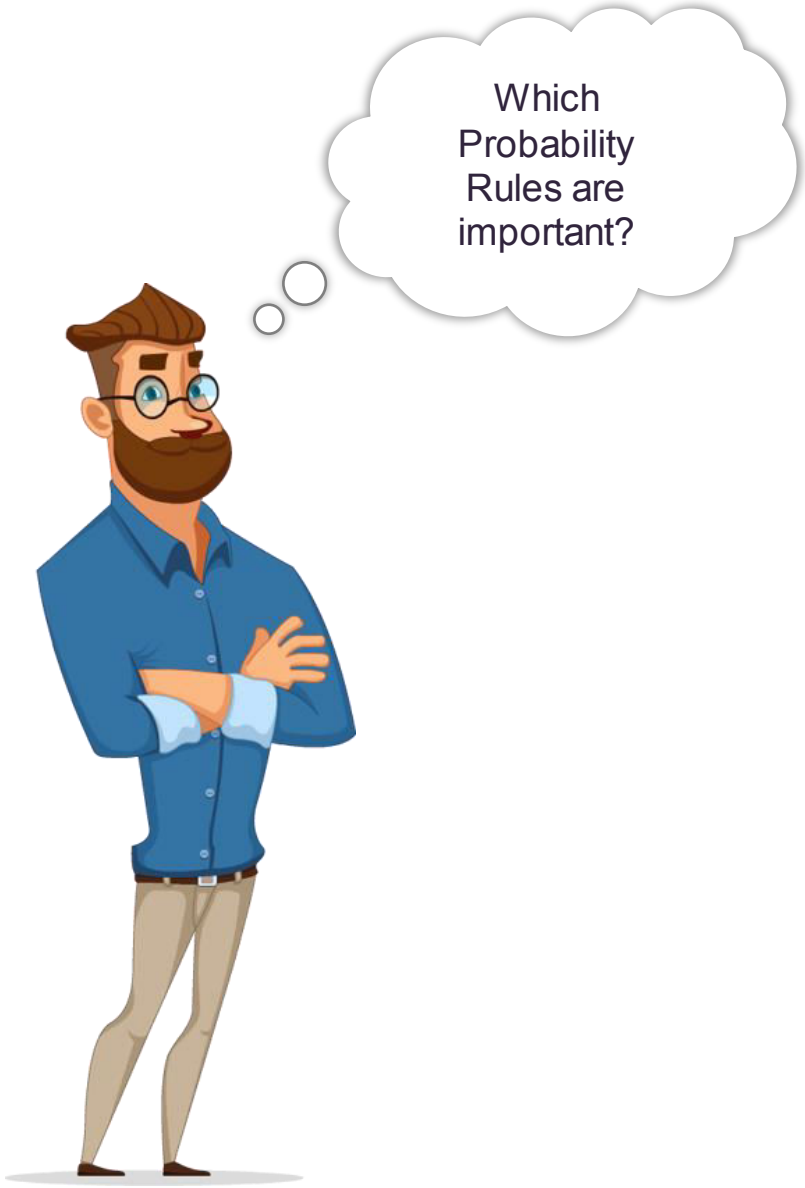
Sample Space

The total number of all the different possible outcomes of a given experiment.



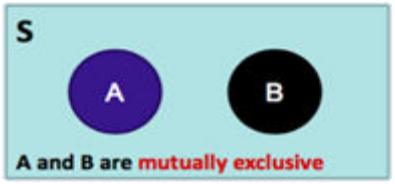
Probability Rules

Probability Rules

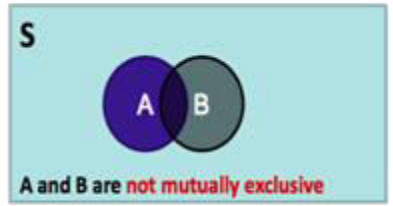


$$P(S) = 1$$

$$0 \leq P(A) \leq 1$$

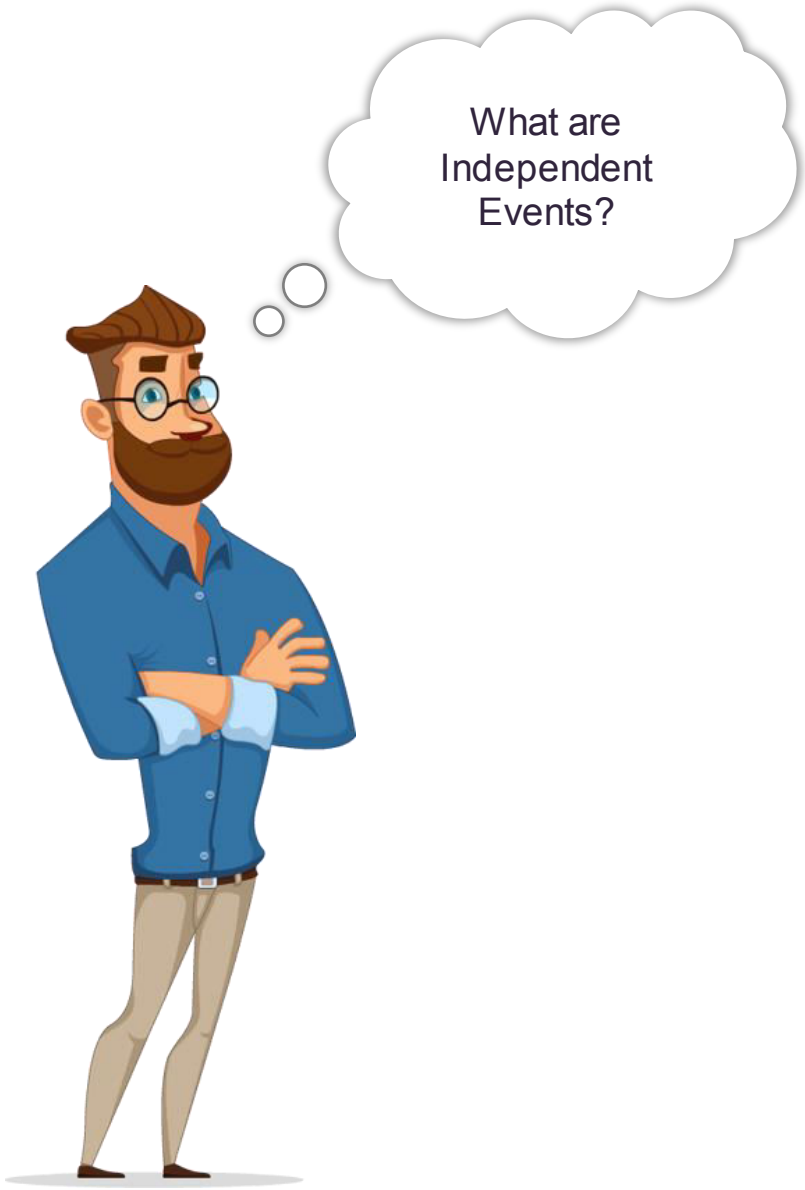


$$P(A \text{ or } B) = P(A) + P(B)$$



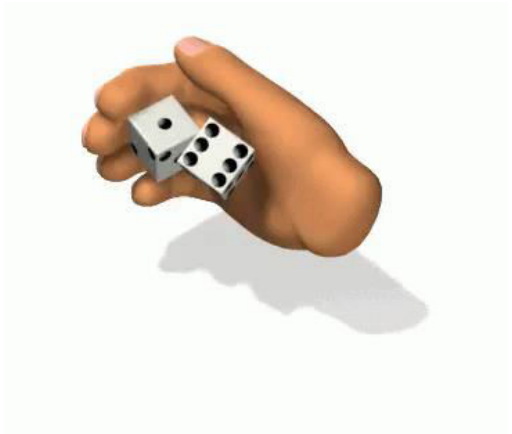
$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

Independent Events



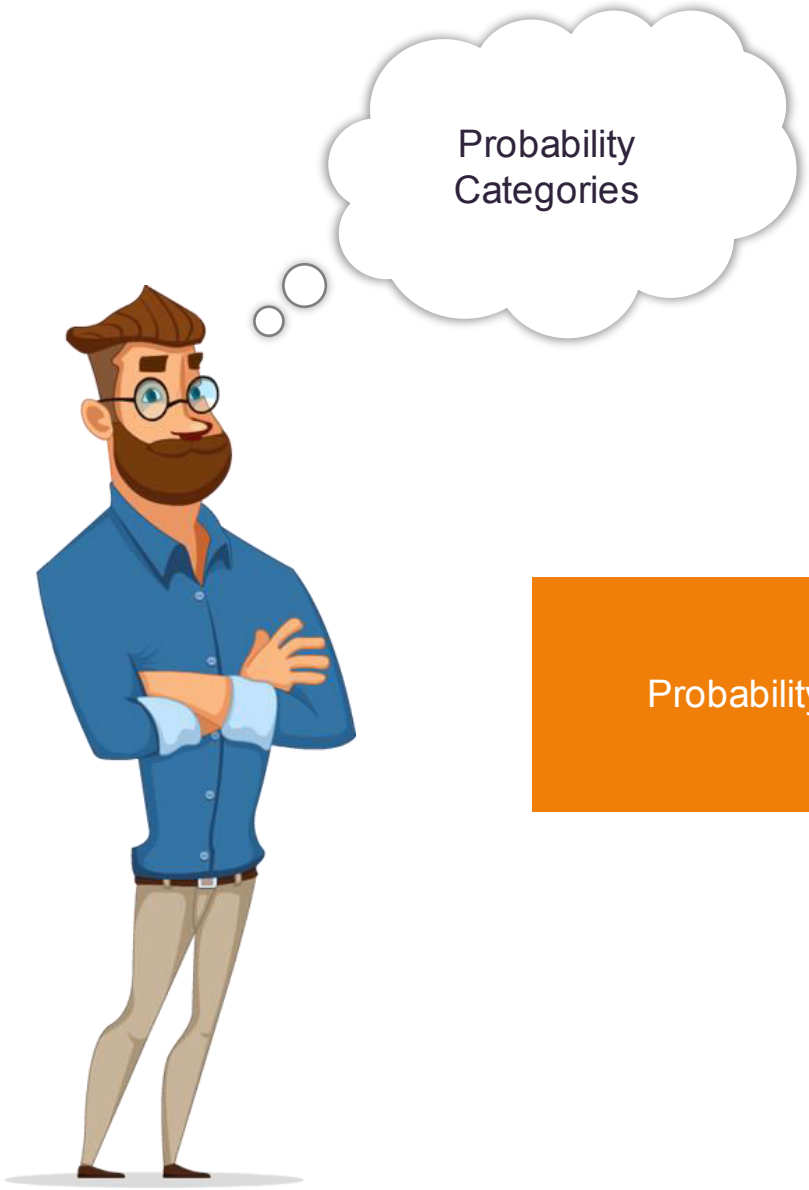
Outcome of event B is not dependent on the outcome of event A.

P(A and B) = P(A) * P(B)



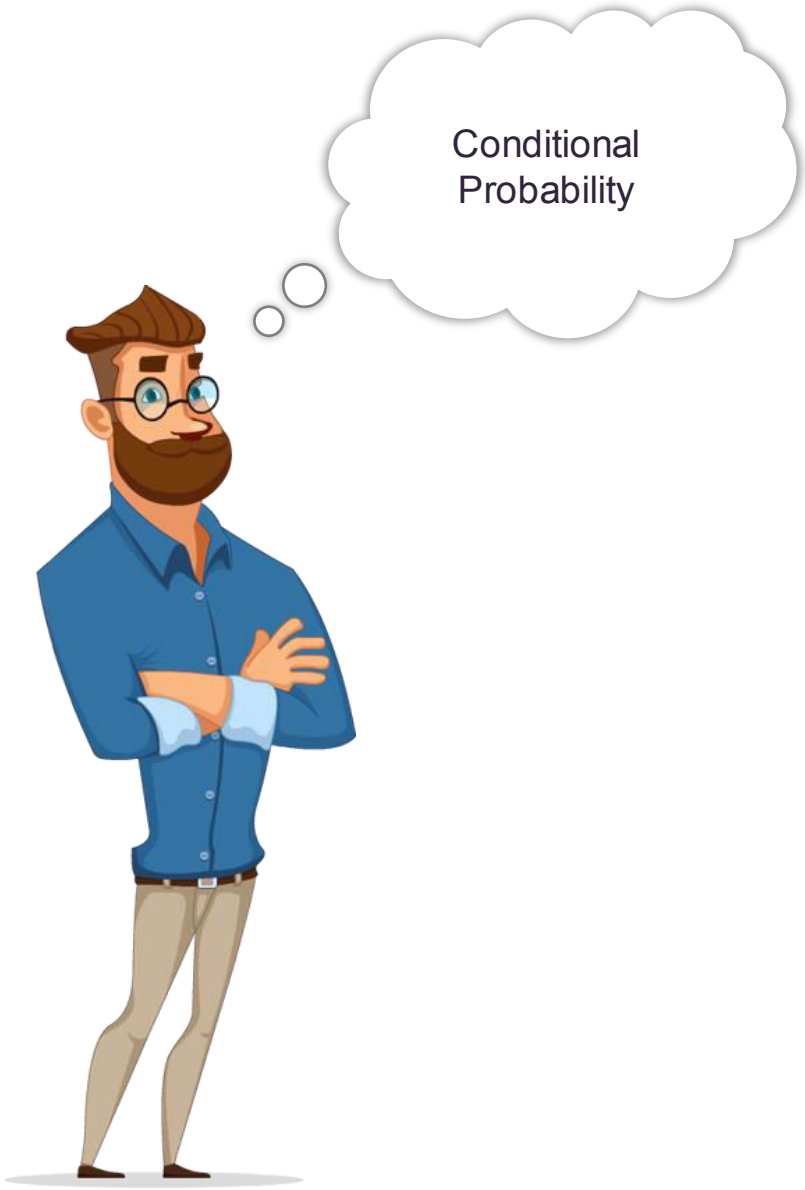
Probability Categories

Probability Categories



- Probability
  - Frequent
    - Find the chance of a particular outcome
  - Conditional
    - Measures the probability of an event if another event has already occurred

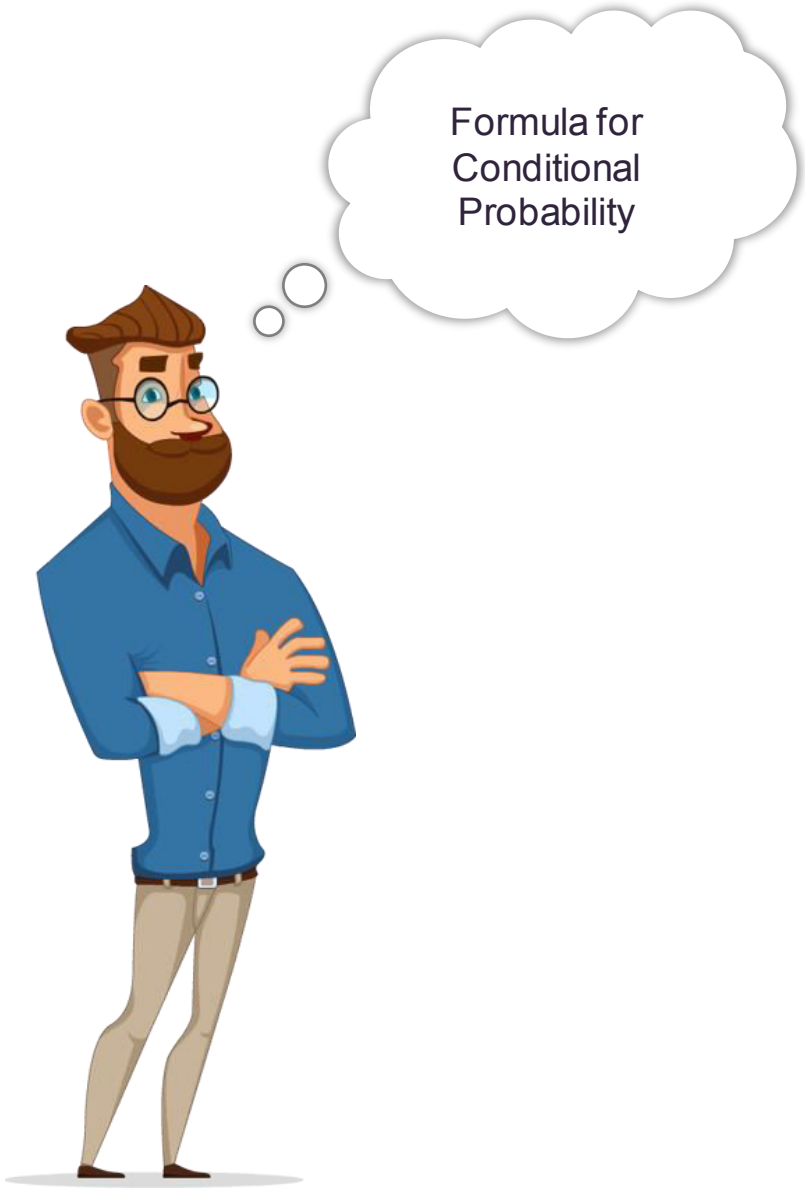
Conditional Probability



$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

$$\text{or } P(A \cap B) = P(A)P(B|A)$$

Conditional Probability



$$P(A|B) = \frac{P(A\ and\ B)}{P(B)} \Rightarrow P(A\ and\ B) = P(B) * P(A|B)$$

Similarly

$$P(B|A) = \frac{P(A\ and\ B)}{P(A)} \Rightarrow P(A\ and\ B) = P(A) * P(B|A)$$

Equating, we get

$$P(A|B) * P(B) = P(A) * P(B|A)$$

$$\therefore P(A|B) = \frac{P(A) * P(B|A)}{P(B)}$$

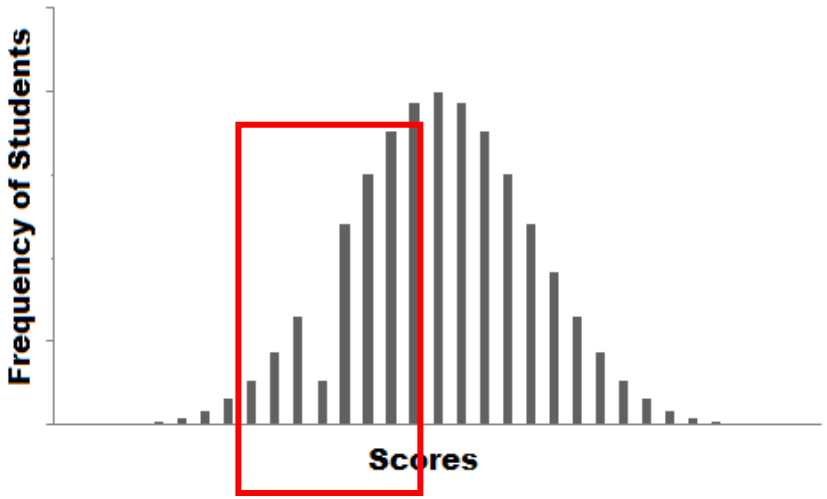
Probability Distribution

Probability Distribution

The probability function for a discrete random variable is the probability mass function.

It shows the exact probabilities for a particular value of the random variable.

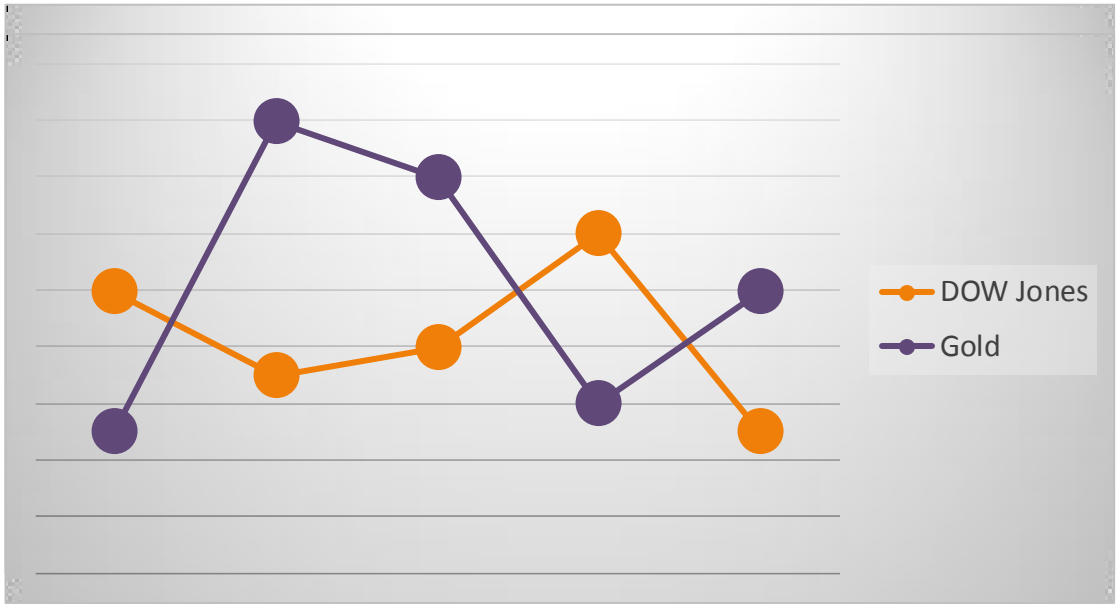
S. No.	Scores
1	25
2	27
3	38
4	42
5	
6	16
7	35
8	46
9	48
10	31



Covariance and Correlation

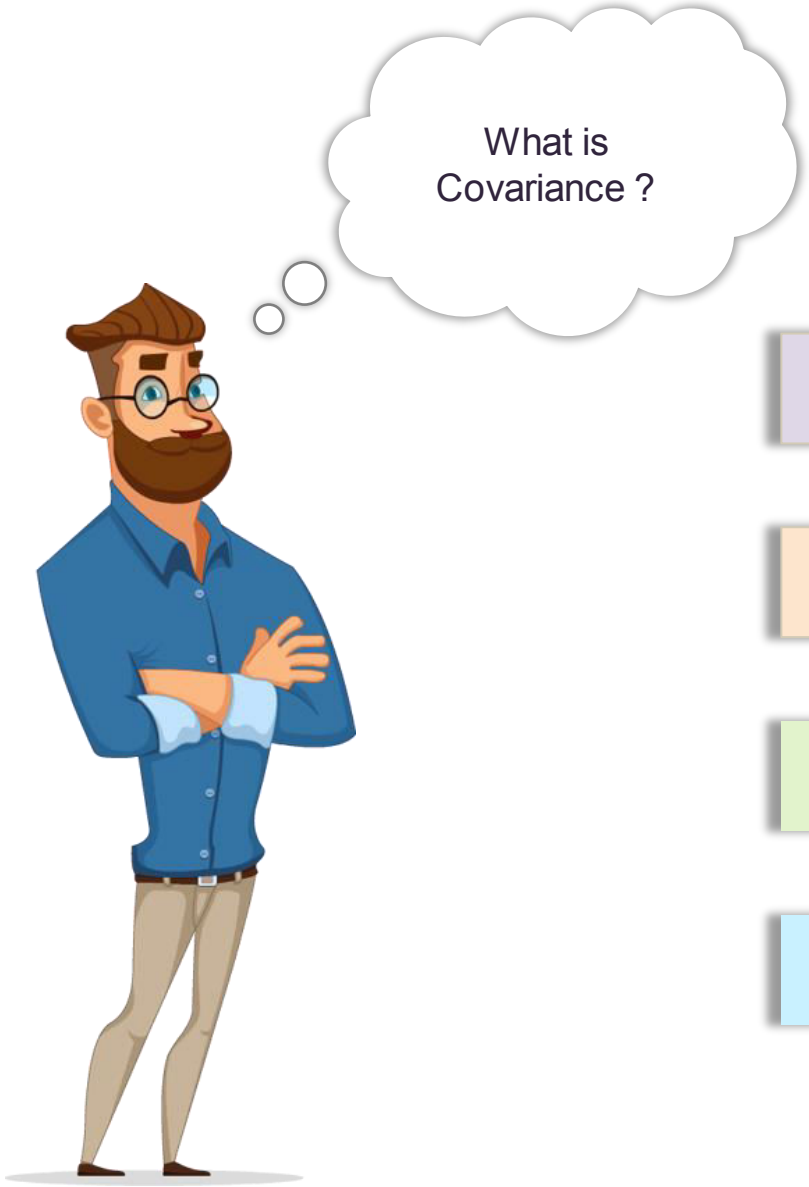
Correlation

When you say that two items correlate, you are saying that the change in one item effects a change in another item. You will always talk about correlation as a range between -1 and 1.



Correlation values are dependent on units of measure of "X" and "Y."

Covariance



If you say that two items tend to vary together then you are talking about the covariance between the two items which can be positive or negative covariance.

Positive covariance indicates that higher than average values of one variable tend to be paired with higher than average values of the other variable.

Negative covariance indicates that higher than average values of one variable tend to be paired with lower than average values of the other variable.

Covariance is not standardized, unlike the correlation coefficient. Therefore, covariance values can range from negative infinity to positive infinity.

Covariance & Correlation

- Covariance
  - Positive Covariance
  - Negative Covariance
- Correlation
  - Positive Correlation
  - Negative Correlation
  - Zero Correlation

Covariance & Correlation

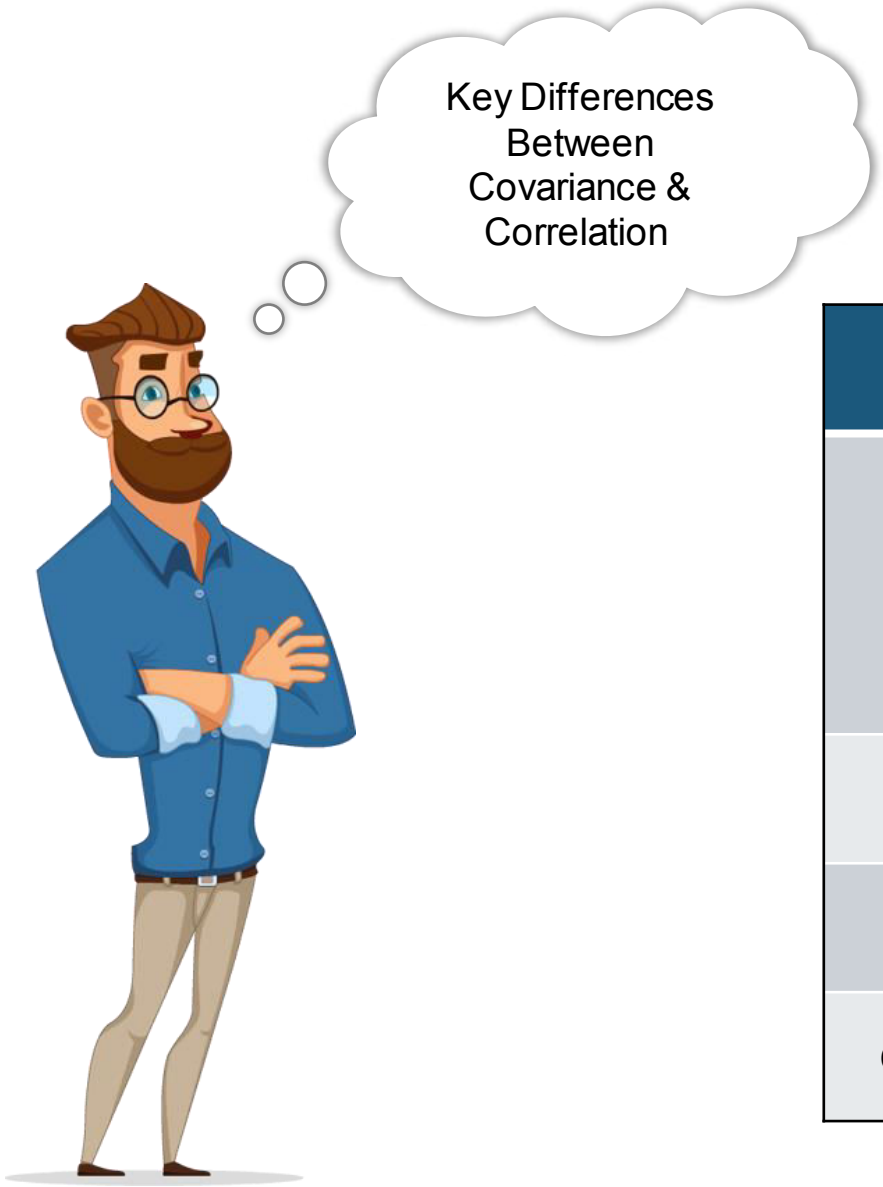
Correlation is dimensionless, i.e. it is a unit-free measure of the relationship between variables. Unlike covariance, where the value is obtained by the product of the units of the two variables.

Covariance & Correlation

$$COV(X,Y) = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{n-1}$$

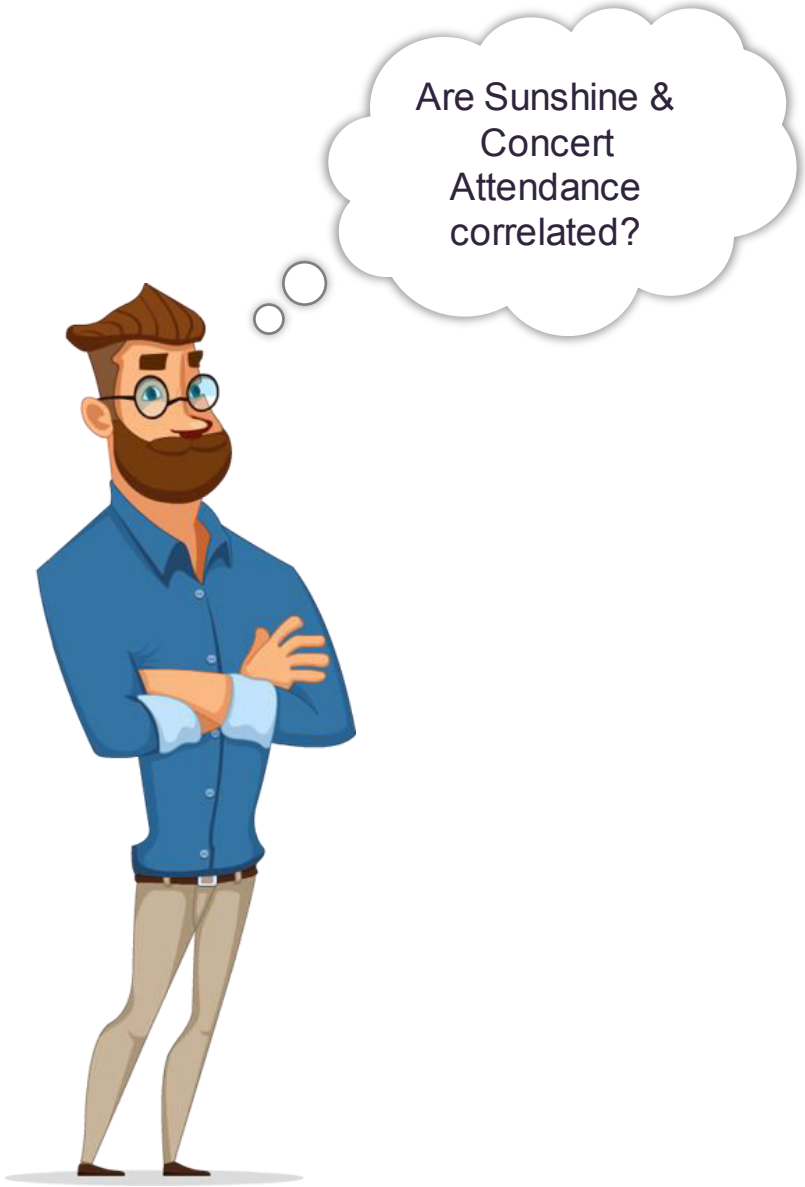
$$Correlation = \rho = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$

Covariance & Correlation



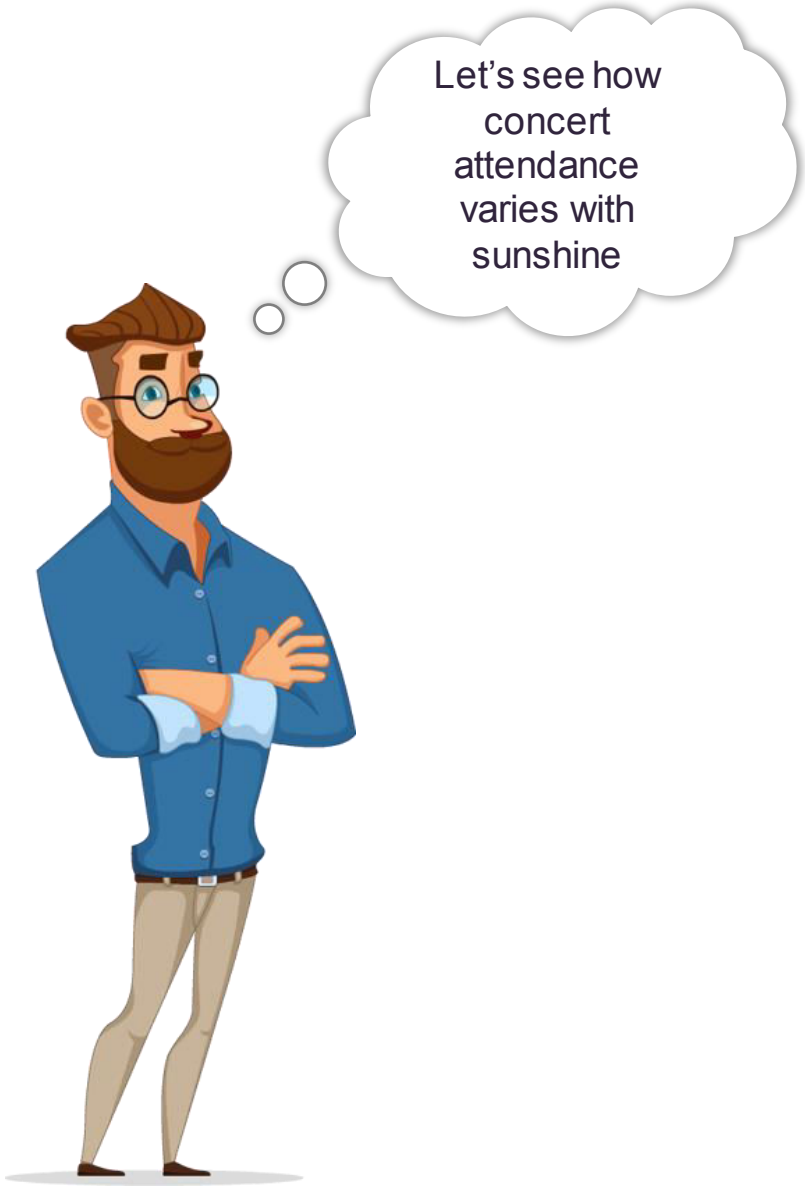
Basis for comparison	Covariance	Correlation
Meaning	A measure indicating the extant to which two random variables change in tandem	A measure indicating that how strongly two random variables are related
What is it?	Measure of correlation	Scaled version of covariance
Values	Lies between -∞ and +∞	Lies between -1 and +1
Change in scale	Affects covariance	Does not affects correlation

Covariance & Correlation



Sunshine (hours)	1.9	2.5	3.2	3.8	4.7	5.5	5.9	7.2
Concert attendance (100s)	22	33	30	42	38	49	42	55

Covariance & Correlation



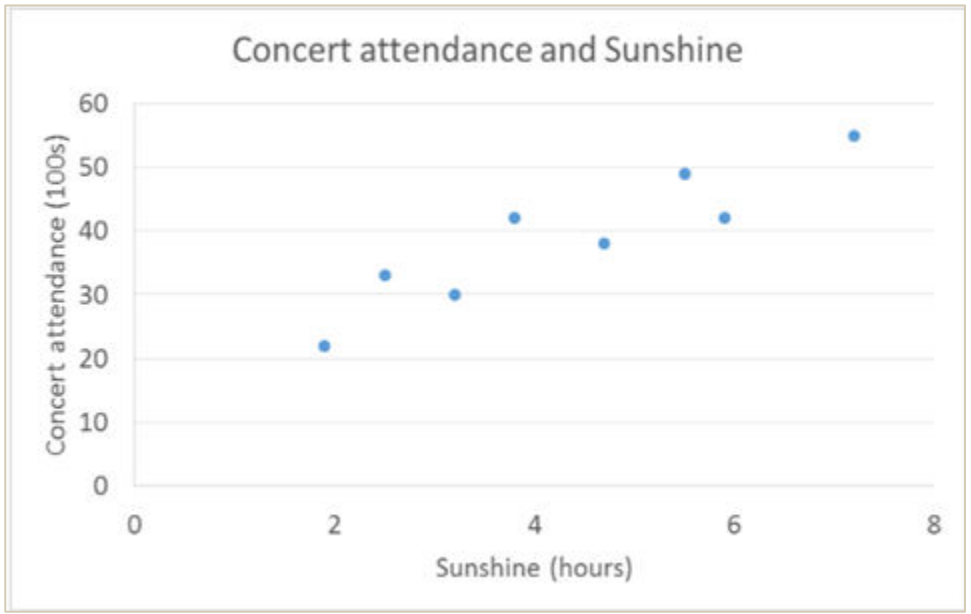
Independent variable (explanatory) – Sunshine – Plotted on X - axis

Dependent variable (response) – Concert attendance – Plotted on Y - axis

Sunshine (hours)	1.9	2.5	3.2	3.8	4.7	5.5	5.9	7.2
Concert attendance (100s)	22	33	30	42	38	49	42	55

Covariance & Correlation

The number of attendees increase with increase in Sunshine



Covariance & Correlation

Positive Linear Correlation

Negative Linear Correlation

No Correlation

Correlation Coefficient

Correlation Coefficient

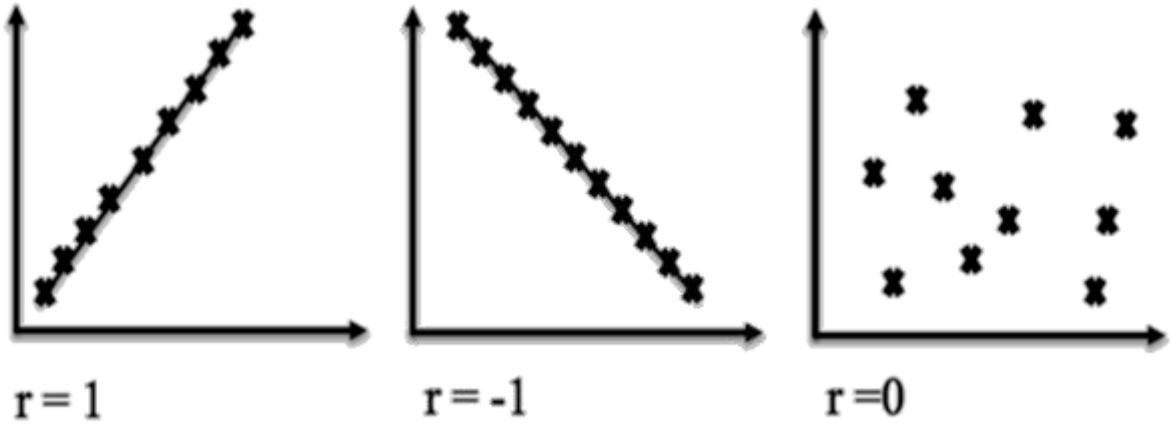
Correlation coefficient are used to find how strong a relationship is between data.

It returns a value between -1 and 1, where:

1	-1	0
indicates a strong positive relationship.	indicates a strong negative relationship.	indicates no relationship at all.

Correlation Coefficient

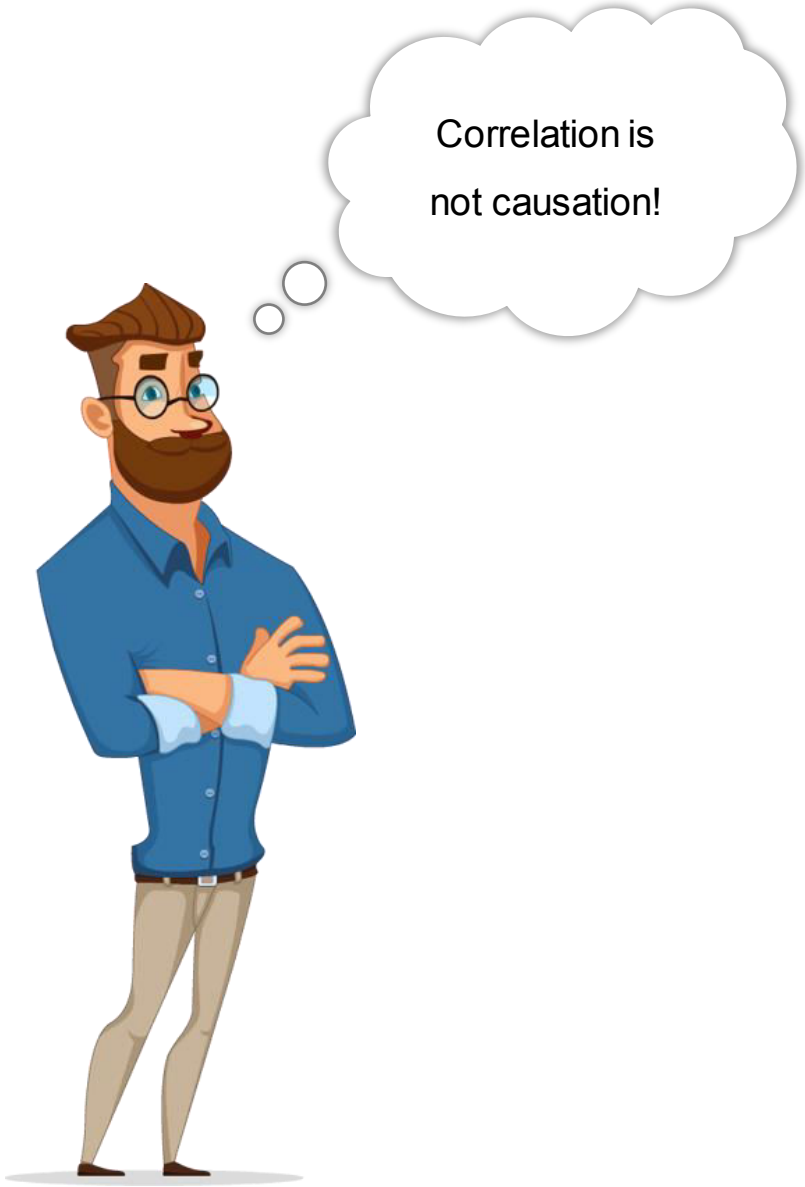
$$r = r_{xy} = \frac{\mathrm{Cov}(x, y)}{S_x \times S_y}$$



Correlation is not Causation!

Because two things correlate does not necessarily mean that one causes the other

Correlation does not cause Causation



Correlation - describes the size and direction of a relationship between two or more variables

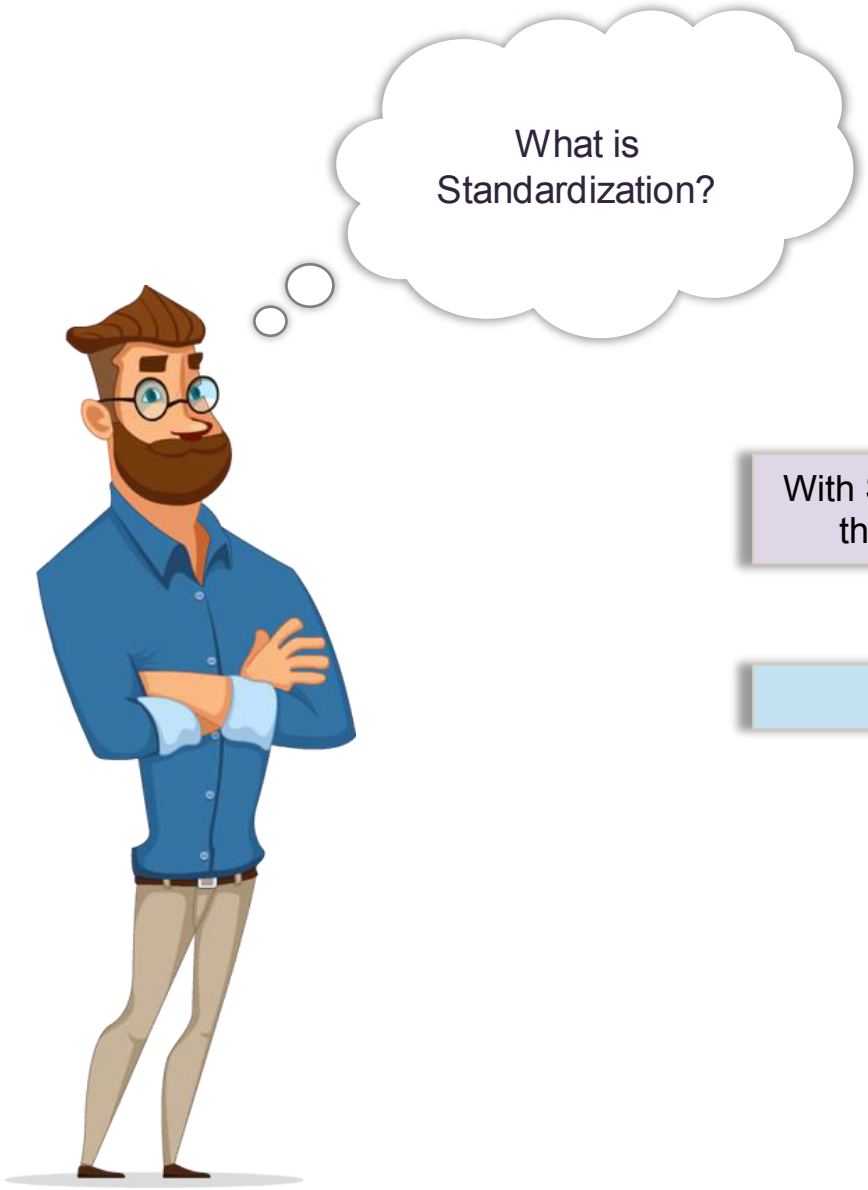
Causation - indicates that one event is the result of the occurrence of the other event

Smoking → Correlate → Alcoholism

Smoking → Causes → The risk of developing Lung Cancer

Standardization & Normalization

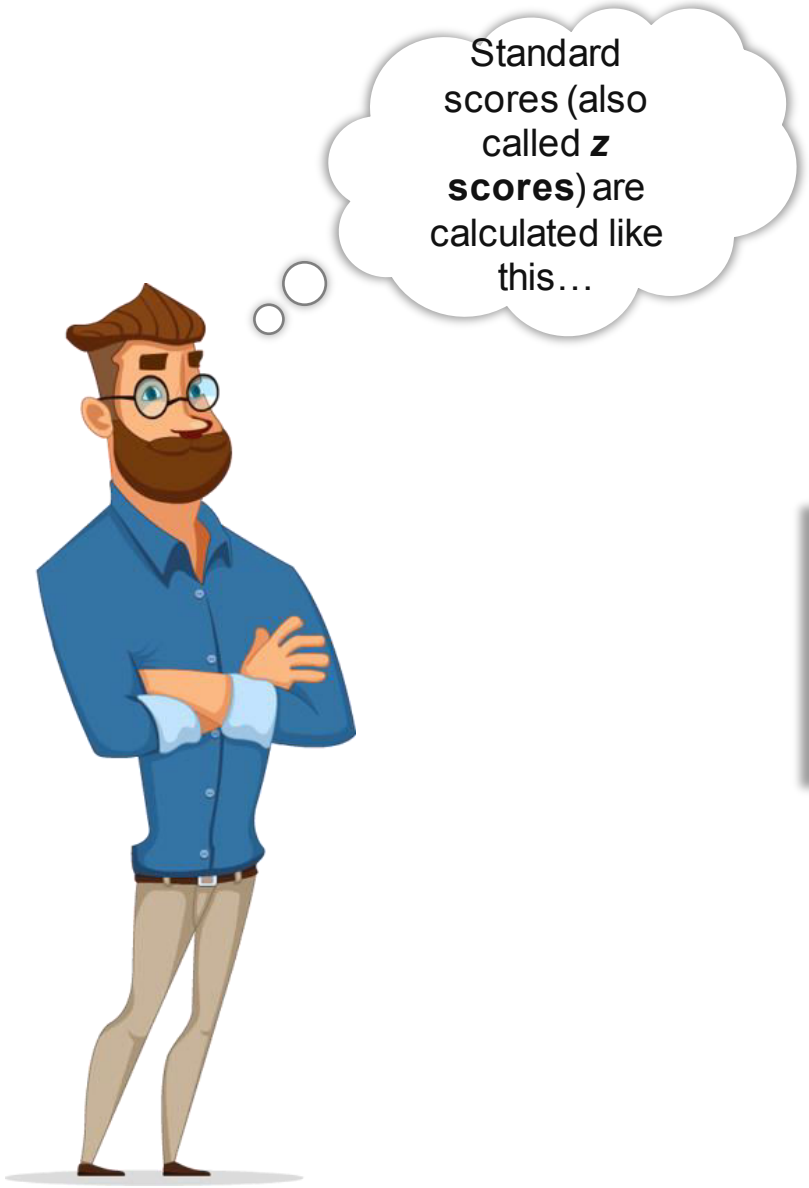
Standardization



With Standardization (or Z-score normalization) the features can be rescaled so that they'll have the properties of a standard normal distribution with μ=0 and σ=1

Where, μ is the mean and σ is the standard deviation from the mean

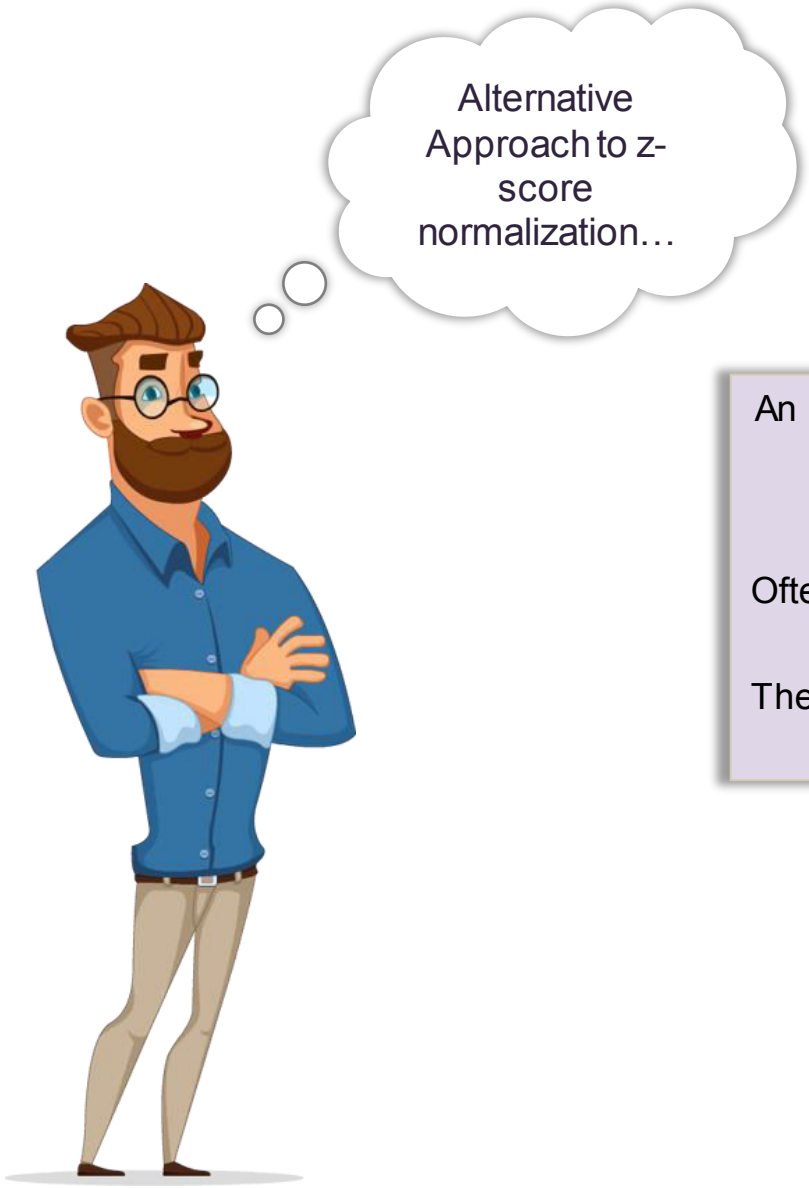
Standardization



$$z = x - \mu / \sigma$$

Standardizing the features so that they are centered around 0 with a standard deviation of 1 is not only important if we are comparing measurements that have different units, but it is also a general requirement for many machine learning algorithms

Normalization

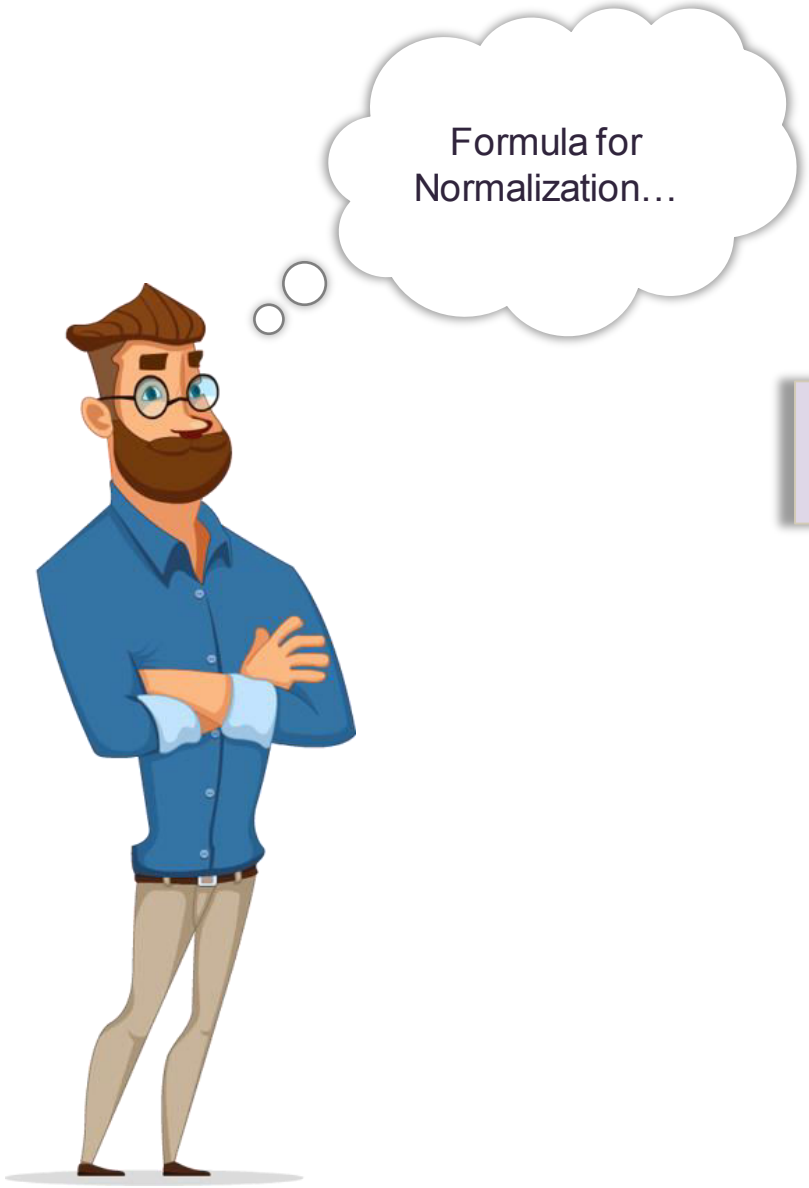


An alternative approach to Z-score normalization (or standardization) is the so-called **Min-Max scaling**

Often also simply called “Normalization” - a common cause for ambiguities.

The data is scaled to a fixed range - 0 to 1.

Normalization



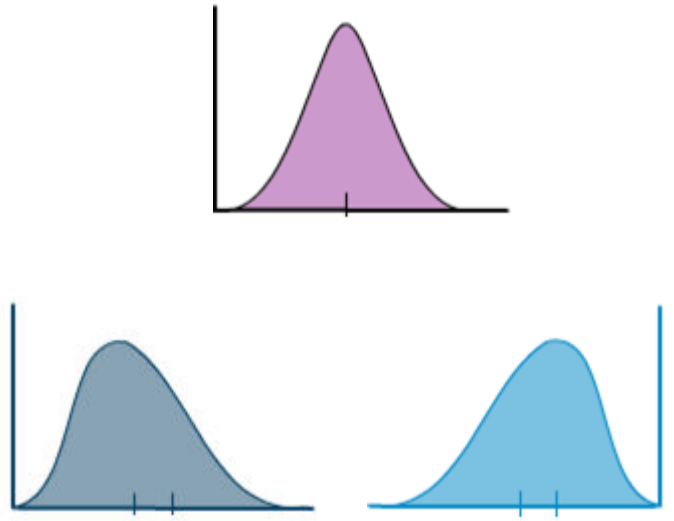
The cost of having this bounded range - in contrast to standardization - is that we will end up with smaller standard deviations, which can suppress the effect of outliers

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

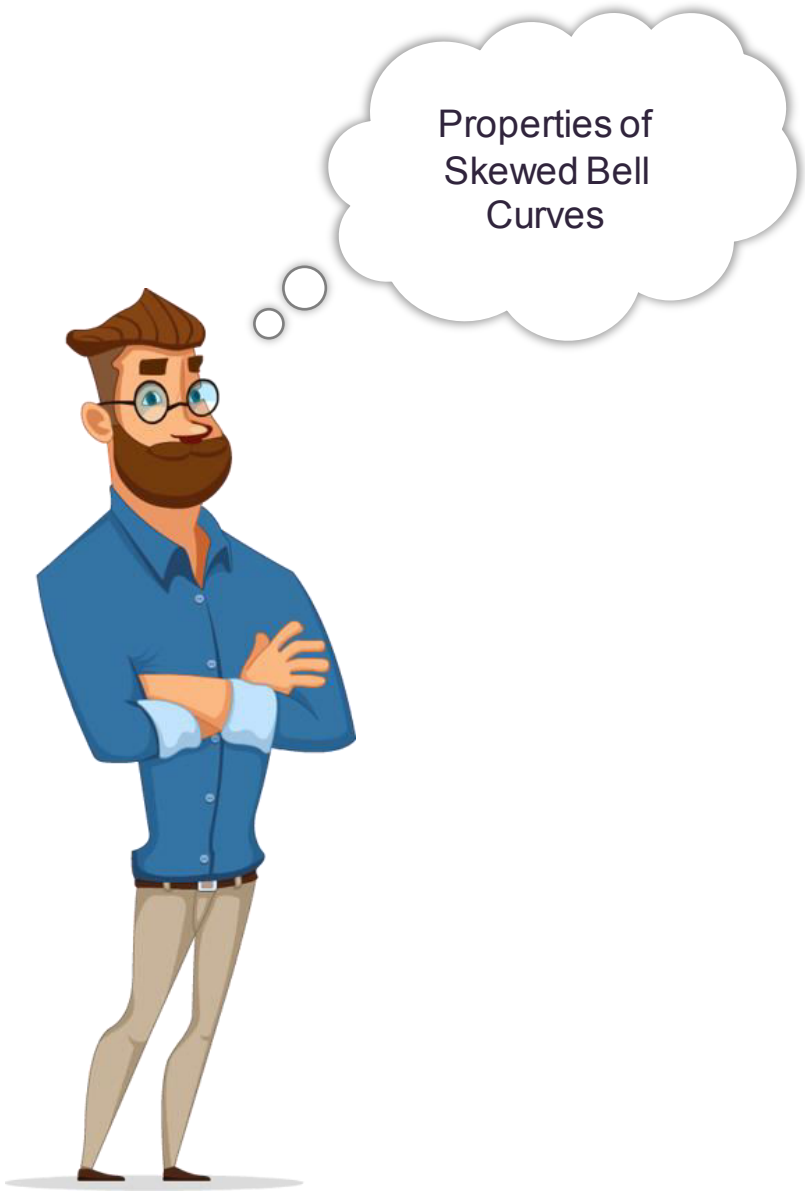
Skewness

Skewness

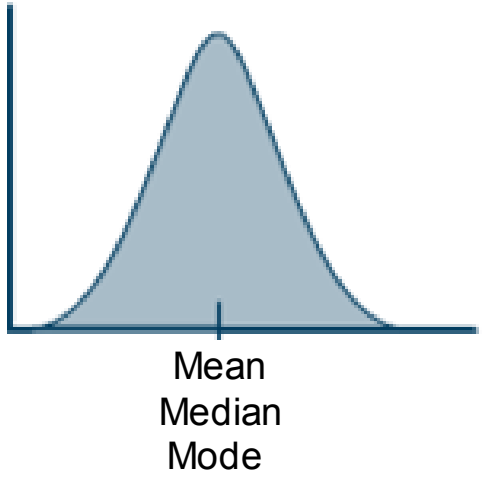
Represents an imbalance and asymmetry from the mean of a data distribution



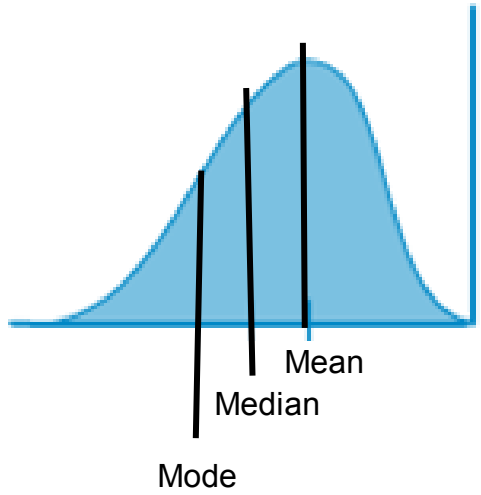
Skewness



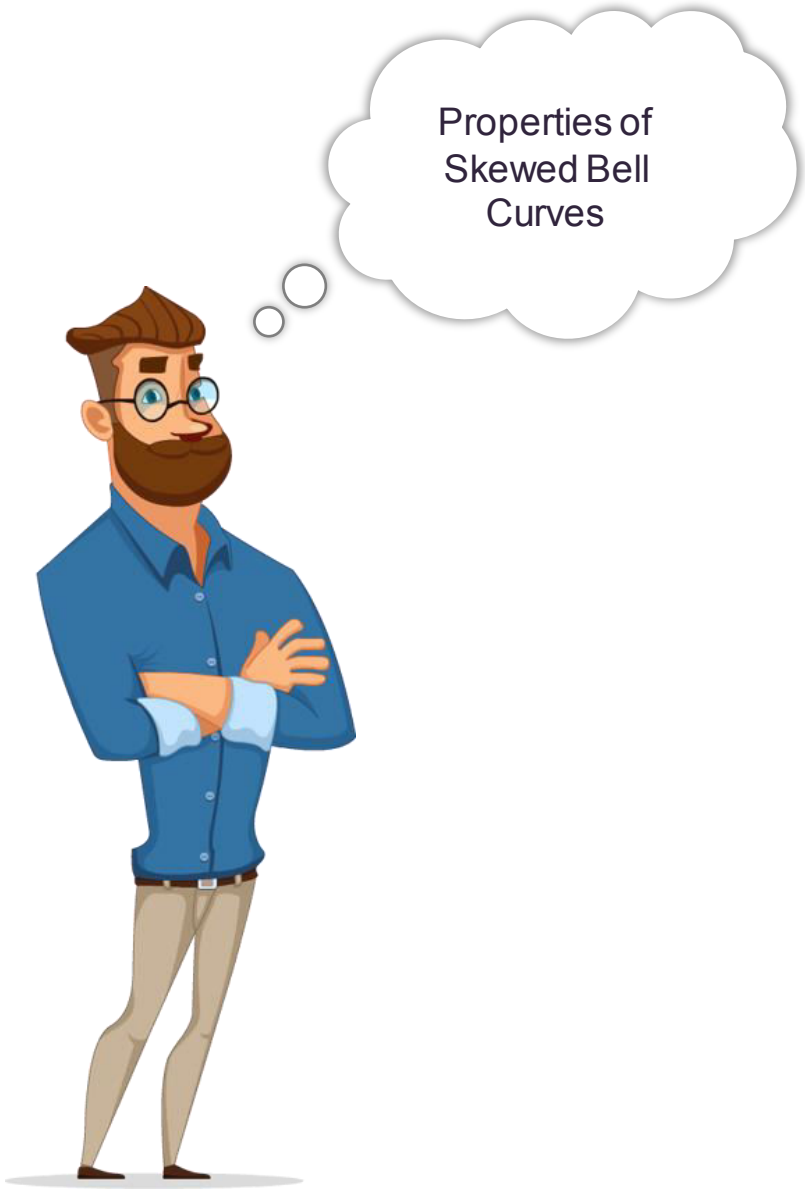
Symmetric Distribution

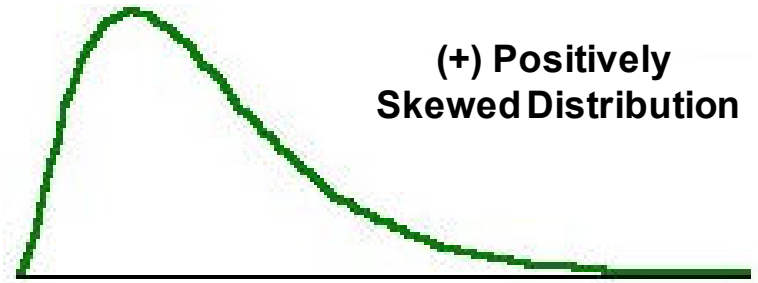


Skewed Distribution



Skewness



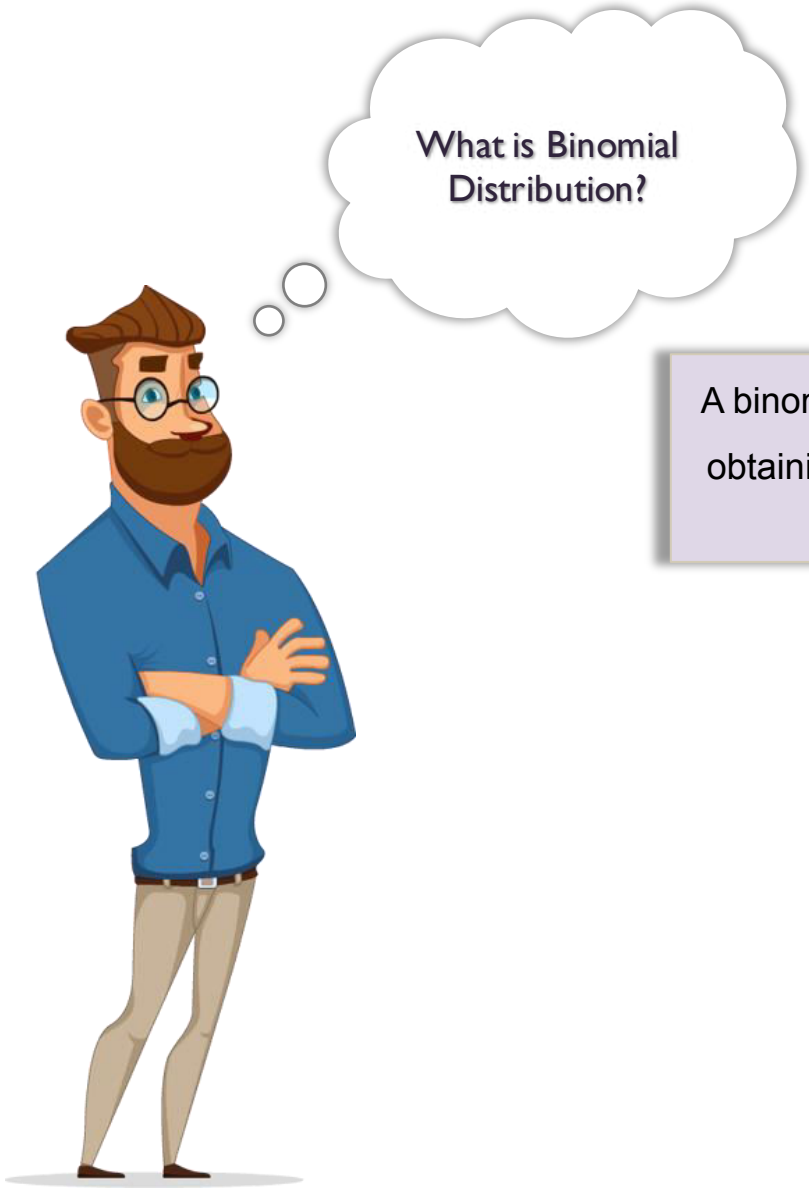


(+) Positively Skewed Distribution

(-) Negatively Skewed Distribution

Binomial Distribution

Binomial Distribution



A binomial distribution is a specific probability distribution. It is used to model the probability of obtaining one of two outcomes, a certain number of times (k), out of fixed number of trials (N) of a discrete random event.

Expected Outcome
Successful Outcome

p

Other Outcome
Failure

$1 - p$

Binomial Distribution

Rule #1

There are only two mutually exclusive outcomes for a discrete random variable

Rule #2

There is a fixed number of repeated trials

Rule #3

Each trial is an independent event

Rule #4

The probability of success for each trial is fixed

Binomial Distribution

The probability of getting x successes in n Bernoulli trials

$$P(x) = \frac{n!}{(n-x)!\,x!}\,p^x q^{n-x}$$

This starts the count of number of ways event can occur

This is the probability of success for x trials

This is the probability of success for x trials

This ends the count of number of ways event can occur

This deletes duplications

Geometric Distribution

Geometric Distribution

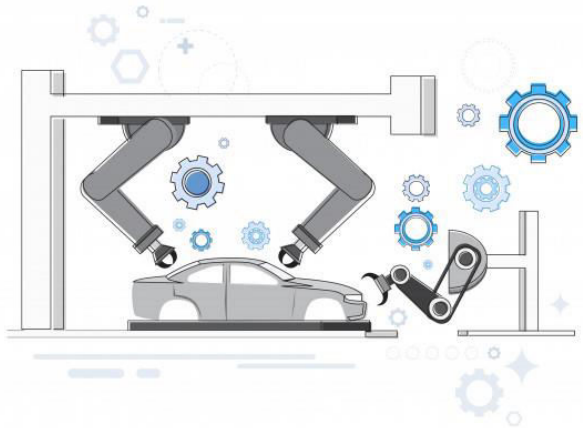
The geometric distribution represents the number of failures before you get a success in a series of Bernoulli trials

$$f(x) = (1 - p)^{x-1}p$$

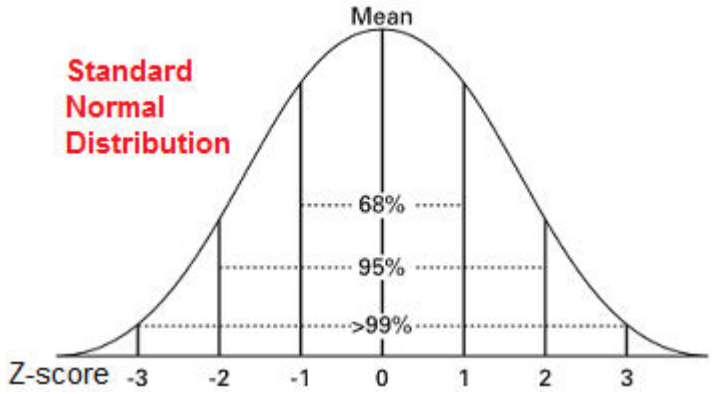
Normal Distribution

Normal Distribution

84th percentile or above

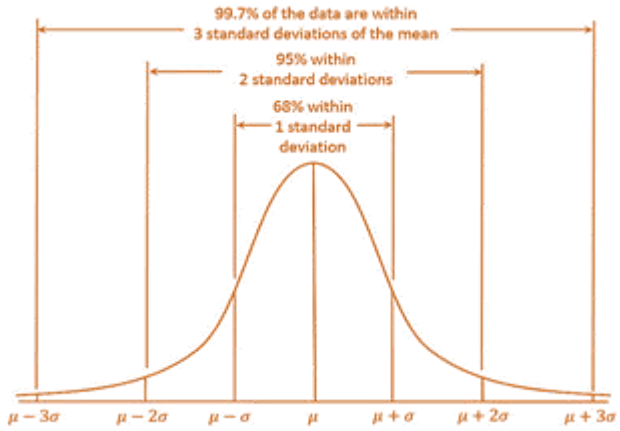


95% of the lithium ion batteries



Normal Distribution

- Symmetric bell shape
- Mean and median are equal
- The area under the normal curve is equal to 1.0
- Normal distributions are denser in the center and less dense in the tails
- 68% of the area of a normal distribution is within one standard deviation of the mean
- Approximately 95% of the area of a normal distribution is within two standard deviations of the mean



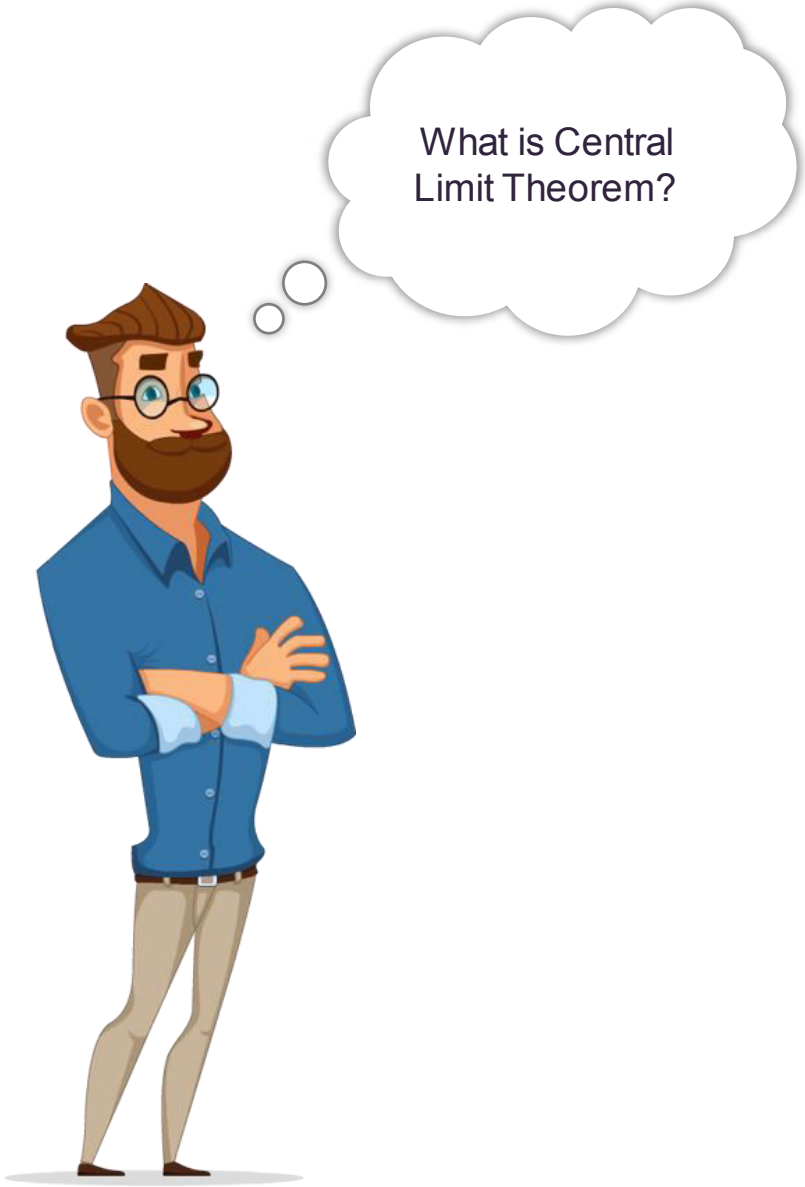
Central Limit Theorem

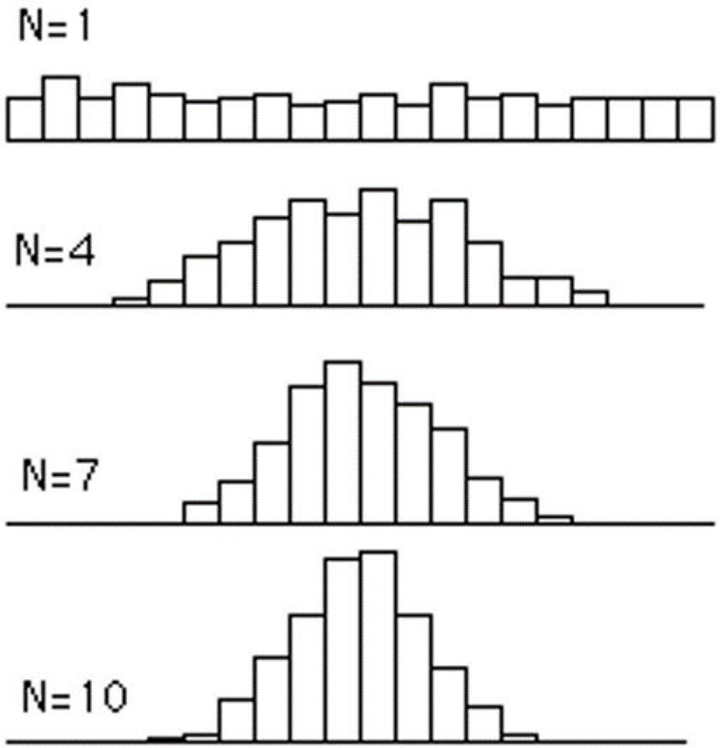
Central Limit Theorem

The Central Limit Theorem states that the sampling distribution of the sample means approaches a normal distribution as the sample size gets larger — no matter what the shape of the population distribution.

This fact holds especially true for sample sizes over 30.

Central Limit Theorem



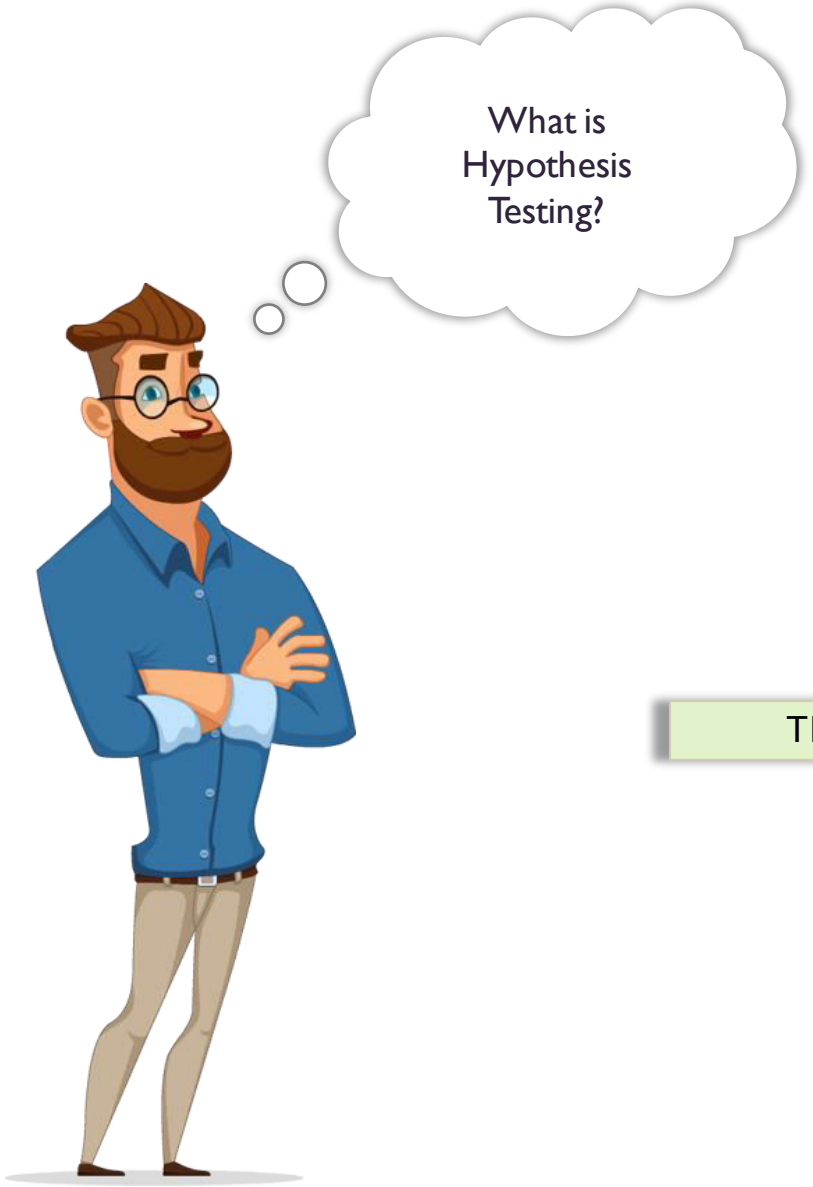


When n increases:

1. the distributions becomes more and more normal.
2. the spread of the distributions decreases.

Hypothesis Testing

Hypothesis Testing

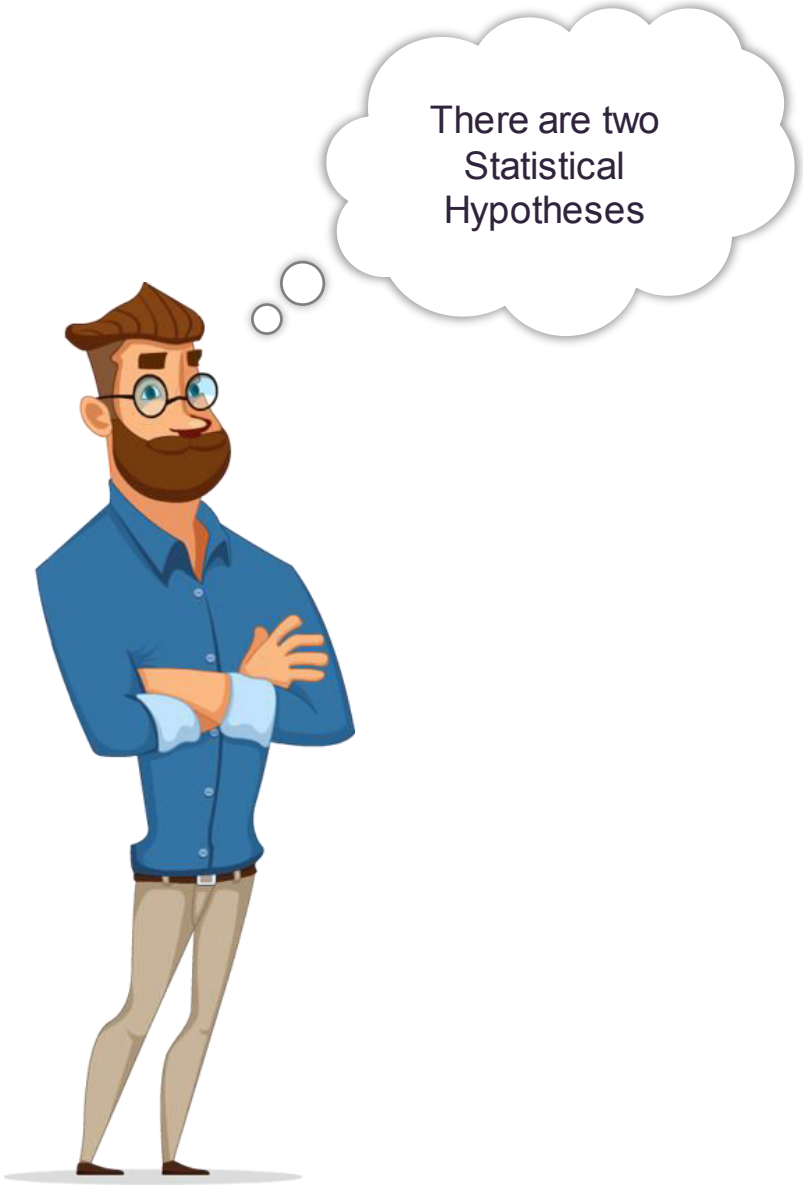


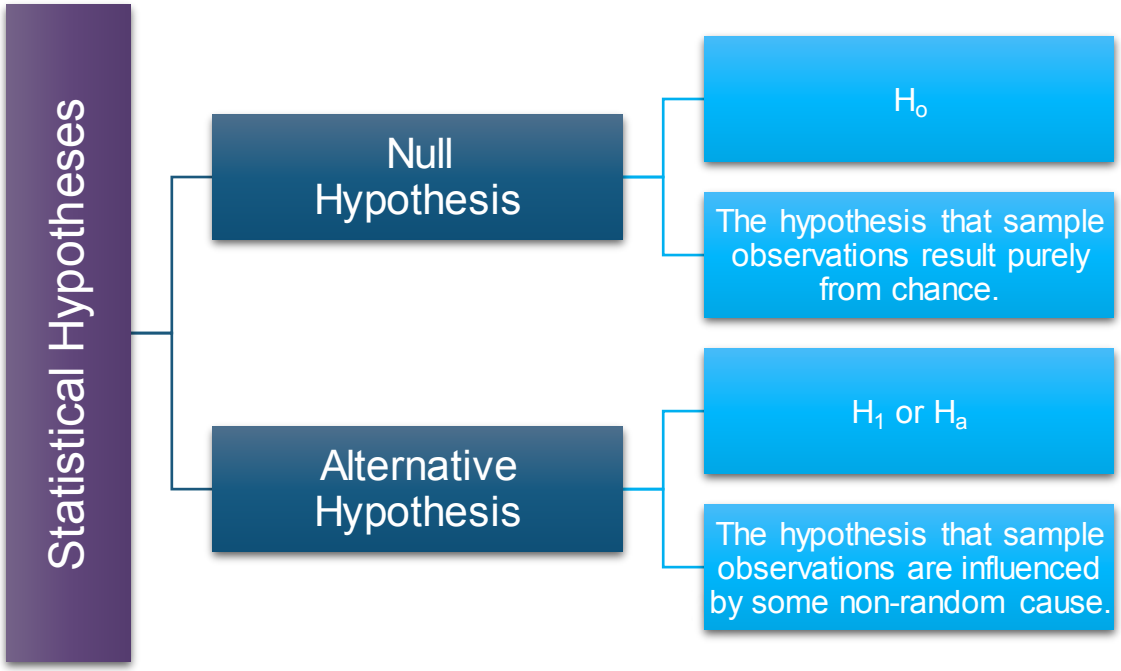
An assumption about a population parameter

This assumption may or may not be true

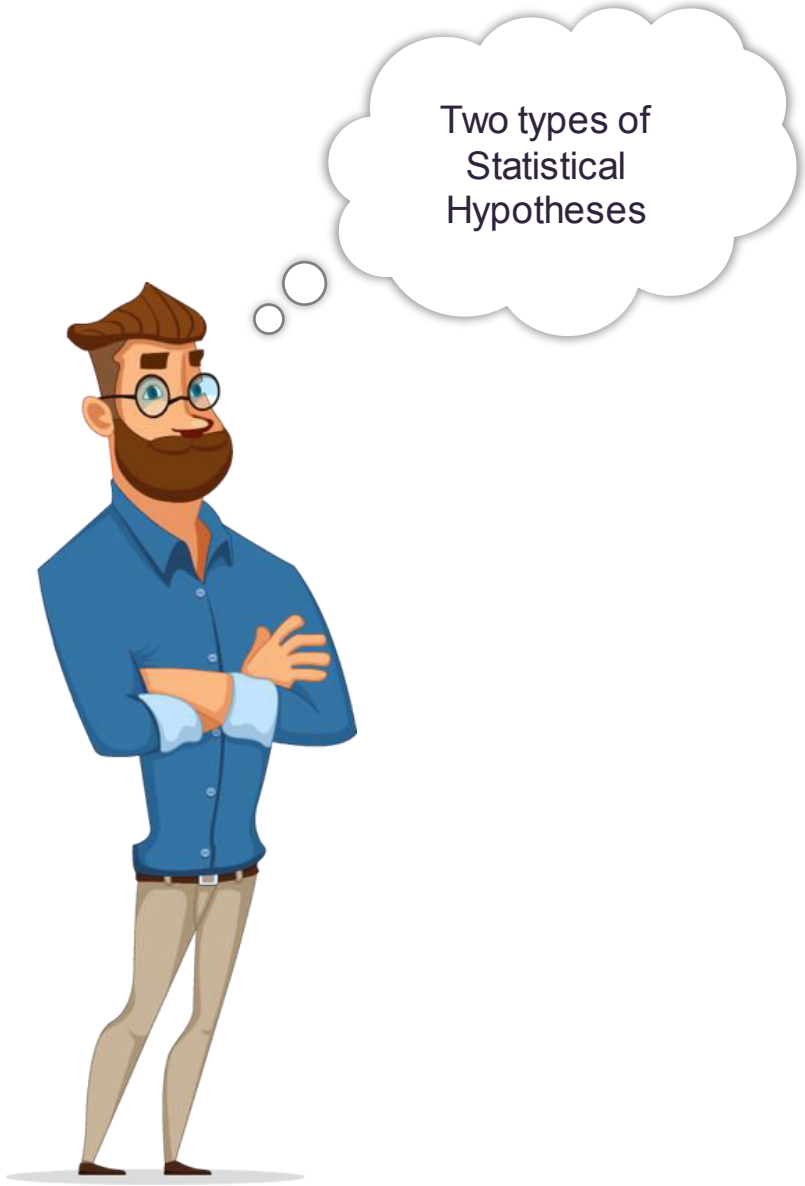
The formal procedures used by statisticians to accept or reject statistical hypotheses

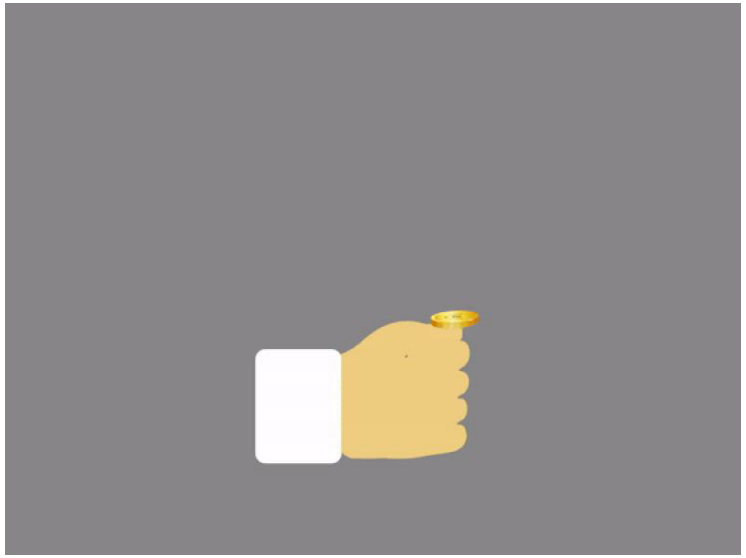
Hypothesis Testing





Hypothesis Testing



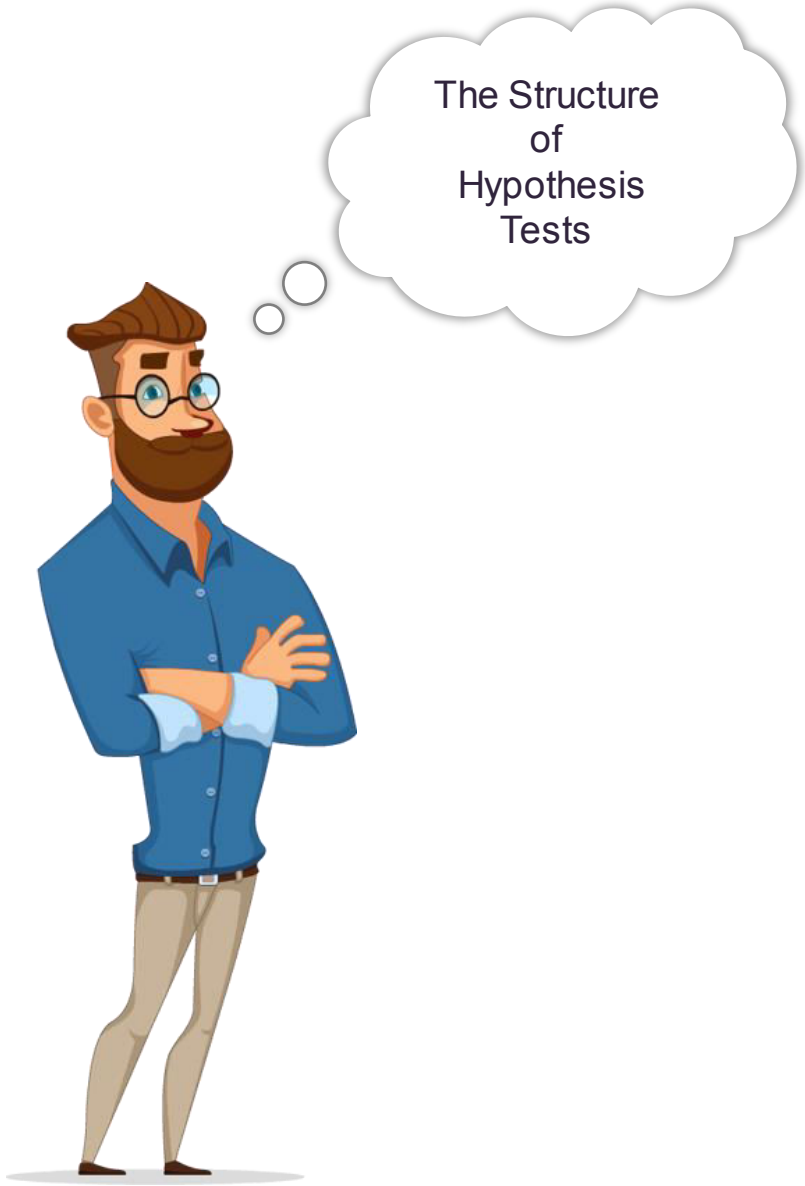


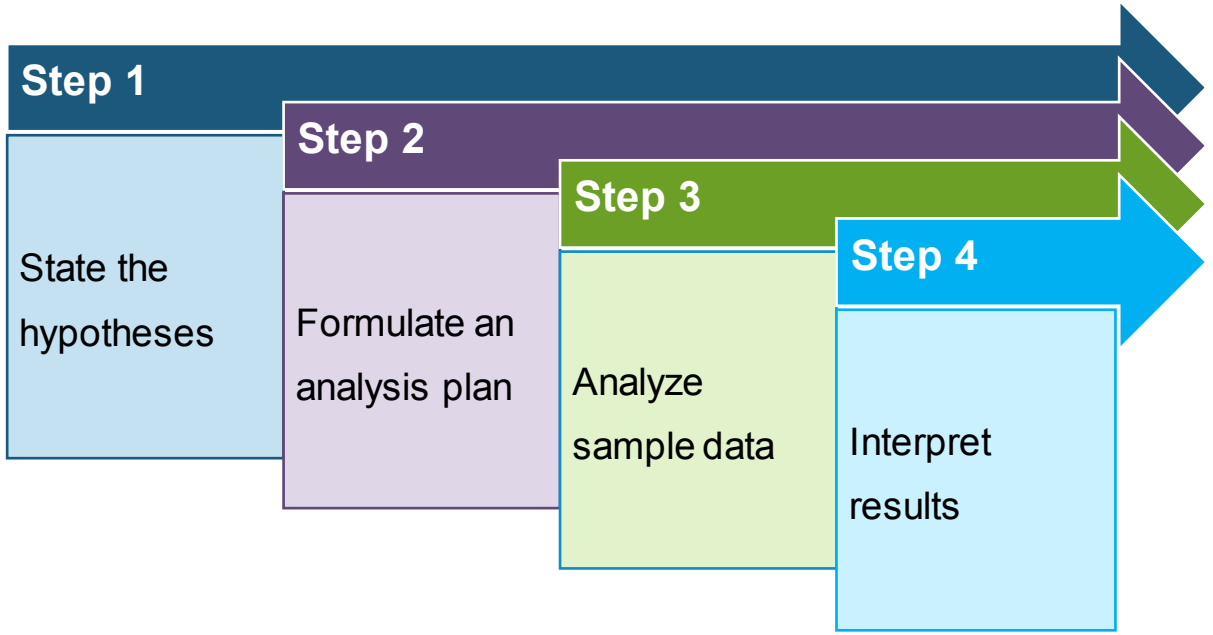
H_o: P = 0.5
H_1: P ≠ 0.5

Coin Flip = 50 times → 40 Heads and 10 Tails

Reject the Hypothesis

Hypothesis Testing



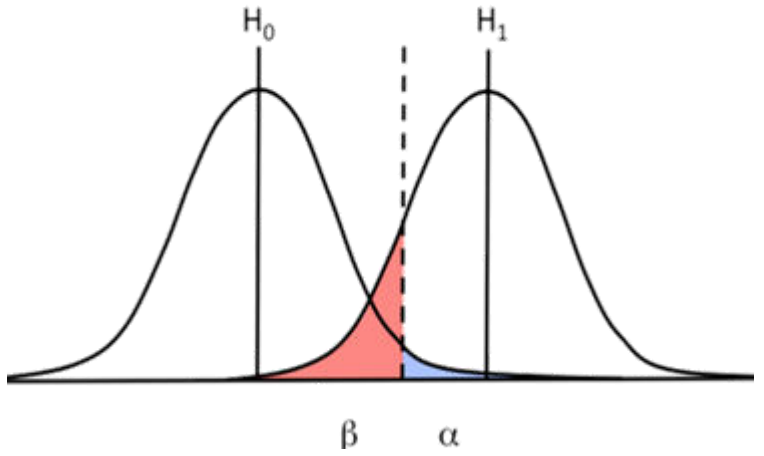


Decision Error

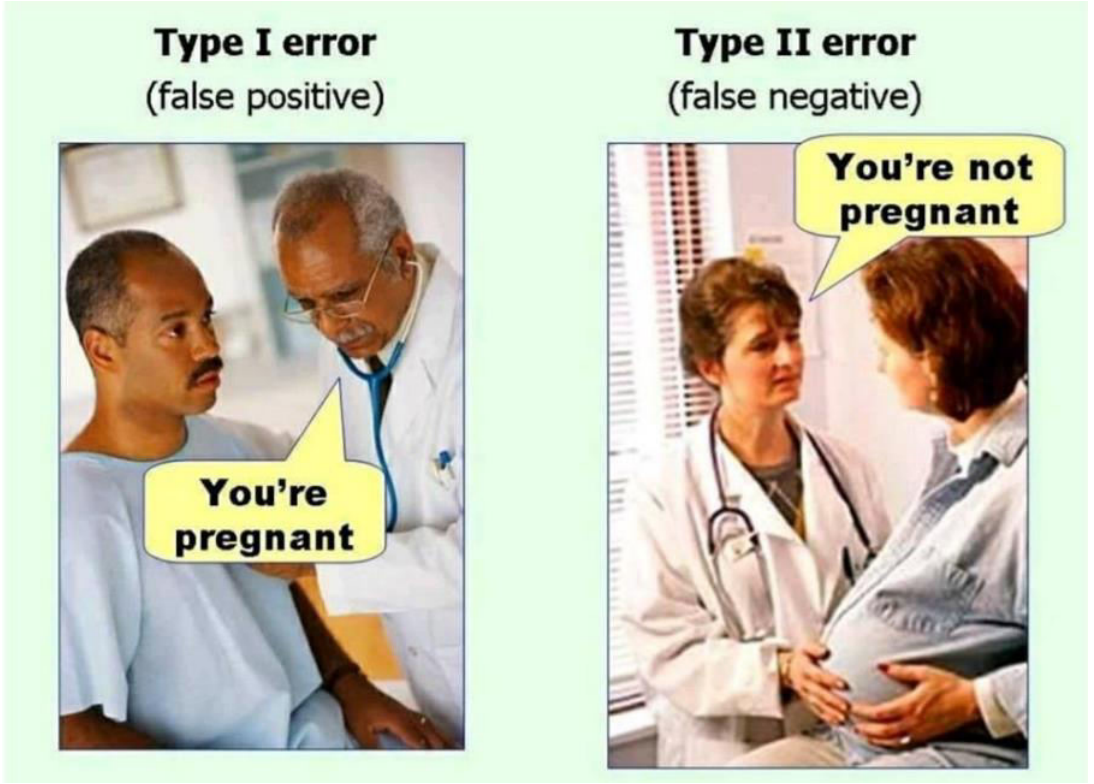
Decision Error

Type II Error

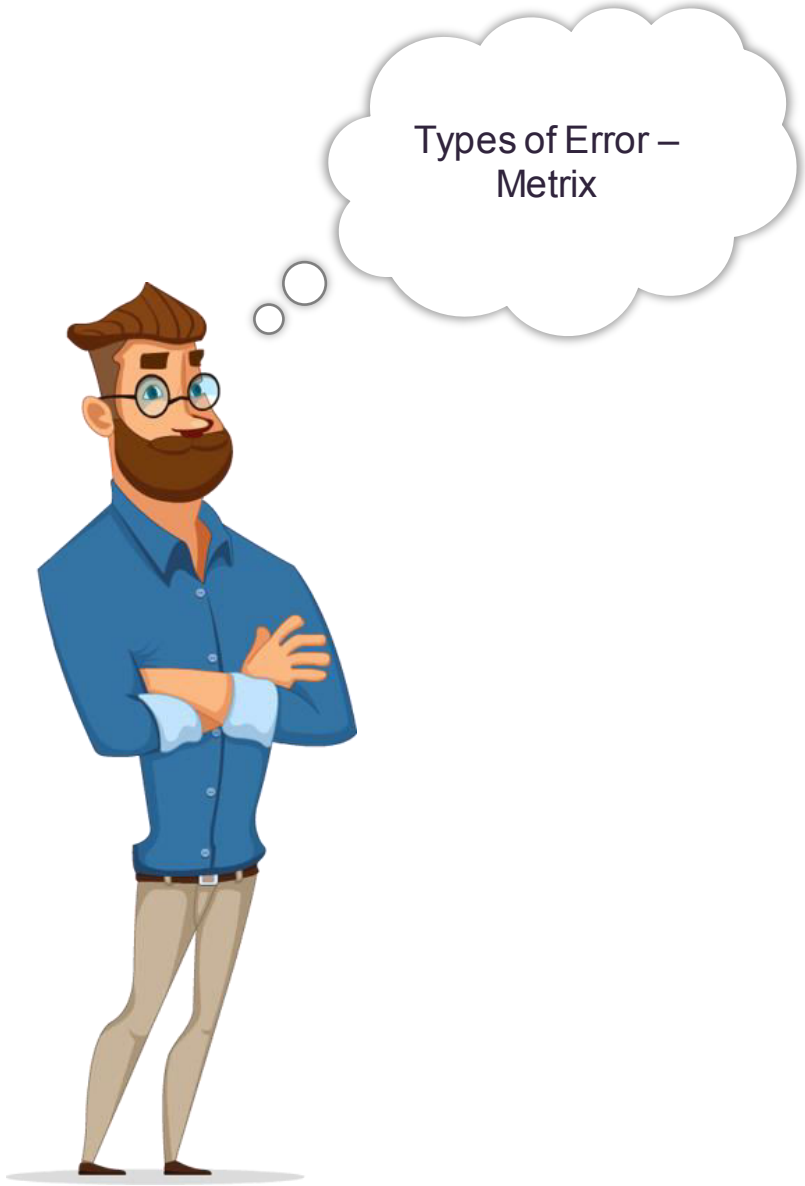
Type I Error



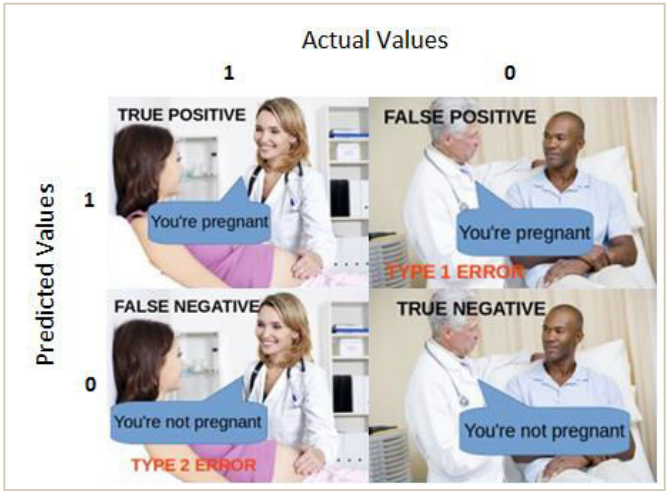
Decision Error



Decision Error

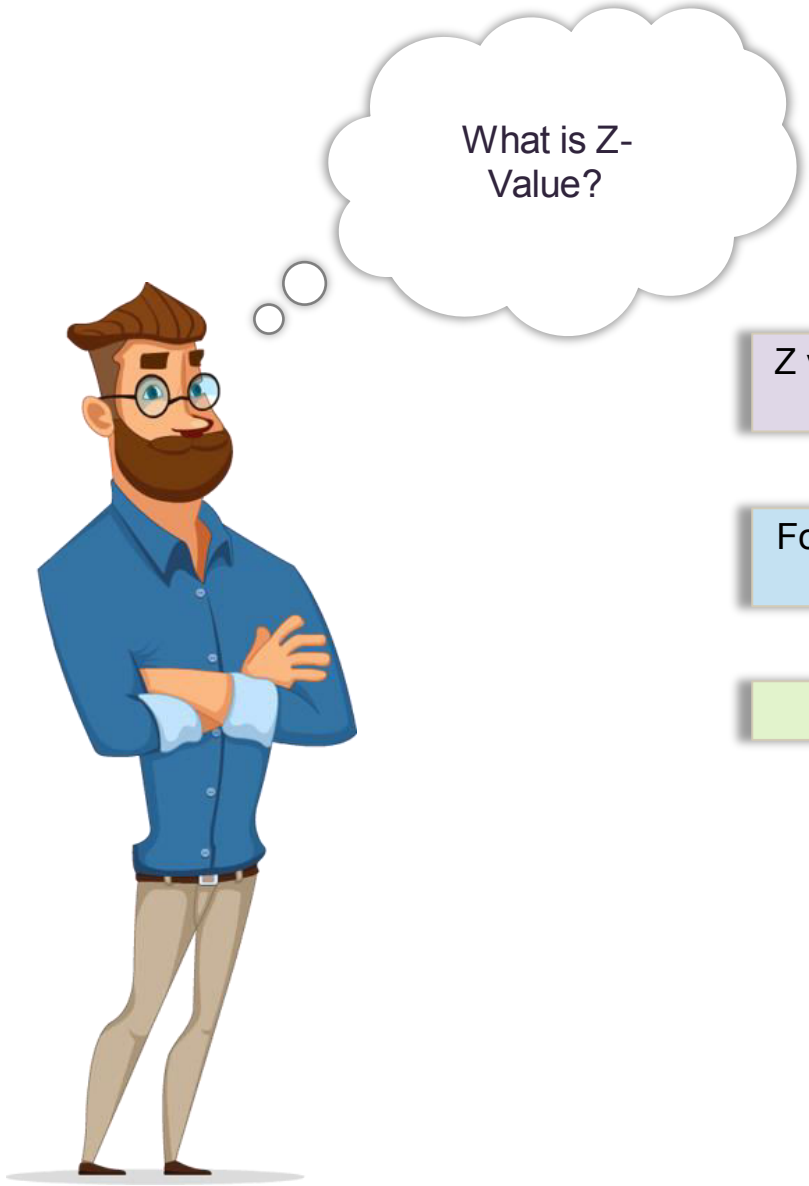


		Condition	
		Condition Positive	Condition Negative
Test Outcome	Test Outcome Positive	**True Positive**	**False Positive** (Type I error)
	Test Outcome Negative	**False Negative** (Type II error)	**True Negative**



Decision Rules

Decision Rules



Z value is a measure of standard deviation i.e. how many standard deviation away from mean is the observed value.

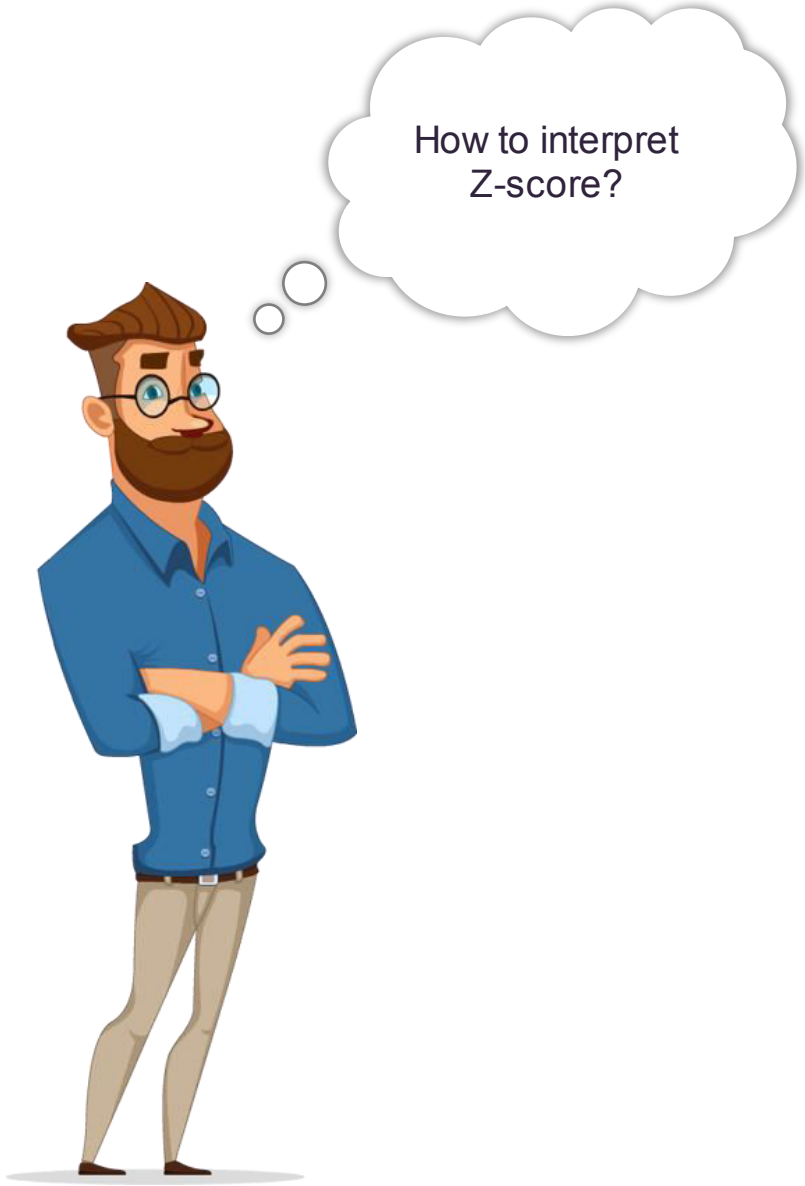
For example, the value of z value = +1.8 can be interpreted as the observed value is +1.8 standard deviations away from the mean.

P-values are probabilities

$$z = \frac{(X - \mu)}{\sigma}$$

$$z = \frac{(190 - 150)}{25} = 1.6$$

Decision Rules



$z-score$<0 … an element less than the mean.

$z-score$>0 … an element greater than the mean.

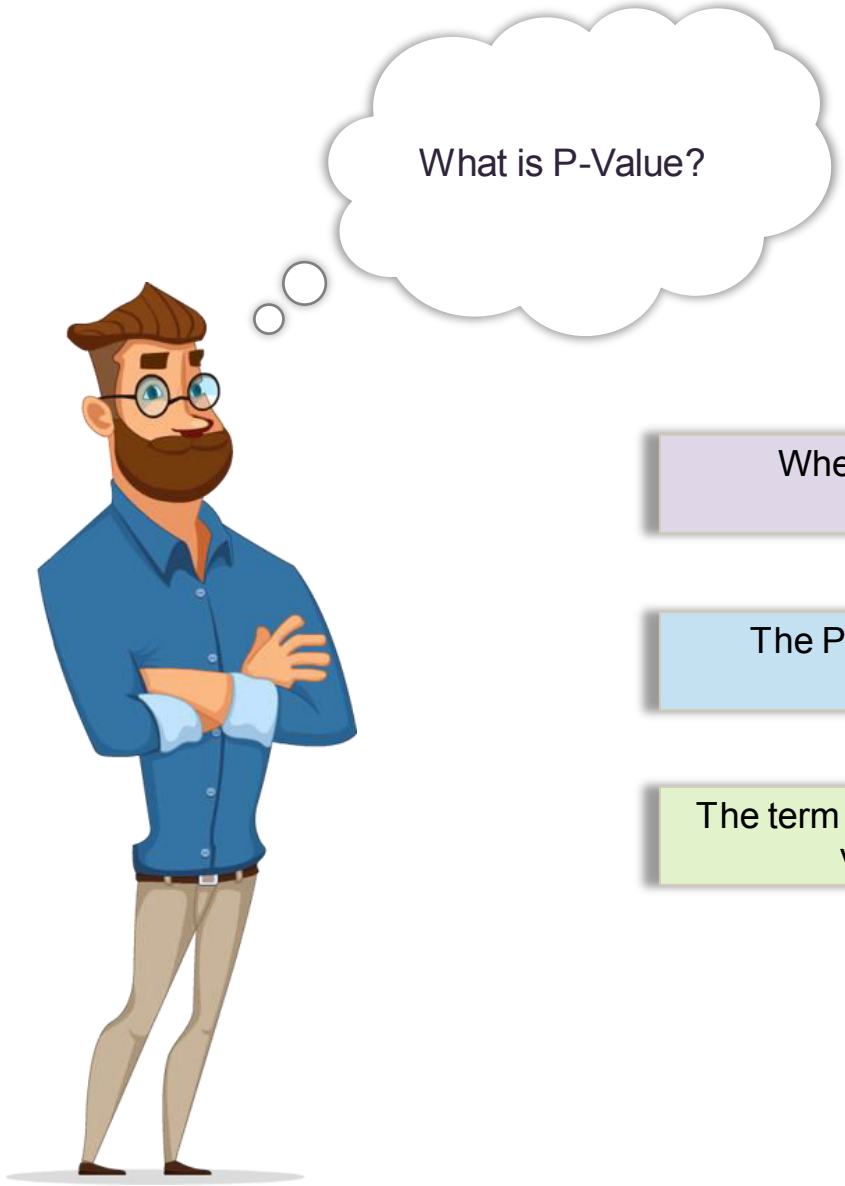
$z-score$=0 … an element equal to the mean.

$z-score$=1 … an element that is 1 standard deviation greater than the mean; $z-score$ =2 , 2 standard deviations greater than the mean; etc.

$z-score$=−1 … an element that is 1 standard deviation less than the mean; $z-score$ =−2 , 2 standard deviations less than the mean; etc.

If the number of elements in the set is large, about 68% of the elements have a z-score between -1 and 1; about 95% have a z-score between -2 and 2; and about 99% have a z-score between -3 and 3.$z-score$ >3 … an element is outlier

Decision Rules

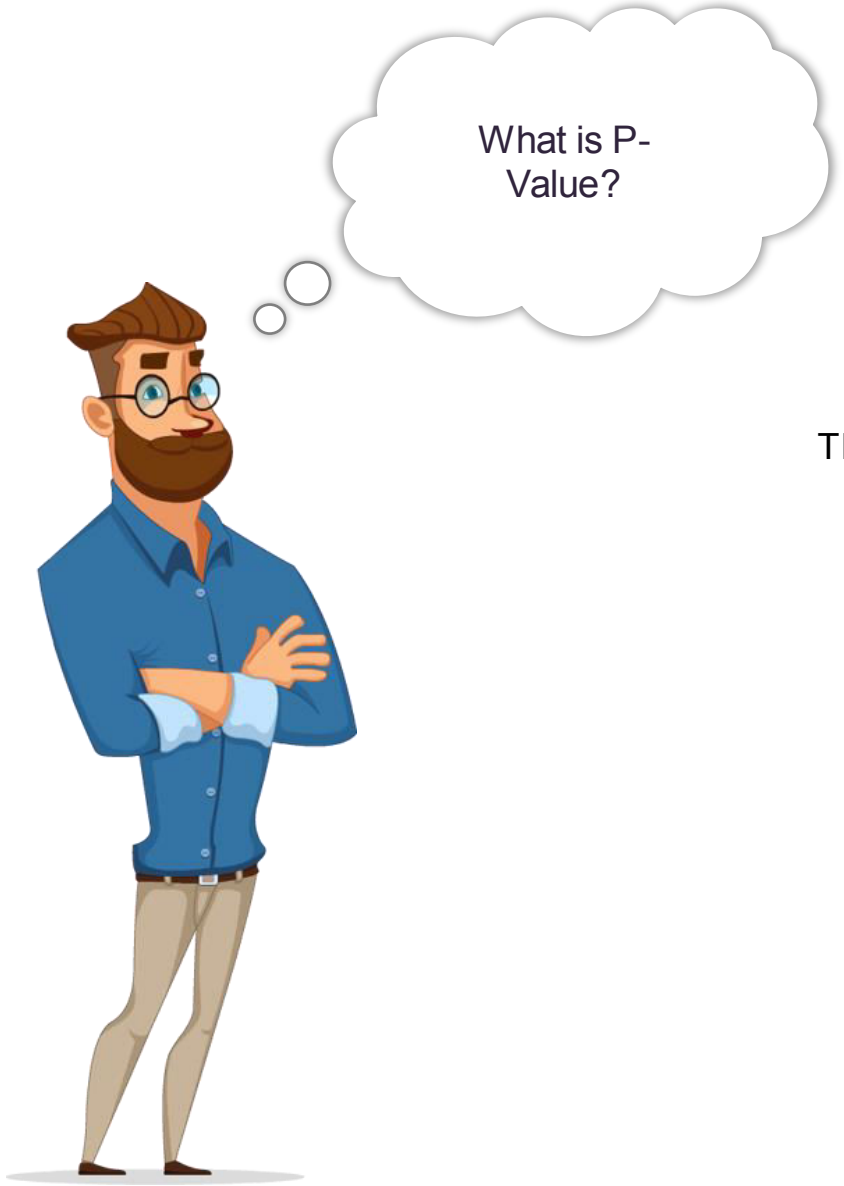


When you perform a hypothesis test in statistics, a p-value helps you determine the significance of your results.

The P value, or calculated probability, is the probability of finding the observed, or more extreme, results when the null hypothesis (H0) of a study question is true.

The term significance level (alpha) is used to refer to a pre-chosen probability and the term "P value" is used to indicate a probability that you calculate after a given study.

Decision Rules



The p-value is a number between 0 and 1 and interpreted in the following way:

- A small p-value (typically ≤ 0.05) indicates strong evidence against the null hypothesis, so you reject the null hypothesis.
- A large p-value (> 0.05) indicates weak evidence against the null hypothesis, so you fail to reject the null hypothesis.
- p-values very close to the cutoff (0.05) are considered to be marginal (could go either way). Always report the p-value so your readers can draw their own conclusions.

Decision Rules

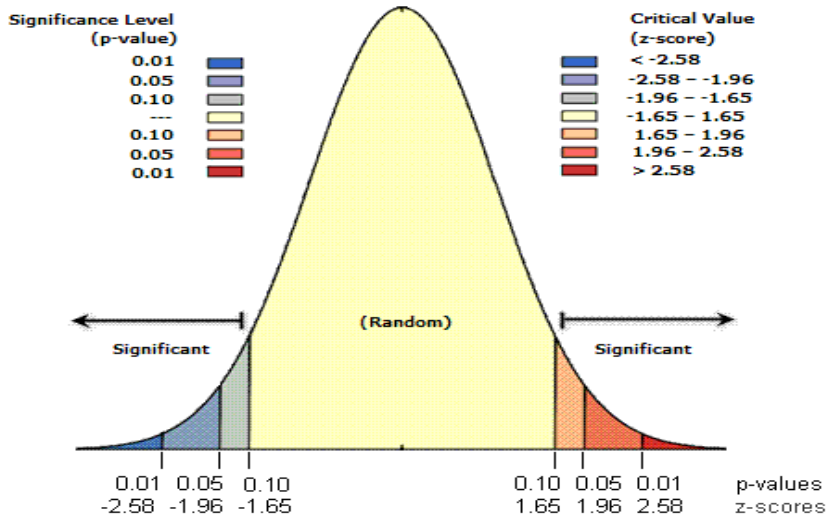
Pizza Delivery

30 MINUTES OR LESS ON AVERAGE

Mean Delivery Time is 30 minutes max, is incorrect

p-value turns out to be 0.001, which is much less than 0.05

Decision Rules



z-score (Standard Deviations)	p-value (Probability)	Confidence level
< -1.65 or > +1.65	< 0.10	90%
< -1.96 or > +1.96	< 0.05	95%
< -2.58 or > +2.58	< 0.01	99%

When the absolute value of the z-score is large and the probabilities are small

Quiz

Quiz

Which of the following statement describes descriptive statistics?

a. Descriptive statistics involves organizing, displaying, and describing data

b. Descriptive statistics use a random sample of data taken from a population to describe and make inferences about the population

c. All of the above

d. None of the above

Quiz

Which of the following statement describes discrete data?

a. Discrete data are continuous

b. Discrete data are whole numbers, and are usually a count of objects

c. Discrete data have labels

d. All of the above

Quiz

Which of the following are four measurement scales?

a. arrange , mutate, summary , ordinal

b. mutate, nominal , interval , arrange

c. nominal , ordinal , interval , ratio

d. All of the above

Quiz

Which of the following are the properties of normal distribution?

a. mean = median = mode

b. Symmetry about the center

c. The total area under the curve is 1

d. All of the above

Quiz

Which of the following is the formula for Z Score?

a. $z = x - \mu / \sigma$

b. $z = x - \mu / s$

c. $z = x / \sigma$

d. None of the above

Thank You