Statistics

It helps us to understand the data better.

What is Statistics ?

Statistics is the science concerned with developing and studying methods for collecting, analysing, interpreting and presenting data.

Statistical Measures:

1. Range
2. Mean
3. Standard Deviation

Correlation

statistics helps us to find the relationships in the data better. such as work experience and salary is in positive correlation.

syllabus:

Topics covered in this module:

1. Basics of Statistics
2. Types of Statistics
3. Population & Sample
4. Central Tendencies
5. Percentiles & Dispersion
6. Statistics Implementation with Python -1
7. Range, Sample Variance & Standard Deviation

8. Correlation & Causation
9. Hypothesis Testing
10. Statistics Implementation with Python -2

1. BASICS OF STATISTICS

Why we need Statistics ?

Statistics is a tool that helps us to extract information & Knowledge from data

Sidd

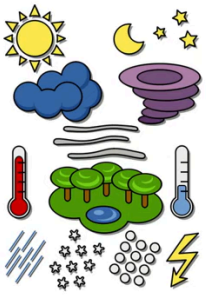
Who is the best Batsman in the world in the time period 2010 to 2020?

Application of statistics - weather forecast, businesses, SIX SIGMA, clinical trial of medicines.

Few Applications of Statistics

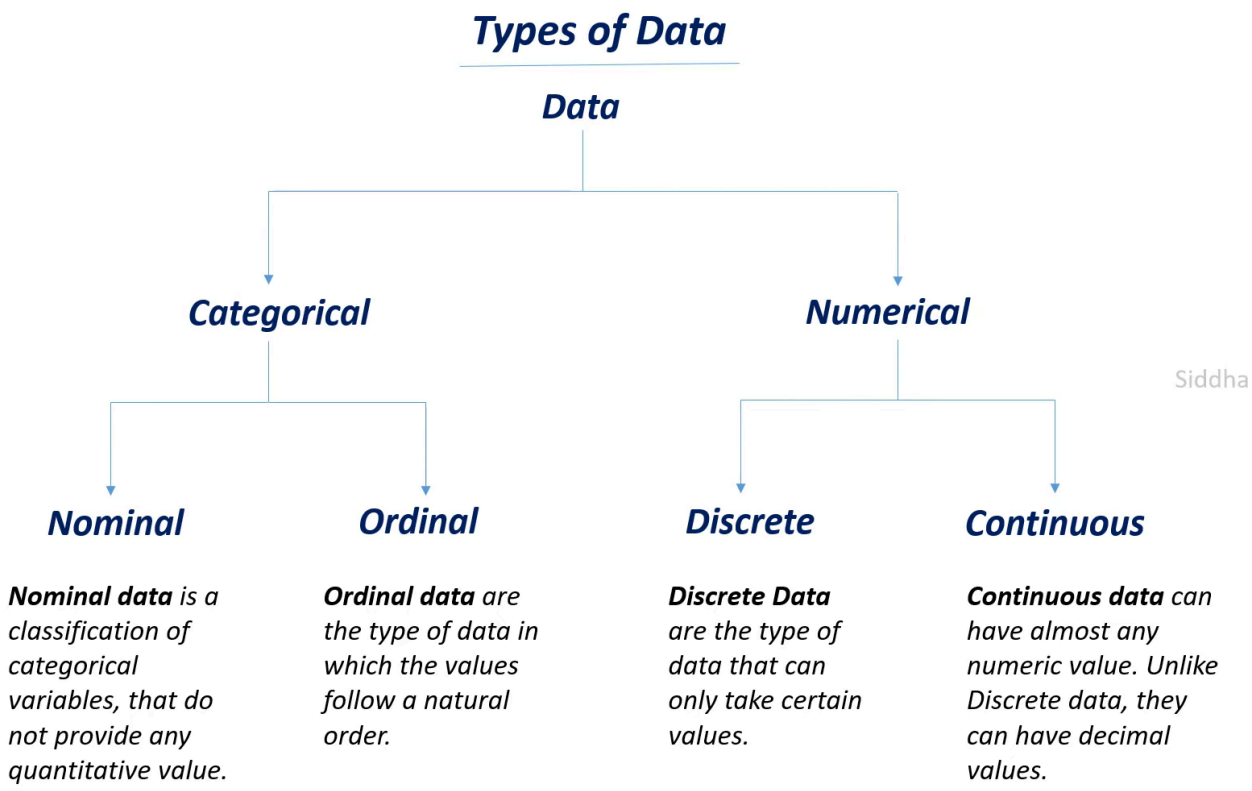
Si

Business



Weather Forecast

Clinical Trial of Medicines



Nominal data = the two genders. there is no significant info in which we can order the data to show which is more important.

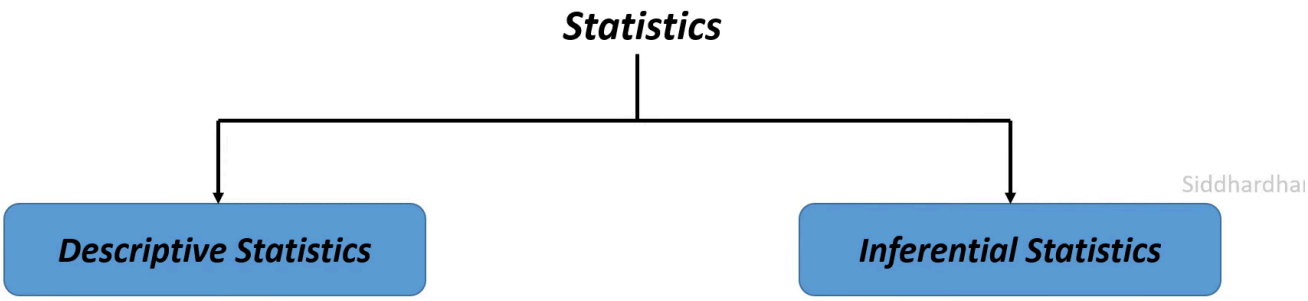
ordinal = review of smartphone as good, bad , very bad. thus we can order this to show more significance.

Discrete = number of students in a class. we cannot give values like 60.6, 40.66.

continuous = we can give decimal values like 30.5 grams of sugar.

TYPES OF STATISTICS

Types of Statistics



- Describe the data
- Data Analysis

- Get inferences about the data
- Data Science

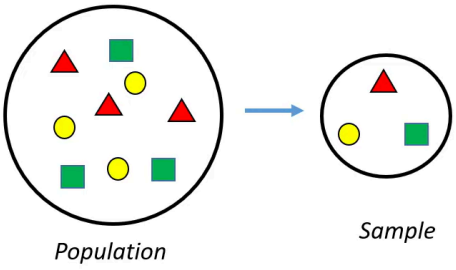
Types of Statistics

1. Descriptive Statistics:

Descriptive statistics *are used to describe the basic features of the data in a study. They provide simple summaries about the sample and the measures.*

2. Inferential Statistics:

Inferential statistics *takes data from a sample and makes inferences and predictions about the larger population from which the sample was drawn.*



Descriptive Statistics

2 important measures of Descriptive Statistics:

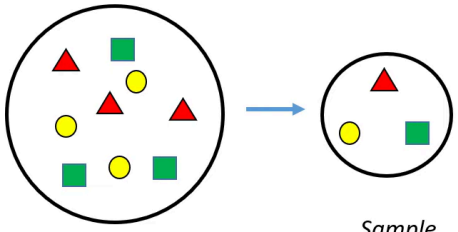
1. Measure of Central Tendencies (Mean, Median, Mode)
2. Measure of Variability (Range, Standard Deviation, Variance)

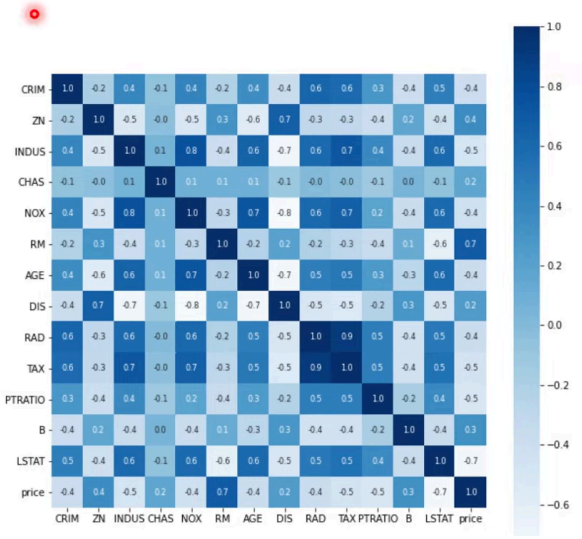
Descriptive Statistics of House Price Dataset

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	price
count	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000
mean	3.613524	11.363636	11.136779	0.069170	0.554695	6.284634	68.574901	3.795043	9.549407	408.237154	18.455534	356.674032	12.653063	22.532806
std	8.601545	23.322453	6.860353	0.253994	0.115878	0.702617	28.148861	2.105710	8.707259	168.537116	2.164946	91.294864	7.141062	9.197104
min	0.006320	0.000000	0.460000	0.000000	0.385000	3.561000	2.900000	1.129600	1.000000	187.000000	12.600000	0.320000	1.730000	5.000000
25%	0.082045	0.000000	5.190000	0.000000	0.449000	5.885500	45.025000	2.100175	4.000000	279.000000	17.400000	375.377500	6.950000	17.025000
50%	0.256510	0.000000	9.690000	0.000000	0.538000	6.208500	77.500000	3.207450	5.000000	330.000000	19.050000	391.440000	11.360000	21.200000
75%	3.677083	12.500000	18.100000	0.000000	0.624000	6.623500	94.075000	5.188425	24.000000	666.000000	20.200000	396.225000	16.955000	25.000000
max	88.976200	100.000000	27.740000	1.000000	0.871000	8.780000	100.000000	12.126500	24.000000	711.000000	22.000000	396.900000	37.970000	50.000000

Inferential Statistics

Inferential statistics *takes data from a sample and makes inferences and predictions about the larger population from which the sample was drawn.*

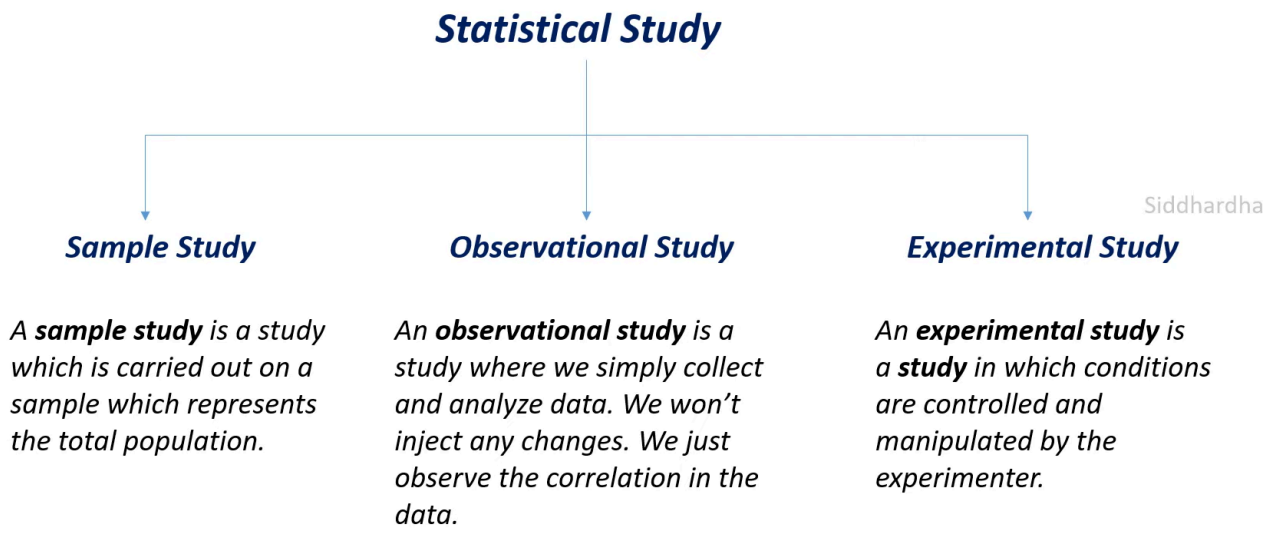




Correlation of House Price Data

TYPES OF STATISTICAL STUDIES

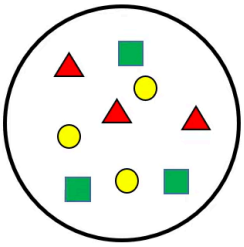
Types of Statistical Studies



1. sample study - finding the average blood sugar level of a Nation by taking a small population.

2. Observational Study

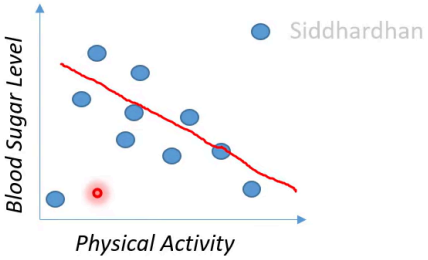
*An **observational study** is a study where we simply collect and analyze data. We won't inject any changes. We just observe the correlation in the data.*



Population

Relation between:

1. Blood Sugar Level
2. Physical Activity



Inference: Blood Sugar Level & Physical Activity are Negatively Correlated

3. Experimental Study

*An **experimental study** is a **study** in which conditions are controlled and manipulated by the experimenter.*

People following healthy food habits & doing regular exercises

Group 1

Group 2

People with unhealthy food habits & doing very minimal exercises

Control Groups

Conclusive Result

(People in Group 1 have optimum blood sugar level)

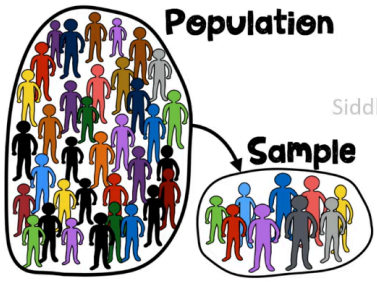
Experimental study is more accurate.

POPULATION AND SAMPLE SAMPLING TECHNNIQUES

Types of Sampling Techniques

Sampling Techniques:

- Simple Random Sampling
- Systematic Sampling
- Stratified Random Sampling
- Cluster Sampling



(Probability Sampling Techniques)

(Non-Probability Sampling Techniques)

Simple Random Sampling

In **Simple Random Sampling**, the sample is randomly picked from a larger population. Hence, all the individual datapoints has an equal probability to be selected as sample data.

Example: Employee survey in a company

Pros:

1. No sample Bias
2. Balanced Sample
3. Simple Method of sampling
4. Requires less domain knowledge

Cons:

1. Population size should be high
2. Cannot represent the population well sometimes

Systematic Sampling

In **Systematic Sampling**, the sample is picked from the population at regular intervals. This type of sampling is carried out if the population is homogeneous and the data points are uniformly distributed

Example: Selecting every 10^{th} member from a population of 10,000

Pros:

1. Quick & easy
2. Less bias
3. Even distribution of data

Cons:

1. Data manipulation risk
2. Requires randomness in data
3. Population should not have patterns.

Stratified Random Sampling

In **Stratified Random Sampling**, the population is subdivided into smaller groups called **Strata.** Samples are obtained randomly from all these strata.

Example: Smartphone sales in all the states

Pros:

1. Finds important characteristics in the population
2. High precision can be obtained if the differences in the strata is high

Cons:

1. Cannot be performed on populations that cannot be classified into groups.
2. Overlapping data points

Cluster Sampling

Cluster Sampling is carried out on population that has inherent groups. This population is subdivided into **clusters** and then random clusters are taken as sample.

Example: Smartphone sales in randomly selected states

Pros:

1. Requires only fewer resources
2. Reduced Variability
3. Advantages of both Random sampling and Stratified Sampling

Cons:

1. Cannot be performed on populations without natural groups
2. Overlapping data points
3. Can't provide a general insight for the entire population

In stratified we take all the strata but in cluster we take selected strata. this is a simple difference.

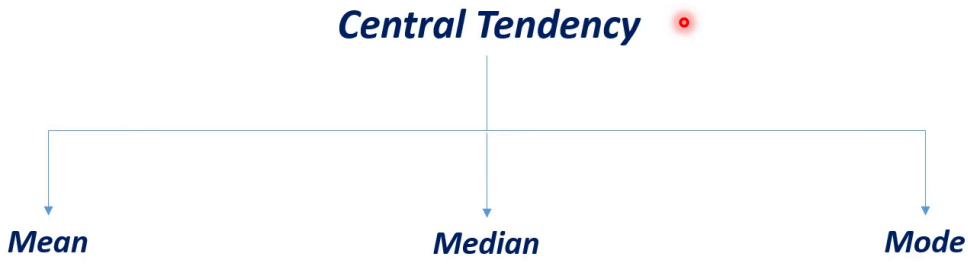
CENTRAL TENDENCIES - Mean, Median and Mode

Important topic for ML

Central Tendency

Central Tendency:

*A measure of **central tendency** is a value that represents the center point or typical value of a dataset. It is a value that summarizes the data.*



Central Tendencies

Mean

Mean *or arithmetic mean is the sum of values divided by the number of values.*

$$M = \frac{\Sigma x}{N}$$

Heights

160
172
165
168
174

$$\frac{160+172+165+168+174}{5}$$

Mean = 167.8

Median

*The **median** is the **middle** value in the list of numbers. To find the median, the numbers have to be listed in numerical order from smallest to largest.*

160 165 168 172 174

160 165 168 172 174 176

$$\frac{168+172}{2} = 170$$

Median = 170

Mode

*The **mode** is the value that occurs most often. If no number in the list is repeated, then there is no mode for the list.*

Heights

160
172
160
168
174

Mode = 160

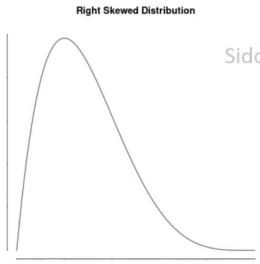
Central Tendencies in Data Pre-Processing

Central Tendencies are very useful in **handling the missing values** in a dataset

Mean : Missing values in a dataset can be replaced with **mean** value, if the data is uniformly distributed.

Median : Missing values in a dataset can be replaced with **median** value, if the data is skewed.

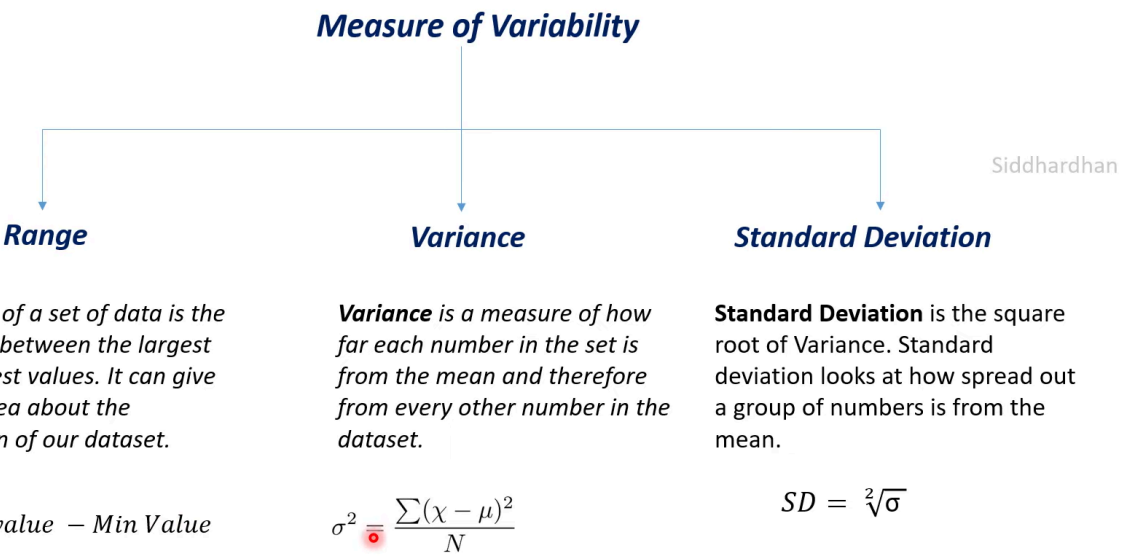
Mode : Missing values in a dataset can be replaced with **mode** value, if the data is skewed. Missing categorical values can also be replaced with **mode** value.



we can't find mean, median for categorical values such as color of a ball.

MEASURE OF VARIABILITY - Range, Variance and Standard Deviation

Measure of Variability



if variance is small then the numbers are closer to the mean. Therefore, more the variance then more the distance of the numbers from the mean.

Range ; Variance ; Standard Deviation

-5, 0, 5, 10, 15,

$Mean = \frac{-5+0+5+10+15}{5} = 5$

Range = 15 – (-5) = 20

$Variance = \frac{(-5-5)^2+(0-5)^2+(5-5)^2+(10-5)^2+(15-5)^2}{5}$

$Variance = 50$

$Standard\ Deviation = 7.1$

3, 4, 5, 6, 7

$Mean = \frac{3+4+5+6+7}{5} = 5$

Range = 7 – 3 = 4

Siddhardhan

$Variance = \frac{(3-5)^2+(4-5)^2+(5-5)^2+(6-5)^2+(7-5)^2}{5}$

$Variance = 2$

$Standard\ Deviation = 1.4$

if all the numbers are the same then variance and stand. deviation will be zero. Thus, these both terms can be greater than zero or equal to zero but can't be negative.

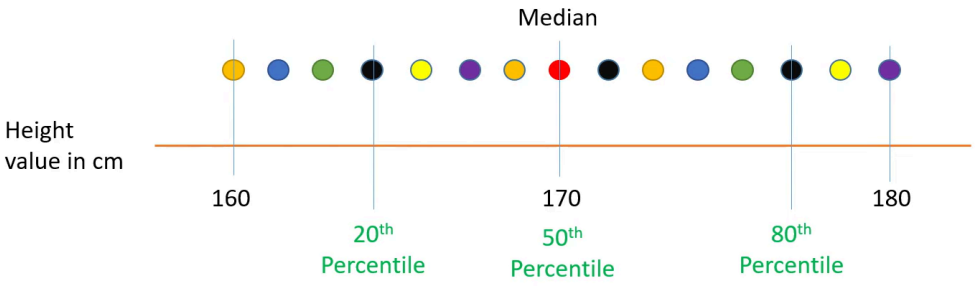
PERCENTILES AND QUANTILES

They help to understand the distribution of data points in a dataset.

Percentiles

Percentile is a value on a scale of 100 that indicates the percent of a distribution that is equal to or below it.

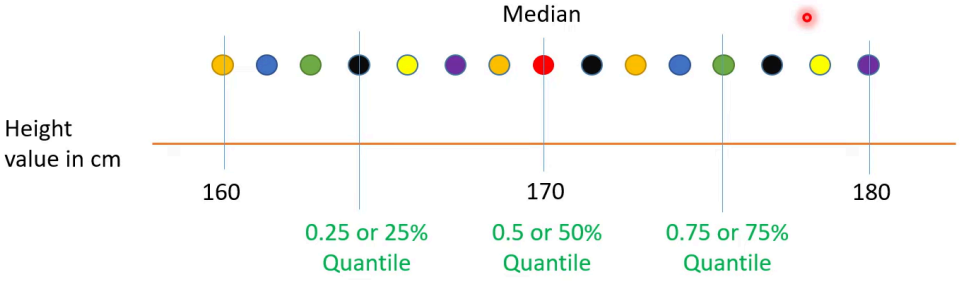
Dataset with Height of 15 people



Quantiles

Quantile is a measure that tells how many values in a dataset are above or below a certain limit. It divides the members of the dataset into equally-sized subgroups.

Dataset with Height of 15 people

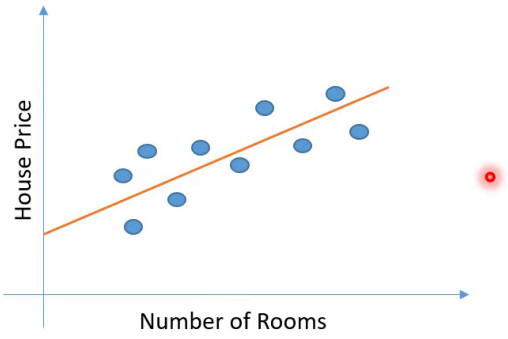


CORRELATION AND CAUSATION

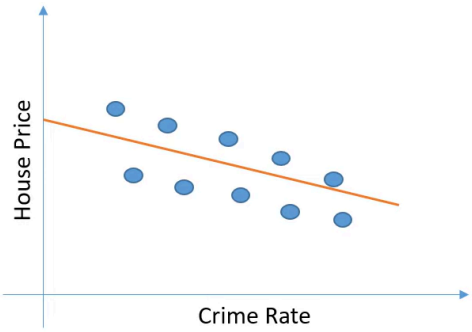
Correlation

Correlation is a measure that determines the extent to which two variables are related to each other in a dataset. But it doesn't mean that one event is the cause of the other event.

Positive Correlation

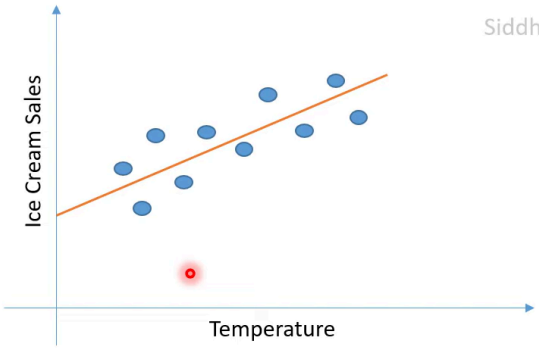


Negative Correlation



Causation

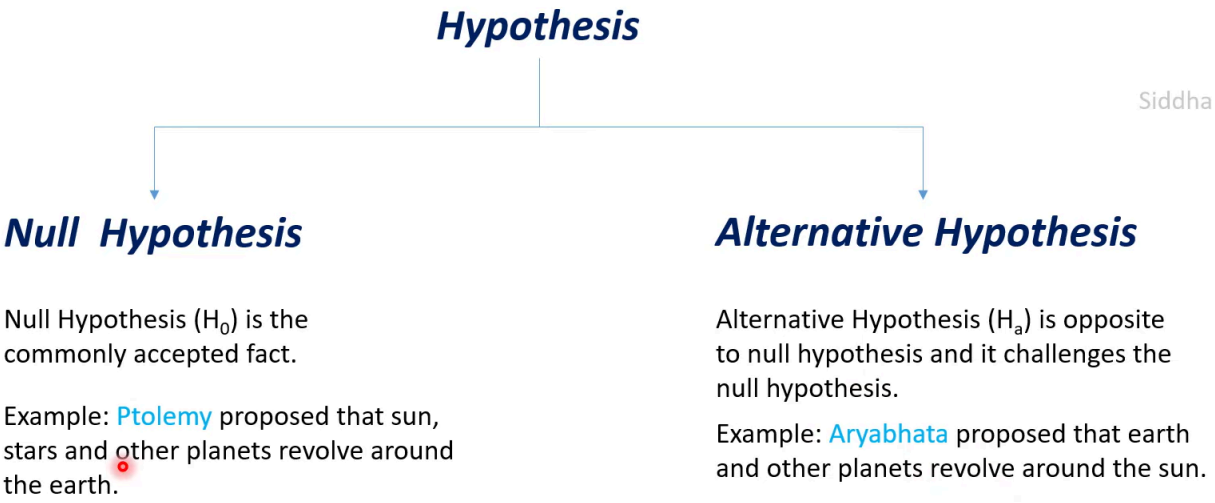
In statistics, Causation means that one event causes another event to occur. Thus, there is a cause and effect relationship between the two variables in a dataset.



Hypothesis testing - Null Hypothesis and Alternate Hypothesis

Hypothesis

Hypothesis is an assumption that is made based on the observations of an experiment.



Hypothesis testing is done to see which of the two is correct.

Example:

we take a drug A and drug B and perform statistical studies on two groups of people to see which drug cures headache faster.

drug a = 10 min
drug b = 15 min

hence, NULL HYPOTHESIS = drug A cures headache faster than drug B

now, some changes are made to the formula of drug B.

ALTERNATIVE HYPO. = drug B is faster than drug A

now , hypothesis testing will be done. so many tests will be performed to see which is correct.

Hypothesis Testing

NULL HYPOTHESIS (H_0): Drug A is more quicker than Drug B.

ALTERNATIVE HYPOTHESIS (H_a): Drug B is more quicker than Drug A.

Sidd

Possible Outcomes of Hypothesis Testing:

- Reject the Null Hypothesis
- Fail to reject the Null Hypothesis

Drug B

Drug A