STATISTICS FOR

Machine Learning

Math for Machine Learning

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

BASICS OF STATISTICS

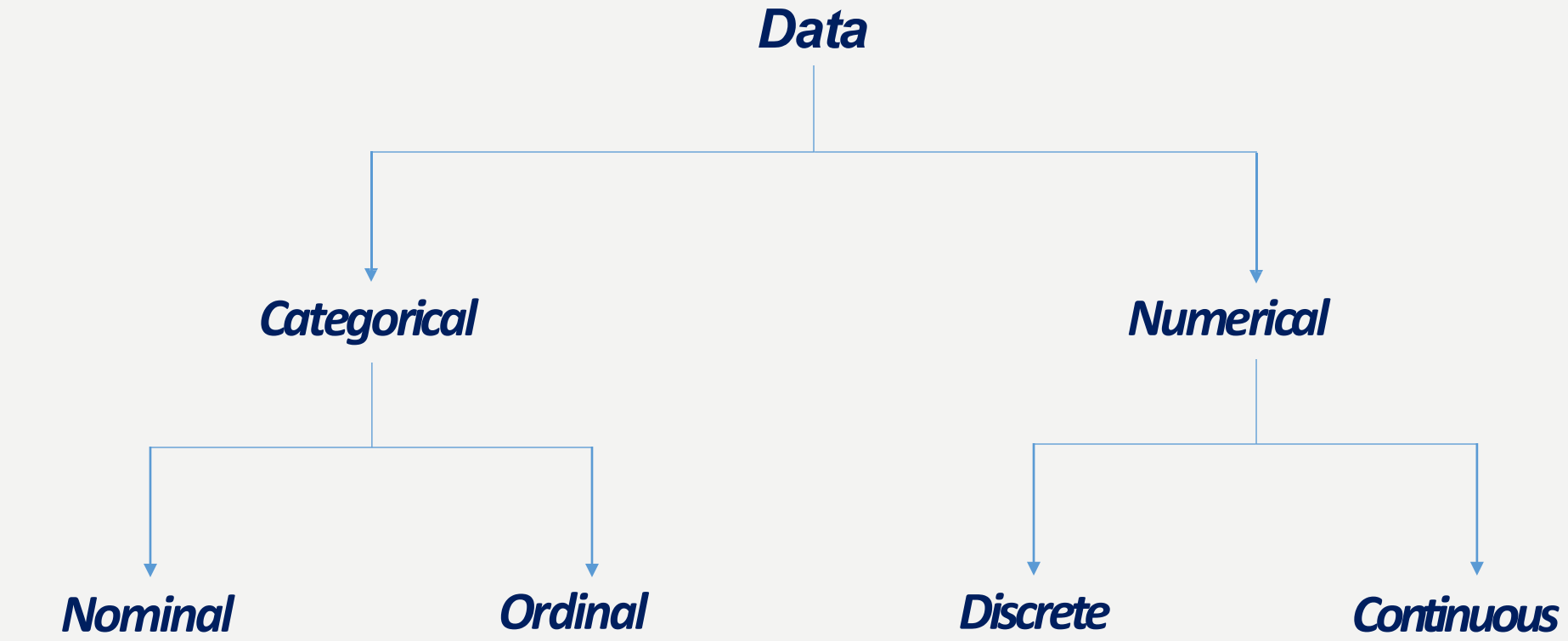
1. Why we need Statistics?
2. Applications of Statistics
3. Types of Data

WHY WE NEED STATISTICS ?

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

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

TYPES OF DATA



Nominal data *is a classification of categorical variables, that do not provide any quantitative value.*

Ordinal data *are the type of data in which the values follow a natural order.*

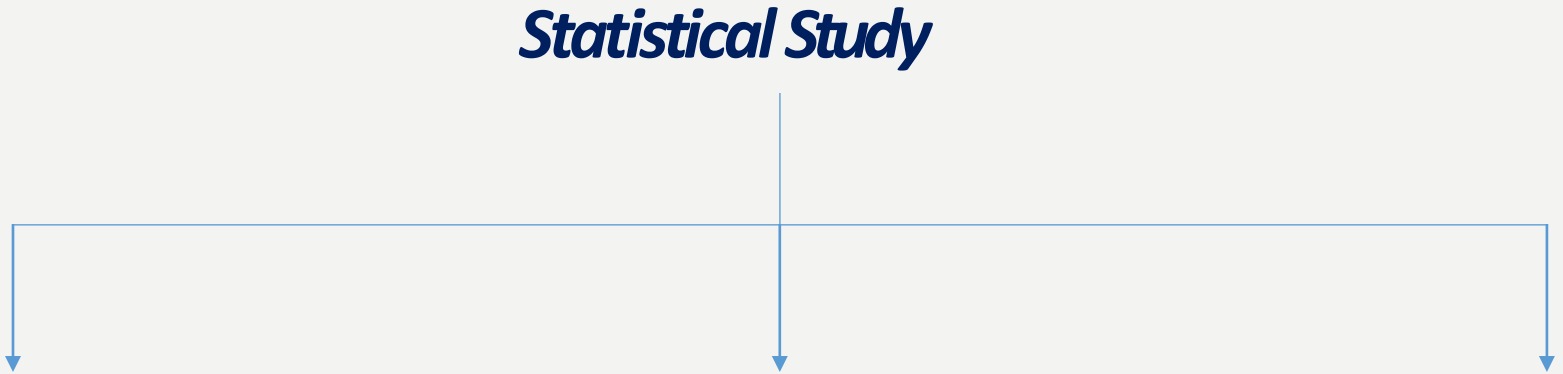
Discrete Data *are the type of data that can only take certain values.*

Continuous data *can have almost any numeric value. Unlike Discrete data, they can have decimal values.*

TYPES OF STATISTICAL STUDIES

Math for Machine Learning

TYPES OF STATISTICAL STUDIES



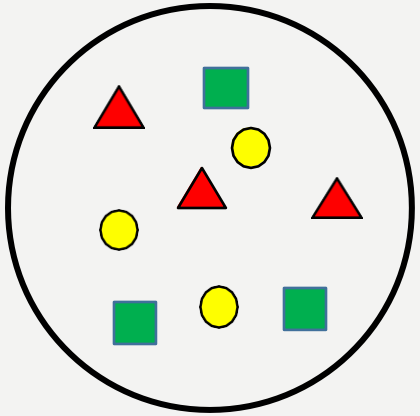
*A **sample study** is a study which is carried out on a sample which represents the total population.*

*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.*

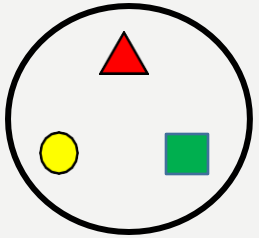
*An **experimental study** is a **study** in which conditions are controlled and manipulated by the experimenter.*

1. SAMPLE STUDY

*A **sample study** is a study which is carried out on a sample which represents the total population.*



Population

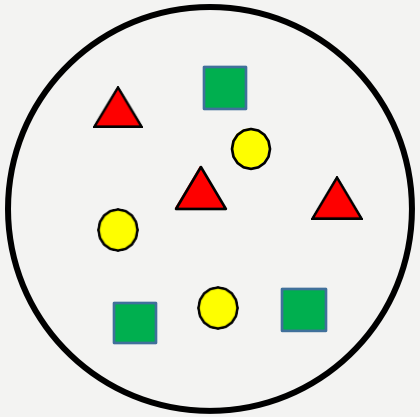


Sample

Average Blood Sugar Level = ?

2. OBSERVATIONAL STUDY

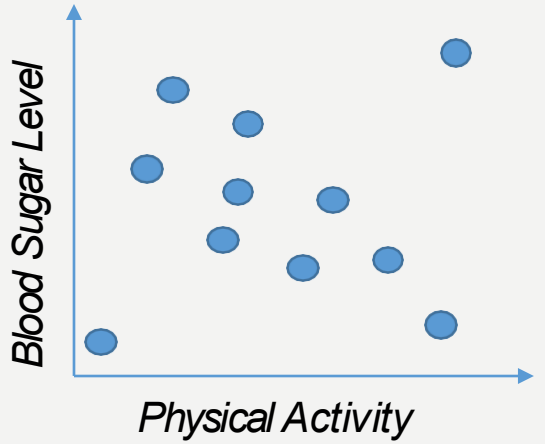
*An **observational study** is a study where we simply collect and analyze data. We won't inject any changes. We just observe the correlation in the data.*



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

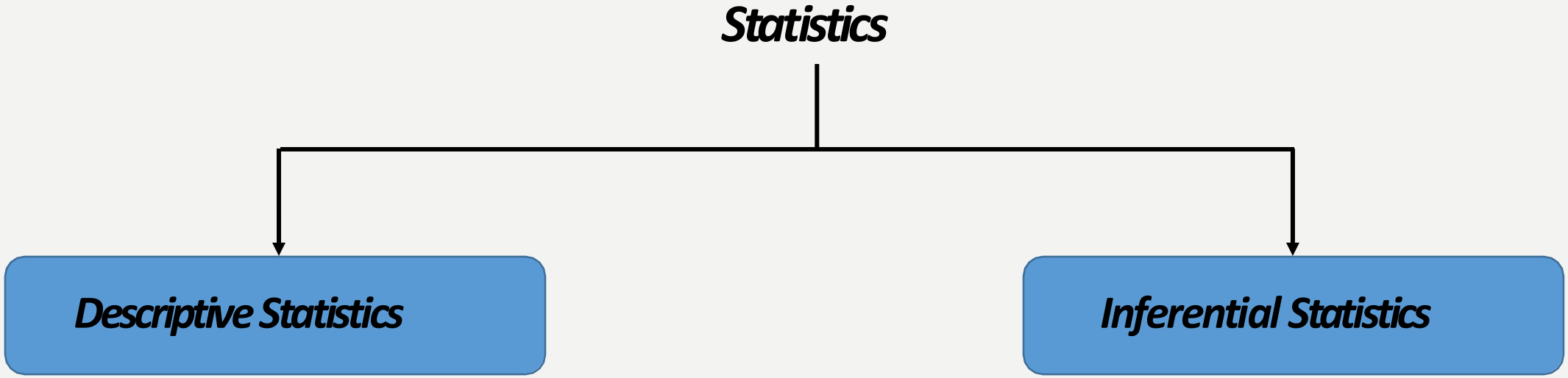
People with unhealthy food habits & doing very minimal exercises

Control Groups

Conclusive Result

(People in Group 1 have optimum blood sugar level)

TYPES OF STATISTICS



- Describe the data
- Data Analysis

- Get inferences about the data
- Data Science

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

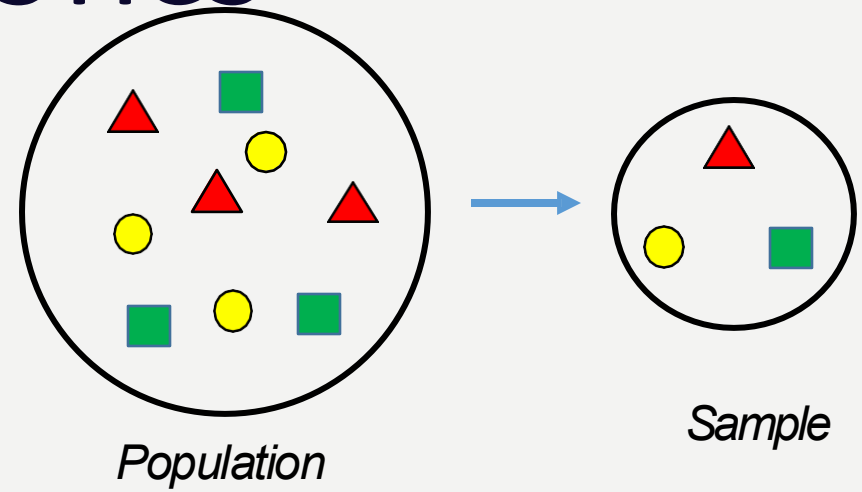
BOSTON HOUSE PRICE DATASET

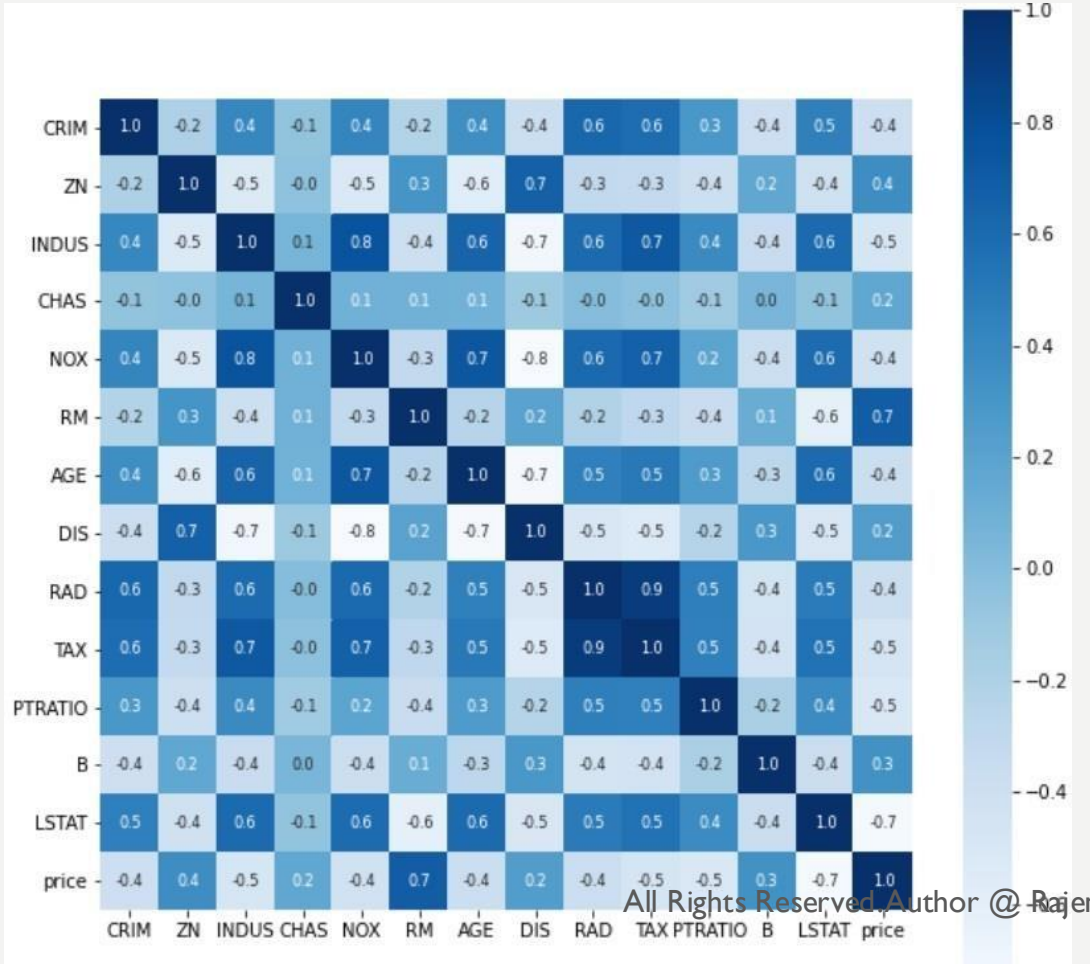
The dataset used in this project comes from the UCI Machine Learning Repository. This data was collected in 1978 and each of the 506 entries represents aggregate information about 14 features of homes from various suburbs located in Boston.

crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	b	lstat	price
0.00632	18	2.31	0	0.538	6.575	65.2	4.09	1	296	15.3	396.9	4.98	24
0.02731	0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.9	9.14	21.6
0.02729	0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
0.03237	0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4

INFERENTIAL STATISTICS

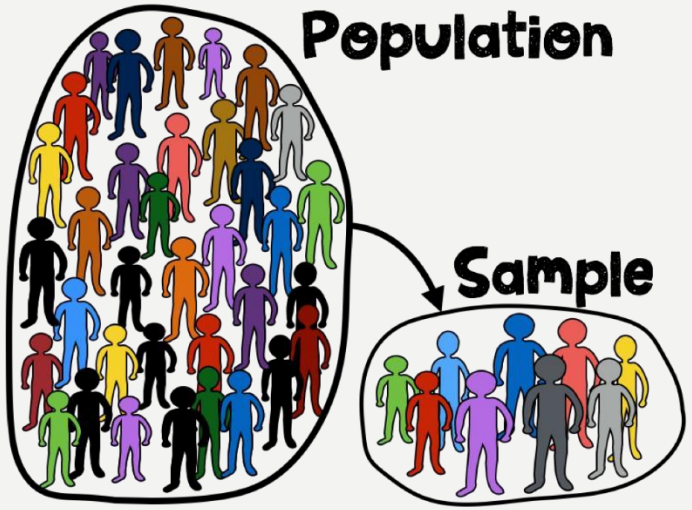
Inferential statistics *takes data from a sample and makes inferences and predictions about the larger population from which the sample was drawn.*





Correlation of House Price Data

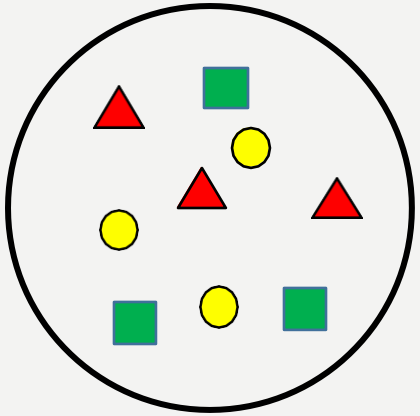
- Population & Sample
- Sampling Techniques



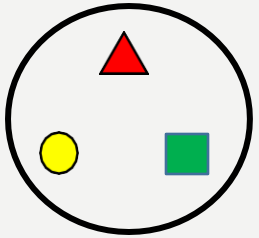
Math for Machine Learning

1. SAMPLE STUDY

*A **sample study** is a study which is carried out on a sample which represents the total population.*



Population



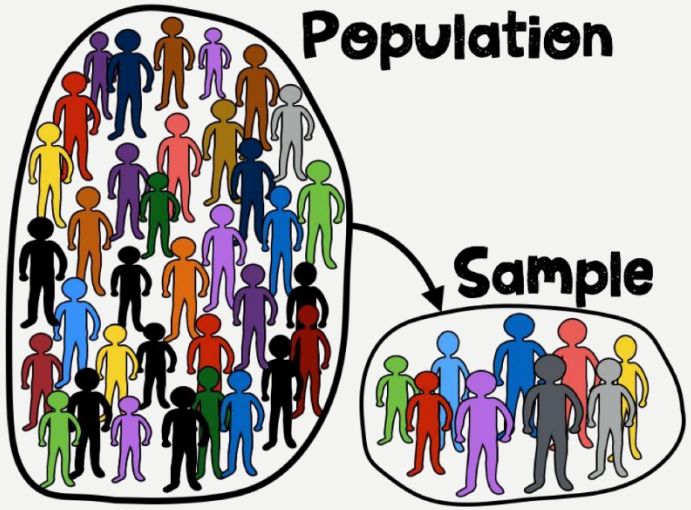
Sample

Average Blood Sugar Level = ?

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

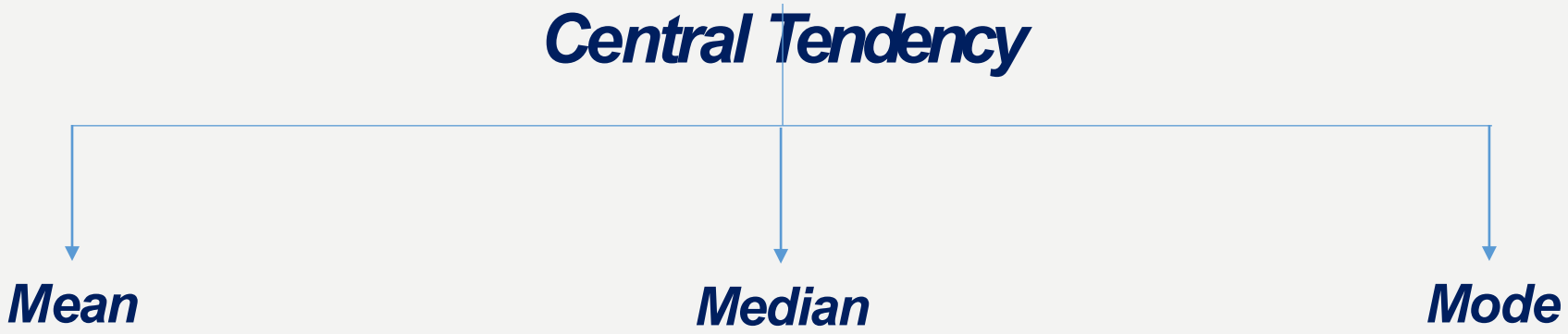
- Measure of
- Central Tendencies Mean, Median & Mode

Math for Machine Learning

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.*

Heights

160
172
165
168
174

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

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

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

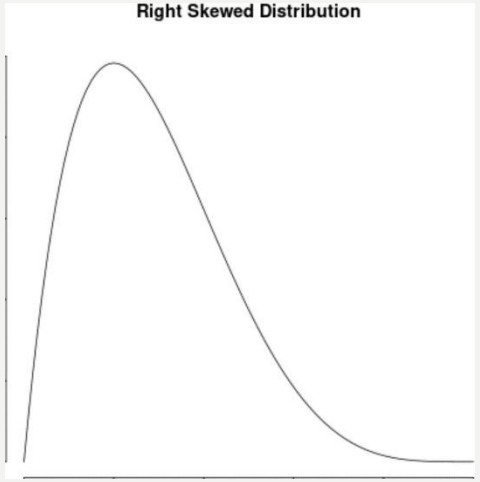
CENTRAL TENDENCIES IN DATA PRE-PROCESSING

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.

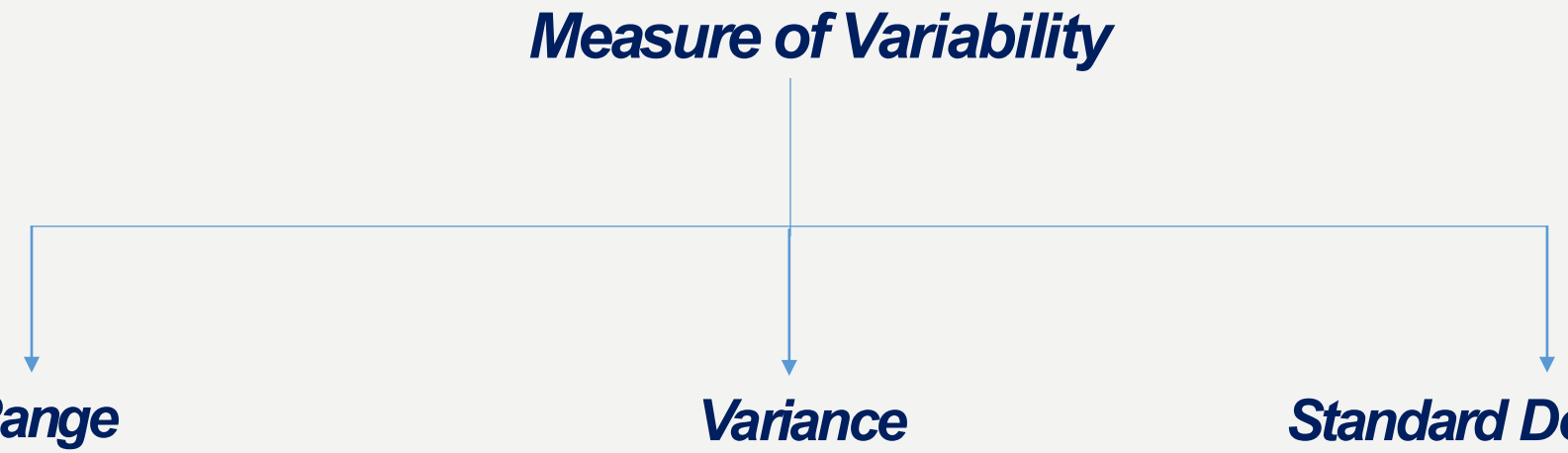
Mode : Missing values in a dataset can be replaced with **mode** value, if the data is skewed. Missing categorical values can also be replaced with **mode** value.



MEASURE OF VARIABILITY: RANGE, VARIANCE & STANDARD DEVIATION

Math for Machine Learning

MEASURE OF VARIABILITY



*The **range** of a set of data is the difference between the largest and smallest values. It can give a rough idea about the distribution of our dataset.*

$$Range = Max\ value\ - Min\ Value$$

***Variance** is a measure of how far each number in the set is from the mean and therefore from every other number in the dataset.*

$$\sigma^2 = \frac{\sum(\chi - \mu)^2}{N}$$

Standard Deviation is the square root of Variance. Standard deviation looks at how spread out a group of numbers is from the mean.

$$SD = \sqrt[2]{\sigma}$$

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

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

$$Range = 7 - 3 = 4$$

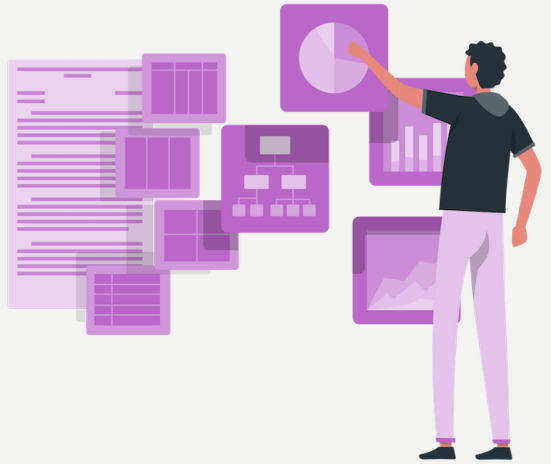
$$Variance = \frac{(3-5)^2 + (4-5)^2 + (5-5)^2 + (6-5)^2 + (7-5)^2}{5}$$

$$Variance = 2$$

$$Standard\ Deviation = 1.4$$

PERCENTILES & QUANTILES

Math for Machine Learning



PURPOSE OF THESE MEASUREMENTS

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