02
Statistics and Probability

AGENDA

- WHAT IS DATA?
- CATEGORIES OF DATA
- WHAT IS STATISTICS?
- BASIC TERMINOLOGIES IN STATISTICS
- SAMPLING TECHNIQUES
- TYPES OF STATISTICS
- DESCRIPTIVE STATISTICS
- PROBABILITY
- INFERENTIAL STATISTICS

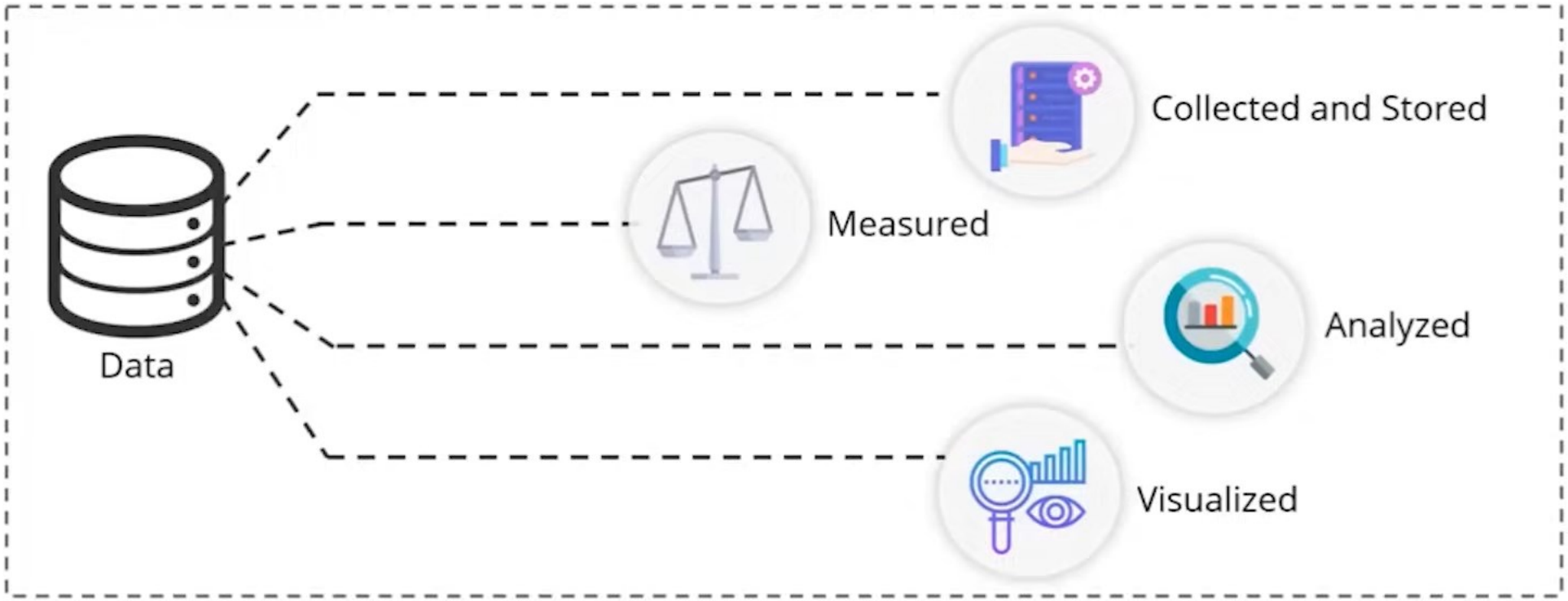
edureka!

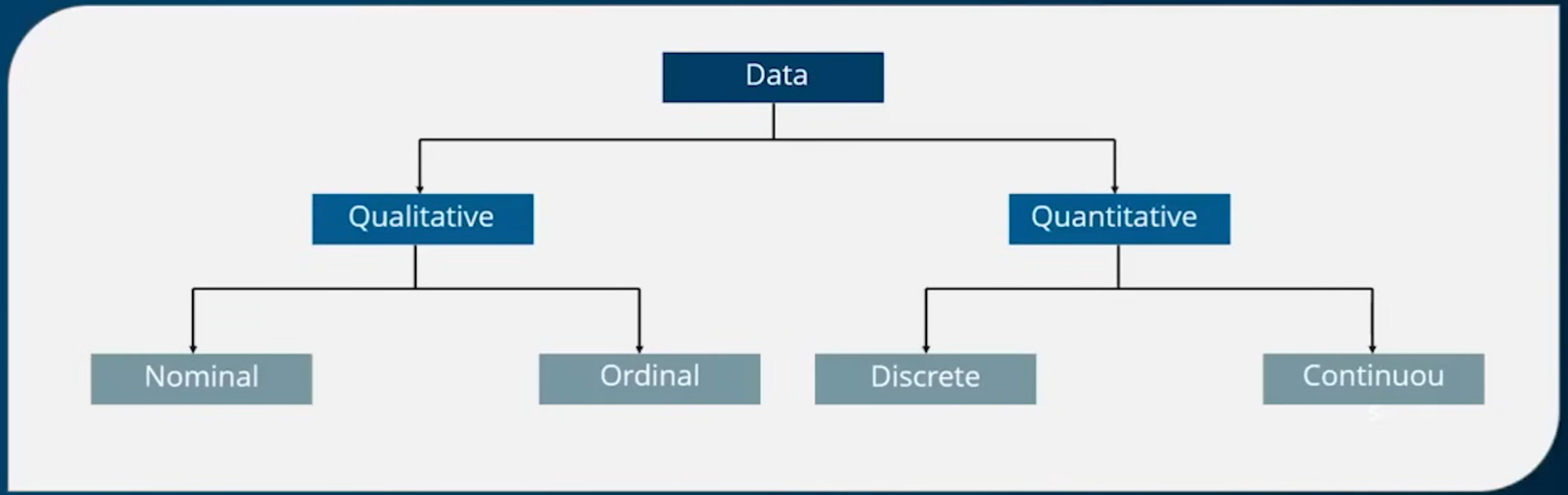
01

WHAT IS DATA?

WHAT IS DATA?

Data refers to facts and statistics collected together for reference or analysis.





Types Of Data

edureka!

QUALITATIVE DATA

Qualitative data deals with characteristics and descriptors that can't be easily measured, but can be observed subjectively.

Nominal Data

Data with no inherent order or ranking such as gender or race, such kind of data is called Nominal data

Gender
Male
Female
Male
Male

Ordinal Data

Data with an ordered series, such as shown in the table, such kind of data is called Ordinal data

Customer ID	Rating
001	Good
002	Average
003	Average
004	Bad

edureka!

QUANTITATIVE DATA

Quantitative data deals with numbers and things you can measure objectively.

Discrete Data

Also known as categorical data, it can hold finite number of possible values.

Example: Number of students in a class

Continuous Data

Data that can hold infinite number of possible values.

Example: Weight of a person

03

WHAT IS STATISTICS?

WHAT IS STATISTICS?

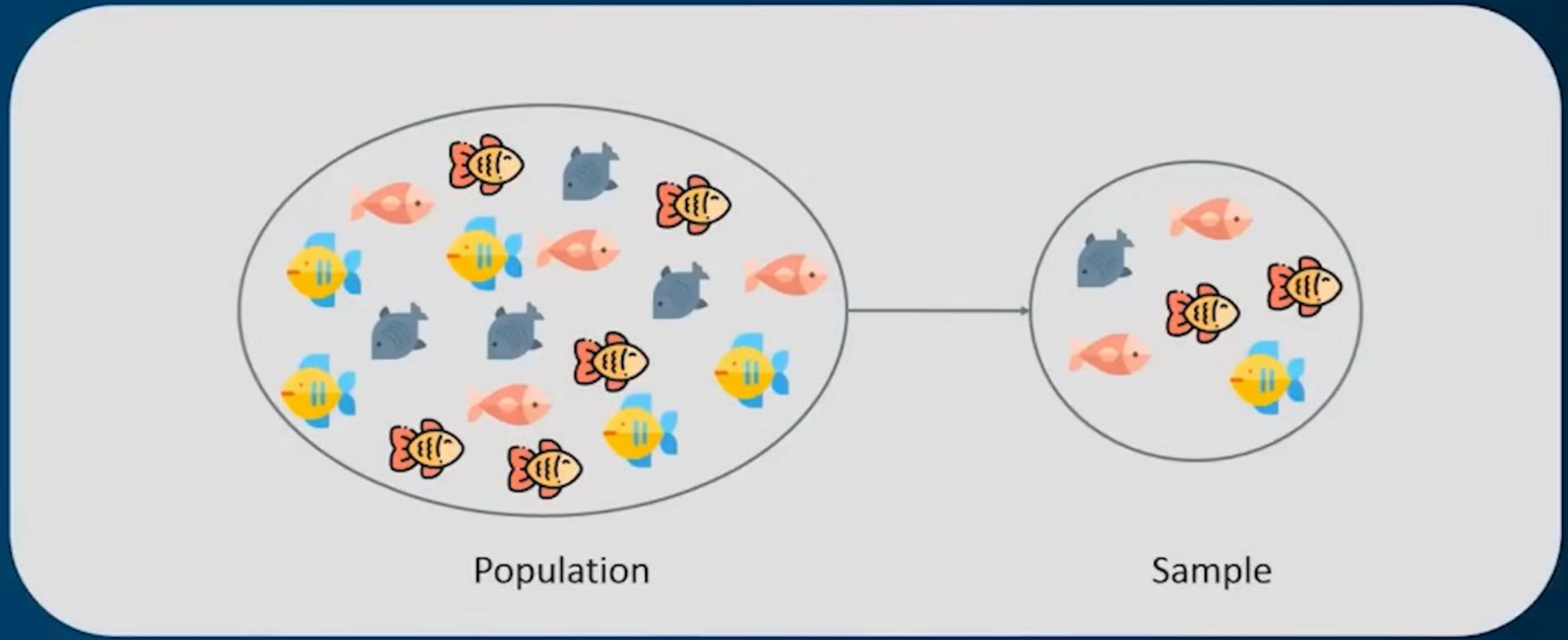
Statistics is an area of applied mathematics concerned with the data collection, analysis, interpretation and presentation.

The latest sales data have just come in, and your boss wants you to prepare a report for management on places where the company could improve its business. What should you look for? What should you not look for?

https://www.kaspersky.com/blog/tip-of-the-week-how-to-get-rid-of-unwanted-emails/3005/

04

BASIC TERMINOLOGIES IN STATISTICS



Statistics Terminologies

Population: A collection or set of individuals or objects or events whose properties are to be analyzed.

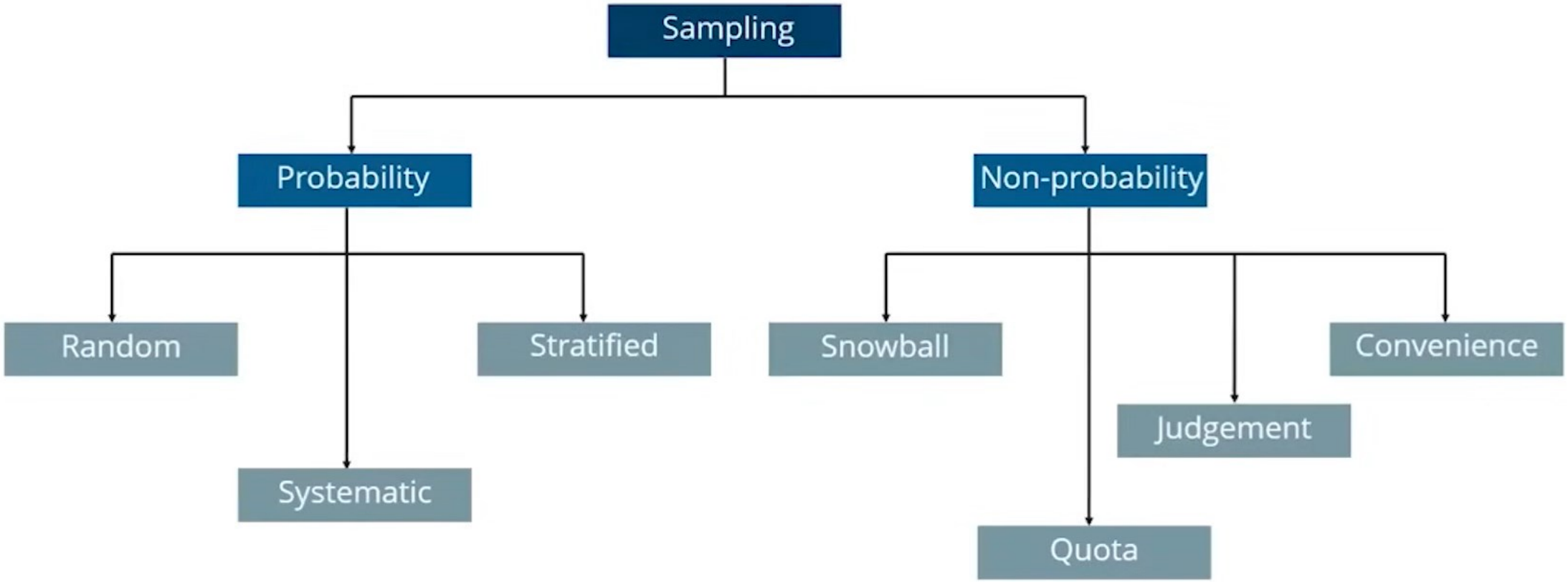
Sample: A subset of population is called 'Sample'. A well chosen sample will contain most of the information about a particular population parameter

05

SAMPLING TECHNIQUES

edureka!

SAMPLING TECHNIQUES



RANDOM SAMPLING

- Random Sampling
- Systematic Sampling
- Stratified Sampling

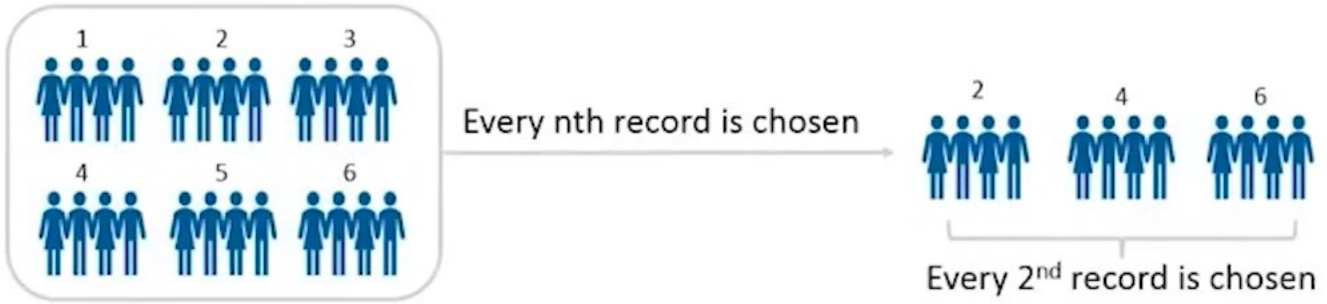
Each member of the population has equal chance of being selected in the sample.

SYSTEMATIC SAMPLING

Random Sampling

Systematic Sampling

Stratified Sampling



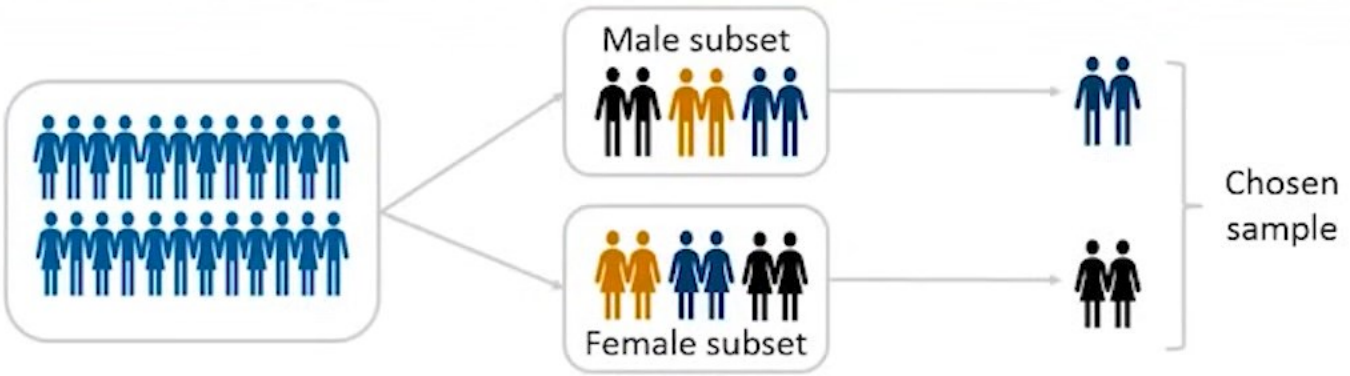
In Systematic sampling every nth record is chosen from the population to be a part of the sample.

STRATIFIED SAMPLING

Random Sampling

Systematic Sampling

Stratified Sampling



- *A stratum is a subset of the population that shares at least one common characteristic, in this case it's gender.*
- *Random sampling is used to select a sufficient number of subjects from each stratum.*

06

TYPES OF STATISTICS

DESCRIPTIVE STATISTICS

Descriptive statistics *uses the data to provide descriptions of the population, either through numerical calculations or graphs or tables.*

Maximum

Average

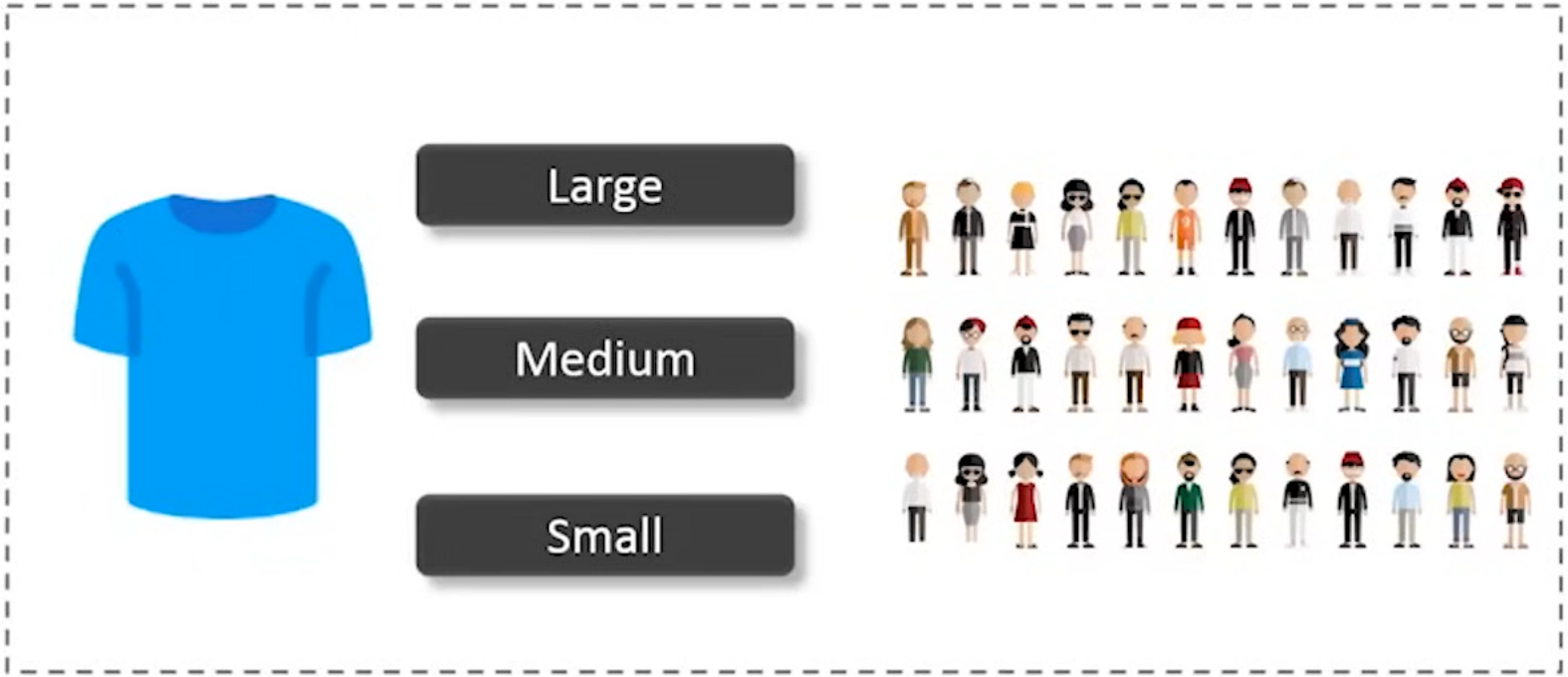
Minimum

Descriptive Statistics is mainly focused upon the main characteristics of data. It provides graphical summary of the data.

https://www.kaspersky.com/blog/tip-of-the-week-how-to-get-rid-of-unwanted-emails/3005/

INFERENTIAL STATISTICS

***Inferential statistics makes inferences and predictions about a population based on a sample** of data taken from the population in question.*



Inferential statistics, generalizes a large dataset and applies probability to draw a conclusion. It allows us to infer data parameters based on a statistical model using a sample data.

https://www.kaspersky.com/blog/tip-of-the-week-how-to-get-rid-of-unwanted-emails/3005/

07

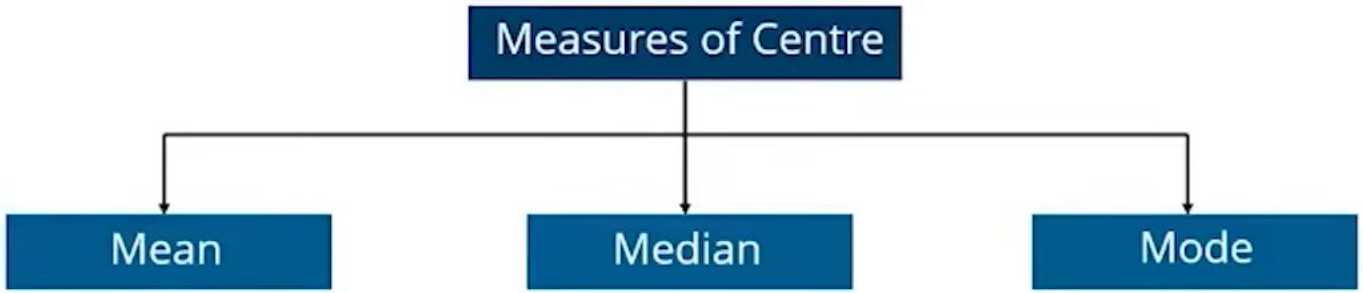
DESCRIPTIVE STATISTICS

DESCRIPTIVE STATISTICS

Descriptive statistics is a method used to describe and understand the features of a specific data set by giving short summaries about the sample and measures of the data.

Descriptive statistics are broken down into two categories:

- **Measures of Central tendency**
- Measures of Variability (spread)



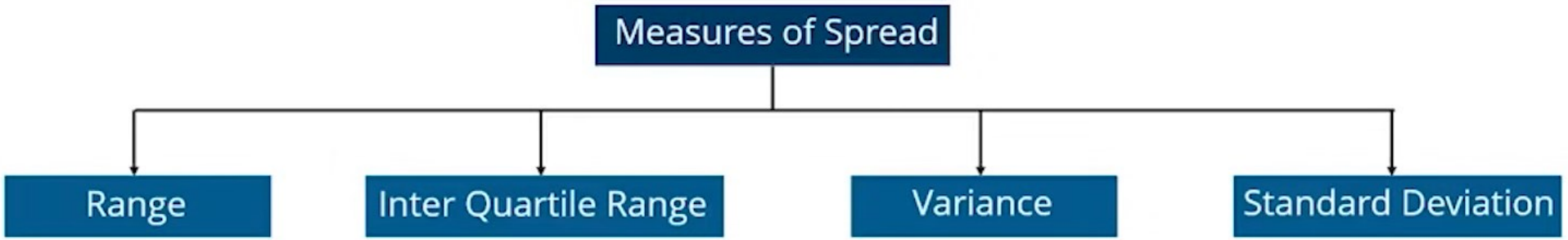
edureka!

DESCRIPTIVE STATISTICS

Descriptive statistics is a method used to describe and understand the features of a specific data set by giving short summaries about the sample and measures of the data.

Descriptive statistics are broken down into two categories:

- Measures of Central tendency
- **Measures of Variability (spread)**



08

MEASURES OF CENTRE

Here is a sample dataset of cars containing the variables:

- *Cars,*
- *Mileage per Gallon(mpg)*
- *Cylinder Type (cyl)*
- *Displacement (disp)*
- *Horse Power(hp)*
- *Real Axle Ratio(drat)*

Cars	mpg	cyl	disp	hp	drat
MazdaRX4	21	6	160	110	3.9
MazdaRX4_WAG	21	6	160	110	3.9
Datsun_710	22.8	4	108	93	3.85
Alto	21.3	6	108	96	3
WagonR	23	4	150	90	4
Toyata_ 11	23	6	108	110	3.9
Honda_12	23	4	160	110	3.9
Ford_11	23	6	160	110	3.9

Mean

Measure of average of all the values in a sample is called Mean.

Here is a sample dataset of cars containing the variables:

- *Cars,*
- *Mileage per Gallon(mpg)*
- *Cylinder Type (cyl)*
- *Displacement (disp)*
- *Horse Power(hp)*
- *Real Axle Ratio(drat)*

Cars	mpg	cyl	disp	hp	drat
MazdaRX4	21	6	160	110	3.9
MazdaRX4_WAG	21	6	160	110	3.9
Datsun_710	22.8	4	108	93	3.85
Alto	21.3	6	108	96	3
WagonR	23	4	150	90	4
Toyata_ 11	23	6	108	110	3.9
Honda_12	23	4	160	110	3.9
Ford_11	23	6	160	110	3.9

Mean

To find out the average horsepower of the cars among the population of cars, we will check and calculate the average of all values:

$$\frac{110 + 110 + 93 + 96 + 90 + 110 + 110 + 110}{8} = \mathbf{103.625}$$

Here is a sample dataset of cars containing the variables:

- *Cars,*
- *Mileage per Gallon(mpg)*
- *Cylinder Type (cyl)*
- *Displacement (disp)*
- *Horse Power(hp)*
- *Real Axle Ratio(drat)*

Cars	mpg	cyl	disp	hp	drat
MazdaRX4	21	6	160	110	3.9
MazdaRX4_WAG	21	6	160	110	3.9
Datsun_710	22.8	4	108	93	3.85
Alto	21.3	6	108	96	3
WagonR	23	4	150	90	4
Toyata_ 11	23	6	108	110	3.9
Honda_12	23	4	160	110	3.9
Ford_11	23	6	160	110	3.9

Median

Measure of the central value of the sample set is called **Median.**

Here is a sample dataset of cars containing the variables:

- *Cars,*
- *Mileage per Gallon(mpg)*
- *Cylinder Type (cyl)*
- *Displacement (disp)*
- *Horse Power(hp)*
- *Real Axle Ratio(drat)*

Cars	mpg	cyl	disp	hp	drat
MazdaRX4	21	6	160	110	3.9
MazdaRX4_WAG	21	6	160	110	3.9
Datsun_710	22.8	4	108	93	3.85
Alto	21.3	6	108	96	3
WagonR	23	4	150	90	4
Toyata_ 11	23	6	108	110	3.9
Honda_12	23	4	160	110	3.9
Ford_11	23	6	160	110	3.9

Median

To find out the center value of mpg among the population of cars, arrange records in *Ascending order, i.e.,* ***21, 21, 21.3, 22.8, 23, 23, 23, 23***

In case of even entries, take average of the two middle values, i.e. (22.8+23)/2 = 22.9

Here is a sample dataset of cars containing the variables:

- *Cars,*
- *Mileage per Gallon(mpg)*
- *Cylinder Type (cyl)*
- *Displacement (disp)*
- *Horse Power(hp)*
- *Real Axle Ratio(drat)*

Cars	mpg	cyl	disp	hp	drat
MazdaRX4	21	6	160	110	3.9
MazdaRX4_WAG	21	6	160	110	3.9
Datsun_710	22.8	4	108	93	3.85
Alto	21.3	6	108	96	3
WagonR	23	4	150	90	4
Toyata_ 11	23	6	108	110	3.9
Honda_12	23	4	160	110	3.9
Ford_11	23	6	160	110	3.9

Mode

The value most recurrent in the sample set is known as Mode.

Here is a sample dataset of cars containing the variables:

- *Cars,*
- *Mileage per Gallon(mpg)*
- *Cylinder Type (cyl)*
- *Displacement (disp)*
- *Horse Power(hp)*
- *Real Axle Ratio(drat)*

Cars	mpg	cyl	disp	hp	drat
MazdaRX4	21	6	160	110	3.9
MazdaRX4_WAG	21	6	160	110	3.9
Datsun_710	22.8	4	108	93	3.85
Alto	21.3	6	108	96	3
WagonR	23	4	150	90	4
Toyata_ 11	23	6	108	110	3.9
Honda_12	23	4	160	110	3.9
Ford_11	23	6	160	110	3.9

Mode

To find the most common type of cylinder among the population of cars, check the value which is repeated most number of times, i.e., *cylinder type 6*

09

MEASURES OF SPREAD

MEASURES OF SPREAD

A measure of spread, sometimes also called a measure of dispersion, is used to describe the variability in a sample or population.

Range | Inter Quartile Range | Variance | Standard Deviation

Range is the given measure of how spread apart the values in a dataset are.

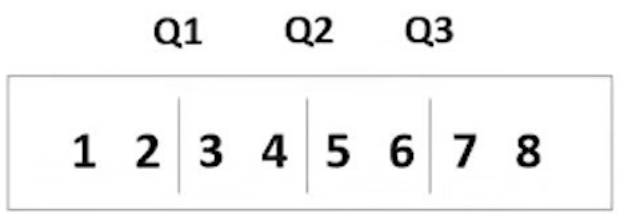
$$Range = Max(x_i) - Min(x_i)$$

MEASURES OF SPREAD

A measure of spread, sometimes also called a measure of dispersion, is used to describe the variability in a sample or population.

Range | Inter Quartile Range | Variance | Standard Deviation

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.



MEASURES OF SPREAD

Consider the marks of the 100 students below, ordered from the lowest to the highest scores

Order	Score	Order	Score	Order	Score	Order	Score	Order	Score
1st	35	21st	42	41st	53	61st	64	81st	74
2nd	37	22nd	42	42nd	53	62nd	64	82nd	74
3rd	37	23rd	44	43rd	54	63rd	65	83rd	74
4th	38	24th	44	44th	55	64th	66	84th	75
5th	39	25th	45	45th	55	65th	67	85th	75
6th	39	26th	45	46th	56	66th	67	86th	76
7th	39	27th	45	47th	57	67th	67	87th	77
8th	39	28th	45	48th	57	68th	67	88th	77
9th	39	29th	47	49th	58	69th	68	89th	79
10th	40	30th	48	50th	58	70th	69	90th	80
11th	40	31st	49	51st	59	71st	69	91st	81
12th	40	32nd	49	52nd	60	72nd	69	92nd	81
13th	40	33rd	49	53rd	61	73rd	70	93rd	81
14th	40	34th	49	54th	62	74th	70	94th	81
15th	40	35th	51	55th	62	75th	71	95th	81
16th	41	36th	51	56th	62	76th	71	96th	81
17th	41	37th	51	57th	63	77th	71	97th	83
18th	42	38th	51	58th	63	78th	72	98th	84
19th	42	39th	52	59th	64	79th	74	99th	84
20th	42	40th	52	60th	64	80th	74	100th	85

The first quartile (Q1) lies between the 25th and 26th.
Q1 = (45 + 45) ÷ 2 = 45

The second quartile (Q2) between the 50th and 51st.
Q2 = (58 + 59) ÷ 2 = 58.5

The third quartile (Q3) between the 75th and 76th.
Q3 = (71 + 71) ÷ 2 = 71

MEASURES OF SPREAD

A measure of spread, sometimes also called a measure of dispersion, is used to describe the variability in a sample or population.

Range | Inter Quartile Range | Variance | Standard Deviation

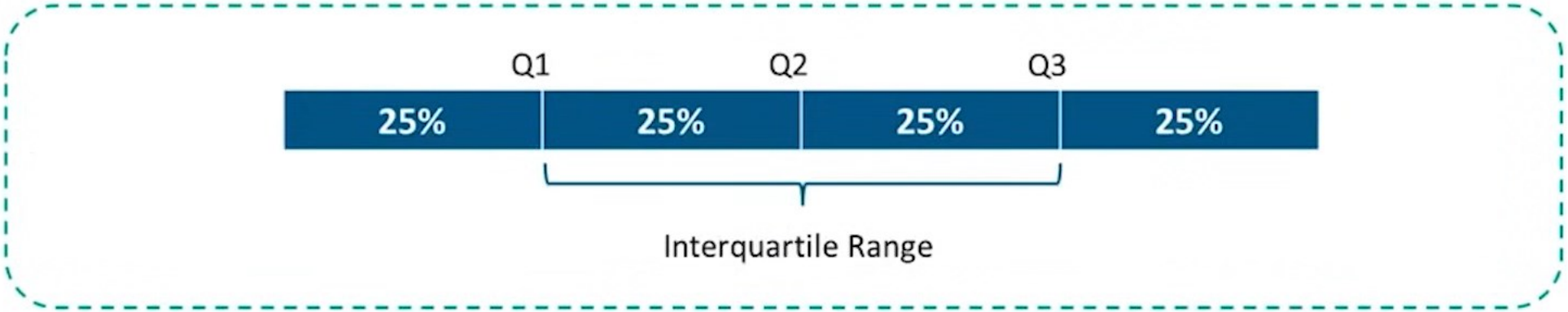
Inter Quartile Range(IQR) is the measure of variability, based on dividing a dataset into quartiles.

- *Quartiles divide a rank-ordered data set into four equal parts, denoted by Q1, Q2, and Q3, respectively*
- *The interquartile range is equal to Q3 minus Q1, i.e.. IQR = Q3 - Q1*

MEASURES OF SPREAD

A measure of spread, sometimes also called a measure of dispersion, is used to describe the variability in a sample or population.

Range | Inter Quartile Range | Variance | Standard Deviation



edureka!

MEASURES OF SPREAD

A measure of spread, sometimes also called a measure of dispersion, is used to describe the variability in a sample or population.

Range | Inter Quartile Range | Variance | Standard Deviation

Variance describes how much a random variable differs from its expected value.
It entails computing squares of deviations.

$$s^2 = \frac{\sum_{1}^{n=1} (x_i - \bar{x})^2}{n}$$

x : Individual data points
n : Total number of data points
$\bar{x}$: Mean of data points

MEASURES OF SPREAD

A measure of spread, sometimes also called a measure of dispersion, is used to describe the variability in a sample or population.

Range | Inter Quartile Range | Variance | Standard Deviation

Deviation is the difference between each element from the mean.

$$Deviation = (x_i - \mu)$$

edureka!

MEASURES OF SPREAD

A measure of spread, sometimes also called a measure of dispersion, is used to describe the variability in a sample or population.

Range | Inter Quartile Range | Variance | Standard Deviation

Population Variance is the average of squared deviations.

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} = (x_i - \mu)^2$$

MEASURES OF SPREAD

A measure of spread, sometimes also called a measure of dispersion, is used to describe the variability in a sample or population.

Range | Inter Quartile Range | Variance | Standard Deviation

Sample Variance is the average of squared differences from the mean.

$$s^2 = \frac{1}{(n-1)} \sum_{i=1}^{N} = (x_i - \bar{x})^2$$

MEASURES OF SPREAD

A measure of spread, sometimes also called a measure of dispersion, is used to describe the variability in a sample or population.

Range | Inter Quartile Range | Variance | Standard Deviation

Standard Deviation is the measure of the dispersion of a set of data from its mean.

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

edureka!

STANDARD DEVIATION

Standard Deviation Use Case: Daenerys has 20 Dragons. They have the numbers 9, 2, 5, 4, 12, 7, 8, 11, 9, 3, 7, 4, 12, 5, 4, 10, 9, 6, 9, 4. Work out the Standard Deviation.

STEP 1

Find out the mean for your sample set.

The Mean is:

$$\frac{9+2+5+4+12+7+8+11+9+3+7+4+12+5+4+10+9+6+9+4}{20}$$

$\mu=7$

STANDARD DEVIATION

Standard Deviation Use Case: Daenerys has 20 Dragons. They have the numbers 9, 2, 5, 4, 12, 7, 8, 11, 9, 3, 7, 4, 12, 5, 4, 10, 9, 6, 9, 4. Work out the Standard Deviation.

STEP 2

Then for each number, subtract the Mean and square the result.

$(x_i - \mu)^2$

(9-7)²= 2^2=4
(2-7)²= $(-5)^2$=25
(5-7)²= $(-2)^2$=4
And so on...

We get the following results:
4, 25, 4, 9, 25, 0, 1, 16, 4, 16, 0, 9, 25, 4, 9, 9, 4, 1, 4, 9

edureka!

STANDARD DEVIATION

Standard Deviation Use Case: Daenerys has 20 Dragons. They have the numbers 9, 2, 5, 4, 12, 7, 8, 11, 9, 3, 7, 4, 12, 5, 4, 10, 9, 6, 9, 4. Work out the Standard Deviation.

STEP 3

Then work out the mean of those squared differences.

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N} =(x_i - \mu)^2}$$

$$\frac{4+25+4+9+25+0+1+16+4+16+0+9+25+4+9+9+4+1+4+9}{20}$$

$\sigma^2 = 8.9$

STANDARD DEVIATION

Standard Deviation Use Case: Daenerys has 20 Dragons. They have the numbers 9, 2, 5, 4, 12, 7, 8, 11, 9, 3, 7, 4, 12, 5, 4, 10, 9, 6, 9, 4. Work out the Standard Deviation.

STEP 4

Take square root of σ^2.

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N} =(x_i - \mu)^2}$$

σ = 2.983

10

INFORMATION GAIN & ENTROPY

edureka!

INFORMATION GAIN & ENTROPY

Entropy

Entropy measures the impurity or uncertainty present in the data.

$$H(S) = -\sum_{i=1}^{N} p_i \log_2 p_i$$

where:

- ***S – set of all instances in the dataset***
- ***N – number of distinct class values***
- ***pi – event probability***

Information Gain (IG)

IG indicates how much "information" a particular feature/ variable gives us about the final outcome.

$$Gain(A,S) = H(S) - \sum_{j=1}^{v} \frac{|S_j|}{|S|} \cdot H(S_j) = H(S) - H(A,S)$$

where:

H(S) – entropy of the whole dataset S

- ***|Sj| – number of instance with j value of an attribute A***
- ***|S| – total number of instances in dataset S***
- ***v – set of distinct values of an attribute A***
- ***H(Sj) – entropy of subset of instances for attribute A***
- ***H(A, S) – entropy of an attribute A***

11

USE CASE

USE CASE

Day	Outlook	Humidity	Wind	Play
D1	Sunny	High	Weak	No
D2	Sunny	High	Strong	No
D3	Overcast	High	Weak	Yes
D4	Rain	High	Weak	Yes
D5	Rain	Normal	Weak	Yes
D6	Rain	Normal	Strong	No
D7	Overcast	Normal	Strong	Yes
D8	Sunny	High	Weak	No
D9	Sunny	Normal	Weak	Yes
D10	Rain	Normal	Weak	Yes
D11	Sunny	Normal	Strong	Yes
D12	Overcast	High	Strong	Yes
D13	Overcast	Normal	Weak	Yes
D14	Rain	High	Strong	No

USE CASE

Yes – 9
No – 5

Day	Outlook	Humidity	Wind
D1	Sunny	High	Weak
D2	Sunny	High	Strong
D8	Sunny	High	Weak
D9	Sunny	Normal	Weak
D11	Sunny	Normal	Strong

Yes – 2
No – 3

Day	Outlook	Humidity	Wind
D3	Overcast	High	Weak
D7	Overcast	Normal	Strong
D12	Overcast	High	Strong
D13	Overcast	Normal	Weak

Yes – 4
No – 0

Day	Outlook	Humidity	Wind
D4	Rain	High	Weak
D5	Rain	Normal	Weak
D6	Rain	Normal	Strong
D10	Rain	Normal	Weak
D14	Rain	High	Strong

Yes – 3
No – 2

USE CASE

From the total of 14 instances we have:

- *9 instances "yes"*
- *5 instances "no"*

The Entropy is:

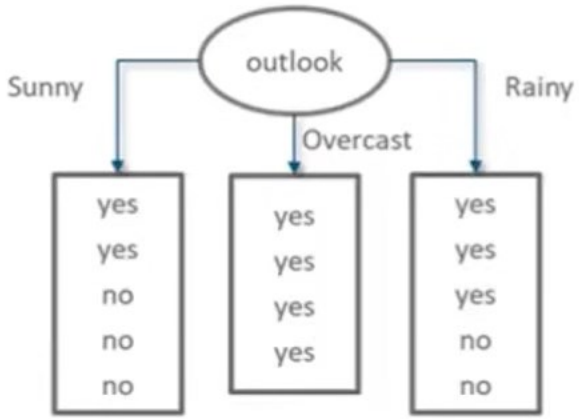
$$H(S) = -\sum_{i=1}^{N} p_i \log_2 p_i$$

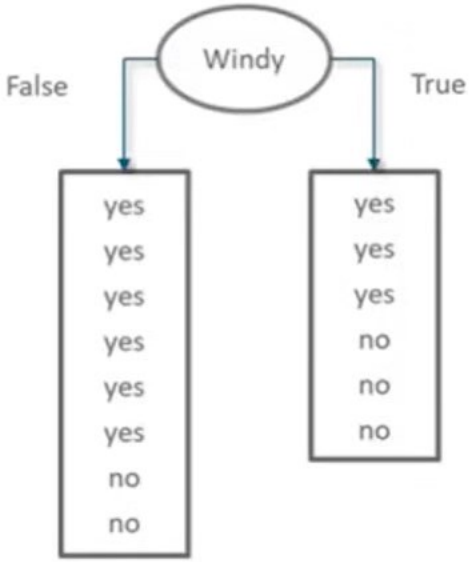
$$H(S) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.940$$

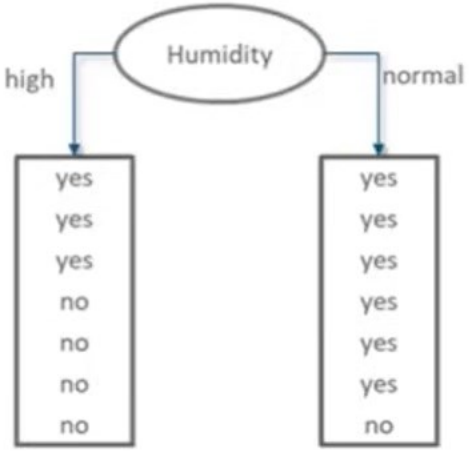
edureka!

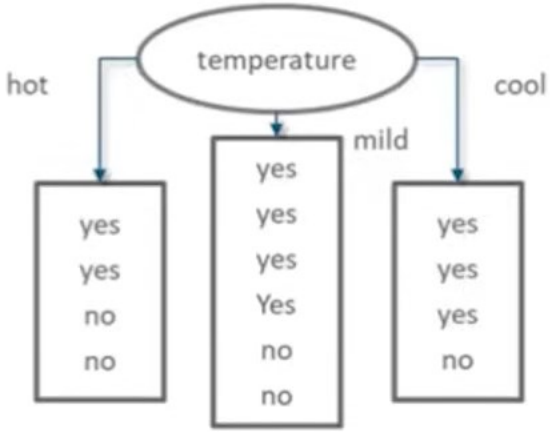
USE CASE

Selecting the root variable









USE CASE

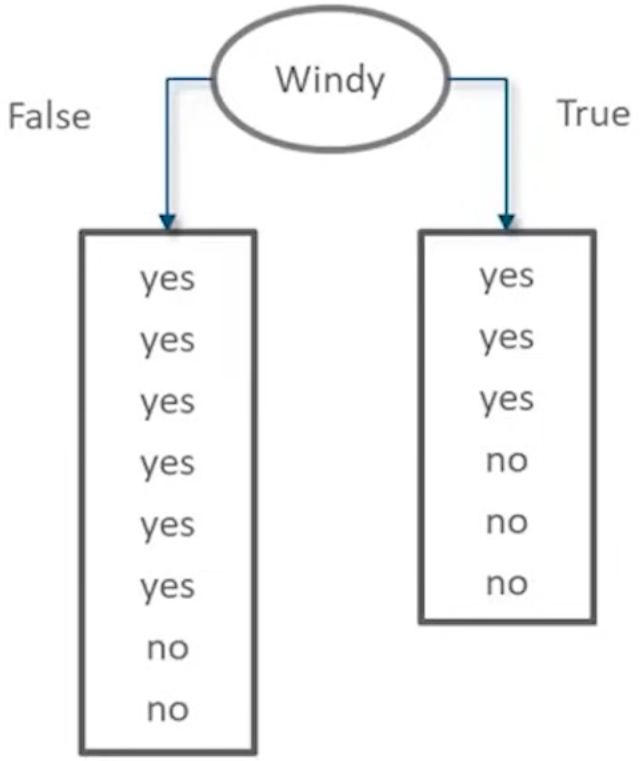
Information Gain of attribute "windy"

From the total of 14 instances we have:

- *6 instances "true"*
- *8 instances "false"*

$$Gain(A,S) = H(S) - \sum_{j=1}^{v} \frac{|S_j|}{|S|} \cdot H(S_j)$$

$$Gain(A_{Windy}, S) = 0.940 - \frac{8}{14} \cdot \left(-\left(\frac{6}{8} \cdot \log_2 \frac{6}{8} + \frac{2}{8} \cdot \log_2 \frac{2}{8} \right) \right) + \frac{6}{14} \cdot \left(-\left(\frac{3}{6} \cdot \log_2 \frac{3}{6} + \frac{3}{6} \cdot \log_2 \frac{3}{6} \right) \right) = 0.048$$



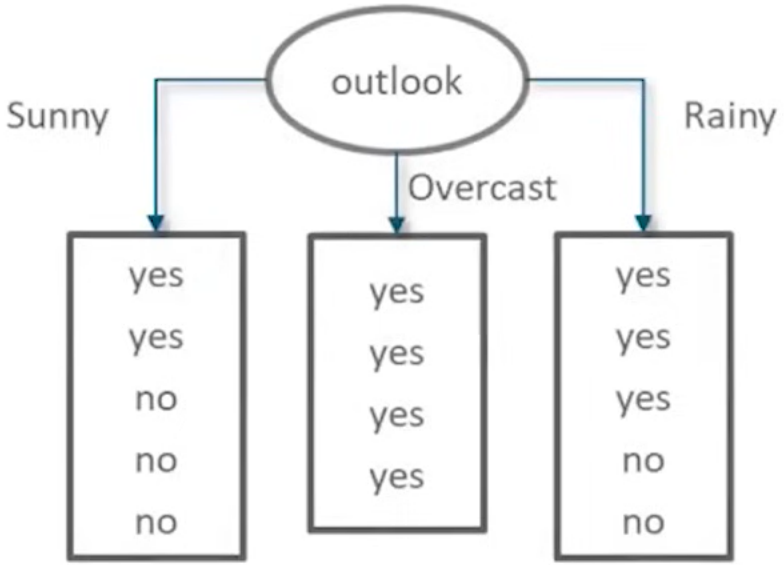
USE CASE

Information Gain of attribute "outlook"

From the total of 14 instances we have:

- *5 instances "sunny"*
- *4 instances "overcast"*
- *5 instances "rainy"*

$$Gain(A_{Outlook}, S) = 0.940 - \frac{5}{14} \cdot \left(-\left(\frac{2}{5} \log_2 \frac{2}{5} + \frac{3}{5} \log_2 \frac{3}{5} \right) \right) + \frac{4}{14} \cdot \left(-\left(\frac{4}{4} \log_2 \frac{4}{4} \right) \right) + \frac{5}{14} \cdot \left(-\left(\frac{3}{5} \cdot \log_2 \frac{3}{5} + \frac{2}{5} \cdot \log_2 \frac{2}{5} \right) \right) = 0.247$$



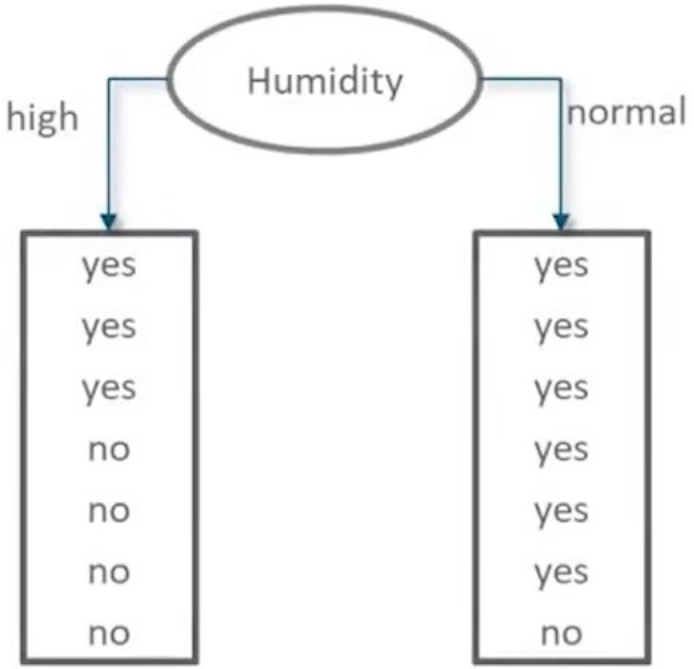
USE CASE

Information Gain of attribute "humidity"

From the total of 14 instances we have:

- *7 instances "high"*
- *7 instances "normal"*

$$Gain(A_{Humidity}, S) = 0.940 - \frac{7}{14} \cdot \left(-\left(\frac{3}{7} \cdot \log_2 \frac{3}{7} + \frac{4}{7} \cdot \log_2 \frac{4}{7} \right) \right) + \frac{7}{14} \cdot \left(-\left(\frac{6}{7} \cdot \log_2 \frac{6}{7} + \frac{1}{7} \cdot \log_2 \frac{1}{7} \right) \right) = 0.151$$



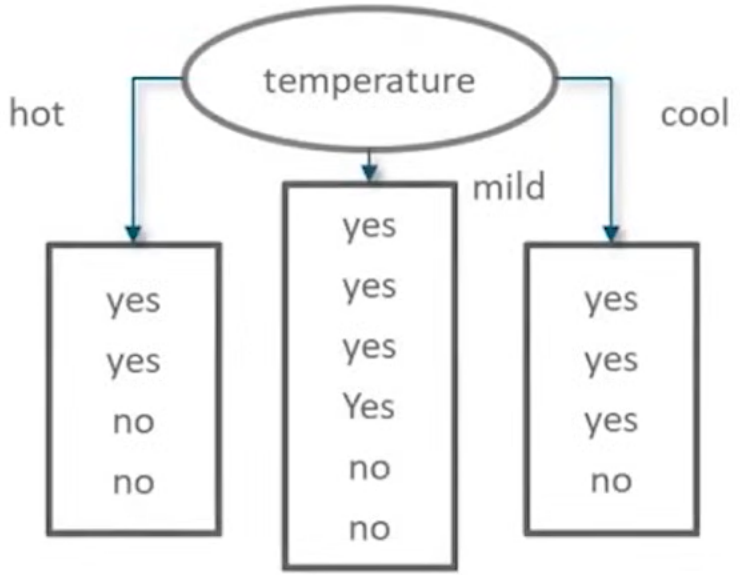
USE CASE

Information Gain of attribute "temperature"

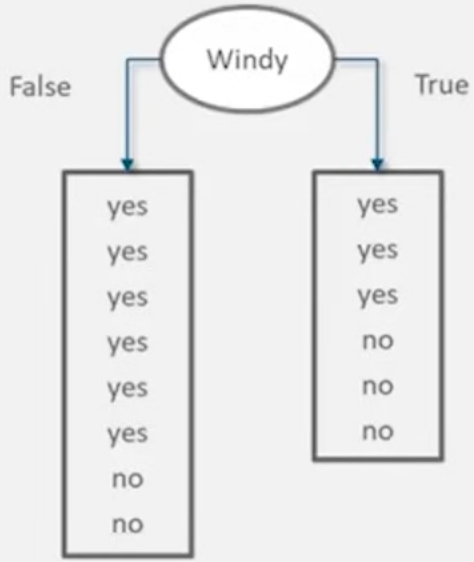
From the total of 14 instances we have:

- *4 instances "hot"*
- *6 instances "mild"*
- *4 instances "cool"*

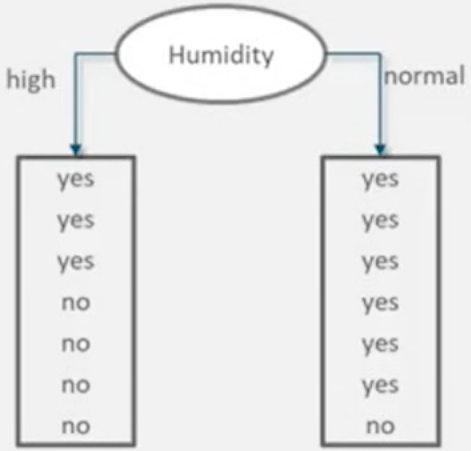
$$Gain(A_{Temperature}, S) = 0.940 - \frac{4}{14} \cdot \left(-\left(\frac{2}{4} \cdot \log_2 \frac{2}{4} + \frac{2}{4} \cdot \log_2 \frac{2}{4} \right) \right) + \frac{6}{14} \cdot \left(-\left(\frac{4}{6} \cdot \log_2 \frac{4}{6} + \frac{2}{6} \cdot \log_2 \frac{2}{6} \right) \right) + \frac{4}{14} \cdot \left(-\left(\frac{3}{4} \cdot \log_2 \frac{3}{4} + \frac{1}{4} \cdot \log_2 \frac{1}{4} \right) \right) = 0.029$$



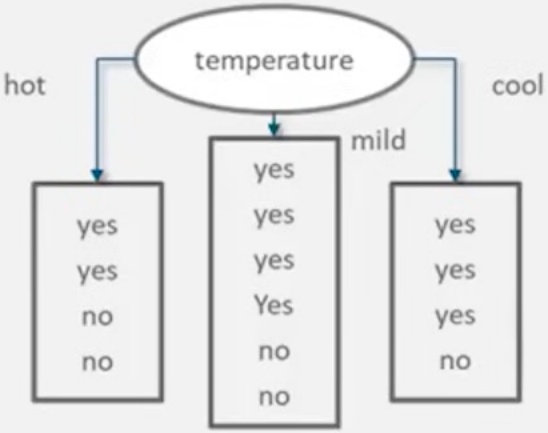
Gain = 0.247



Gain = 0.048

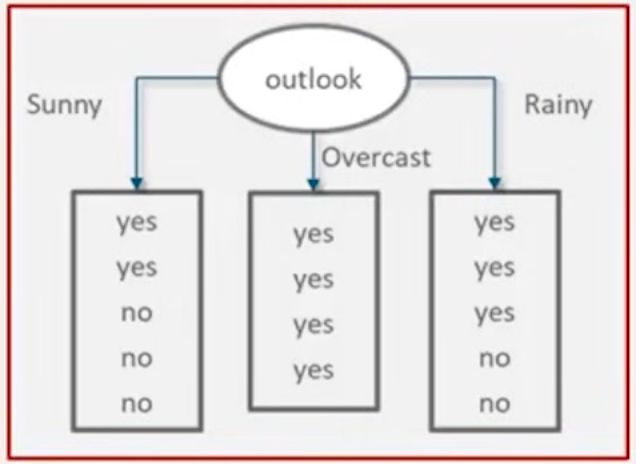


Gain = 0.151

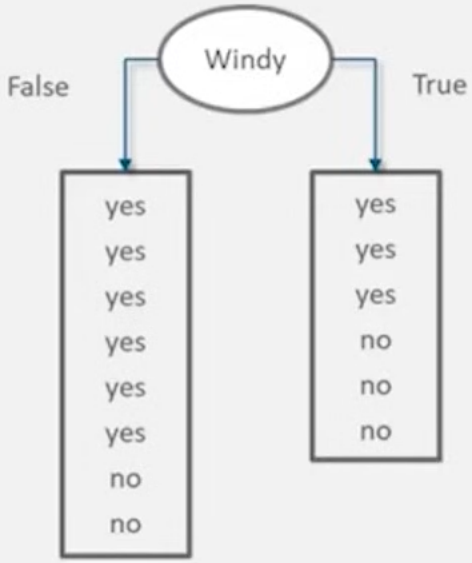


Gain = 0.029

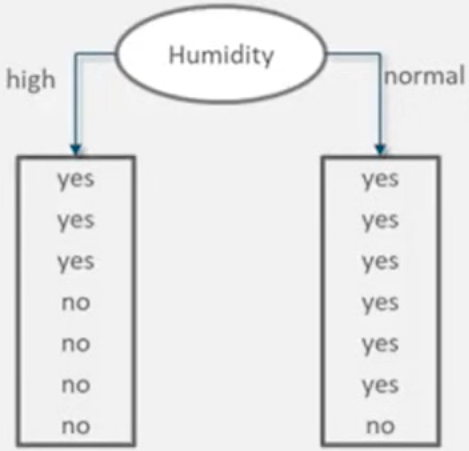
The variable with the highest IG is used to split the data at the root node.



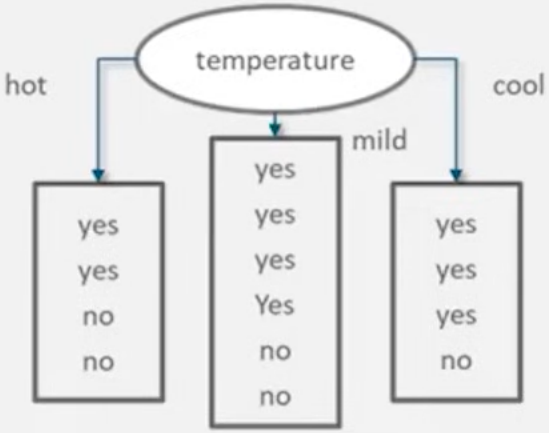
Gain = 0.247



Gain = 0.048



Gain = 0.151



Gain = 0.029

The variable with the highest IG is used to split the data at the root node. The 'Outlook' variable has the highest IG, therefore it can be assigned to the root node.

12

CONFUSION MATRIX

CONFUSION MATRIX

A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known.

Confusion Matrix represents a tabular representation of Actual vs Predicted values
You can calculate the accuracy of your model with:

$$\frac{\text{True Positives} + \text{True Negatives}}{\text{True Positives} + \text{True Negatives} + \text{False Positives} + \text{False Negatives}}$$

edureka!

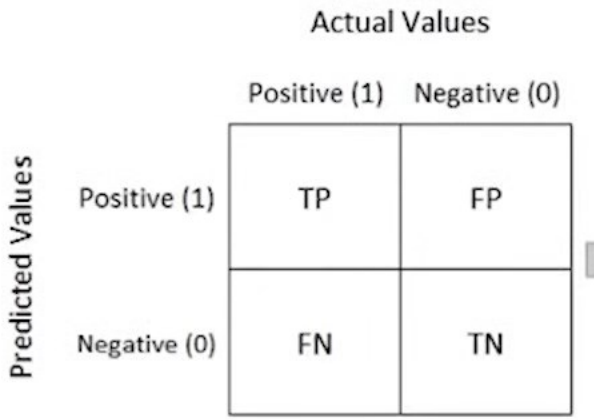
CONFUSION MATRIX

- There are two possible predicted classes: "yes" and "no"
- The classifier made a total of 165 predictions
- Out of those 165 cases, the classifier predicted "yes" 110 times, and "no" 55 times
- In reality, 105 patients in the sample have the disease, and 60 patients do not

n=165	**Predicted: NO**	**Predicted: YES**
Actual: NO	50	10
Actual: YES	5	100

edureka!

CONFUSION MATRIX



true positives (TP): These are cases in which we predicted yes (they have the disease), and they do have the disease

true negatives (TN): We predicted no, and they don't have the disease

false positives (FP): We predicted yes, but they don't actually have the disease. (Also known as a "Type I error.")

false negatives (FN): We predicted no, but they actually do have the disease. (Also known as a "Type II error.")


5 seconds
13
PROBABILITY

13

PROBABILITY

WHAT IS PROBABILITY?

Probability is the measure of how likely an event will occur.

- Probability is the ratio of desired outcomes to total outcomes:

(desired outcomes) / (total outcomes)

- Probabilities of all outcomes always sums to 1

Example:

- On rolling a dice, you get 6 possible outcomes
- Each possibility only has one outcome, so each has a probability of 1/6
- For example, the probability of getting a number '2' on the dice is 1/6

14

TERMINOLOGIES IN PROBABILITY

edureka!

TERMINOLOGIES IN PROBABILITY

Random Experiment

An experiment or a process for which the outcome cannot be predicted with certainty

Sample Space

The entire possible set of outcomes of a random experiment is the sample space (S) of that experiment

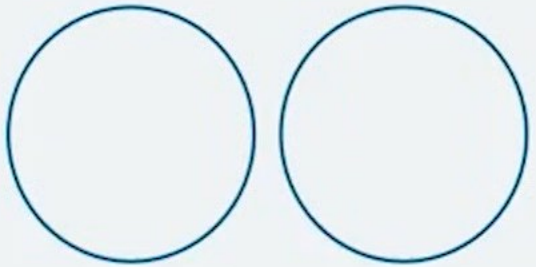
Event

One or more outcomes of an experiment. It is a subset of sample space(S)

TYPES OF EVENTS

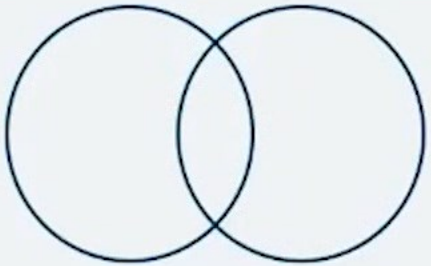
Disjoint Events *do not have any common outcomes.*

- The outcome of a ball delivered cannot be a sixer and a wicket
- A single card drawn from a deck cannot be a king and a queen
- A man cannot be dead and alive



Non-Disjoint Events *can have common outcomes*

- A student can get 100 marks in statistics and 100 marks in probability
- The outcome of a ball delivered can be a no ball and a six

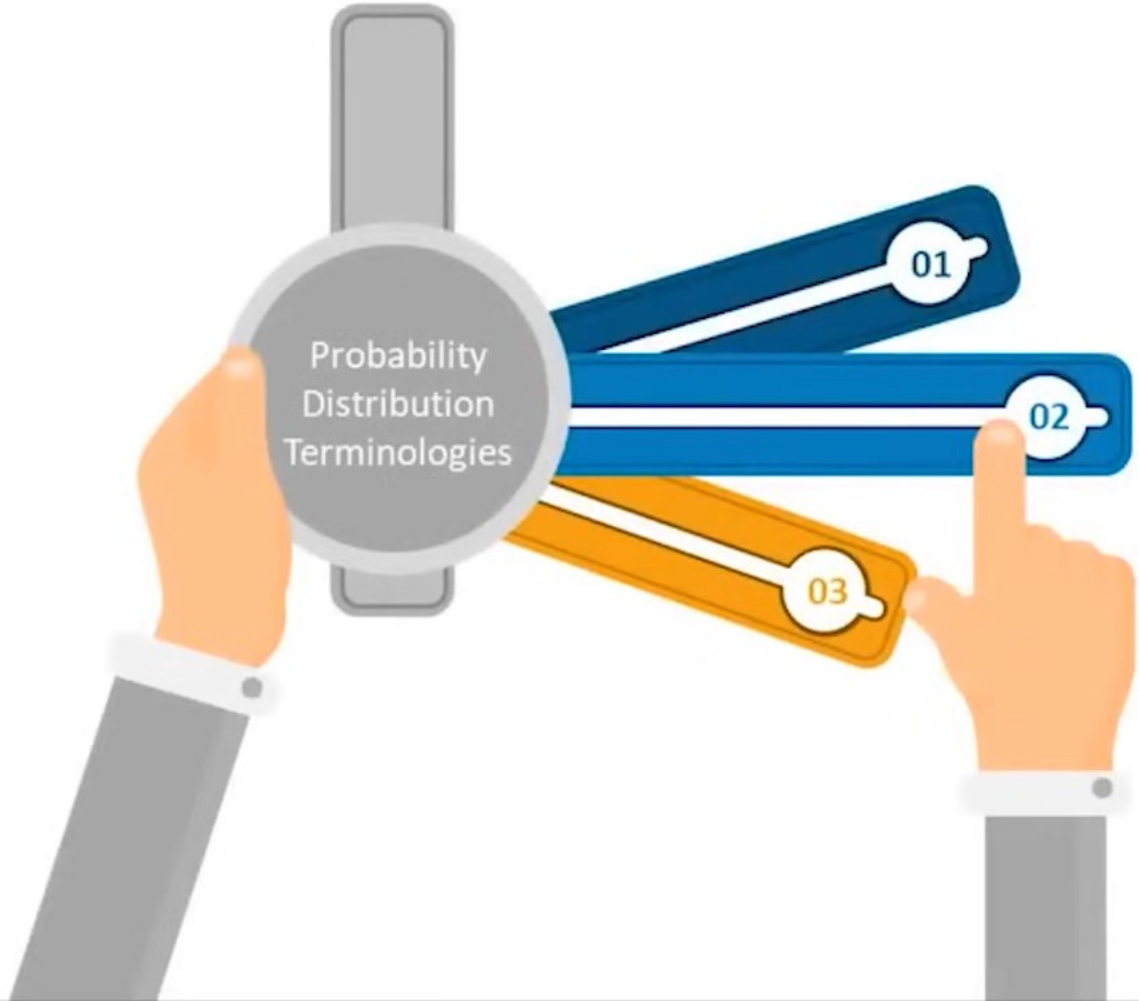


15

PROBABILITY DISTRIBUTION

edureka!

PROBABILITY DISTRIBUTION



Probability Density Function 01

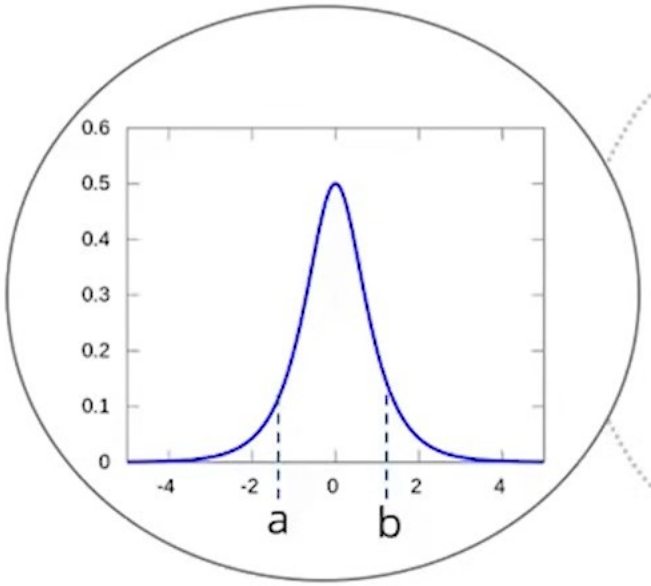
Normal Distribution 02

Central Limit Theorem 03

edureka!

PROBABILITY DENSITY FUNCTION

The equation describing a continuous probability distribution is called a Probability Density Function



Property 01

Graph of a PDF will be continuous over a range

Property 02

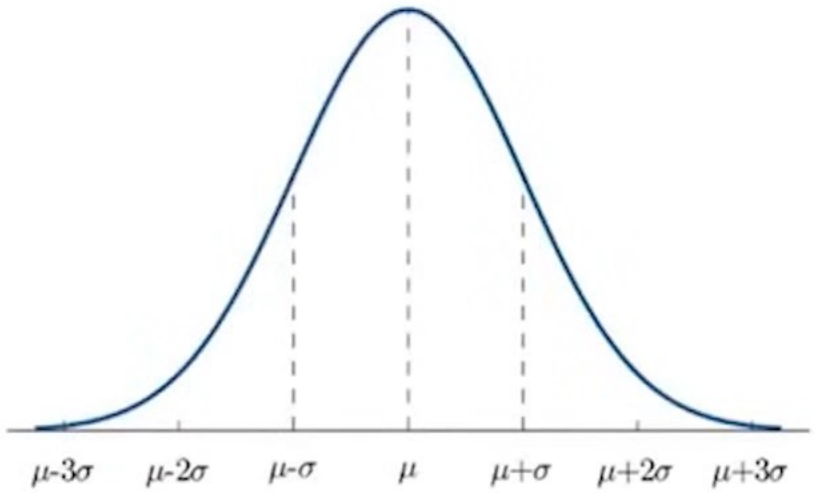
Area bounded by the curve of density function and the x-axis is equal to 1

Property 03

Probability that a random variable assumes a value between a & b is equal to the area under the PDF bounded by a & b

NORMAL DISTRIBUTION

The Normal Distribution is a probability distribution that associates the normal random variable X with a cumulative probability



$$Y = [\ 1/\sigma * \text{sqrt}(2\pi)\ ] * e^{-(x - \mu)2/2\sigma2}$$

Where,
- *X* is a normal random variable
- μ is the mean and
- σ is the standard deviation

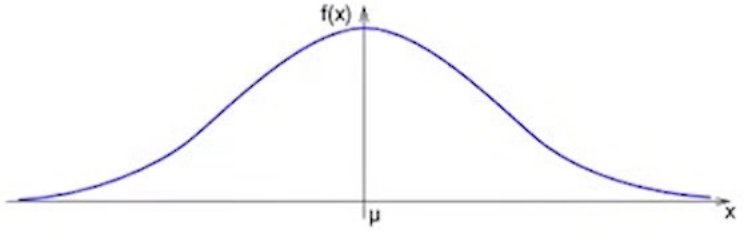
Note: Normal Random variable is variable with mean at 0 and variance equal to 1

edureka!

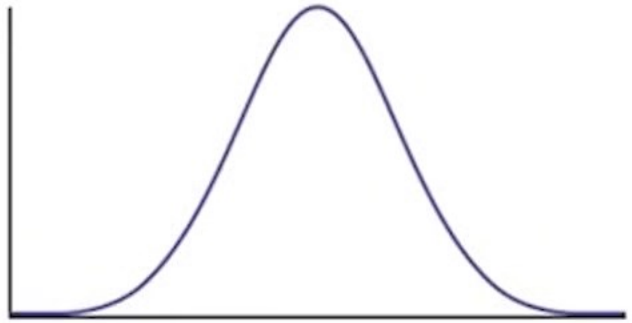
STANDARD DEVIATION & CURVE

The graph of the Normal Distribution depends on two factors: the *Mean* and the *Standard Deviation*

- **Mean:** *Determines the location of center of the graph*
- **Standard Deviation:** *Determines the height of the graph*



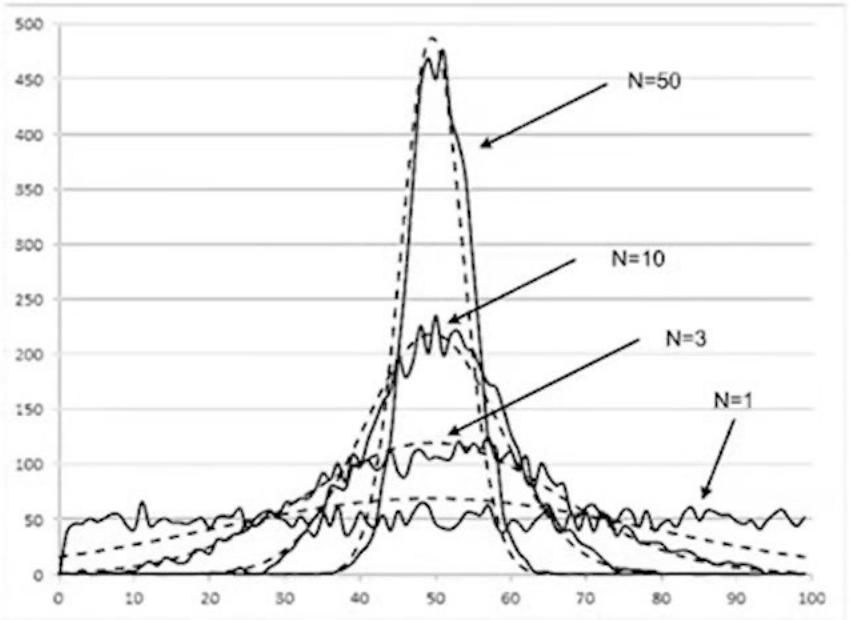
If the standard deviation is large, the curve is short and wide.



If the standard deviation is small, the curve is tall and narrow.

CENTRAL LIMIT THEOREM

The ***Central Limit Theorem*** states that the sampling distribution of the mean of any independent, random variable will be normal or nearly normal, if the sample size is large enough

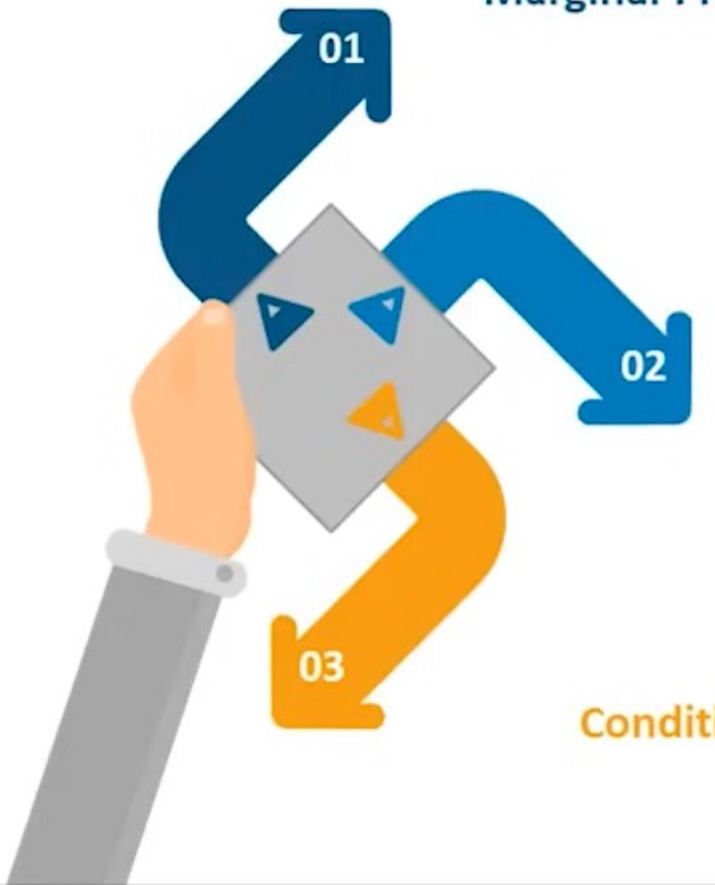


16

TYPES OF PROBABILITY

edureka!

TYPES OF PROBABILITY



Marginal Probability

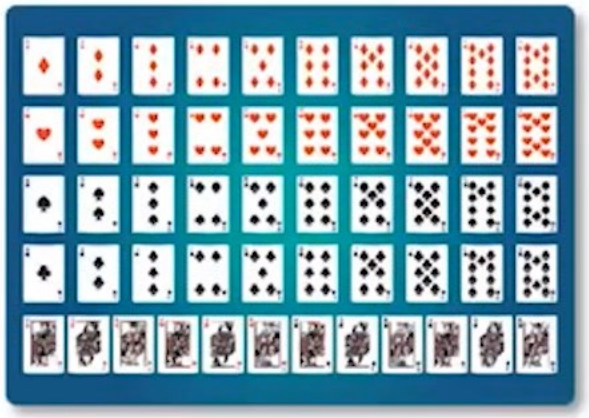
Joint Probability

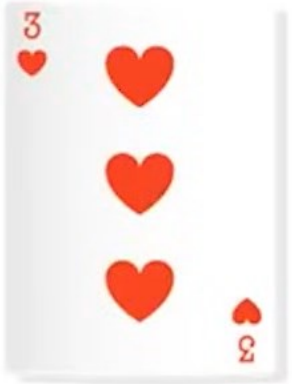
Conditional Probability

edureka!

MARGINAL PROBABILITY

Marginal Probability is the probability of occurrence of a single event.





Marginal Probability = $\frac{13}{52}$

It can be expressed as: P (A) = $\sum_{i=1}^{k} P(x_i)$

edureka!

JOINT PROBABILITY

Joint Probability is a measure of two events happening at the same time

Example: The probability that a card is an Ace of hearts = P (Ace of hearts)
(There are 13 heart cards in a deck of 52 and out of them one in the Ace of hearts)

CONDITIONAL PROBABILITY

- *Probability of an event or outcome based on the occurrence of a previous event or outcome*
- *Conditional Probability of an event B is the probability that the event will occur given that an event A has already occurred*

If A and B are dependent events then the expression for conditional probability is given by:

$$P(B|A) = P(A \text{ and } B) \ / \ P(A)$$

If A and B are independent events then the expression for conditional probability is given by:

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

17

USE CASE

USE CASE

EDUREKA'S TRAINING STUDY
TRAINING AND SALARY PACKAGE OF CANDIDATES Study examines salary package and training undergone by candidates
Sample: 60 candidates without training and 45 candidates with edureka's training
Task to do: Assessment of training with salary package

USE CASE – DATA SET

Results		Training		
		Without Edureka Training	With Edureka Training	Total
Salary Package obtained by participant	Very Poor Package	5	0	5
	Poor Package	10	0	10
	Average Package	40	10	50
	Good Package	5	30	35
	Excellent Package	0	5	5
	Total	60	45	105

MARGINAL PROBABILITY

The probability that a candidate has undergone Edureka's training

P(Edu.Training) = 45 /105 ≈ 0.42

Results		Training		
		Without Edureka Training	With Edureka Training	Total
Salary Package obtained by participant	Very Poor Package	5	0	5
	Poor Package	10	0	10
	Average Package	40	10	50
	Good Package	5	30	35
	Excellent Package	0	5	5
	Total	60	45	105

JOINT PROBABILITY

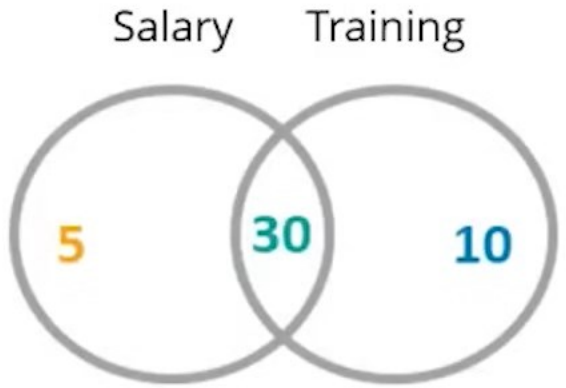
Finding the probability that a candidate has attended Edureka's training and also has good package.

Results		Training		
		Without Edureka Training	With Edureka Training	Total
Salary Package obtained by participant	Very Poor Package	5	0	5
	Poor Package	10	0	10
	Average Package	40	10	50
	Good Package	5	30	35
	Excellent Package	0	5	5
	Total	60	45	105

edureka!

JOINT PROBABILITY

Results		Training		
		Without Edureka Training	With Edureka Training	Total
Salary Package obtained by participant	Very Poor Package	5	0	5
	Poor Package	10	0	10
	Average Package	40	10	50
	Good Package	5	30	35
	Excellent Package	0	5	5
	Total	60	45	105



P (Good Package & Edu.Training) =
30 /105 ≈ 0.28

edureka!

CONDITIONAL PROBABILITY

Finding the probability that a candidate has a good package given that he has not undergone training

Results		Training		
		Without Edureka Training	With Edureka Training	Total
Salary Package obtained by participant	Very Poor Package	5	0	5
	Poor Package	10	0	10
	Average Package	40	10	50
	Good Package	5	30	35
	Excellent Package	0	5	5
	Total	60	45	105

edureka!

CONDITIONAL PROBABILITY

Results		Training		
		Without Edureka Training	With Edureka Training	Total
Salary Package obtained by participant	Very Poor Package	5	0	5
	Poor Package	10	0	10
	Average Package	40	10	50
	Good Package	5	30	35
	Excellent Package	0	5	5
	Total	60	45	105

P (Good Package | Without Edureka) = 5 / 60 ≈ 0.08

18

BAYES' THEOREM

edureka!

BAYES' THEOREM

Shows the relation between one conditional probability and its inverse

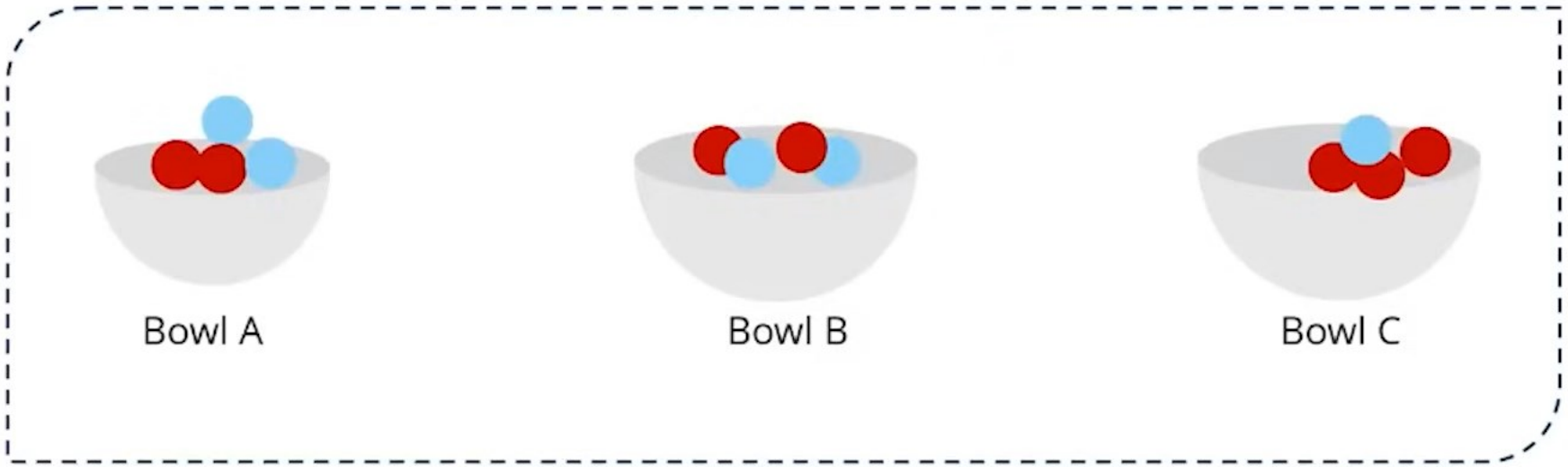
P (B | A) is referred to as *likelihood ratio* which measures the probability (given event A) of occurrence of B

P (A) is referred to as *Prior* which represents the actual probability distribution of A

P (A | B) is referred to as *posterior* which means the probability of occurrence of A given B

BAYES' THEOREM EXAMPLE

" Consider 3 bowls. Bowl A contains 2 blue balls and 4 red balls; Bowl B contains 8 blue balls and 4 red balls, Bowl C contains 1 blue ball and 3 red balls. We draw 1 ball from each bowl. What is the probability to draw a blue ball from Bowl A if we know that we drew exactly a total of 2 blue balls? "



BAYES' THEOREM EXAMPLE

- Let *A* be the *event of picking a blue ball* from bag A, and let *X* be the *event of picking exactly two blue balls*
- We want Probability(A|X), i.e. probability of occurrence of event A given X

By the definition of *Conditional Probability*,

$$\Pr(A|X) = \frac{Pr(A \cap X)}{\Pr(X)}$$

- We need to find the two probabilities on the right-side of equal to symbolop

BAYES' THEOREM EXAMPLE

Steps to execute the problem

Step 1:

First find Pr(X). This can happen in three ways:

(i) white from A, white from B, red from C

(ii) white from A, red from B, white from C

(iii) red from A, white from B, white from C

Next we find Pr(A∩X).

This is the sum of terms (i) and (ii) above

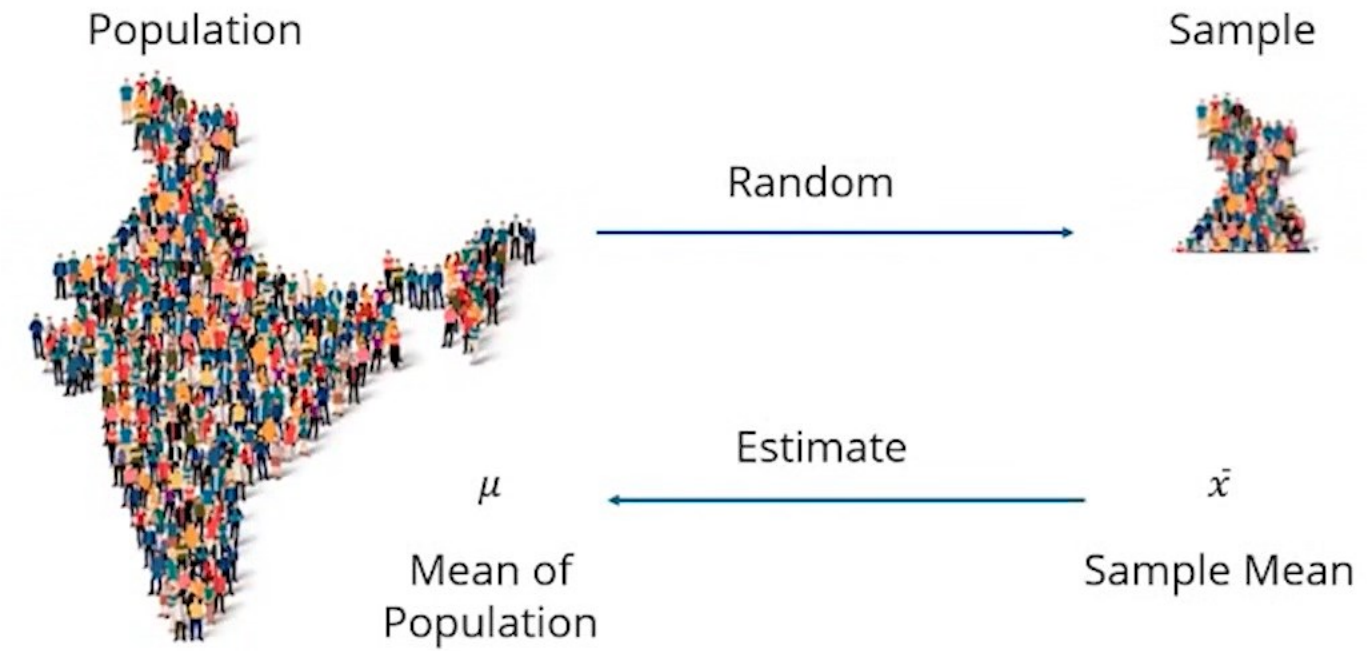
Step 2:

19

INFERENTIAL STATISTICS

POINT ESTIMATION

Point Estimation is concerned with the use of the sample data to measure a single value which serves as an approximate value or the best estimate of an unknown population parameter.



edureka!

FINDING THE ESTIMATES

Method of Moments

Estimates are found out by equating the first k sample moments to the corresponding k population moments

Maximum of Likelihood

Uses a model and the values in the model to maximize a likelihood function. This results in the most likely parameter for the inputs selected

Bayes' Estimators

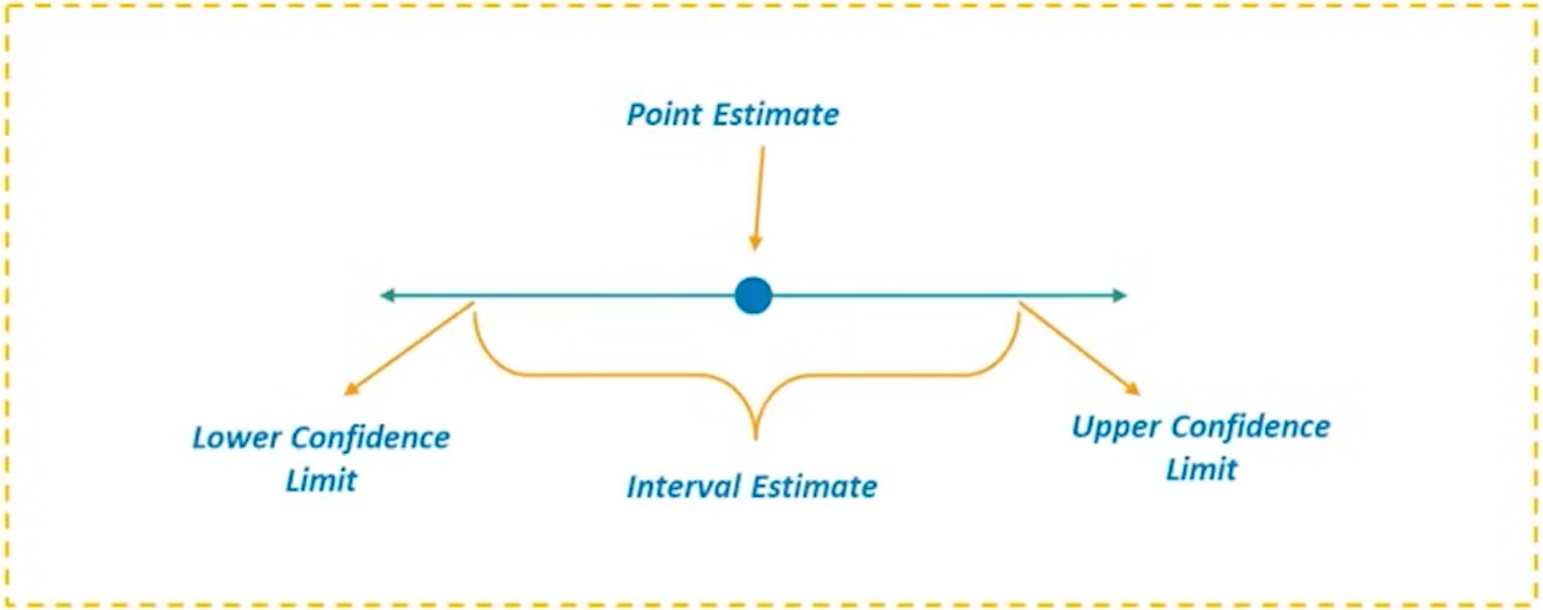
Minimizes the average risk (an expectation of random variables)

Best Unbiased Estimators

Several unbiased estimators can be used to approximate a parameter (which one is "best" depends on what parameter you are trying to find)

INTERVAL ESTIMATE

An Interval, or range of values, used to estimate a population parameter is called Interval Estimate.



edureka!

CONFIDENCE INTERVAL

Confidence Interval is the measure of your confidence, that the interval estimate contains the population mean, μ

Statisticians use a confidence interval to describe the amount of uncertainty associated with a sample estimate of a population parameter

Technically, a range of values so constructed that there is a specified probability of including the true value of a parameter within it

MARGIN OF ERROR

- Difference between the point estimate and the actual population parameter value is called the **Sampling Error**
- When μ is estimated, the sampling error is the difference $\mu - \bar{x}$

Margin of Error E, for a given level of confidence is the greatest possible distance between the point estimate and the value of the parameter it is estimating

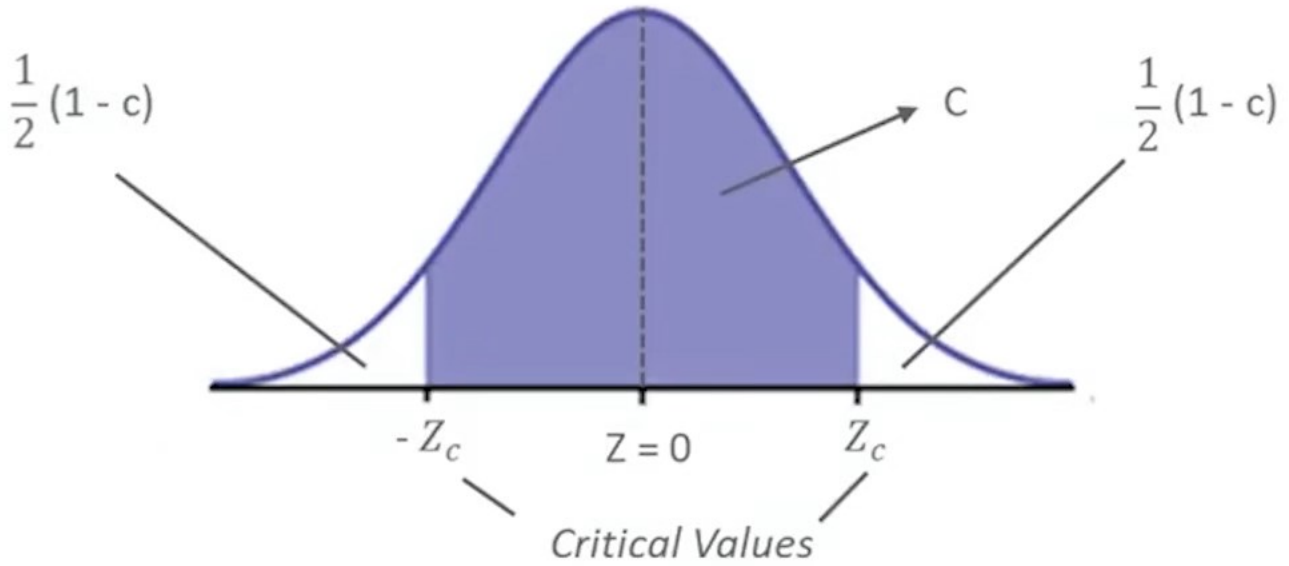
$$E = Z_c \frac{\sigma}{\sqrt{n}}$$

22

ESTIMATING LEVEL OF CONFIDENCE

ESTIMATING LEVEL OF CONFIDENCE

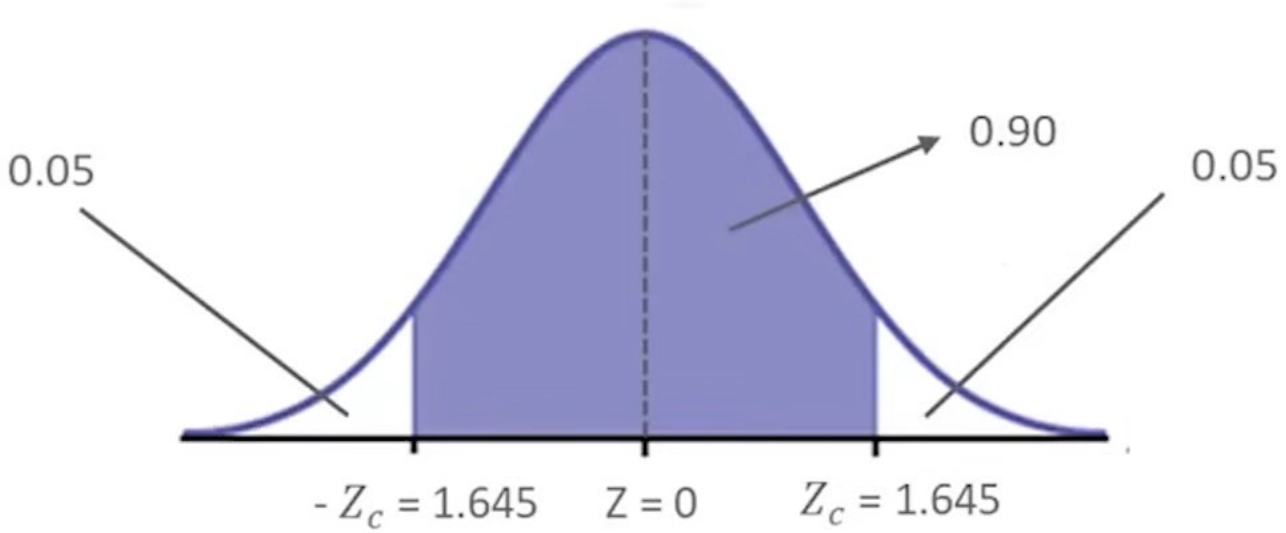
The level of confidence c, is the probability that the interval estimate contains the population parameter.



C is the area beneath the normal curve between the critical values
Corresponding Z score can be calculated using the standard normal table

ESTIMATING LEVEL OF CONFIDENCE

If the level of confidence is 90%, this means that you are 90% confident that the interval contains the population mean, μ.



The Corresponding Z – scores are ± 1.645

23

USE CASE

MARGIN OF ERROR – USE CASE

A random sample of 32 textbook prices is taken from a local college bookstore. The mean of the sample is $\bar{x}$ = 74.22, and the sample standard deviation is S = 23.44. Use a 95% confidence level and find the margin of error for the mean price of all textbooks in the bookstore

You know by formula,

$$E = Z_c \frac{\sigma}{\sqrt{n}}$$

$$E = 1.96 * (23.44/\sqrt{32}) \approx 8.12$$

24

HYPOTHESIS TESTING

HYPOTHESIS TESTING

Statisticians use hypothesis testing to formally check whether the hypothesis is accepted or rejected.

Hypothesis testing is conducted in the following manner:

- **State the Hypotheses –** This stage involves stating the null and alternative hypotheses.
- **Formulate an Analysis Plan –** This stage involves the construction of an analysis plan.
- **Analyse Sample Data –** This stage involves the calculation and interpretation of the test statistic as described in the analysis plan.
- **Interpret Results –** This stage involves the application of the decision rule described in the analysis plan.

25

UNDERSTANDING HYPOTHESIS TESTING

edureka!

HYPOTHESIS TESTING EXAMPLE