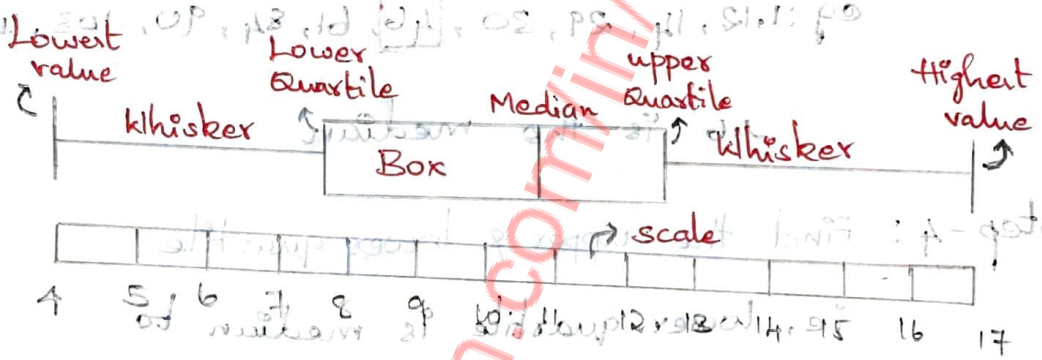
Box and Whisker plots and the five number summary :

Box and Whisker plot :

Box plot summarizes data using the minimum, lower quartile, median, upper quartile and maximum values. It allows to see important data at a glance.



The Five number summary

The five number summary is the numerical representation of box and whisker plot.

The Five number summary consists of,

- ⇒ The minimum value in a dataset.
- ⇒ The 1st Quartile (lower quartile / 25%).
- ⇒ The Median (2nd Quartile / 50%).
- ⇒ The 3rd Quartile (upper quartile / 75%).
- ⇒ The maximum value in a dataset.

Note: The Box and Whisker plot should be constructed with the scale.

steps to construct box & Whisker plot: (odd number)

step-1: Take the set of given number / data points.

eg: 1, 46, 29, 84, 12, 14, 103, 61, 90, 30, 110

step-2: Place them in Ascending order.

eg: 1, 12, 14, 29, 30, 46, 61, 84, 90, 103, 110

step-3: Find the median

ie., Median is the middle value of a data set

eg: 1, 12, 14, 29, 30, [46], 61, 84, 90, 103, 110

46 is the median

step-4: Find the upper & lower quartile

i.e, lower quartile is median to the left of actual median and upper quartile is median to the right of actual median.

eg: (1, 12, [14], 29, 30), 46, (61, 84, [90], 103, 110)

14 is lower Quartile

90 is upper Quartile

step-5: Find the minimum & maximum values.

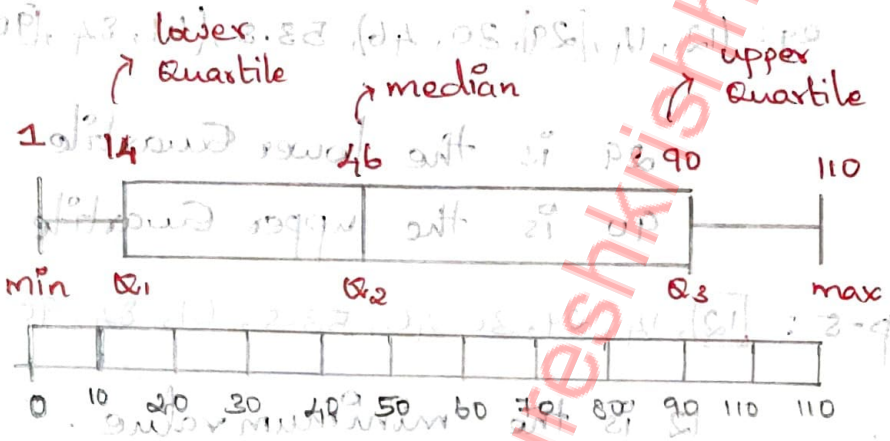
eg: [1], 12, 14, 29, 30, 46, 61, 84, 90, 103, [110]

1 is the minimum

110 is the maximum

step-6: construct the box plot with available data.

Example:



The Five number summary is,

→ Minimum - 1
→ lower Quartile - 14
→ median - 46
→ upper Quartile - 90
→ Maximum - 110

steps to construct box plot (Even number):

step-1: 46, 29, 84, 12, 14, 103, 61, 90, 30, 110

step-2: 12, 14, 29, 30, 46, 61, 84, 90, 103, 110

step-3: Find the median:

i.e., For even number dataset, the median would be the mean of middle two numbers.

eg: 12, 14, 29, 30, 46, 61, 84, 90, 103, 110

mean = (46+61)/2

median = 53.5

53.5 is the median

step-4 : Find the lower and upper Quartile by placing the median in the dataset.

eg: (12, 14, [29], 30, 46), 53.5, (61, 84, [90], 103, 110)

29 is the lower Quartile.

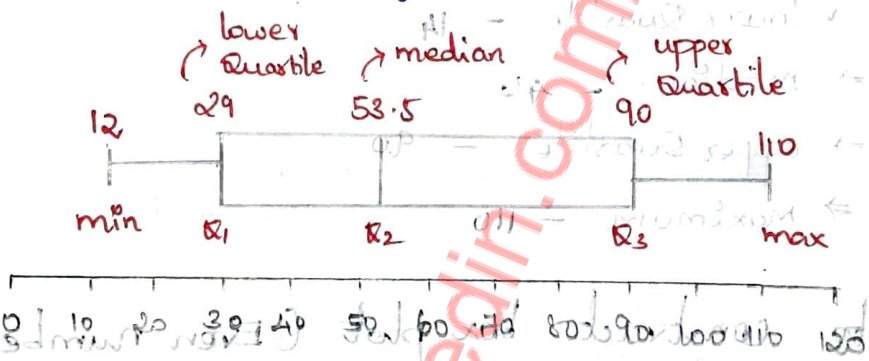
90 is the upper Quartile

step-5 : [12], 14, 29, 30, 46, 53.5, 61, 84, 90, 103, [110]

12 is the minimum value.

110 is the maximum value.

step-6 : constructing box plot :



The Five number summary is,

⇒ Minimum – 12

⇒ lower Quartile – 29

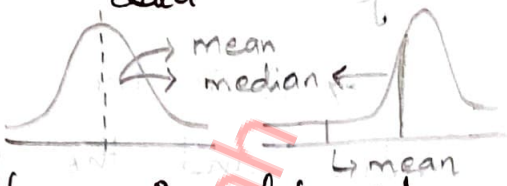
⇒ Median – 53.5

⇒ Upper Quartile – 90

⇒ Maximum – 110

Central tendency:

Mean and Median in symmetric & Asymmetric data

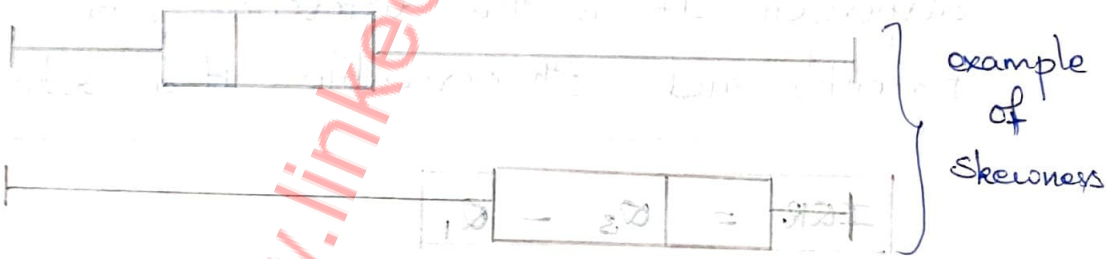


* Mean
* Median
* Mode → more frequently occuring datapoint.

→ In general mean is affected by the outliers (datapoints that differs significantly from other observations). But median is not affected by outliers. In box-plot median can only be plotted.

⇒ When there is no significant difference in the value of mean & median, then the dataset has no outliers.

⇒ Asymetrical boxplot represents the presence of outliers in the dataset with skewness.



Dispersion:

Dispersion represents the distribution of dataset as its stretched or squeezed. The dispersion can be characterized by

* Range
* Inter Quartile Range (IQR)
* variance (σ^2)
* standard deviation (σ)

Range :

→ It is the distance between the minimum value and the maximum value of the dataset.

$$\text{Range} = \underset{\text{value}}{\{\text{Maximum}\}} - \{\text{Minimum value}\}$$

→ It gives sense of overall spread of data, but it doesnt gives sense of dispersion about central values as the range is affected by outliers.

Inter Quartile Range (IQR) :

→ IQR is also called as mid spread / H-spread, that is a measure of statistical dispersion. It is the difference between 75th percentile and 25th percentile of a dataset.

$$IQR = Q_3 - Q_1$$

$$\text{Lower fence} = Q_1 - 1.5\,(IQR)$$

$$\text{Upper fence} = Q_3 + 1.5\,(IQR)$$

→ Theoritically the datapoints below lower fence and above upper fence are considered to be the outliers.

Variance:

variance is the measure of degree of spread of datapoints from the mean. It is denoted by σ^2. smaller the variance closes the individual data points are from the mean.

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

variance is expressed in squared units.

Example: ds = 102, 128, 131, 98, 140, 93, 110

$$\mu = \left\{ \frac{102 + 128 + 131 + 98 + 140 + 93 + 110}{7} \right\}$$

$$\mu = 114.57 \quad , \quad n = 7$$

$$\{\text{sum of squares}\} = (102 - 114.57)^2 + (128 - 114.57)^2 + (131 - 114.57)^2 + (98 - 114.57)^2 + (140 - 114.57)^2 + (93 - 114.57)^2 + (110 - 114.57)^2$$

$$SS = 2015.714$$

$$\{\text{variance for population}\} = SS/N = \frac{2015.714}{7} = 287.96$$

$$\{\text{variance for sample}\} = SS/n-1 = \frac{2015.714}{7-1} = 335.95$$

Population : Entire group of datapoints that is used to draw conclusion.

sample : A smaller (but hopefully representative) collection of unit from population used to analyse truth about population.

Standard deviation:

When the mean of 2 dataset is same we use variance & SD to find the best dataset.

⇒ It is the measure of average deviation of observation / datapoints from the mean. standard deviation converts variance into some meaningful value.

⇒ standard deviation, is the square root of the variance, denoted by σ.

standard deviation is expressed in same units as the provided data.

From previous example:

{standard deviation of population} ⇒ $\sqrt{287.96}$ ⇒ 16.97

{standard deviation of sample} ⇒ $\sqrt{335.95}$ ⇒ 18.328

similarity and differences between variance and standard deviation?

⇒ Both are used to observe the spread of datapoints.

⇒ variance is expressed in squared units but SD is expressed in actual units.

⇒ The calculation of variance uses squares to prevent difference above the mean to cancel those below.

⇒ SD helps in identifying datapoint within 1 SD

statistics

- Descriptive statistics
  - Measure of central tendency
    - mean.
    - median.
    - mode.
  - Measure of dispersion
    - Range.
    - IQR.
    - variance.
    - standard deviation.
  - symmetry
    - skewness
- Inferential statistics

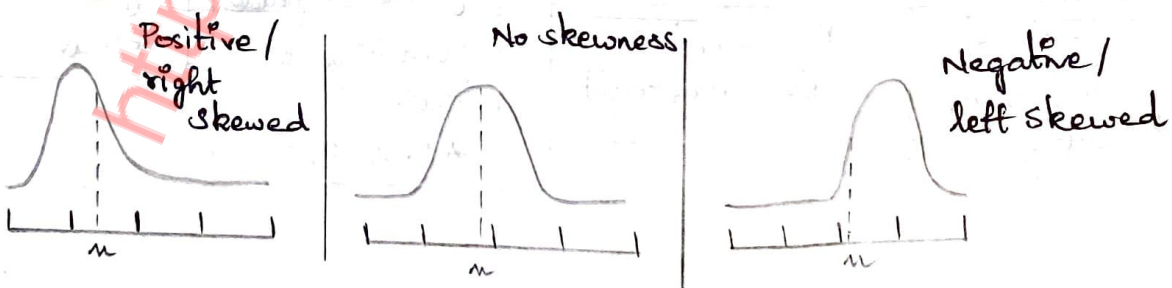
Descriptive statistics :

⇒ The descriptive statistics is used to characterizise and draw insights about the overall behaviour of the data.

⇒ This is used in all stages of exploratory data analysis (EDA).

symmetry :

Data is said to be skewed if it is not evenly distributed about the mean.



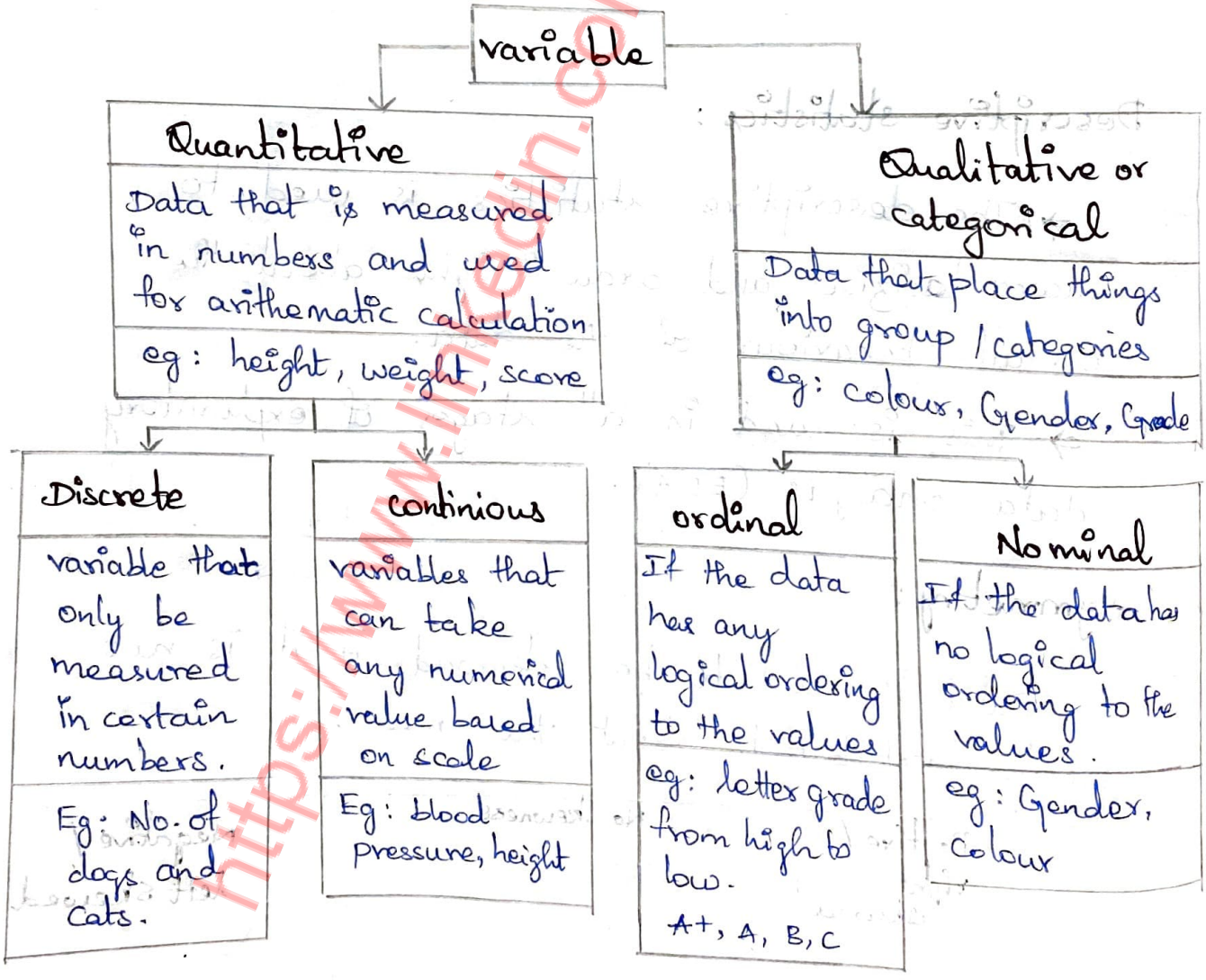
Inferential statistics:

→ This is used to infer information about the population based on what we know from the sample data

→ Inferential statistics always deals with the sample data.

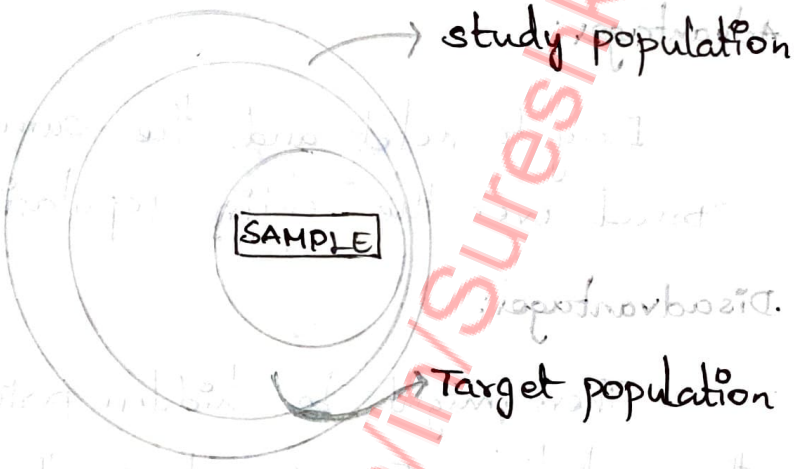
Variable:

In statistics what we examine is a variable, they can be measurable, countable and categorized.



Sampling:

Sample is a "smaller" (but hopefully a representative) collection of units from a population used to determine truths about the population.



Type of sampling:

* Simple Random sampling.
* Systematic sampling.
* Stratified sampling.
* Cluster sampling.
* Convenience / haphazard sampling.

Simple Random sampling:

In simple random sampling each subset of the frame are given equal probability of selection. This is the most simple and less time consuming sampling type. It is widely used and cost effective sampling method.

disadvantages - sample may be biased and there are chances that smaller subgroup from popn may not present in sample.

Systematic sampling:

Systematic sampling involves random start. Then proceeds with the selection of 'k'th element from the population. Example k = 3, it pics all the third datapoint from the population / sample

Advantages:

Easy to selet and the sample is evenly spread over the entire population.

Disadvantages:

There might be hidden periodicity in the selection. Example k = 3, then what if every 3rd datapoint selected in female / male.

Stratified sampling:

In stratified sampling we create strata based on the features. collection of data with similar kind of characters is called as strata.

Example:

Gender	Occupation	State
Male	Private	TN
Female	Govt	KL
Female	Govt	TN
Male	Govt	KL
Female	Private	KL
Male	Private	TN

Male ★		Female ★	
Govt ★	Private ★	Govt	Private
KL	TN	KL	KL
TN	KL	TN	TN

★ - strata

* Now within each strata simple random sampling or systematic sampling can be done.
* When we create strata and if we do sampling from ~~the~~ each strata, the sample created will be unbiased.

Disadvantages:

* It is the most time consuming sampling method and space consuming.
* By making more strata, we complicate the design of sampling.
* stratified sampling requires a larger study population than other sampling method.

When ever in interview sampling Question is asked we have to explain stratified sampling and then random & systematic. Note

Cluster sampling:

Cluster sampling is a two stage sampling. First the sample of area is chosen, second stage a sample of respondents within the area is choosen. Its a small representation of sample.

Advantages:

cost effective and most feasible type of sampling method.

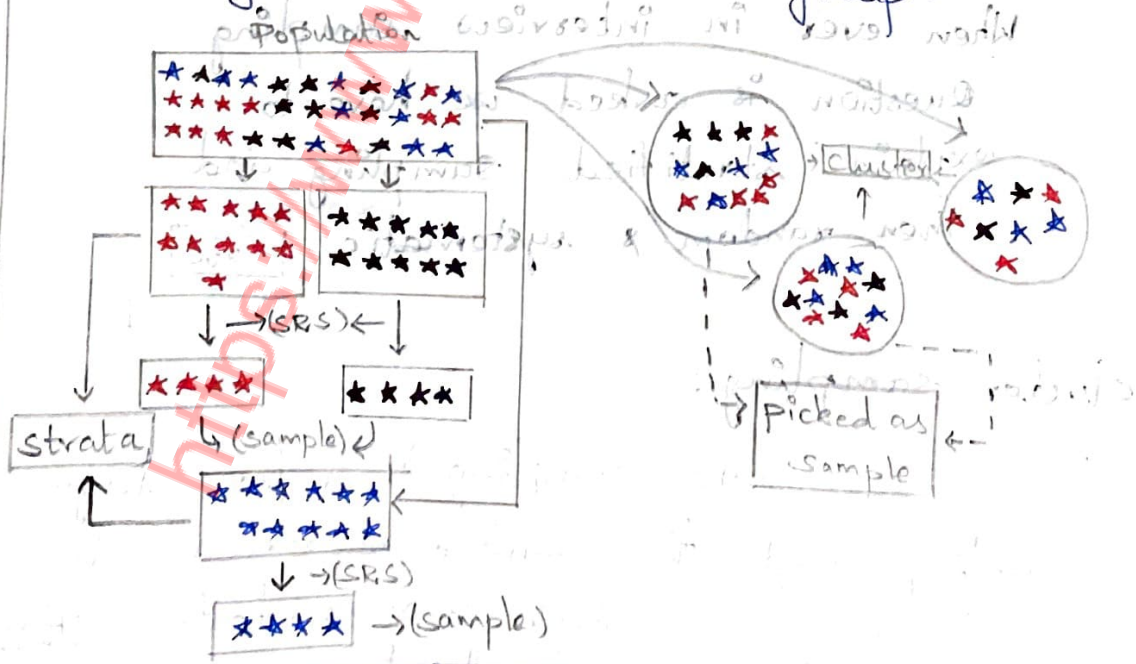
Disadvantages:

This method is prone to biases and has chances of high sampling errors.

What is the difference between stratified sampling and cluster sampling ?

⇒ In stratified sampling technique, the sample is created out of random selection of elements from all the strata. In cluster sampling all units of randomly selected clusters forms a sample.

→ stratified sampling is homogeneous within the ~~within~~ the group and hetrogeneous between the group. cluster sampling is hetrogeneous within the group and homogeneous between the group.

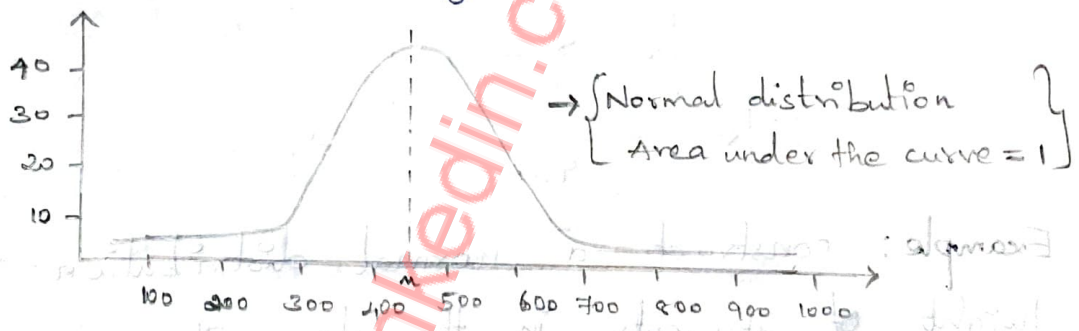


convenience sampling:

The sample being drawned from the part of population which is close to hand, readily available and convinient. This sample is not a representative of entire population and has a high sampling error and bias.

Normal distribution:

Normal distribution is a type of density curve, where the total area under the curve is always equal to 1. By its structure it is also called as bell curve and it has no skewness in general.



⇒ We say that the data is "normally distributed" when,

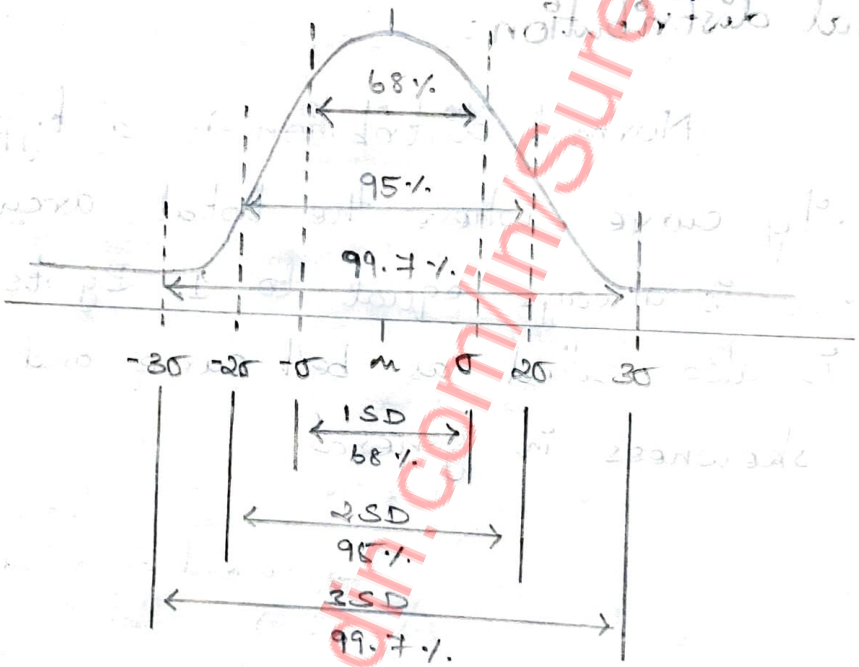
i) mean = media = mode.
ii) symmetric about the centre.
iii) It follows 68 - 95 - 99.7 % rule.

→ Mean characterizes the position of normal distribution, SD characterizes the spread of a normal distribution.

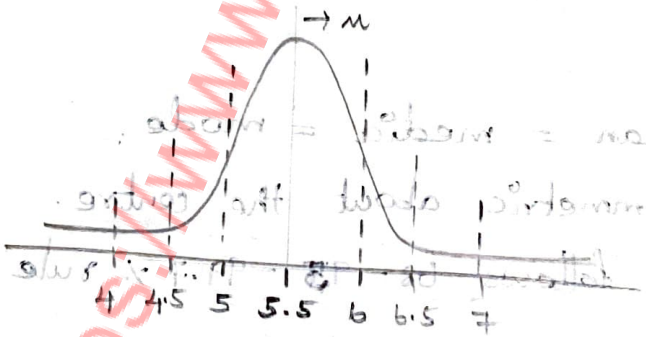
larger σ = More spread of distribution = flatter curve
smaller σ = Less spread of distribution = taller curve.

68 - 95 - 99.7 %. Rule:

In any normal distribution 68 % of data is distributed or lies between one standard deviation $-\sigma$ to σ. 95 % of data lies between -2σ to 2σ and 99.7 % of data lies between -3σ to 3σ. (Data points after 95 % or two standard deviation is considered as outliers).



Example: construct a normal distribution of height of students in the class. The mean height is 5.5 and standard deviation is 0.5

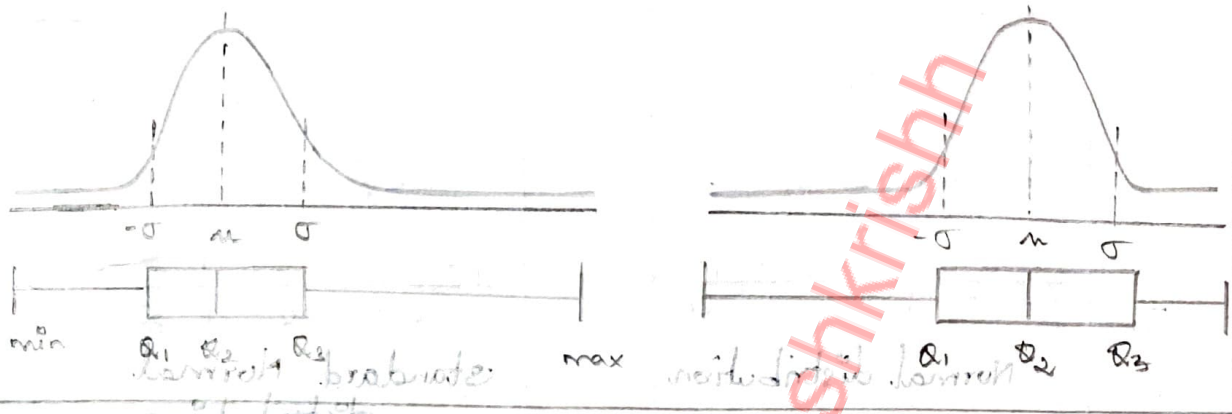


68 % ($\mu+\sigma$ and $\mu-\sigma$) = 5 to 6 (one SD)

95 % ($\mu+2\sigma$ and $\mu-2\sigma$) = 4.5 to 6.5 (two SD)

99.7 % ($\mu+3\sigma$ and $\mu-3\sigma$) = 4 to 7 (three SD)

Box plot may give a general idea of how a distribution of data looks like,



Standard Normal distribution:

* The process of converting normal distribution to standard normal distribution is called as standardisation. The standard normal distribution has a mean of 0 and standard deviation of 1 always.

* To convert normal distribution to standard normal distribution we use z-score formula.

z-score:

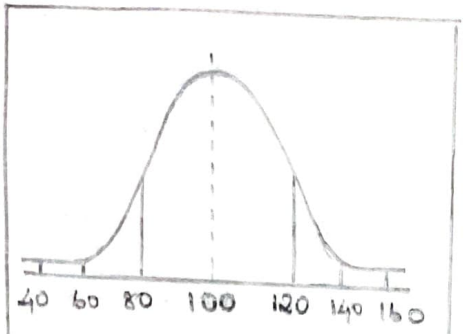
$$z = \frac{x - \mu}{\sigma/\sqrt{n}} \Rightarrow \text{sample.}$$

$$z = \frac{x - \mu}{\sigma} \rightarrow \text{population.}$$

* Corresponding z-score will help us to find the area associated to the left of the curve using z-score table.

* Standard normal distribution makes the data unitless.

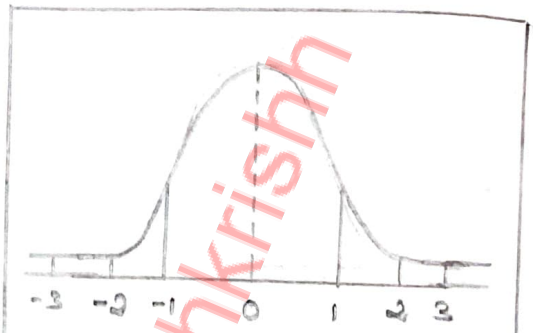
Difference between normal distribution and standard normal distribution?



Normal distribution

μ = some number

σ = some number



Standard Normal distribution

μ = always 0

σ = always 1

Example: consider normal distribution with the mean of 60 and SD of 10. Find the proportion of student scoring less than 49.

$$x = 49, \quad \mu = 60, \quad \sigma = 10$$

$$z = \frac{x - \mu}{\sigma} = \frac{49 - 60}{10} = -1.1$$

From z score table $z(-1.1) = 0.1357 = 13.57\%$

Example: For a group of students mean height is 5.5 feet and SD is 0.5 feet. Find the proportion of students between 5.81 and 6.3?

$$x_1 = 5.81, \quad x_2 = 6.3, \quad \mu = 5.5, \quad \sigma = 0.5$$

$$z_1 = \frac{5.81 - 5.5}{0.5} \qquad z_2 = \frac{6.3 - 5.5}{0.5}$$

$$z_1 = 0.62 \qquad z_2 = 1.6$$

$$P(z_1 < 0.62) = 0.732 \qquad P(z_2 < 1.6) = 0.945$$

$$P(0.732 < z < 0.945) = 0.945 - 0.732 = 0.213 = 21.3\%$$

T- distribution : (used when SD is unknown)

* T-distribution is also called as students t-distribution invented by william Gosset. This distribution is used when the sample size is less than 30.

* This is identical to normal distribution. similar to z-score in normal distribution we have t-score for t-distribution.

* Interval estimate of t-distribution can be calculated by the aid of degree of freedom.

$$t = \frac{x - u}{s}, \quad dof = n - 1$$

s - estimated SD
n - sample size

* T-score can be calculated from t-table using the degree of freedom.

When will t-distribution behave like a z-distribution ?

When we keep increasing dof (which will increase sample size), after a threshold t-distribution will behave like z-distribution.

How to choose the distribution based on the sample size ?

When,

$n > 30 \Rightarrow$ z - distribution
$n < 30 \Rightarrow$ t - distribution
$n = 30 \Rightarrow$ either of distribution

How much is a adequate sample size?

⇒ In most of the cases adequate sample size is 30.

⇒ When the distribution is skewed or if it contains outliers sample size should be 50 or more.

⇒ If population is not normally distributed but is roughly symmetric sample size of 15 is sufficient.

⇒ If the distribution is approximately normal a sample size less than 15 can be used.

Hypothesis Testing:

Hypothesis is an assumption about the population parameter. It is a procedure to accept or reject a hypothesis. A random sample from the population is used for testing.

Types of hypothesis:

* Null hypothesis (H_0): It is a tentative and true assumption about the parameter.
* Alternative hypothesis (H_a): It is opposite of what is stated in null hypothesis.

Example: Label on soft drinks says it contains 67.6 fluid ounces.

H_0 : The label is correct $\mu \geq 67.6$ ounces

H_1 : The label is incorrect $\mu < 67.6$ ounces.

Hypothesis decision errors:

Two types of error can occur from the hypothesis testing. They are:

* Type-I error: This error occurs when we reject the null hypothesis when it true. Probability of commiting this error is called a significance level. It is denoted by α.

* Type-II error: This error occur when we accept the null hypothesis when its false. It is denoted by β.

		ACTUAL	
		H_0 is true	H_0 is False
PREDICTED	Accept H_0	True positive correct decision	False Negative Type-II error (β)
	Reject H_0	False positive Type-I error (α)	True negative correct decision

* Type-I error ⇒ False positive (α)

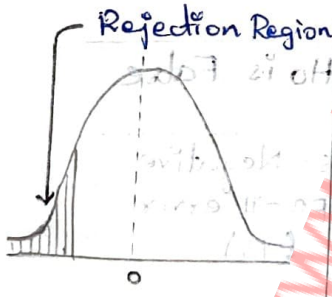
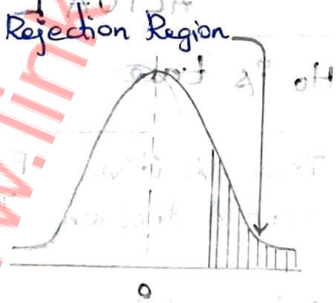
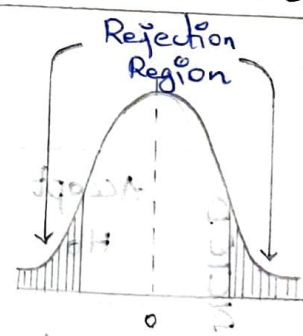
* Type-II error ⇒ False Negative (β)

one tailed and two tailed test:

→ **one tailed test:** When the region of rejection is on only one side of the distribution it is one tailed test. It is further classified into,

* **Left tailed :** When the alternative hypothesis states that mean is less than a value.
* **Right tailed :** When the alternative hypothesis states that the mean is greater than a value.

→ **two tailed test :** When the region of rejection is on both sides of the distribution then it is two tailed test. This happens when the null hypothesis value is equal to the mean.

LEFT - TAILED	RIGHT - TAILED	ON BOTH SIDES
		
$H_0 : \mu \geq \mu_0$ $H_a : \mu < \mu_0$	$H_0 : \mu \leq \mu_0$ $H_a : \mu > \mu_0$	$H_0 : \mu = \mu_0$ $H_a : \mu \neq \mu_0$
ONE - TAILED		TWO - TAILED

Statistical Inference:

The process of guessing the truth about the population from a sample.

hypothesis decision rules:

Null hypothesis can be rejected by two ways - i) reference to p-value, ii) reference to the region of acceptance.

* p-value ⇒ if p-value $\leq \alpha$, we reject H_0.
  if p-value $> \alpha$, we accept H_0.
* Region of acceptance ⇒ if the test statistics falls with the region of acceptance, then we accept the H_0, else we reject the H_0.

steps of Hypothesis testing:

step-1: Develope the null and Alternative hypothesis

step-2: specify the level of significance α.

step-3: collect sample data and compute the value of test statistics. ⇒ critical value approach.

step-4: use the value of test statistic to compute the p-value ⇒ p-value approach.

step-5: Reject H_0, if p-value $\leq \alpha$.
(or) Reject H_0, if $z_\alpha \leq z$

What is p-value ?

p-value is used in hypothesis testing, that helps to accept or reject the null hypothesis. p-value is the evidence against the null hypothesis. smaller the p-value, stronger the evidence to reject H_0.

Problem - 1 :

The response time for a random sample of 40 medical emergencies were tabulated. The sample mean is 13.25 min. The population standard deviation is 3.2 mins. Perform a hypothesis test with 0.05 level of significance to determine whether the service goal of 12 minutes or less is being achieved.

Solution :

Given :

$$n = 40, \quad \bar{x} = 13.25 \text{ mins}, \quad \sigma = 3.2 \text{ mins}$$

$$\alpha = 0.05, \quad \mu = 12 \text{ mins}.$$

Step-1: Develope null and alternate hypothesis.

$H_0 : \mu \leq 12$ mins

$H_a : \mu > 12$ mins $\Rightarrow$ Right-tailed test

Step-2: Level of significance : (α)

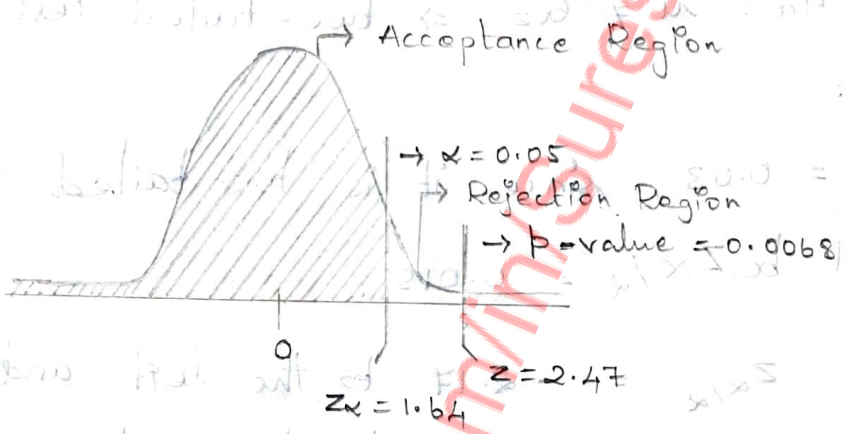
$$\alpha = 0.05, \quad z_{0.05} = 1.64$$

step-3 : computing the value of test statistics:

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

$$= \frac{13.25 - 12}{3.2/\sqrt{40}} \Rightarrow \frac{1.25}{0.5059} = 2.47$$

$Z = 2.47$, corresponding p-value $= 0.0068$ [step-4]



step-5:

There is sufficient statistical evidence, that service goal of 12 minutes or less cant be acheived. so we reject null hypotheris.

Problem -2:

sample of 30 toothpaste tubes, provides a sample mean of 6.1 oz. The population standard deviation is 0.2 oz. Perform a hypotheris test with 0.03 level of significance, to determine the filling process should continue or to be stopped for correction. The mean filling weight of the population of the toothpaste tubes is 6 oz. otherwise the process will be adjusted.

Solution:

Given:

$$n = 30 \;,\; \bar{x} = 6.1oz \;,\; \sigma = 0.2oz$$

$$\alpha = 0.03 \;,\; \mu = 6oz$$

step-1:

$H_0 : \mu = 6oz$

$H_a : \mu \neq 6oz \Rightarrow$ two-tailed test.

step-2:

$\alpha = 0.03$, since it is two tailed test $Z\alpha$ will be $Z\alpha/2 = 0.015$.

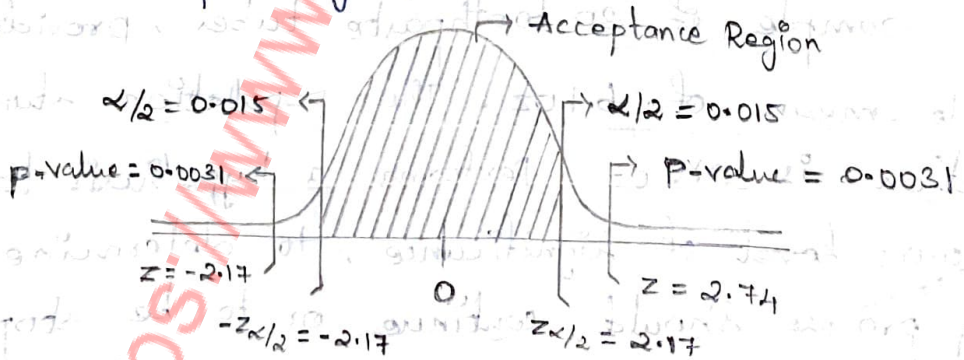
$z_{\alpha/2} = -2.17$ to the left and 2.17 to the right.

step-3:

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{6.1 - 6}{0.2/\sqrt{30}} = 2.74$$

step-4:

corresponding p-value for z score = 0.0031



step-5:

since the p-value is less than α and the critical z-value falls in rejection region we reject null hypothesis.

Problem -3 :

A highway patrol samples vehicle speed at various location on particular roadway. The sample vehicle speed used to test the hypothesis $H_0 : \mu \leq 65$ mph.

The location where H_0 is rejected are deemed the best location for radar traps. At location 'F', a sample of 64 vehicles shows a mean speed of 66.2 mph with sample standard deviation of 4.2 mph. use $\alpha = 0.05$ to test the hypothesis.

Solution :

Given :

$$n = 64, \quad \bar{x} = 66.2 \text{ mph} \quad s = 4.2 \text{ mph}$$

$$\mu = 65 \text{ mph}, \quad \alpha = 0.05$$

step - 1 : $H_0 : \mu \leq 65$ mph

$H_a : \mu > 65$ mph $\Rightarrow$ right tailed test.

step - 2 :

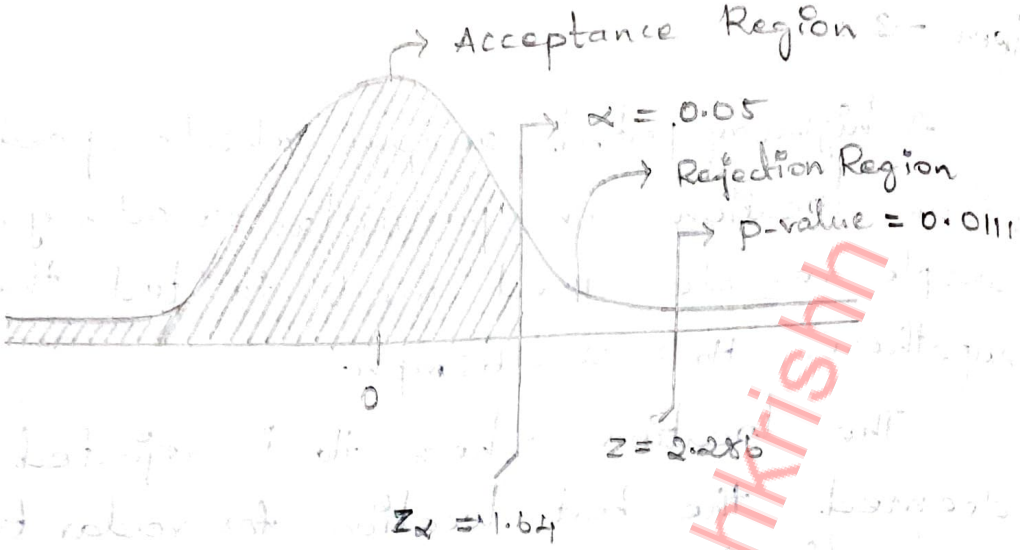
$$\alpha = 0.05, \quad z_{0.05} = 1.64$$

step - 3 :

$$z = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{66.2 - 65}{4.2/\sqrt{64}} = 2.286$$

step - 4 :

corresponding p-value for z score is calculated as 0.0111



Step-5:

By statistical evidence we can reject null hypothesis and we are atleast 95% confidence that the mean speed of vehicle at location 'F' is greater than 65 mph.

Test for population proportion:

$$z = \frac{\bar{P} - P_0}{\sigma_p}$$

Where,

$$\sigma_p = \sqrt{\frac{P_0(1-P_0)}{n}}$$

Problem-4:

For a christmas and newyear week, the national council estimated that 500 people would be killed & 25,000 will be injured on road accident. NSC claimed 50% of accidents will be caused by drunk driving. sample of 120 accidents shows 67 caused by drunk driving. Test NSC claim with $\alpha = 0.05$

solution:

Given:

$n_1 = 120$, $n_2 = 67$, $\alpha = 0.05$, $p_0 = 0.5$

step-1:

$H_0 : p = 0.5$

$H_a : p \neq 0.5 \Rightarrow$ two tailed test.

step-2:

$\alpha = 0.05$, $Z_{\alpha/2} = 1.96$ ($\because$ two-tailed)

step-3:

$$\sigma_p = \sqrt{\frac{p_0(1-p_0)}{n}} = \sqrt{\frac{0.5(1-0.5)}{120}}$$

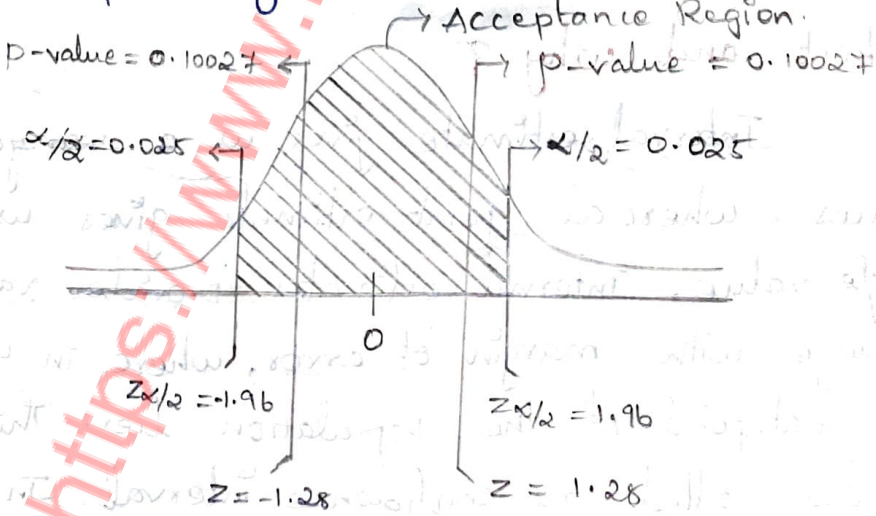
$$\sigma_p = 0.0456$$

$$Z = \frac{\bar{P} - p_0}{\sigma_p} \Rightarrow \frac{(67/120) - 0.5}{0.0456}$$

$$Z = 1.28$$

step-4:

Corresponding p-value for z score is 0.10027



step-5:

By statistical evidence 50% of accidents are caused by drunk driving, we accept H_0.

Point estimator:

It uses the sample data to calculate a single value which is to serve as a best guess or best estimate for an unknown population parameter.

Interval estimator:

It uses the sample data to calculate an interval (a range) of possible values of an unknown population parameter.

Designation	Population	Sample
Mean	μ	$\bar{x}$
standard deviation	σ	s

Point estimate or interval estimate which is best and why ?

Interval estimate gives us a range of values, whereas point estimate gives us a single value. Interval estimates provides values in range with margin of error, where in which the datapoint of the population lies. This is also called as confidence interval. In this way interval estimate provides more information and prefered when making inferences.

Point estimate = $\bar{x}$

Interval estimate = $\bar{x} \pm$ margin of error

Problem – 1:

For sample size $n = 36$, the sample mean income is 41,100. The population is not highly skewed. The population standard deviation is estimated to be 4500 which has $\alpha = 0.05$. Consider interval estimate for one tailed test.

Solution:

Given:

$n = 36$, $\bar{x} = 41,100$, $\sigma = 4500$, $\alpha = 0.05$

$$\left\{\begin{matrix}\text{Interval} \\ \text{estimate}\end{matrix}\right\} = \bar{x} \pm (\alpha)\,(\sigma/\sqrt{n})$$

$$\left\{\begin{matrix}\text{Interval} \\ \text{estimate}\end{matrix}\right\} = 41,100 \pm (\alpha_{0.05})\left(4500/\sqrt{36}\right)$$

$$= 41,100 \pm (1.64)(750)$$

$$= 39,870 \text{ to } 42330$$

Problem – 2:

A sample of 16 apparatments has a sample mean of 750 $ /month, and sample standard deviation of 55 $. calculate interval estimate with $\alpha = 0.05$, considering two tail test.

solution:

Given: $n = 16$ (we use t-distribution ie $n < 30$)

$\bar{x} = 750$, $s = 55$, $\alpha = 0.05$

dof $= n - 1$

$= 16 - 1$

$= 15$

Interval estimate $= \bar{x} \pm (t_{\alpha/2})(\sigma/\sqrt{n})$

$= 750 \pm (t_{0.025})(55/\sqrt{16})$

$= 750 \pm (2.1314)(13.75)$

$= 720.693$ to 779.306.

Sample size for interval estimate:

$$\{\text{Margin of error, } E\} = (Z_\alpha)(\sigma/\sqrt{n})$$

$$\{\text{sample size } n\} = \frac{(Z_\alpha)^2(\sigma^2)}{E^2}$$

Problem-3:

For $\alpha = 0.05$ and margin of error is 500 and has a population standard deviation of 4500. Find the required sample size.

Solution:

$$\{\text{sample size}\} = \frac{(\alpha_{0.05})^2 (4500)^2}{(500)^2}$$

$$= \frac{(1.64)^2 (4500)^2}{(500)^2}$$

$= 217.85 \simeq 220$ (or) 200

Interval estimate for population proportion:

$$\left\{\begin{array}{l}\text{Interval estimate}\\ \text{for population}\\ \text{proportion}\end{array}\right\} = \bar{p} \pm Z_\alpha \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

Problem-4:

For an election campaign, PSI found that 220 registered voters, out of 500 favours a particular party. Develope 95% confidence interval for population proportion.

Given:

$n = 500$, $\bar{p} = 220/500 = 0.44$

$\alpha = 1 - \text{confidence interval}$

$= 1 - 0.95$

$= 0.05$

$$\left\{\begin{array}{l}\text{Interval}\\ \text{estimate}\end{array}\right\} = \bar{p} \pm \sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \times Z_\alpha$$

$$= 0.44 \pm \sqrt{\frac{0.44(1-0.44)}{500}} \times (Z_{0.05})$$

$$= 0.44 \pm (0.0222)(1.64)$$

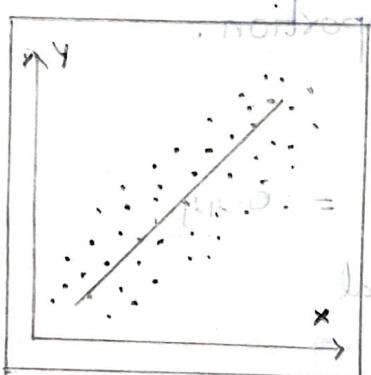
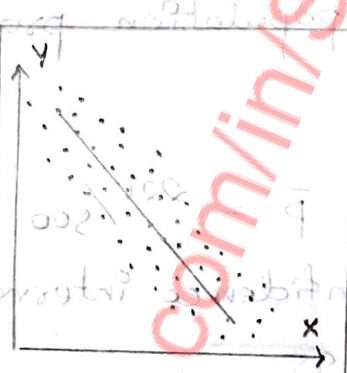
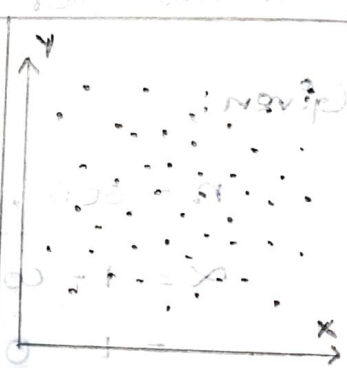
$$= 0.404 \text{ to } 0.476$$

Correlation:

Correlation tells the direction and strength of linear relationship shared by two quantitative variable. It is expressed using the scatter plot.

> Degree of linear Relationship between two variables.

* Direction:

		
Positive Correlation	Negative Correlation	No correlation
x increases, Y also increases	x increases, Y decreases	No linear pattern for both x and Y

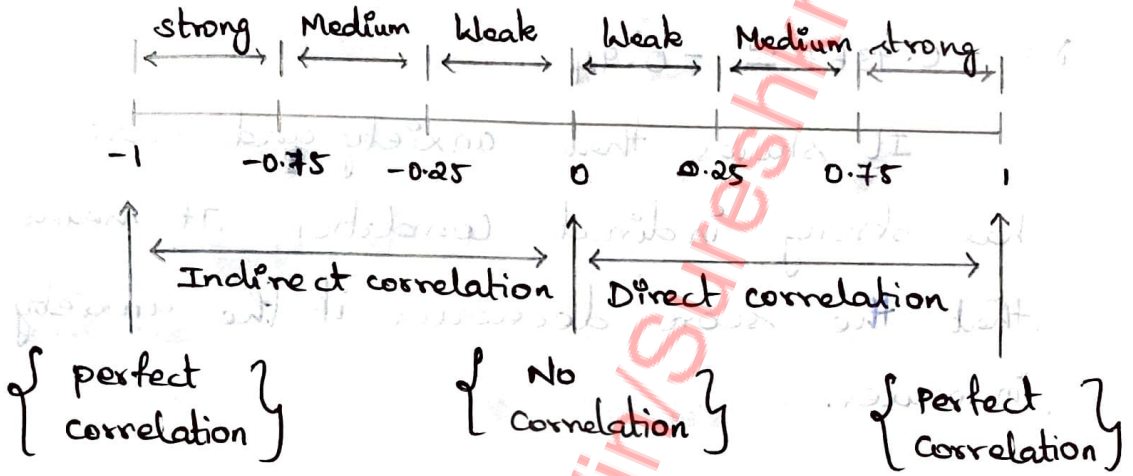
* Correlation coefficient:

It is the statistic showing the degree of ~~freedom~~ relationship between two variables. Pearson's correlation coefficient is used to measure the nature and strength of two quantitative variables.

⇒ Sign of 'r' denotes the nature of association

⇒ value of 'r' denotes the strength of association

r has values ranging between -1 and 1. As the value of r moves towards the extreme point in either sides the strength of the correlation increases.



simple correlation coefficient equation: (r)

$$r = \frac{\Sigma_{xy} - \frac{\Sigma x \, \Sigma y}{n}}{\sqrt{\left\{\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right\}\left\{\Sigma y^2 - \frac{(\Sigma y)^2}{n}\right\}}}$$

Example: Relationship between Anxiety & test score

Anxiety (x)	scores (y)	(x^2)	(y^2)	(xy)
10	2	100	4	20
8	3	64	9	24
2	9	4	81	18
1	7	1	49	7
5	6	25	36	30
6	5	36	25	30
$\Sigma x = 32$	$\Sigma y = 32$	$\Sigma x^2 = 230$	$\Sigma y^2 = 204$	$\Sigma xy = 129$

$$r = \left\{(129) - \frac{(32 \times 32)}{6}\right\} \Big/ \sqrt{\left\{230 - \frac{(32)^2}{6}\right\}\left\{204 - \frac{(32)^2}{6}\right\}}$$

$$r = \left\{(-41.66) / 44.4722\right\}$$

$$r = -0.9367 \simeq -0.94$$

It shows that anxiety and score has strong indirect correlation. It means that the score decreases if the anxiety increases.

ANOVA :

ANOVA stands for Analysis of variance. Hypothesis of one way ANOVA includes both null and alternative hypothesis.

Why do we use ANOVA ?

t and z - test can handle only one or two sample test. For the problems involving more than two samples we use ANOVA test.

Hypothesis of one way ANOVA :

$H_0 : \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_n$

H_a : Not all population mean are same atleast one of them is different

eg :- $H_1 : \mu_1 = \mu_2 \neq \mu_3 \neq \mu_4 = \mu_5$

The ratio of variance follows an F- distribution

What is the use of F- distribution ?

F- distribution is used to find if the variance among the group are same or not.

$$\frac{\sigma^2_{between}}{\sigma^2_{within}} \sim F_{n,m}$$

F-test tests the hypothesis that two variances are equal. Variance is said to be equal when the F-test value is close to 1.

$$H_0 : \sigma^2_{between} = \sigma^2_{within}$$
$$H_a : \sigma^2_{between} \neq \sigma^2_{within}$$

ANOVA table:

source of variation	dof	sum of squares	Mean sum of squares	F-stats	p-value
Between (k- group)	$k-1$	(SSB) sum of squared deviation of group mean from grand mean	$\frac{SSB}{k-1}$	$\frac{(SSB)/k-1}{(SSW)/nk-1}$	From chart
Within (n-individual per group)	$nk-k$	(SSW) sum of squared deviation of observation from their group mean	$\frac{SSW}{nk-k}$		
Total variation	$nk-1$	TSS (or) SST	$TSS = SSB + SSW$		

coefficient of determination:

The amount of variation in the outcome variable (dependent) that is explained by the predictor (independent). Denoted by R^2

$$R^2 = \frac{SSB}{SSW + SSB} \quad \text{or} \quad \frac{SSB}{TSS}$$

How much variation in the data, that the model is able to understand.

$R^2 = 20\% \Rightarrow$ model can understand only 20% of the data variation.

$R^2 = 80\% \Rightarrow$ model can understand only 80% of the data variation.

In regression based machine learning models R^2 is considered as accuracy.

ANOVA Example:

Treatment-1	Treatment-2	Treatment-3	Treatment-4
67	50	48	47
42	52	49	67
67	43	50	54
56	67	55	67
62	67	56	68
64	59	61	65
59	67	61	65
72	64	60	56
71	63	59	60
60	65	64	65

Step-1:

Calculate sum of squares between (SSB) the group

Mean for group-1 : 62.0

Mean for group-2 : 59.7

Mean for group-3 : 56.3

Mean for group-4 : 61.4

Grand mean $= (62 + 59.7 + 56.3 + 61.4) / 4$

$= 59.85$

$SSB = [(62 - 59.85)^2 + (59.7 - 59.85)^2 + (56.3 - 59.85)^2 + (61.4 - 59.85)^2] \times n$

$SSB = 19.65 \times 10 = 196.5$

Step-2:

Calculate sum of square within (SSW) the groups

$SSW = (67 - 62)^2 + (42 - 62)^2 + \ldots + (50 - 59.7)^2 + \ldots$

$+ \ldots$ (sum of 40 squared deviation)

$SSW = 2060.5$

Step-3:

$$\left\{\begin{array}{c}\text{Mean sum of} \\ \text{Square between} \\ \text{the group}\end{array}\right\} = \frac{SSB}{k-1} = \frac{196.5}{4-1} = 65.5$$

$$\left\{\begin{array}{c}\text{Mean sum of} \\ \text{square within} \\ \text{the group}\end{array}\right\} = \frac{SSW}{nk-k} = \frac{2060.5}{(10\times4)-4} = 57.2$$

step-4:

Calculating F-statistics score:

$$F_{n,m} \sim \frac{\sigma^2_{between}}{\sigma^2_{within}}$$

$$= \frac{65.5}{57.2}$$

F-statistics = 1.14.

The corresponding p-value for F-statistics 1.14 and degree of freedom 3 and 36 is 0.344.

step-5: ANOVA table

Source of variation	dof	sum of squares	Mean sum of squares	F-statistic	P-value
Between (k-groups)	3	196.5	65.5	1.14	0.344
Within (n-individual per group)	36	2060.5	57.2	-	-
Total variation	39	2257	-	-	-

$$R^2 = \frac{196.5}{2257} = 0.0870 \times 100 = 8.70\%$$

Since F-statistics value is close to 1 and p-value > α, we accept null hypothesis.

⟷ Probability ⟷

Numerical description of how likely an event is to occur, or how likely it is that a proportion is true. (equally likely)

* Lowest possible value of probability is 0 and this means that event wont occur
* Highest possible value of probability is 1 and this means that is a sure event.

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

Probability is a study of chances.

Random variable:

Random variable x is a variable whose value depends on the outcome of an random phenomenon.

Eg: Getting a value (random variable) after throwing a dice. The probability a value occuring from dice is 1/6.

Random variable is classified into:

* Discrete random variable.
* Continuous random variable.

Example:

⇒ Discrete - dead/alive, dice, count, 1/0, etc.,

⇒ Continuous - BP, weight, real number between 1-6

Probability Function : $p(x)$

Probability Function maps the value of x against their respective probabilities of occurances $p(x)$.

Example:

$x = \{$ Number of heads after flipping a fair coin three times. $\}$

Possibility : $\{$ HHH, HHT, HTH, THH, THT, TTH, HTT, TTT $\}$

Probability function:

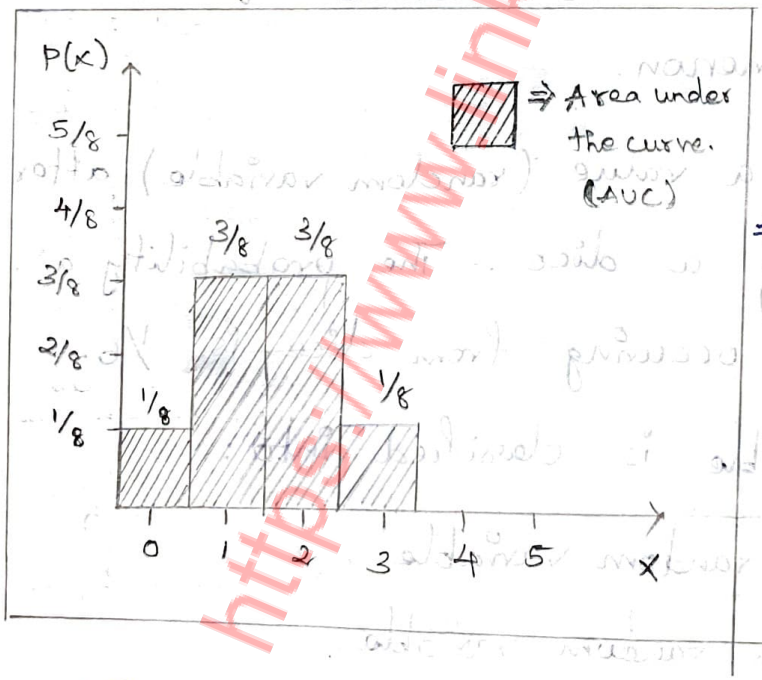
$P(x=0) = 1/8$

$P(x=1) = 3/8$

$P(x=2) = 3/8$

$P(x=3) = 1/8$

$\Rightarrow$ [Area under the curve]



$\Rightarrow$ { probability distribution of variable x. }

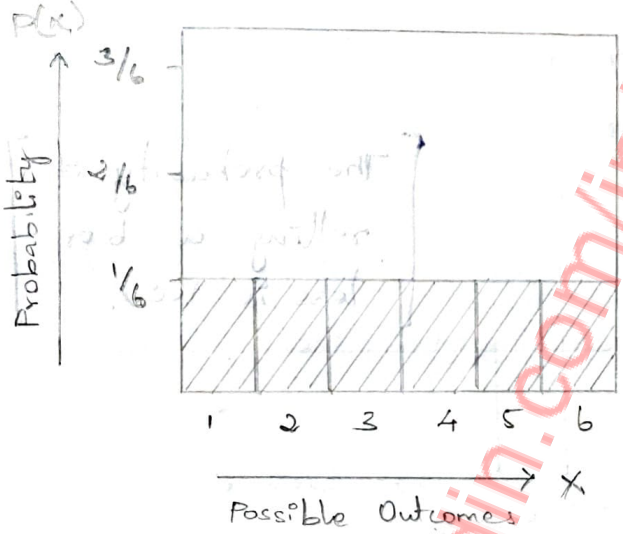
$\rightarrow$ The Area under the probability function is always equal to 1.

Probability mass function : (PMF)

It is a function that gives probabilities that a discrete random variables is exactly equal to some values. It is also called as density of discrete function or discrete density function.

Example:

⇒ Rolling a dice and probability of possible outcomes



[Here the possibility of getting the outcome is equally likely.]

* Probability mass Function (PMF) is generally represented by histograms

* The sum of probability of outcome is 1.

PMF of discrete random variable 'x' assigns probabilities to the possible value of a random variable. The process of assigning specific value to discrete random variable (DRV) is PMF

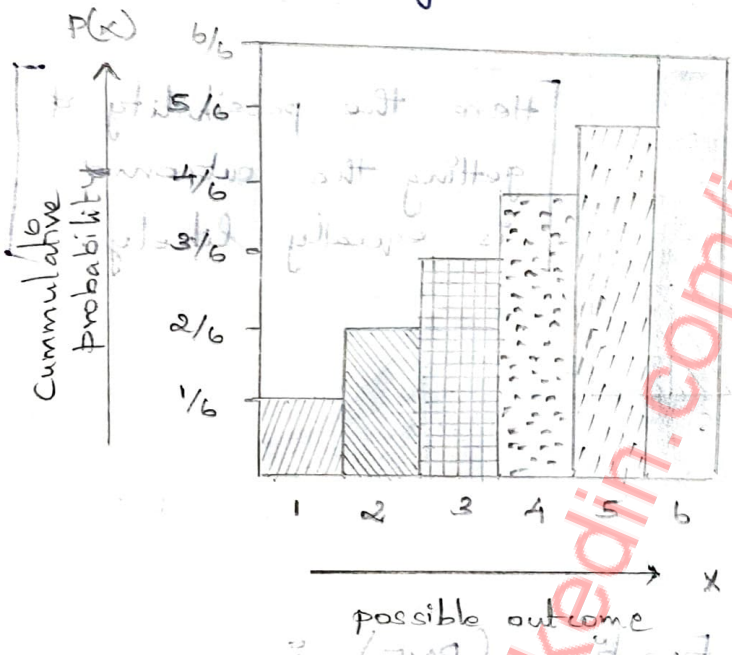
Cumulative distribution Function (CDF):

cumulative distribution of a random variable x is a real value / real Number.

CDF is a uniform distribution adding up the previous value as it goes along.

Example:

Rolling a dice and representing the outcome by cumulative distribution function.



[The probability of rolling a 6 or less is 100%.]

Event, Non-event, sample space:

Event: An event is a set of outcome of an experiment to which probability is assigned.

{sample space}: set of all possible outcomes of an experiment.

{Event space}: It is a set of event that contains all set of outcomes and subset of sample space.

{Favourable event}: collection of all events which are of interest to the problem we are working
↳ Always problem centric.

Types of probability:

1. **A priori classical probability:** The probability of an event is based on prior knowledge of the process involved.

   eg: throwing dice, picking a card from deck, tossing a coin.

2. **Emprical classical probabilites:** The probability of an event is based on observed data. This data is widely used in ML models.

   eg: credit lending, fraud detection etc.,

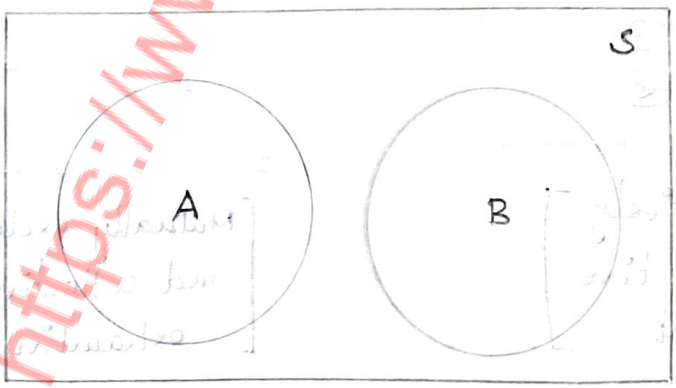
3. **subjective probability:** The probability of an event is determined by an individual, based on that person's past experience, personal opinion, and/or analysis of particular situation.

   eg: Based on anyone's experience on part.

Types of event:

⇒ **Mutually exclusive:** Events that doesnot have any common outcomes. (cannot overlap)

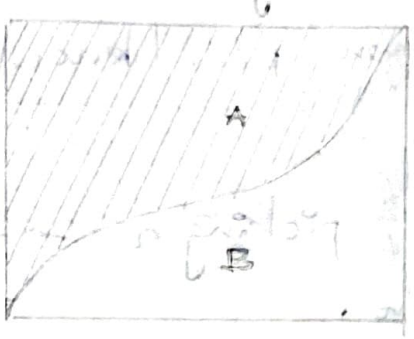
Event A:
P(At least one 5)
Event B:
P(sum < 4)



Example: getting atleast one 5 and getting sum less than 4. Use two six sided dice.

In this case A and B are disjoint and mutually exclusive. Means $P(A \cap B) = 0$

⇒ collectively exhaustive: Events that contains all outputs or outcomes in the sample space.



Exhaustive events occupy all the possible event of the sample space.

Example:

While throwing two six sided dice.

$P(\text{At least one 6}) = 11/36$

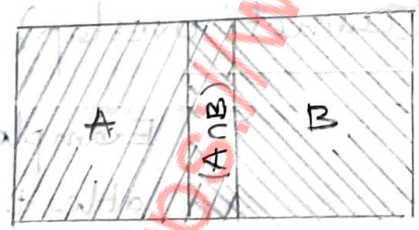
$P(\text{sum} \leq 11) = 35/36$

This is called as exhaustive event because it contains all the outcomes in the sample space. $P(A \cup B) = 1$

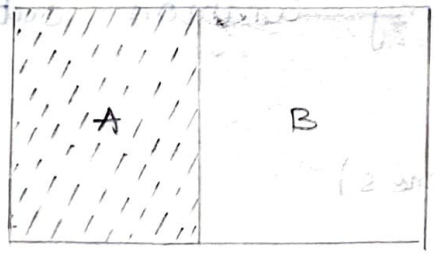
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$= 11/36 + 35/36 - 10/36$$

$$= 36/36 \Rightarrow 1$$



[collectively exhaustive event]



[Mutually exclusive and collectively exhaustive]

Joint and Marginal probability:

		Event		
		B_1	B_2	Total
Event	A_1	$P(A_1 \text{ and } B_1)$	$P(A_1 \text{ and } B_2)$	$P(A_1)$
	A_2	$P(A_2 \text{ and } B_1)$	$P(A_2 \text{ and } B_2)$	$P(A_2)$
Total		$P(B_1)$	$P(B_2)$	1

$P(A_1 \text{ and } B_1)$, $P(A_1 \text{ and } B_2)$
$P(A_2 \text{ and } B_1)$, $P(A_2 \text{ and } B_2)$ $\Rightarrow$ Joint probability.

$P(A_1)$, $P(A_2)$, $P(B_1)$, $P(B_2)$ $\Rightarrow$ Marginal probability.

conditional probability:

Conditional probability is the probability of an event given another event has occured.

* $P(A|B) = \dfrac{P(A \text{ and } B)}{P(B)}$ } Probability of A given B has occured.

* $P(B|A) = \dfrac{P(A \text{ and } B)}{P(A)}$ } Probability of B given that A has occured.

From the above equation we have

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

Example 1:

Find the probability of selecting a male or stats student from the population.

	Stats	Not stats	Total
Male	84	145	229
Female	76	134	210
Total	160	279	439

solution:

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

$$= (229 + 160 - 84)/439$$

$$= 305/439$$

$$= 69.4\ \%$$

Example 2:

Of all the cars on used car lot, 70% have air conditioning (AC) and 40% have CD player (CD). 20% of the cars have both. What is the probability that a car has CD player, give it has AC?

solution:

$$P(CD|AC) = \frac{P(CD \text{ and } AC)}{P(AC)}$$

$$\Rightarrow \frac{20}{70} \Rightarrow 0.2857$$

$$\Rightarrow 28.57\ \%$$

Statistical Independence:

⇒ Two events are said to be independent if and only if,

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

⇒ Event A and B are said to be independent when the probability of one event is not affected by the other event.

Multiplication Rule:

Multiplication rule is used when the two events takes place together. This is referred as intersection and denoted by P(A and B) or $P(A \cap B)$

⇒ For dependent event:

$$\left.\begin{matrix} P(A \text{ and } B) \\ (\text{or}) \\ P(A \cap B) \end{matrix}\right\} = P(A|B) \times P(B)$$

⇒ For independent event:

$$\left.\begin{matrix} P(A \text{ and } B) \\ \text{or} \\ P(A \cap B) \end{matrix}\right\} = P(A) \times P(B)$$

Example:

A city council has 5 democrats, 4 republic and 3 independents. Find the probability of randomly selecting a democrat followed by an independent candidate.

Solution: $P(I \text{ and } D) = P(D) \times P(I|D)$

$= 5/12 \times 3/11$

$= 0.113 \Rightarrow 11.3\%$

Bayes theorem:

It describes the probability of the event based on the prior knowledge of conditions that might be related to the events.

⇒ Theorem:

We know that:

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{P(A \cap B)}{P(B)} \quad \text{— ①}$$

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)} = \frac{P(A \cap B)}{P(A)} \quad \text{— ②}$$

From equation ① and ②, we have

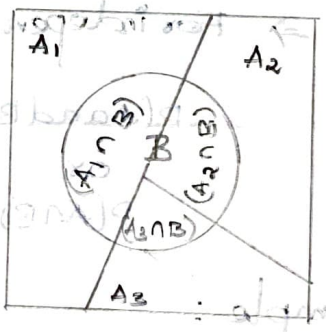
$$P(A \cap B) = P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$$

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \quad \text{— ③}$$

From the diagram,

$$P(B) = P(A_1 \cap B) + P(A_2 \cap B) + P(A_3 \cap B)$$

$$\Rightarrow \left.\begin{aligned} & P(B|A_1) \cdot P(A_1) + \\ & P(B|A_2) \cdot P(A_2) + \\ & P(B|A_3) \cdot P(A_3) \end{aligned}\right\} \quad \text{— ④}$$



Applying eqn ④ in ③, we have

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{\left.\begin{aligned} & P(B|A_1) \cdot P(A_1) + \\ & P(B|A_2) \cdot P(A_2) + \\ & P(B|A_3) \cdot P(A_3) \end{aligned}\right\}} \Rightarrow \frac{P(B|A_i) \cdot P(A_i)}{\sum_{i=1}^{n} P(B|A_i) \cdot P(A_i)}$$

Example:

A_1 - lowest risk, A_2 - medium risk, A_3 - high risk. Risk means borrower might fail to pay.

Historical data - 30% customers are A_1, 60% are A_2 and 10% are A_3. It is found that 1% that are rated A_1, 10% that are rated A_2 and 18% that are rated A_3 became defaulters. What is the probability of picking a defaulter from the pool, who has received A_1 rating?

solution:

$P(D)$ = probability of defaulter.

$P(A_1) = 0.30$ | $P(D|A_1) = 0.01$

$P(A_2) = 0.60$ | $P(D|A_2) = 0.10$

$P(A_3) = 0.10$ | $P(D|A_3) = 0.18$

$$P(A_1|D) = \frac{P(D|A_1) \cdot P(A_1)}{\{P(D|A_1) \cdot P(A_1) + P(D|A_2) \cdot P(A_2) + P(D|A_3) \cdot P(A_3)\}}$$

$$= \frac{(0.01 \times 0.30)}{\{(0.01 \times 0.30) + (0.10 \times 0.60) + (0.18 \times 0.10)\}} = \frac{(0.01 \times 0.30)}{(3 \times 10^{-3} + 0.06 + 0.018)}$$

$$= 0.037$$

$P(\text{rating } A_1 \mid \text{defaulter}) = 0.037 \Rightarrow 3.7\%$

Binomial probability distribution:

It is a probability distribution that summarizes the likelihood that a value will take one of two independent value under the given set of parameter. some rules are,

* Independent & fixed no. of. trails.
* Two potential outcomes per trail.
* P(success) should be same across all trails.

Formula:

$$\binom{n}{x} (p^x)(1-p)^{n-x}$$

- n = no. of trails
- x = success out of n trails
- p → probability of success
- $(1-p)$ → probability of failure

$$\binom{n}{x} \Rightarrow {}_nC^x = \frac{n!}{(n-x)! \cdot (x)!}$$

Example 1:

(To find the probability of outcome). Find the probability of getting 0, 1, 2, 3, 4, 5 number of heads by flipping a coin 5 times.

X = # of heads (H) from flipping a coin

Solution:

$$\left\{\begin{matrix}\text{Possible outcome} \\ \text{from 5 flips}\end{matrix}\right\} = 2 \times 2 \times 2 \times 2 \times 2$$
$$= (2)^5 \Rightarrow 32 \text{ outcomes.}$$

$$P(x=0) = \frac{5C^0}{32} = \left\{\frac{5!}{(5-0)!\,(0)!}\right\} / 32 = 1/32$$

$$P(x=1) = \frac{5C^1}{32} = \left\{\frac{5!}{(4!)\,(1!)}\right\} / 32 = 5/32$$

$$P(x=2) = \frac{5C^2}{32} = \left\{\frac{5!}{(3!)\,(2!)}\right\} / 32 = 10/32$$

$$P(x=3) = \frac{5C^3}{32} = \left\{\frac{5!}{(2!)\,(3!)}\right\} / 32 = 10/32$$

$$P(x=4) = \frac{5C^4}{32} = \left\{\frac{5!}{(1!)\,(4!)}\right\} / 32 = 5/32$$

$$P(x=5) = \frac{5C^5}{32} = \left\{\frac{5!}{(0!)\,(5!)}\right\} / 32 = 1/32$$

Example 2 :

In a basket ball match, considering a free throw. Probability of missing a free throw is 30 %. Let x = exactly 2 scores in 6 attempts. Find the probability of x.

Solution :

$$6C^2 = \left\{\frac{6!}{(6-2)!\,(2!)}\right\} = 15$$

$$P(x=r) = nC^r\,(p)^r\,(1-p)^{n-r}$$

$$P(x=2) = 15\,(0.7)^2\,(0.3)^4$$

$$= 15 \times 0.49 \times 0.0081$$

$$= 0.05953 \simeq 0.06$$

$$P(x=2) = 6\ \%$$

Poisson distribution:

Poisson distribution helps us to predict the probability of certain event from happening when we know how often the event has occured.

> "It gives the probability of a given number of events happening in a fixed interval of time"
>
> DISTRIBUTION OF RARE EVENTS AND # OF EVENTS IN FIXED TIME

For poisson distribution the mean and variance is always equal. [Note: Gamma distribution also has same mean & variance].

μ, mean $= \lambda$

σ^2, variance $= \lambda$

λ - expected no. of hits in given time period.

Formula: (distribution)

$$P(x = n) = \frac{(\lambda)^n (e)^{-\lambda}}{n!}$$

λ - mean / expected no. of outcomes.

n - No. of occurances.

e - Euler's Number (2.718 - constant)

$n!$ - factorial fo no. of occurances.

Formula: (In given time) - poisson process.

$$P(x=n) = \frac{(rt)^n (e)^{-rt}}{n!}$$

rt = r no. of. events per unit time ($\lambda = rt$)

n = no. of. occurances.

e = Euler's number (2.718 - constant)

$n!$ = factorial for no. of. occurances.

Example-1: New cases occuring 2 per month in India what is the probability that 0, 1, 2, 3, 4 ... cases will occur next month.

n	$P(x=n)$
0	$\frac{(2)^0 (e)^{-2}}{0!} = 0.135$
1	$\frac{(2)^1 (e)^{-2}}{1!} = 0.27$
2	$\frac{(2)^2 (e)^{-2}}{2!} = 0.27$
3	$\frac{(2)^3 (e)^{-2}}{3!} = 0.18$
4	$\frac{(2)^4 (e)^{-2}}{4!} = 0.09$

Example-2: (poisson process)

New cases occuring at a rate of 3 per month. Find the probability that exactly 6 cases will occur in next 3 months.

sol:

$$P(x=6) = \frac{(3\times3)^6 (e)^{-(3\times3)}}{6!} = \frac{(531441)(1.235\times10^{-4})}{720}$$

$$= 0.091 \Rightarrow 9.1\%$$