UNIT 2: REGRESSION- PART - 1

Points to be covered in this topic

- 1. Introduction
- 2. Curve fitting by the method of least squares
- 3. Fitting the lines y = a + bx and x = a + by
- 4. Multiple regression
- 5. Standard error of regression

INTRODUCTION

- Regression is to determine **statistical relationship between two or more variables.** Using regression analysis it is possible to **estimate or predict value of one variable from the given value of the other variable.**
- For example, **temperature and drug dissolved in solvent are correlated.** It can help to find out amount of drug dissolved at particular temperature.
- Regression analysis estimates or **predicts the value of one variable**, if the value of other variable is known.
- Thus, regression analysis is a statistical method used in various disciplines that attempts to determine the strength and character of the **relationship between one dependent variable (y)** and a number of **other independent variables (X1, X2, X3 … etc).**

Properties of Regression:

1. Regression is illustration of response variable as a **function of predictor variable.**
2. It is estimated when there is **significant correlation between the response and predictor variable.**
3. It predicts only a **probable value** of the response on a known value of predictor.
4. It can be worked out in two ways: **variable x on variable y** and **variable y on variable x.**
5. Regression coefficient is used to work out **regression equation.**

Methods to Study Regression:

There are two methods used to study regression namely:
Graphical method and Mathematical method.

- In graphical method, values of variables are plotted on graph as points that give a scatter plot. Usually, **independent variable is placed on x-axis and dependent variable on y-axis.**
  The regression line is drawn through point to balance points on each side of line.
- In mathematical method, regression line is used to show **regression of x on y** and **regression of y on x** to estimate correct correlation coefficient.

Types of Regression:

- There are two basic types of regression namely;

  ✓ **Simple linear regression**

  ✓ **Multiple linear regression**

- The former uses one **independent variable to explain or predict** the outcome of the dependent variable, whereas the later uses two or more independent variables to predict the outcome.
- The linear regression is **represented using graphs by straight line** however; **non-linear relationship between variables forms a curve** and hence is called as **curvilinear regression.**
  In case of more complicated data and for its analysis a non-linear regression method is used.

Linear Regression

- The relationship between **two or more variables** is estimated by plotting the **individual observations as points on scatter plot.**
  The significance of relationship between variables is determined on the basis of **nature and direction of point on the plot.**
- To estimate relationship between variables a straight line is drawn **approaching as close as possible through all the points on the plot.**
- **Relationship between two variables is presented by a regression line.**
  It gives an average value of one variable (x) from any given value of other variable (y).
- There are always two regression lines to **demonstrate the relationship between x and y variables.**
- One line demonstrates **regression of y on x** and other line demonstrates **regression of x on y.**

- In case of **perfect correlation (+1)**, both the lines coincide to become single line. If both these lines are close to each other it is an indication of **strong correlation between both the variables, x and y**.
- When these lines are farther from each other is an **indication of weaker correlation between variables**.
- The **correlation value 0** is an **indication that both the variables are independent of each other** and their line will intersect with each other at some location on the plot.

Regression Equation:

- The equation of regression line is called as **'regression equation'**.
- The two variables are supposed to have **linear relationship** if change in one variable reflects certain amount of effect on other variable.
- The usefulness of regression analysis is used to predict the value of one variable from the **known values of other variables**. For two lines their equations are:

The linear regression equation for x upon y is:

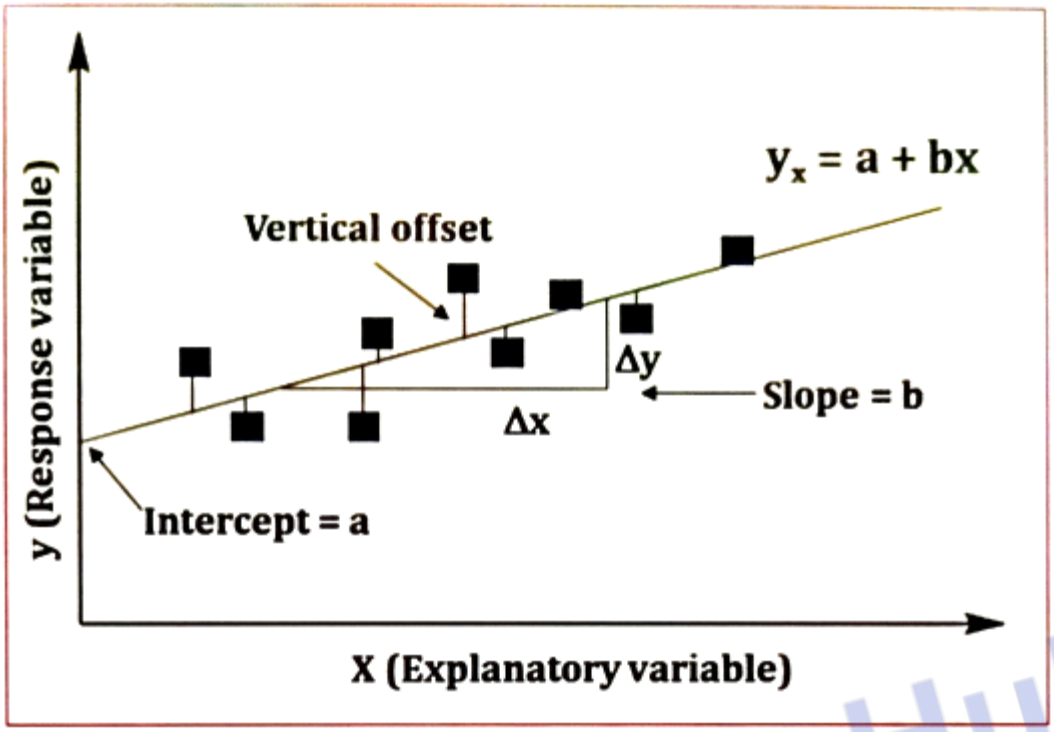
$$x_Y = a + by$$

The linear regression equation for y upon x is:

$$y_X = a + bx$$

- Where, x and y are two variables and; a and b are unknown constants.
- These equations **estimates the value of y for the known value of x and x for the known value of y**.
- The value of **a** is the point of **intercept of regression line on y-axis**.

- The **value of b**, also called as **regression coefficient**, is computed as slope of the regression line. The regression coefficient indicates relation between **every unit change of x with the corresponding change in y**.



Regression Line for y on x Showing its Regression Equation

The constants **a and b** can also be obtained from their **respective mean, y and x**, using following formulae:

$$b = \frac{\sum dxdy}{\sum dx^2} = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum dx^2} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} \quad \text{.....(2.3)}$$

$$a = \bar{y} - b\bar{x} \quad \text{.....(2.4)}$$

The general forms of linear regression are:

(a) **Simple linear regression**:

$$\hat{y} = a + \beta x + C \qquad \text{.....(2.5)}$$

(b) **Multiple linear regression**:

$$\hat{y} = a + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + ... + \beta_n x_n + C \qquad \text{.....(2.6)}$$

- Where, **y** is predicting or dependent variable, **x** is variable used to predict y, the independent variable, **a** is an intercept, **β** is slope (regression coefficient) and **C** is regression residual.
- The symbol $\hat{y}$ denotes the estimated value of **y** for a given value of **x**. This equation is known as the **regression equation of y on x**.
- This also represents a regression line of **y on x** when drawn on a graph. This means each unit change in x produces a change of **β in y**. A positive change indicates direct relationship whereas the negative change indicates **inverse** relationships between x and y.
- Regression is employed to determine whether estimated value of **regression coefficient (β)** deviates significantly from the ideal zero value.

Example:

In an experiment, data recorded for two variables such as **disintegrant concentration (DC)** and **disintegration time (DT)** is given.

Data for experiment

Test No.	DC (%)	DT (min)	Test No.	DC (%)	DT (min)	Test No.	DC (%)	DT (min)
1	2	40	**6**	13	26	**11**	21	20
2	4	35	**7**	14	25	**12**	24	19
3	7	32	**8**	16	25	**13**	25	19
4	8	30	**9**	19	23	**14**	27	15
5	10	29	**10**	20	22	**15**	30	10
						n= 15	$\sum X = 240$	$\Sigma Y = 370$

Plot the scatter graph and determine coefficient of correlation, r. With the data given , obtain the two regression equations. Using regression equation plot estimated regression line to estimate its r.

Solution: Given: n = 15, ΣX = 240 and ΣY = 370 Thus on computation of data we get

$\sum x2=4886$

$\sum y2=9976$

$\sum xy=4945$

$\bar{x}=16$

$\bar{y}=24.67$

The value of r for the regression line in graph is **0.98**

Regression equation of y on x is:

$$y_x = a + bx$$

$$b = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x^2)}{n}}$$

$$= \frac{-925}{1046} = -0.88432$$

The two regression equations for the given data are:

$$a = \bar{y} - b\bar{x}$$

$$= 24.67 - (-0.88 \times 16) = 38.81581$$

$$y_x = a + bx$$

$$= 38.81581 + (-0.88432)x$$

The values of y after subsisting values of x in equation Yx = 38.81581 + (0.88432)x can be estimated as:

X = 0 $\hat{y}$ **= 38.815**

X = 5 $\hat{y}$ **= 34.039**

X = 10 $\hat{y}$ **= 29.965**

X = 15 $\hat{y}$ **= 25.540**

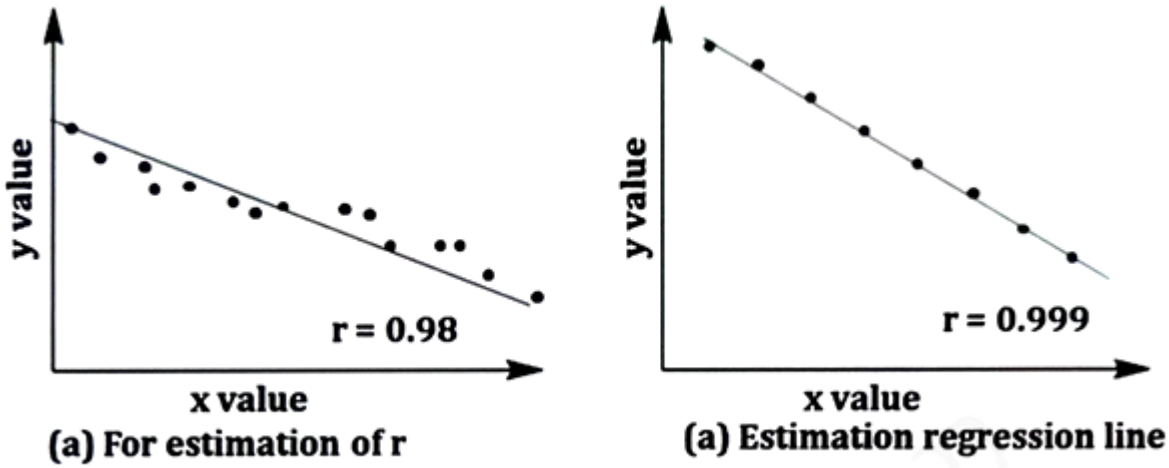
X = 20 $\hat{y}$ **= 21.115**

X = 25 $\hat{y}$ **= 16.690**

X = 30 $\hat{y}$ **= 12.265**

X = 35 $\hat{y}$ **= 07.840**

On plotting these values on a graph we get a straight line, known as estimated regression line, The estimated regression line shows **r = 0. 999**



Scatter plot

Curve Fitting – Method of Least Squares

- **Least-square method** is a mathematical procedure for **finding the best-fit curve** to a given set of data points by **minimising the sum of the squares of the offsets** (residuals) of the points from the curve.
- The sum-of-squares form is used because it treats residuals as a **continuous differentiable quantity**.
- However, using squares of the offsets, outlying points can have a **disproportionate effect on the fit.**
- If x and y variables shows a linear relationship between them, **a line is obtained that best fits this linear relationship.**
- This line has equation y = bx + a This equation gives a y-value for any x-value. The least squares method is a **more accurate statistical procedure to find the best fit for a set of data** points by minimizing the sum of the offsets or residuals of points front the plotted curve.

- It is called a least squares because the best line of fit minimizes the variance.
- The variance is the sum of squares of the errors .
- The aim of best line of fit that minimizes the variance is to **find the equation that fits the points as closely as possible.**
- This provides a fitting function for the **independent variable x that estimates y for a given x.** It also allows uncertainties of the data points along the x- and y-axes to be included, and also provides a much simpler analyticai form for the fitting parameters.

Linear Least Squares Fitting:

- This technique is the simplest and commonly applied form of **linear regression** because it provides an **answer to the problem of finding the best fit straight line** using a set of data points.
- If the relationship between the **two parameters that are plotted is known, it becomes easy to transform the data in suitable form** that the resulting line is a straight line.
- This can be demonstrated with an example wherein to **analyze the K as a function of T, plotting K vs. VT** is better than that of plotting K vs. T. Thus, standard forms for exponential, logarithmic and power laws are often used to easily compute the data.

Non-linear Least Squares Fitting:

- For non-linear least squares fitting to a **number of unknown parameters**, a linear least squares fitting may be applied repetitively to a linearized form of the function until **convergence is obtained**.

- It is possible to **linearize a non-linear function** and **use linear methods for determining fit parameters** without resorting to repetitive procedures.
- The **non- linear fit may have good or poor convergence** properties depending on the type of fit and initial parameters chosen.
- If uncertainties (error) are given for the points, points can be weighted differently to give the high-quality points a more weightage.

Vertical Least Squares Fitting:

Vertical least-squares fitting is applied by **finding the sum of the squares of the vertical deviations R2** of a set of n data points from a function f.

$$R^2 = \sum \left[y - f(xa_1, a_2, \ldots, a_n) \right]^2 \qquad \text{.....(2.7)}$$

- This procedure does **not** minimise the actual distances from the line. In addition, although the simple sum of distances seems to be more appropriate quantity to minimize, using absolute value, results in discontinuous derivatives that is difficult to treat analytically.
- Thus, instead the square deviations from each point are summed, and the resulting residual is minimized to find the best fit line. This procedure results in far away points being given undue large weightage. The condition for R^2 to be a minimum is presented as

$$\frac{\partial (R)^2}{\partial a} = 0 \qquad \text{.....(2.8)}$$

For a **linear fit**:

$$f(a,b) = a + bx \qquad \text{.....(2.9)}$$

Thus,

$$R^2(a,b)=\sum\left[y-(a+bx)\right]^2 \qquad \text{.....(2.10)}$$

$$\frac{\partial(R)^2}{\partial a}=-2\sum\left[y-(a+bx)\right]=0 \qquad \text{.....(2.11)}$$

$$\frac{\partial(R)^2}{\partial b}=-2\sum\left[y-(a+bx)\right]x=0 \qquad \text{.....(2.12)}$$

The equations (2.11) and (2.12) lead to the equations

$$na+b\sum x=\sum y \qquad \text{.....(2.13)}$$

$$a\sum x+b\sum x^2=\sum xy \qquad \text{.....(2.14)}$$

Matrix form

$$\begin{bmatrix} n & x \\ x & x^2 \end{bmatrix}\begin{bmatrix} a \\ b \end{bmatrix}=\begin{bmatrix} \sum y \\ \sum xy \end{bmatrix}$$

$$\begin{bmatrix} a \\ b \end{bmatrix}=\begin{bmatrix} n & x \\ x & x^2 \end{bmatrix}\begin{bmatrix} \sum y \\ \sum xy \end{bmatrix}$$

The 2 x 2 matrix inverse is:

$$\begin{bmatrix} a \\ b \end{bmatrix} = \frac{1}{b\sum x^2 - (\sum x)^2} \begin{bmatrix} \sum y \sum x^2 - \sum x \sum xy \\ n\sum xy - \sum x \sum y \end{bmatrix} \quad \text{.....(2.15)}$$

$$a = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x^2)} \quad \text{.....(2.16)}$$

$$b = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2} \quad \text{.....(2.17)}$$

To find a linear regression equations first determine if there is any relationship between the two variables. In fact, this is some prediction 111ade by the researcher. It also requires independent and dependent variables data in column format.

Multiple Regression

- In studies when **there are two or more independent variables**, the analysis describing relationship between them is called **'multiple correlations'** and the equation that describes such relationship is known as the **'multiple regression equation'.**
- It explains the relationship between **two or more independent variables and one dependent variable.** Since this **regression uses two or more independent variables it is called 'multiple regression'.**
- Multiple regression has two types namely; **multiple linear regression** and **multiple non-linear regression.**

- Multiple linear regression analysis is a set of techniques used to **study the straight-line relationships between two or more variables.**
- It is used to predict values of the **dependent variable indicating ' independent variables that have a major effect on the dependent variable**. It can be used when there are three or more independent variables.

Assumptions of Multiple Regression:

1. All variables are **continuous measurement variables**.
2. Variable distributions are **unimodal and approximately symmetrical**.
3. There is a **linear relationship** between **predictor** and **response** variables.

Computation of Multiple Linear Regression:

- The purpose of a multiple regression is to find an equation that best predicts the **y variable as a linear function of x variables.**
- Multiple regression estitnates the P's in the equation. The nntltiple linear regression equation is presented in the following general for1n:

$$y = \beta_o + \beta_1 x_{1j} + \beta_2 x_{2j} + \ldots\ldots \beta_n x_{nj} + C_j \qquad \ldots\ldots(2.18)$$

- Where, **x's are the independent variables** and **y is the dependent variable**.
- The subscript **j** represents the observation number. The **β's (i = 1 , 2... n) are the unknown regression coefficients**, indicating how the criterion variable changes when the predictor variable changes.
- Their estimates are denoted **b's**. **Each β represents the unknown parameter,** while b is its estimate**. The Cj is the error of observation j.**
- As an example, let's say that, the **tablet hardness of a batch of product is dependent on factors like type of binder, amount of moisture and the amount**

of compressional force applied. Using hardness test one can estimate the appropriate relationship among these factors.

- The most commonly used method for regression problem is **least square regression analysis**. In this method, the h's are selected to minimize the sum of the squared residuals. This set of h's is not necessarily the set that is desired, since they may be distorted by outliers. The sample multiple regression equation is presented as:

$$\bar{y}_j = b_o + b_1 x_{1j} + b_2 x_{2j} + \ldots + b_n x_{nj} \qquad \text{.....(2.19)}$$

- **If n = 1, the model is called simple linear regression**. It should be noted that the number of normal equations would **depend upon the number of independent variables.**
- **In case of 2 independent variables, there are 3 equations,** if there are 3 P01 independent variables then 4 equations and so on, are used. The intercept, β_o, is the point at which the regression plane intersects the y-axis.
- In multiple regression analysis**, the regression coefficients,** β_1, β_2, **become less reliable** when the degree of correlation between the **independent variables, x1, x2, increases.**
- When there is a high degree of correlation between independent variables, it is then known as the **problem of multicollinearity.**
- In such condition. it is suitable to use only one **set of the independent ' variable to make an estimate.**
- Actually, adding a **second variable (x2) and correlating it with the first variable (x1)** alters the values of the regression coefficients.
- However, the predictions for the dependent variables can be made even when **multicollinearity is present.**
- When there is more than one independent variable, a difference between the **collective effect of the two independent variables and the individual effect of**

each variable is taken separately. The collective effect is given by the coefficient of multi pie correlations (Ry , x1x2].

$$R_{y.x_1x_2} = \sqrt{\frac{\beta_1 \sum x_{1i} y_1 + \beta_2 \sum x_{2i} y_1}{\sum y_i^2}}$$

Where

$\bar{x}_{1i}$ is $(\bar{x}_{1i} - \bar{x}_1)$

$\bar{x}_{2i}$ is $(\bar{x}_{2i} - \bar{x}_2)$

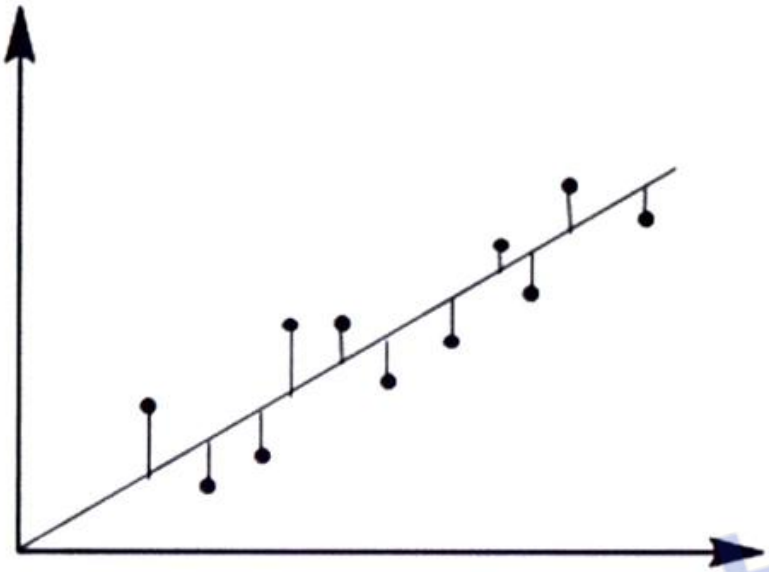
$\bar{y}_i$ is $(\bar{y}_i - \bar{y})$

β_1 and β_2 are the regression coefficient

Standard Error of Regression

- The **standard error(s) of the regression coefficient** is the SD of the estimate.
- Used for **hypothesis testing or confidence limits**.
- The S of the regression represents **the average distance that the observed values fall from the regression line.**
- The S gives information about **how wrong the regression model is on using the uni.ts of the response variable.**
- The S becomes smaller when the **data points are closer to the line.**
- The S of the regression can be used to **assess the precision of the predictions.**

- Approximately **95% of the observations are supposed to falI within ±2 S of the regression** from the regression line, which is a quick approximation of a 95% prediction interval.
- If regression model is used to make predictions, **assessing the S of the regression might be more important than assessing R2**



Standard Error of Regression

UNIT 2: PROBABILITY- PART - 2

Points to be covered in this topic

1. Definition of probability
2. Binomial distribution
3. Normal distribution
4. Poisson's distribution
5. Sample, Population
6. Large sample, small sample
7. Null hypothesis, alternative hypothesis
8. Sampling, Essence of sampling
9. Types of sampling
10. Error-I type, Error-II type, Standard error of mean (SEM)

∇ INTRODUCTION

- Probability is the science of uncertainty that provides **precise mathematical rules** for understanding and **analyzing researcher ignorance**.
- It gives a framework for working with limited knowledge and for making sensible decisions based on **actions taken and those not known**.
- The probability of any outcome of a **random phenomenon is the number of times** the outcome would occur in a **series of repetitions**.

- It is the measure of the chance that an **event will occur in a random experiment**. It is quantified as a number **between 0 and 1**, where, **0 indicates impossibility** and **1 indicates certainty**.
- The higher is the probability of an event, the **more chance that the event will occur**. In simple words, the probability of a single event occurring is to divide the number of events by the number of possible outcomes.
- The concept of probability can be understood through an example, wherein a **pharmaceutical company tested a new drug** and for that the company **injected new drug to 100 patients and obtained information**.
- Based on previous observations company **obtained probability using the data that there was no any side effects 72 times**, **mild side effects 25 times** and **severe side effects 3 times**.
- Thus, the probability of severe side effect occurrences is obtained as **ratio of number of times severe side effects** observed to the number of times experiments performed. Thus, the **ratio 3/100 = 0.03 or 3% is the probability**.
- Objectives of probability testing is to calculate probabilities by counting the outcomes in a sample space, use **counting formulas to compute probabilities**, **understand how probability theory is used** in pharmaceuticals and to understand the relationship between probability and odds.

Notations in Probability:

Let A and B denote two events -

1. **A ∪ B** is the event that, either A or B or both occur.
2. **A ∩ B** is the event that both A and B occur simultaneously.

The complement of A is denoted by $\bar{A}$. The $\bar{A}$ is the event that A does not occur. Thus,

$$P(A) = 1 - P(A)$$

Events A and B are mutually exclusive if both cannot occur at the same time. A and B are independent events, if and only if;

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

Law of Probability:

1. **Multiplication law**: If $A_1, A_2, A_3, \ldots\ldots, A_n$ are independent events, then

$$P(A_1 \cap A_2 \cap \ldots\ldots \cap A_n) = P(A_1)\ P(A_2) \ldots\ldots P(A_n)$$

2. **Addition Law**: If A and B are any events, then

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

This law can be extended to **more than 2 events**.

Conditional Probability:

- It is a measure of the probability of an event given that, another event has already occurred. **If the event of interest is A and the event B is known** (or assumed) to have occurred, the **conditional probability of A given B**, is written as **P(A|B)**.

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

A and B are independent events, if and only if;

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

Example: A drug test for an illegal drug is such that it is **98% accurate in the case of a user of that drug**. This means, it produces a positive result with probability 0.98 in the case that the tested individual uses the drug. It is **90% accurate in the case of a non-user of the drug**. This means, it is negative with probability 0.9 in the case the person does not use the drug.

Suppose it is known that 10% of the entire population uses this drug.

(a) What is the probability that the **tested individual uses this illegal drug**?

(b) What is the probability of a **false positive** with this test (for example, the probability of **obtaining a positive drug test given the person tested is a non-user**)?

(c) What is the probability of obtaining a **false negative** for this test (for example, **the probability that the test is negative, but the individual tested is a user**)?

Solution:

Let's consider that:

- **+** = Event that the **drug test is positive** for an individual.
- **–** = Event that the drug test is **negative** for an individual.
- **A** = Event that the person tested **does use the drug** being tested for.

We want to find:

P[A I+], P[+I A] (false positive), and **P[-1 A]** (false negative). We know that, **P[A] = 0.1 P[+I A]=0.98,** and **P[-1 A]=0.9** and from this **P[+IA]=0.9** which allows us to find **P[A I+]** using **Bayes Theorem:**

$$P[A|+]=\frac{P[A|+]\ P[A]}{P[+]}$$

$$P[A|+]=\frac{P[A|+]\ P[A]}{P[A|+]\ P[A]+P[+|\bar{A}]\ P[\bar{A}]}$$

$$=\frac{0.98\times0.1}{(0.98\times0.1)+0.9\times0.1}=0.52$$

Thus, the probability of tested individual using illegal drug is 0.52

The probability of a false positive is:

$$P[A|+] = 1 - P[-IA]$$

$$= 1 - 0.9 = 0.1$$

Therefore, the probability of obtaining a **positive drug test given the person tested is a non-user is 0.1.**

The probability of a false negative is:

$$P[-IA] = 1 - P[A|+]$$

$$= 1 - 0.98 = 0.02$$

The probability that the test is negative, but **the individual tested is a user is 0.02.**

- **Independence**:
  - Two events are said to be independent of each other, **if the probability that one event occurs in no way affects** the probability of the other event occurring.
  - This means that, if the **observation about one event is known**, **it does not affect the probability of the other**. Thus, for independent events A and B below is true:

$$P(A,B)=P(A) \times P(B)$$

Where,

$$P(A) \neq 0 \text{ and } P(B) \neq 0$$

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

For example, **a die is rolled and a coin is flipped**.
The probability of getting any number face on the die in no way influences the probability of getting a head or a tail on the coin.

Conditional Independence:

- According to the formula for conditional probability the events A and B are said to be independent if, **P (A) = P(AIB) or alternatively if P(A, B) = P(A) P(B).**
- This means that **two events A and B are conditionally independent when given a third event C accurately**, if the occurrence of A and the occurrence of B, are independent events in their conditional probability distribution given C.

In other words, **A and B are conditionally independent given C**, if and only if, given knowledge that C has already occurred,
knowledge of whether A occurs provides no additional information on the probability of B occurring, and knowledge of whether B occurs provides no additional information on the probability of A occurring. This relation can be mathematically presented as:

$$P(A|C, B|C)\ P(A|C)\ P(BIC)$$

where, $P(A|C) \neq 0$ and $P(B \mid C) \neq 0$

Probability Distributions:

- A probability distribution is a **mathematical function** that describes the likelihood of obtaining all the **possible outcome values** that a random variable can assume.

- In other words, the values of the variable vary based on the **underlying probability distribution**.
- Probability distributions are of two types namely;
  - ✅ **Continuous probability distributions** and
  - ✅ **Discrete probability distributions**.
- This classification is based on whether they define probabilities for **continuous variables** or **discrete variables**.
- Other types of probability distributions include:
  - ✅ **Conditional Probability Distribution**,
  - ✅ **Joint Probability Distribution**, and
  - ✅ **Factorial Distribution**.

Probability distribution depends on the random variable X, whether it is discrete or continuous.

BINOMIAL DISTRIBUTION

- A binomial distribution is the **probability of a success or failure outcome in an experiment** that is repeated several times.
- The binomial distribution has two possible outcomes and thus is **prefixed with "bi"**.In a situation, when there are **more than two distinct outcomes**, a multinomial probability model might be appropriate distribution.
- The binomial distributions has certain assumptions such that, **the number of trials (n) must be fixed in advance,** the probability that the event occurs (P) must be same for **all trials and the trials must be independent.**

- As an example, let us consider if **a new drug is introduced to cure a disease,** it either cures the disease (it's successful) or it does not cure the disease (it's a failure).
- Basically, anything we can think of that can only be a success or a failure; can be represented by a binomial distribution. **Thus, while using the binomial distribution, it must be clearly specified that which outcome is the "success" and which is the "failure".**
- Let **n** denote the **number of observations or the number of times the process is repeated,** and **X denote the number of successes or events of interest occurring for n observations.** The probability of success or occurrence of the outcome of interest is indicated by **P.** The binomial equation also uses factorials.

The binomial distribution formula is:

$$b(X; n, P) = C_x^n \times P^x \times (1- P)^{n-x}$$

Where, **b is binomial probability**, **X** is total number of successes, **P** is **probability of success** on an individual trial and **n is number of trials**.

The binomial distribution formula can also be written as:

$$b(X) = C_x^n \times P^x \times Q^{n-x}$$

$$b(X) = C_x^n \times P^x \times Q^{n-x}$$

$$C_x^n = \frac{n!}{x!(n-x)!}$$

The factorial of a **non-negative integer K is denoted by K!**, which is the product of **all positive integers less than or equal to K.** Using notations of factorial, the binomial distribution model is mathematically expressed as:

$$P(X \text{ "successes"}) = \frac{n!}{x!(n-x)!} \times Px \times (1- P)^{n-x}$$

- The first variable in the binomial formula, n, stands for the number of times the experiment runs. The second variable, **P, represents the probability of any one specific outcome.**

Example:

Suppose that 90 % of adults with allergies report symptomatic relief with a specific medication. If the medication is given to 10 new patients with allergies, what is the probability that it is effective in exactly six patients?

Solution:

First, it must be verified that the three assumptions of the binomial distribution model are satisfied.

(a) **The outcome is relief from symptoms (yes or no). Relief from symptoms is a 'success'.**

(b) **The probability of success for each person is 0.9.**

(c) **The replications are independent.**

In this example: number of observations (n) is **10**, the successes (or events) of interest is (x) **6** and **P is 0.9.**

Therefore, the probability of 6 successes is:

$$P(6 \text{ successes}) = \frac{10!}{6!\,(10-6)!}(0.9)^6\,(1-0.9)^{10-6}$$

This is equivalent to

$$P(6 \text{ successes}) = \frac{10\,(9)(8)(7)(6)(5)(4)(3)(2)(1)}{[6\,(5)(4)(3)(2)(1)]\quad[4(3)(2)(1)]}\,(0.531)\,(0.0001)$$

Thus,

$$P(6 \text{ successes}) = \frac{10\,(9)(8)(7)}{(4)(3)(2)(1)}\,(0.0000531)$$

$$= \frac{5040}{(24)}\,(0.0000531) = 0.01$$

Thus, there is a 1 % probability that exactly 6 of 10 patients will report relief from symptoms when the **probability that any one reports relief is 90 %.**

NORMAL DISTRIBUTION

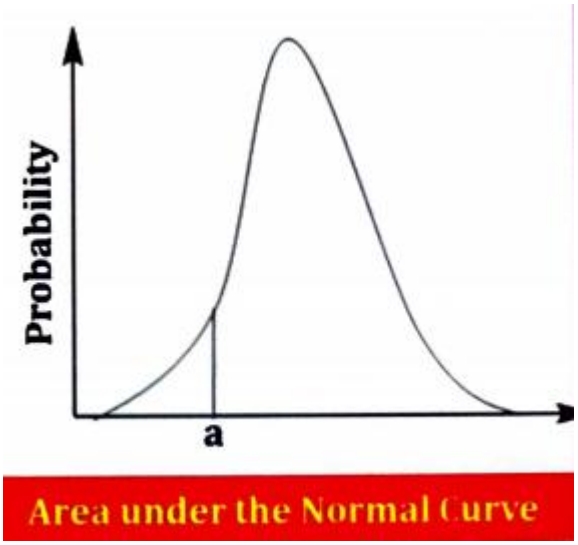
- The normal distribution, also known as the **Gaussian distribution,** describes how the values of a variable are distributed.
- It is a **symmetric distribution** in which most of the observed values **cluster around the central peak.**
- The probabilities for values those farther or away from the **mean are tapered off equally in both directions.**
- As an example, **heights, blood pressure and measurement error are all that follow the normal distribution**. Extreme values at both ends of the distribution are unlikely.
- In graph form, normal distribution appears as a bell shape curve and thus it is also known as the **bell curve.**

(a) Properties of a Normal Distribution:

1. The **mean, mode and median** are equal.
2. The curve is **symmetric at the center** (i.e. around the mean, μ).
3. The observed exactly **half of the values fall to the left of center** and exactly **half the values to the right of center.**
4. The **total AUC is equal to 1.**

5. The probability that a **normal random variable x equals any particular value is 0.**

- The normal distribution is a **continuous probability distribution** and has several other implications for probability.
- The probability that **x will be greater than a equals the area** under the normal curve bounded by a and **plus infinity, a non-shaded area.**
- The probability that x will be less than a equals the area under the normal curve bounded by a and minus **infinity, a shaded area.**



Area under the Normal Curve

In addition, all normal curves, regardless of its mean or SD, conforms to the following rules.

1. About 68% of the AUC falls within 1 SD of the mean.
2. About 95% of the AUC falls within 2 SDs of the mean.
3. About 99.7% of the AUC falls within 3 SDs of the mean.

Altogether, these points are known as the 'Empirical Rule' or the **'68-95-99.7 Rule'**. If given a normal distribution, most outcomes fall within 3 SDs of the mean.

The normal distribution refers to a **family of continuous probability distributions described by the normal equation.**Mathematically, normal distribution for the value of the random variable *y* is defined with the following equation:

$$y = \left\{ \frac{1}{\left[\sigma \times \sqrt{2\pi}\right]} \right\} \times e^{-(x-\mu)/2\sigma^2}$$

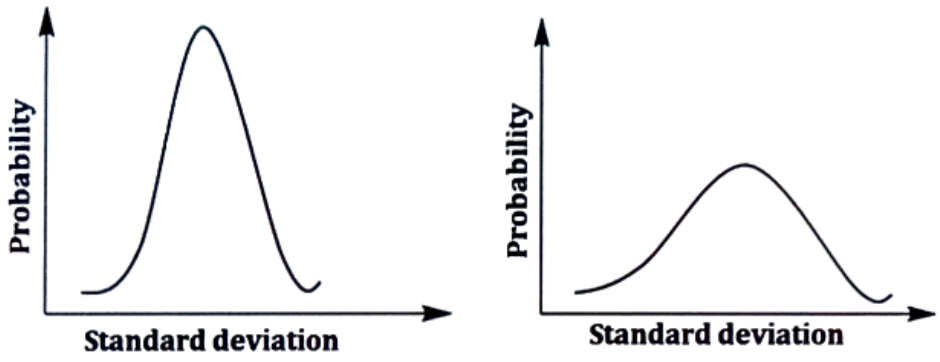
Where,

- x is a normal random variable,
- μ is the mean, σ is the SD,
- π is approximately 3.14159, and
- e is approximately 2.71828.

The normal equation is the probability density function for the normal distribution.

(b) The Normal Curve:

- The parameters for the normal distribution define **its shape and probabilities entirely**. The graph of the normal distribution depends on two factors; the **mean** and the **SD**.
- The **mean** of the distribution determines the **location of the center of the graph**, and the **SD determines the height and width of the graph**.
- All normal distributions look like a **symmetric, bell-shaped curve**. When the SD is small, the curve is tall and narrow; and when the SD is big, the curve is short and wide.
- The normal distribution takes **more than one form** and the **shape changes** are based on the parameter values.



(a) Smaller (b) Bigger

Standard Deviation

(c) Standard Deviation:

- The SD is a measure of variability. **It defines the width of the normal distribution.** The SD determines **how far away from the mean the values tend to fall**.
- It represents the characteristic distance between the **observations and the average**. Changing the SD either **tightens or spreads out the width of the distribution** along the x-axis. Larger SDs produce distributions that are more spread out.

(d) Mean:

- The mean is the **central tendency of the distribution**. The location of peak for normal distributions is **defined by the mean**. Most of the values in this distribution group around the mean.
- In its graphical presentation, **any change in the mean shifts the entire curve** either left or right on the x-axis.

Example:

An average implant manufactured by the pharma company lasts 300 days with a SD of 50 days.
Assuming that implant drug release is normally distributed, what is the probability that an implant will release drug at most 365 days?

Solution:

Given a mean score of 300 days and a SD of 50 days.

- The values given are normal random variable **365 days**, **mean = 300 days**, and **SD = 50 days**.

- To find the cumulative probability that duration over which **implant will release drug ≤ 365 days**.
- We can enter these values into the **Normal Distribution Calculator** and compute the cumulative probability.
- Thus, upon computation the cumulative probability **P(X < 365) is 0.90.** Hence, there is a **90% chance** that an implant will release drug within 365 days.

Example:

In testing tablets for disintegration time the test result is normally distributed. The test has a mean 100 sec and a SD 10 sec.Calculate the probability that tablets tested will show disintegration times between 90 sec and 110 sec.(Given that: cumulative probabilities; P(X < 110) and P(X < 90) are 0.84 and 0.16, respectively.)

Solution:

The probability condition in this example to find the answer is to express this situation as:

$$P(90 < X < 110) = P(X < 110) - P(X < 90)$$

The probabilities for terms on the right side of the above equation are given as follows:

Substituting these values in above equation we get:

$$P(90 < X < 110) = P(X < 110) - P(X < 90)$$
$$P(90 < X < 110) = 0.84 - 0.16$$
$$P(90 < X < 110) = 0.68$$

Thus, about **68% of the test results will fall between 90 sec and 110 sec**.

POISSON'S DISTRIBUTION

- Poisson distribution is a **discrete probability distribution** of the number of events that occurs in a given time period, **under condition that the average number of times** the event occurs over that time period is given.
- It is a model that **describes randomly occurring events**. This distribution is used to find the probability of a number of events in a time period and **finding the probability** of waiting some time until the next event occurs.
- A Poisson random variable is the **number of successes that result from a Poisson experiment.** This variable satisfies the following conditions:
    1. The number of successes in **two disjoint time intervals is independent**.
    2. The probability of a success during a small time interval is **proportional to the entire length** of the time interval.
    3. This variable also applies to **disjoint regions of space**.

Poisson distribution is the probability distribution that is obtained from a Poisson experiment. A **Poisson experiment is a statistical experiment that has following properties:**

1. The experimental outcomes can be categorized as **successes or failures**.
2. The average number of **successes that occurs in a specified region** is known.
3. The probability of success to occur is **proportional to the size of the region**.
4. The probability of success to **occur in a very small region is almost zero**.

- The specified region could be a **length, an area, a volume, a period of time**, etc.
- The probability distribution of a Poisson random variable **X** is given by the formula:

$$P(X;\mu)=\frac{e^{-\mu}\mu^{x}}{x!}$$

Where, **x is 0, 1, 2, 3...; e is 2.71828** and **μ is mean number of successes in the given time interval or region of space**. The **X** denotes random variable, and **x** denotes specific value of this variable.

(a) Mean and Variance of Poisson Distribution:

If **μ** is the average number of successes that occurs in a given time interval or region in the Poisson distribution, then the mean and the variance are equal to **μ**.

$$\text{Mean} = E(X) = \mu$$

$$\text{Variance } V(X) = \sigma^2 = \mu$$

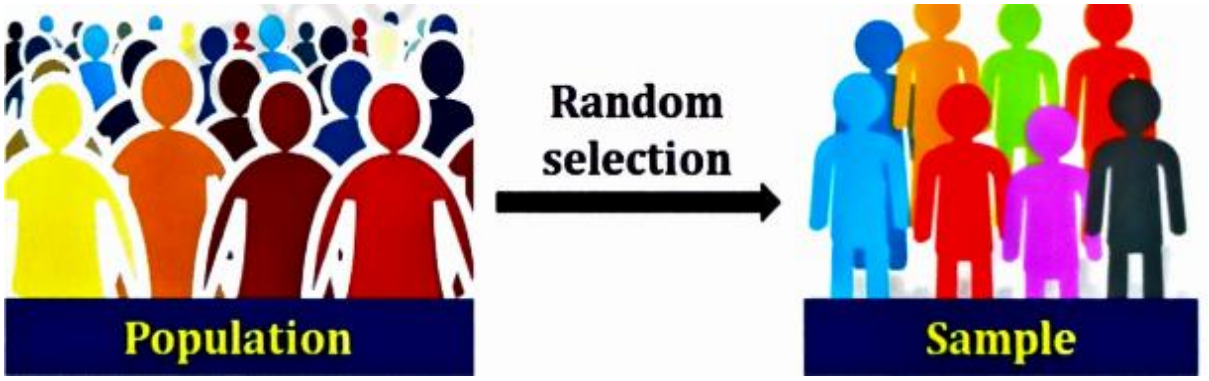
- In this distribution, **μ is needed to determine the probability** of an event.
- The Poisson distribution **PMF gives the probability of observing k events** in a time period along with the length of the period and the average events per time.

SAMPLING AND HYPOTHESIS

❖ Sample and population

- A population is a **complete set of subjects** with a specialized set of characteristics. **A sample is a subset of that population.** This means, it is a group of **phenomena that have something in common**.
- Population sampling is the process of selecting a **subset of subjects which represents the entire population**. The sample must have sufficient size to perform **statistical analysis** as well as it must be selected at random.
- Random sample of a population has an **equal chance of being selected**. The most commonly used sample is one which is selected **by simple random sampling method**.

- A population has a characteristic called **parameter**. A **statistic is a characteristic of a sample.** Thus, inferential statistics is employed to make a guess about a population parameter. **This guess is based on a statistic calculated from a sample randomly selected from population.**



- An example of statistical inference is to know the mean hardness of the tablets of a particular formulation; here hardness is the parameter of tablet population.
- We withdraw a random sample of 20 tablets and determine their mean hardness, for example 5.8 kg/cm2 , which is a statistic. Thus, it can be concluded that the population mean hardness (4) is close to 5.8 kg/cm2 . There are different symbols, given in Table 2.4, used to denote statistics and parameters.

Notations used in sample statistics and population parameters

Notation for	Sample statistics	Population parameters
Mean	$\bar{x}$	μ
Standard deviation	s	σ
Variance	s^2	σ^2

❖ Sample Size:

(a) Large Sample:

- When sample size **n** is **greater than 30 ($n \geq 30$)**, it is known as **large sample**. For large samples, the sampling distributions of statistic are **normal**.
- A study of sampling distribution of statistic for **large sample is known as large sample theory**.
- The primary goal of inferential statistics is to generalize from a sample to a population; thus, it is **less of an inference if the sample size is large**. Generally, selecting larger samples is good.

✔ **The reasons for this are**:

1. Larger samples more closely estimate the population.
2. Large samples are good to eliminate risk of the small sample being selected.
3. It provides accurate mean values, identify outliers and smaller margin of error.
4. The amount of sampling variability is small when the sample size is large.
5. The larger is sample size smaller will be the value of the standard error.
6. If sample sizes are large, a sampling distribution is normally distributed.

(b) Small Sample:

- When the sample size **n is less than 30 ($n < 30$)**, it is known as **small sample**. For small samples, the sampling distributions **are t, F and Chi-square (χ^2)**.
- A study of sampling distributions for small samples is known as **small sample theory**.

- Small samples are bad because we may run at **greater risk as small sample is unusual, just by chance**. If there is an increased probability of one small sample being unusual, that means **if many small samples are drawn, unusual samples are more frequent**. Therefore, there is greater **sampling variability with small samples**.

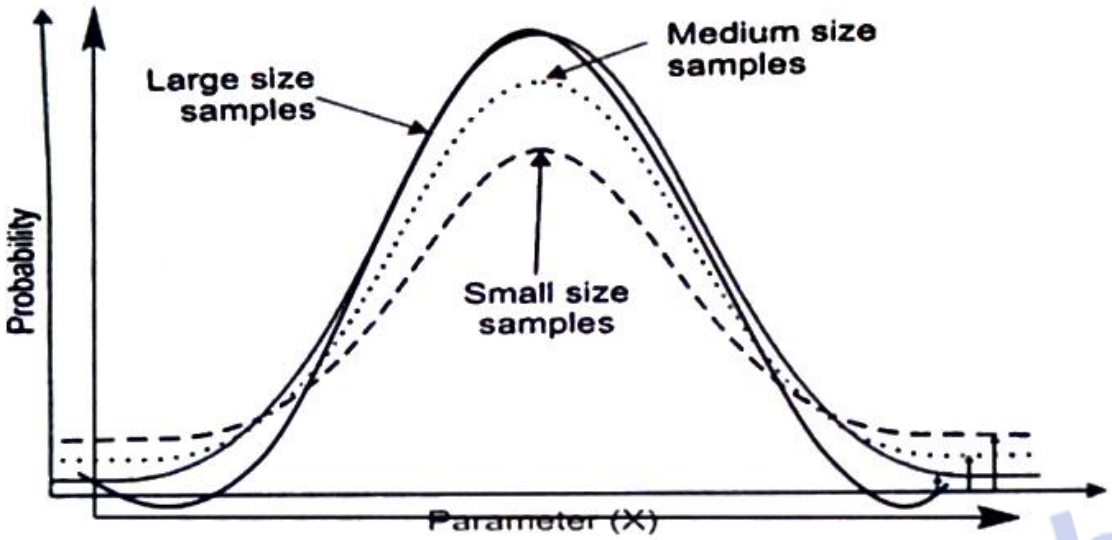


Illustration of the Effect of Sample Sizes

- In the curve with the small size samples, notice that there are **fewer samples with means around the middle value**, and more samples with means out at the extremes.
- Both the **right and left tails of the distribution** are fatter. In the curve with the large size samples, notice that there are **more samples with means around the middle**, therefore closer to the population value, and fewer with sample means at the extremes.
- The differences in the curves represent **differences in the SD of the sampling distribution**. Smaller samples tend to have larger standard errors and larger samples tend to have smaller standard errors.

❖ Hypothesis

- A statistical hypothesis is a **statement or assumption** about the nature of a population parameter. The assumption **may or may not be true**.
- Hypothesis testing is prescribed process to **accept or reject statistical hypotheses**.
- In order to determine whether a **statistical hypothesis is true or not**, the entire population needs to be examined.
  Although examining entire population is difficult, **usually random samples are used from the population**.
  Hypothesis is rejected when sample data is not consistent.

(a) Characteristics of Hypothesis:

1. Hypothesis should be clear and precise so as to **deduce reliable inferences**.
2. It should be **capable of being tested**.
3. It should state **relationship between variables**, if it is relational hypothesis.
4. It should be limited in **scope** and must be **specific**.
5. It should be a **simple statement** to better understand by all concerned.
6. It should be **consistent** with most known facts.
7. It should be open for testing within a **reasonable time**.
8. It must explain the facts that gave rise to the **need for explanation**.

(b) Types of Hypothesis:

1. Null hypothesis:

- The null hypothesis is a statement that we want to test.**There is no significant difference between specified two populations**, as any observed difference is being due to sampling or experimental error. **It is denoted by H_0**.

- In this hypothesis, sample observations results purely by chance.
  For example, **if we measure the granule size from two batches (populations) of same product**.
- The null hypothesis could be that the average granule size in one batch is the same as the average granule size in other batch.
- If we measure the amount of granule tested from both the batches, the **null hypothesis** could be that the **ratio of granules in one batch to other batch is equal to a theoretical expectation** of a **1:1**.

---

2. Alternative hypothesis:

- An alternative hypothesis is the one that **sample observations are influenced by some other non-random cause**. It is denoted by **H_1 or Ho**.
- In this hypothesis, the things are different from each other, or **different from a theoretical expectation**.
- An example of alternative hypothesis is that, **tablets from one batch have a different average tablet size than tablets from other batch of same product**; another would be that, the disintegration time is different from 1:1.

(c) Hypothesis Testing:

- The objective of hypothesis testing is **determination of the likelihood of getting observed results as per null hypothesis**.
- For example, suppose **researchers postulate that drug AB may reduce blood glucose level**.
- A null hypothesis might be that **drug AB does not reduce blood glucose level**. The **alternative hypothesis** might be that the **drug AB does reduce blood glucose level** (this is the researcher wish).

- Thus, these hypotheses would be expressed as **H_0: P = 0.5** and **H_1: P = 0.5**.

  Researcher to test these hypotheses needs two groups of patients**; one who receives drug AB and the second who receives placebo**.

- Upon measuring **blood glucose level for both these groups as treated and control**, if the blood glucose level data show significant difference then he will **reject H_0. This means he accepts H_1.**

- In testing hypothesis, a statisticians **follow appropriate procedures to estimate whether to reject a null hypothesis** based on sample data, or accept it. This process of determining type of hypothesis is called hypothesis testing.

Hypothesis testing consists of following common steps:

1. Statement of the hypothesis:

- This involves making statement for the **null and alternative hypotheses**.
- The hypotheses are stated in such a way that they are mutually exclusive that means, **if one is true, the other must be false**.

---

2. Formulation of analysis plan:

- The **analysis plan** details the processing of sample data to evaluate the null hypothesis.
- The **evaluation is projected around a single test statistic**.

---

3. Analysis of sample data:

- Find the value of the **test statistic** (mean score, proportion, t statistic, z-score, etc.) described in the analysis plan.

---

4. Interpretation of results:

- At this stage, **decision rule** described in analysis plan is used to make **conclusion**.
- On applying null hypothesis to the value, **if the test statistic is doubtful, null hypothesis is rejected**.

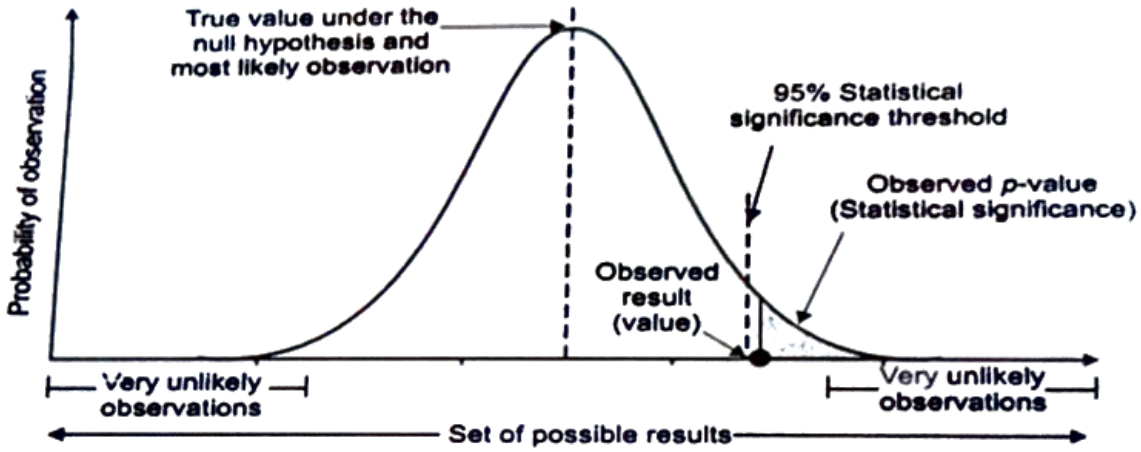
The analysis plan comprises of decision rules to reject null hypothesis.

The decision rules are described with reference to a p-value or to a region of acceptance.

---

(1) p-value:

- It is a value used to **measure the strength of evidence in support of a null hypothesis**. Suppose the test statistic is equal to significance level.
- The **significance level**, denoted as **alpha (α)**, is the probability of rejecting the null hypothesis when it is true.
- For example, a significance level of **0.05 indicates a 5% risk** of concluding that a difference exists when there is no actual difference.
- The statistical significance level is expressed as a **p-value between 0 and 1**.
- The p-value is the probability of observing a test statistic as close as to S, assuming the null hypothesis is true.
- **The smaller the p-value than α, the stronger is the evidence** that we reject the **null hypothesis**.



Significance of Results in Relation to the Null Hypothesis

---

(2) Region of acceptance:

- The region of acceptance is a **range of values**. **Acceptance region** is the **interval within the sampling distribution** of the test statistic that is consistent with the **null hypothesis (H_0)** from hypothesis testing.
- The set of values **outside** the region of acceptance is called the **region of rejection**. **Acceptance region** is the **complementary region** to the rejection region.
- If the test statistic lay within the **region of acceptance**, the **null hypothesis is not rejected**.
- If the test statistic lay **within the region of rejection**, the **null hypothesis is rejected**.
- Thus, it can be said that **the hypothesis has been rejected at a level of significance**. Both these approaches are **equivalent** as some statisticians use the **p-value approach** while others use the **region of acceptance** approach.

❖ Sampling

- Sampling is a statistical procedure employed in **selection of the individual observation** that helps us to make statistical inferences about the sample.
- The significance of sampling is to **provide various types of statistical information** of a qualitative or quantitative nature about the **set of results by examining a few selected subjects (units)**.
- There are various sampling methods used in statistics. **Sampling method is a scientific procedure used for selecting those sampling units** which would give the required estimation with margins of uncertainty from examining **only a part of whole sample** (population).
- The main types of probability sampling methods are **simple random sampling, stratified sampling, cluster sampling, multistage sampling**, and systematic random sampling.
- The objectives of sampling are as follows:
    1. To study the selected units instead of the whole universe.
    2. To get results in a short period of time.
    3. To get essential information at low cost.
    4. To minimize sampling variances.
    5. To get real and error free conclusion.

❖ Essence of Sampling

The essence of sampling is the selection of a part of sample from the whole population in order to make inference about the whole. These essentials include:

1. It must possess characteristics of the **original population of it is representative** of.

2. It should have **similar nature** as that of population.
3. For result **reliability sample should be enough in number**.
4. It must be **adequate** to make inference.
5. It must be **independent** that **same sample would not be selected repeatedly**.

❖ Types of Sampling

A. Simple Random Sampling:

- In this method, each item of the population is **chosen entirely by chance** and each subject (unit) of the **population has an equal chance, or probability**, of being selected.
- The selection is usually made using **random numbers**. As an example, suppose there are **N = 900 tablets in a batch (population)** from which a sample tablets (size) **n = 10** is to be selected.
- Then for selection, tablets are numbered **from 1 to 900**. Since data runs into three digits it is appropriate to use random numbers that contain three digits, for example from **100 to 900**.

B. Systematic Sampling:

- In this method, first population items are **numbered from 1 to N**.
- For example, suppose the **sample size is n**, and then **sampling interval is calculated by dividing N by n**. Then a number between **1 and N/n is generated** and those population items are selected as the sample.
- Other items in the sample are obtained by **adding the sampling interval N/n successively to the random number**.
- The benefit of using this method is that, **the sample is evenly distributed over the whole data**.

- As an example, suppose the batch of tablets is divided into **N = 1000 tablets** which are numbered consecutively.
- A **10% sample of tablet is to be taken**, which gives a sampling interval of **k = 10**. If the **random number between 1 and 10 is 6**, the tablets with the numbers **06, 16, 26, 36, 46, ..., 996** are selected in the sample.

C. Stratified Sampling:

- In this method, the population is first divided into **subgroups that all those subgroups possess a similar characteristic**.
- This method is used when **researchers reasonably expect the measurement of interest to vary between the different subgroups**, and which to ensure representation from all the subgroups.
- For example, in a study of **stroke outcomes Researcher may stratify the population by sex, to ensure equal representation of men and women**. The study sample is then obtained by taking equal sample sizes from each stratum.
- In stratified sampling, it may also be **appropriate to choose non-equal sample sizes** from each section.
- For example, in a study of the health outcomes of nursing staff in a particular state, if there are 3 hospitals each with different numbers of nursing staff say **hospital A has 400 nurses, hospital B has 800 and hospital C has 1600**, then it becomes appropriate to choose the sample numbers from each hospital **proportionally**, such as **10 from hospital A, 20 from hospital B and 40 from hospital C**.

D. Clustered Sampling:

- In a clustered sampling, a group of **sample units is selected together**, rather than **each unit is being selected separately**.

- The population is **divided into subgroups, known as clusters**, which are randomly selected for inclusion in the study.
- Clusters are usually pre-defined. For example, **individual General Practitioners those who treat all common medical conditions** and refer patients to hospitals and other medical services for urgent and specialist treatment could be identified as clusters.
- In single-stage cluster sampling, all members of the **chosen clusters are included in the study**.
- In two-stage cluster sampling, a selection of **individuals from each cluster is randomly made for inclusion**.

E. Quota Sampling:

- This method of sampling is often used by **market researchers**. Interviewers receive a quota of subjects of a specified type, interviewing them to recruit.
- For example, an interviewer might be asked to select **40 adult men, 40 adult women, 20 teenage girls and 20 teenage boys** so that they could interview them about their use of over the counter (**OTC**) products.
- Ideally the quotas chosen would proportionally represent the characteristics of the **underlying population**. The advantage of this method is that, it is **relatively straightforward and potentially representative**; the chosen sample may not be representative of other characteristics that were not considered.

F. Non-probability Sampling Methods:

- When a **sampling frame is not available** or it cannot be created, in such cases a non-probability sampling method is used. This type has two methods of sampling namely: **convenience sampling** and **quota sampling**.
- Convenience sampling is a method by which, for **convenience sake, the study units that happen to be available at the time of data collection** are selected in the sample.

- Quota sampling is a method by which **different categories of sample units are included to ensure that the sample contains units from all these categories**.
- For example, a quota sample of patients from a health center that might include 20 patients with acute respiratory infection, 20 with diarrhea, and 20 with malaria.

❖ Types of Error in Sampling

- Sampling naturally cannot **select units from a population** and therefore error occurs. A sampling error is a **statistical error that occurs when the researcher does not select a sample** that truly represents the whole population.
- In addition, the results found in the sample do not represent the results that would be **obtained from using the whole population**. Sampling error can be **reduced by randomizing sample selection process** and/or by increasing the number of samples.

Type I Error

- A type I error occurs when the **statistician rejects a null hypothesis (H_0)** that is actually true in the population. When this happens, it is type I error.
- A type I error is also known as **false-positive**.
- Hypothesis testing is the method of testing if variation between two sample distributions can just be **explained through random chance or not**. If it has to conclude that **two distributions vary in a meaningful way**, enough precautions need to be taken to see that the **differences are not just through random chance**.
- The process of hypothesis testing is supported by **identifying type of error and by specifying parametric limits** that how much Type I error will be

permitted. Type I error is that, where lot of care is exercised by minimizing the chance of its occurrence.

- Usually, attempts are made to set **Type I error as 0.05 or 0.01**, so that there is only a **5 or 1 in 100 chance that the variation is due to chance**.
- This is called the **'level of significance'**. In addition, there is no guarantee that 5 in 100 is rare enough so significance levels need to be chosen carefully.
- Type I error is generally reported as the **p-value whose power is derived from random sampling**.

Type II Error

- Type II error is when the **statistician accepts a null hypothesis (H_0) that is false**. It means an investigator **fails to reject a null hypothesis** that is actually false in the population.
- The Type II error is also known as **false-negative**.
- The probability of committing a **Type II error is denoted as beta (β)**, also known as **β error**.
- The probability of not committing a **Type II error is called the power of the test**. The Type II error rate is often **denoted as β**.
- The power of a study is **defined as 1 - β** and is the probability of rejecting the null hypothesis when it is false. **The most common reason for Type II errors is that, the study is too small**.
- The concept of power is really **only relevant when a study is being planned**. Although Type II errors can never be avoided entirely, the investigator can **reduce their probability by increasing the sample size**.
- The larger the sample size, the lesser is the probability that it will **differ substantially from the population**.

❖ Standard Error of Mean

- The **Standard Error of Mean (SEM)** is employed to refer to the **SD of mean or median**.
- For example, the **SEM refers to the SD of the distribution** of sample's mean taken from a population.
- Actually SEM measures how far is the **'sample mean'** of the data from the true **'population mean'** and is usually smaller than the SD.
- SEM is used in clinical trial studies to specify the **characteristics of sample data** and to describe statistical analysis results.
- The **SEM is nothing but statistical inference made, based on the sampling distribution**. The standard error of the mean (i.e. $SD_{\bar{x}}$) is calculated by using following equation.

Standard deviation $\sigma = \sqrt{\dfrac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1}}$

Variance $(n-1) = \sigma^2$

$$SEM\left(\bar{\sigma_+}\right) = \frac{\sigma}{\sqrt{n}}$$

Where,

x is the sample's mean and n is the sample size.

SEM is calculated by taking the **SD** and dividing it by the square root of the sample size. The SD formula involves the following steps:

1. 📌 **Take the square of the difference between each data point and the sample mean and find the sum of those values.**
2. 📌 **Divide that sum by the sample size minus one, called as the variance.**
3. 📌 **Calculate the square root of the variance to obtain SD.**

- SEM is employed as a way to **validate the accuracy of a sample or multiple samples** by estimating deviation within the means.
- It expresses **precision level of mean of the sample** against the true mean of the population.
- **As size of the sample increases**, the SEM decreases against the SD and the true mean of the population is known with **greater specificity**.
- On the contrary, **increasing sample size** provides a more specific measurement of the SD.

However, depending on the dispersion of **the additional data added to the sample, the SD may be more or less. SEM** is the considered part of descriptive statistics that it represents the **SD of the mean within a dataset**. This is used as a measure of variation for **random variables, which provides measurement for the spread.** The smaller is the spread, the more accurate data set is.

UNIT – 2: PARAMETRIC TEST- PART - 3

Points to be covered in this topic

1. **Introduction**
2. **t-test**
   - Sample, Pooled or Unpaired and Paired
3. **ANOVA**
   - One way and Two way
4. **Least Significance difference**

PARAMETRIC TEST

- A parameter in statistics is a **characteristic of a population that refers to feature about a sample.** The population mean is a parameter and the **sample mean is a statistic.**
- For example, the drug **nexium** is used to reduce the acid produced in the stomach and heal damage to the esophagus.
- A researcher wants to **estimate the proportion of patients taking nexium that are healed within 6 weeks.** A random sample of **200 patients suffering from acid reflux disease** is obtained, and **180 of those patients were healed after 6 weeks.**
- The **parameter** is the proportion of patients healed by **nexium in 6 weeks**. The **statistic is 180/200 = 0.9**, the proportion healed in the sample.

Parametric methods refer to those in which a set of data has a **normal vs. a non-normal distribution**, respectively.

- A parametric test is a hypothesis testing procedure based on the **assumption that observed data are distributed according to some distributions** of well-known form up to some unknown parameters on which we want to make **inference such as the mean, or the success probability.**
- For example, the **normal family of distributions** all has the **same general shape** and are parameterized by **mean and SD.** That means, if the mean and SD are known and if the distribution is normal, the probability of any future observation falling in a given range is known.
- Suppose that we have a **sample of 99 test scores with a mean of 200 and a SD of 1.** Assume that all 198 test scores are random sample findings from a normal distribution. Thus, it can be predicted that **there is a 1% chance that the 100th test score will be higher than 102.33 (= 100 + 2.33 SD).**
  This is when the 100th test score comes from the same distribution as the others. The 2.33 SD value is computed from parametric statistical methods, for 99 independent observations from the same normal distribution.

t-Test:

- The t-test is a type of inferential statistic used to determine whether there is a **significant difference between the means of two groups**, which may be related in certain features.
- The t-statistic is mostly used **when the data sets follow a normal distribution and may have unknown variances.** A t-test is a tool used for hypothesis testing to test an assumption applicable to a population.
- The two groups of data must be **independent from one another.** The t-statistic is equal to the difference between **the groups means divided by the standard error of the difference between the group means.** The standard error (SE) sampling distribution expressed as: $\mathbf{SE = SD \sqrt{N}}$

$$t = \frac{Sample\ mean\ difference}{Sample\ SE\ difference}$$

- In order to determine the statistical significance **t-test uses t-statistic, t-distribution values, and degrees of freedom (df).** To conduct a test with three or more means **ANOVA** is required to be used.
- A t-test allows **comparing the average values of the two data sets** and determines if they are from the same population. For example, if we were to take a **sample of capsules from batch A and another sample of capsules from batch B**, we would not expect them to have exactly the same mean and SD.
- Similarly, samples taken from the placebo-fed control group and those taken from the drug prescribed group should have a slightly different mean and SD.
- Mathematically, the t-test considers a **sample from each of the two sets and establishes the problem statement by assuming a null hypothesis** that the two means are equal.
- Using formulas, certain values are calculated and compared against the standard values and the applied assumption of null hypothesis is either accepted or rejected. If the null hypothesis qualifies to be rejected, it indicates that **data readings are strong and are probably not by chance.**

The assumptions of t-test are as follows:

1. **Scale of measurement:** The scale of measurement applied to the data collected follows a continuous or ordinal scale.
2. **Simple random sample:** The data is collected from a representative, randomly selected portion of the **total population.**
3. **Data:** When data is plotted, it results in a normal distribution with bell-shaped distribution curve.

4. **Homogeneity of variance:** Homogeneous or equal variance exists when the **SDs of samples are approximately equal.**

- In order to calculate outcome for t-test requires three data values namely: difference between the mean values from each data set (mean difference), SD of each group and the number of data values of each group.
- The outcome of this calculation is t-value. This t-value is compared with a value obtained from a critical value table (**t-distribution table from ANOVA**).
- This comparison is used to determine the effect of **chance alone** on the difference as well as to know whether the difference is outside that chance range. The t-test enquires whether the difference between the groups represents a true difference in the study or if it is possibly a **meaningless random difference.**

There are three main types of t-test that are categorized as dependent and independent t-tests:

1. **An independent samples t-test compares the means for two groups.**
2. **A one sample t-test tests the mean of a single group against a known mean.**
3. **A paired sample t-test compares means from the same group at different times.**

When it comes to testing hypotheses regarding two population means, the most commonly used t-test is the two-sample t-test.

There are two versions of this test:

1. **Pooled test:** It is used when the variances of the two populations are equal.

2. **Unpooled t-test:** It is used when the variances of the two populations are unequal.

One sample t-Test:

- A one sample t-test is a parametric test that determines whether the **sample mean is statistically different from a known** (or hypothesized) population mean.
- This test is also known as **single sample t-test and the variable used in this test is known as test variable.** The test variable is compared against a test value, which is a **known or hypothesized value of the mean** in the population.
- **This test can only compare a single sample mean to a specified constant.** The one-sample t-test can be used when the population variances are equal or unequal, and with large or small samples.

The one sample t-test is commonly employed to test the following:

1. Statistical difference between a sample mean and a known (or hypothesized) value of the mean in the population.
2. Statistical difference between the sample mean and the sample midpoint of the test variable.
3. Statistical difference between the samples mean of the test variable and chance.
4. Statistical difference between a change score and a zero.

The data for one sample t-test must meet the following requirements:

1. Test variable (interval or ratio level) must be continuous.

2. Scores on the test variable need to be independent of observations.
3. Random sampling of data from the population.
4. There must be **normal distribution of the sample and population** on the test variable.
5. There must be **homogeneity of variances.**
6. There must be **no outliers.**

This test can be used as follows

1. Defining hypotheses: The **Table** shows sets of **null and alternative hypotheses**. Each set makes a statement that how the true **population's mean** (μ) is related to hypothesized value M.

Set	H_0	H_1	Number of tail
1	$\mu = M$	$\mu \neq M$	2
2	$\mu \geq M$	$\mu < M$	1
3	$\mu \leq M$	$\mu > M$	1

---

2. Specifying significance level: The significance levels should be between 0 and 1 such as equal to 0.01, 0.05, or 0.1 etc.

3. Finding df: The df is (n - 1) where, n is the number of observations in the sample.

4. Calculating test statistic: The **test statistic** is a **t-statistic** calculated using the following equation.

$$t = \frac{\bar{x} - M}{\frac{SD}{\sqrt{n}}}$$

Where, *x* is the observed sample mean, M is the hypothesized population mean from the null hypothesis, and SD is the standard deviation of the sample

5. Computing p-value: The **p-value** is the probability of observing a sample **t statistic as extreme as the test statistic**. Since the test statistic is a t-statistic, assess the probability associated with the t-statistic, having the degrees of freedom computed above.

---

6. Evaluating null hypothesis: The evaluation involves **comparing the p-value to the significance level**, and rejecting the null hypothesis when the p-value is less than the significance level. The null hypothesis (H_0) and (two-tailed) alternative hypothesis (H_1) of the one sample t-test can be expressed as:

H_0: $\mu = \bar{x}$ (the sample mean is equal to the proposed population mean)

H_1: $\mu \neq \bar{x}$ (the sample mean is not equal to the proposed population mean)

Pooled t-Test:

Pooled test is an activity to compare the means of two independent samples using a technique known as the Pooled t-test. This test requires a number of assumptions.

1. The first sample of size n_1 is drawn from a normal population with mean μ_1 and variance sigma σ^2.
2. The second sample of size n_2 is also drawn from a normal population with a mean μ_2 and variance σ^2.
3. The two samples are **independent**.

The pooled test is used when the variances of the **two populations are equal**.
The pooled test has limited sensitivity because it deviates from the assumptions of equal population variances. But the overall performance of this test is significantly superior to that of the unpooled t-test. When two populations have or assumed to have the same variance, the **t-value and df for equal variance t-test for the equality of the two population means** is calculated as:

$$t = \frac{(\bar{x}_1 - \bar{x}_2)(\mu_1 - \mu_2)}{S_P = \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

In this test, n_1 $\bar{x}_1$ and n_2 $\bar{x}_2$ are the sample sizes and the sample means of the first and the second sample, respectively, whereas μ_1 and μ_2 are the corresponding population means. S_p^2 The quantity is the pooled sample variance (SD) defined by the following equation:

$$S_P^2 = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 + n_2 - 2)}}$$

Where, S_1^2 and S_2^2 are the sample variances of the first and the second sample, respectively. Given the two populations are normal, the t-statistic mentioned above certainly has an exact t-distribution with a df equal to (n_1 + n_2 - 2). If the null hypothesis H_0: $\mu_1 = \mu_2$ is true. That means if the two populations have the same means, then above test statistic reduces to:

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{S_P = \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Paired t-Test:

- The **paired t-test** is performed when the **samples typically consist of matched pairs of similar units**, or when there are cases of repeated measures, or compares two means that are from the same individual, object, or related units.
- For example, there may be instances that **the same patients being tested repeatedly**; before and after receiving a particular treatment.
- In such situations, each patient is being used as a control sample against themselves.This method also applies to cases where the samples are related in some manner or have similar individuality, for example, comparative analysis involving children, parents or siblings.
- Paired t-tests are of a **dependent type**, because the two sets of samples are related.This test is used to determine whether there is statistical evidence that the **mean difference between paired observations on a particular outcome** is significantly different from zero.
- The variable in this test is **measured at two different times or for two related conditions or units**.The paired samples t-test has been commonly used to test the statistical difference between two time points, two conditions, two measurements and matched pair, respectively.
  The test can only compare the means for **two paired units on a continuous outcome that is normally distributed**.

The data for paired t-test must meet the following requirements:

1. Dependent variable is continuous (interval or ratio level).
2. Related samples/groups (dependent observations).
3. Random sample of data are from the population.
4. The sampling distribution of the mean difference between data pairs is approximately normally distributed.

5. No outliers in the difference between the two paired groups.

The hypotheses can be expressed in two different ways that describe same idea and are mathematically equivalent:

- **$H_0: \mu_1 - \mu_2$** (the paired population means are equal)
- **$H_1: \mu_1 - \mu_2$** (the paired population means are not equal)

The formula for calculating the t-value and df for a paired t-test is:

$$t = \frac{\bar{x}_{diff} - 0}{S_{\bar{x}}}$$

$$where, S_{\bar{x}} = \frac{S_{diff}}{\sqrt{n}}$$

$\bar{x}_{diff}$ = **Sample mean of the difference**

n = **Sample size (no. of observation)**

S_{diff} = **Sample standard deviation of the difference**

$S_{\bar{x}}$ **Estimated standard error of the mean** $\frac{SD}{\sqrt{n}}$

The calculated **t-value** is compared to the critical t-value with **df = (n - 1)** from the **t-distribution table** for a chosen confidence level. If the t-value is greater than the critical t-value, the **null hypothesis is rejected** with conclusion that the means are **significantly different**.

Unpaired t-Test:

- An unpaired t-test, also known as an **independent parametric t-test**, is a statistical procedure that compares the **means of two independent or unrelated groups** to determine if there is a significant difference between the two. For example, comparison of average weights of tablets grouped by **size** as **small and large size tablets**, **which are two independent groups**. The independent samples t-test has two different forms:

1. **Standard Student's t-test**: It assumes that the variance of the two groups is equal.
2. **Welch's t-test**: The variance is not assumed to be same in the two groups, which results in the fractional degrees of freedom.

The unpaired t-test hypotheses are similar to those for a paired t-test:

1. The **null hypothesis (H_0)** states that, there is **no significant difference** between the means of the two groups.
2. The **alternative hypothesis (H_1)** states that, there is a **significant difference** between the two population means, and that this difference is unlikely to be caused by sampling error or by chance.

Assumptions of an unpaired t-test are as follows:

1. The dependent variable is **normally distributed**.
2. The observations are **sampled independently**.
3. The dependent variable is measured on an **incremental level**, for example, ratios or intervals.
4. The **variance of data is the same** between groups (same SD).
5. The independent variables must consist of **two independent groups**.

Examples of unpaired t-test are given below:

1. A pharmaceutical study or any other treatment plan, wherein half of the subjects are allocated to **treatment group** and half of subjects are randomly allocated to **control group**.
2. Research during which there are two independent groups, such as **women and men**, which examines whether the **average bone density** is significantly different between the two groups.

- In the case of **unequal variances**, a **Welch's test** should be used.
- In the unpaired t-method tests the null hypothesis is that the population means of two independent random samples from **normal distribution** (approximately) are equal.Considering **equal variances**, the test statistic is calculated as:

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$s^2 = \frac{\sum_{i=1}^{n_1} (x_i - \bar{x}_1)^2 + \sum_{i=1}^{n_2} (x_j - \bar{x}_2)^2}{(n_1 + n_2 - 2)}$$

Where, x1 and x2 are the sample means, S2 is the pooled sample variance, n1 and n2 are the sample sizes and t is a Student t quantile with (n1 + n2 - 2) degrees of freedom i. The test power is calculated as the power achieved with the given sample sizes and variances for finding the observed difference between means with a two-sided type I error probability of (100- CI%) %.

This test should not be used in cases with significant difference **between** the variances of the two samples.

Considering unequal variances, the test statistic can be calculated as follows:

$$d = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

$$df = \frac{\left[\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right]^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}}$$

$$S_1^2 = \frac{\sum_{i=1}^{n_1}(x_i - \bar{x}_1)^2}{(n_1 - 1)}$$

$$S_2^2 = \frac{\sum_{j=1}^{n_2}(x_j - \bar{x}_2)^2}{(n_2 - 1)}$$

ANOVA

- **ANOVA** (analysis of variance) is used to **test the differences between two or more means**.
- The name **ANOVA** is appropriate because conclusions about means are made by **analyzing variance**. This test can be helpful for analyzing the variance of a **single dependent variable across three or more samples or data sets**.
- It shows few Type-I errors and is appropriate for a variety of issues. **ANOVA groups the differences by comparing means of each group** and includes spreading out the variance into different sources. There are two main types of ANOVA namely; **one-way ANOVA** and **two-way ANOVA**.

One-way ANOVA:

- **One-way ANOVA** is between groups and is used when we want to **test two groups to see if there is a difference between them**.
- This method compares the **means of two or more independent groups** and determines whether there is any statistical evidence that the associated population means are **significantly different**.
- One-way ANOVA is a **parametric test** and is also known as **one-factor ANOVA**, one-way ANOVA, between subjects ANOVA.
  The variables used in this test are known as dependent variable and independent variable which divides cases into two or more mutually exclusive levels, or groups.
- The one-way ANOVA is used to analyze data from field studies, **experiments and quasi-experiments**.
- It is frequently used to test the **statistical differences among the means of two or more groups**, among the means of two or more interventions and among the means of two or more change scores, respectively.As one-way ANOVA can compare the means for two groups, it also can compare the means **across three or more groups**.

The data for one way ANOVA must meet the following requirements:

1. It must have dependent variable that is continuous (interval or ratio level).
2. It must have independent variable that is categorical (two or more groups).
3. It should have cases with values on both the dependent and independent variables.
4. There should not be any relationship between the subjects in each sample.
5. The sampling method from the population must be random data sampling.

6. There should be normal distribution of the dependent variable for each group for each level of the factor.
7. The population must have homogeneity of variances.
8. There should be no outliers.

The null and alternative hypotheses of one-way ANOVA can be expressed as:

- **Ho: μ1 μ2 μ3 = ... = μk** (all k population means are equal).
- **H_1:** At least one **μ_i** must be different (at least one of the K population means is not equal to the others).

Where, **μ_i** is the population mean of the i^{th} group (i = 1, 2,...,k).

Two-way ANOVA:

- Two-way ANOVA is used **with or without replicates when there is one group** and we wish to double-test that same group.
- A **two-way ANOVA** is an extension of the one-way ANOVA as it is a **hypothesis-based test**. It is used to determine the effect, or testing differences, of **two nominal independent variables on a continuous dependent variable**.
- This test is used when **we have one measurement variable (quantitative) and two nominal variables**.For example, suppose we want to find out if there is **any interaction between amount of binder and lubricant for hardness level of tablets**.
- The **hardness level** is the outcome variable that can be measured. Binder and lubricants are the two categorical variables.**These categorical variables are also called the independent variables or factors**.

- The factors can be divided into levels.In the above example, lubricant level could be split into three levels: **low, middle and high amount**.Binder could be divided into three levels: **natural, semisynthetic, and synthetic**.
- **Treatment groups** are all possible combinations of the factors.
  In this example, there would be 2 x 2 = 4 or 3 x 3 = 9 treatment groups.

  Usually random factors are considered to have no statistical influence on a data set, while systematic factors have a statistical significance.
  Using ANOVA, a statistician is able to establish whether the variability of the outcomes is due to chance or due to the factors in the analysis.
- ANOVA test results are used in an **F-test on the significance of the regression** formula overall.
- The two-way ANOVA result calculates a **main effect** as well as an **interaction effect**.The **main effect** is similar to that of one-way ANOVA and an individual factor's effect is measured separately.
- For the two-way interaction effects all factors are measured at the same time. If there is more than one observation in each cell then interaction effects between the factors are easy to test.
- Multiple scores for factor could be entered into cells and in such cases the number in each cell must be equal.

Two null hypotheses are tested by placing one observation in each category. In such case, those hypotheses are:

- **H_{01}:** All the lubricant strength groups have equal hardness.
- **H_{02}:** All the binder groups have equal hardness.

For multiple observations in these categories, a third hypothesis can be tested

H_{03}: Factors are independent or the interaction effect does not exist between them.

The F-statistic is computed for testing all those three hypotheses.

The assumptions of Two-way ANNOVA are as follows:

1. **The population must be close to a normal distribution** that each sample is taken from a normally distributed population.
2. **Samples must be independent** that each sample has been drawn independently of the other samples.
3. **Population variances must be equal** that the variance of data in the different groups should be the same.
4. **Groups must have equal sample sizes.**
5. **Dependent variable should be measured on a scale** which is subdivided using increments.
6. **Two independent variables should be in categorical, independent groups.**

◇ Least Significance Differences

- The least significant difference (LSD) test is used in the context of the ANOVA, when the F-ratio suggests rejection of the null hypothesis.
- This means when the **difference between the population means is significant**. LSD is basically an extended **Student's t-test** which uses a more comprehensive estimate of the error in the data for making statistically valid mean comparisons.
- The basic idea of this test is to compare the populations taken in pairs. It is then used to proceed in a **one-way or two-way ANOVA**.
- **LSD is the value at a particular level of statistical probability**, for example, **$P \leq 0.01$; means with 99% accuracy**, when exceeded by the difference between two category means for a particular characteristic, then the two categories are said to be different for that characteristic at that or lesser levels of probability.

The difference between two means is stated significant at any desired level of significance if it exceeds the value derived from the following formula:

$$LSD = \frac{r\left(s\sqrt{2}\right)}{\sqrt{n}}$$

- LSD procedure being a **two-step testing procedure**, it is used for pairwise comparisons of several treatment groups.
- In the first step, a **global test is performed for the null hypothesis** that the expected means of all treatment groups under study are equal.
- If this global null hypothesis can be rejected at the pre-specified level of significance, then in the **second step all pairwise comparisons are performed at the same level of significance**.
- LSD is known to protect the experiment-wise **Type I error rate** at the nominal level of significance, if the number of treatment groups is three. Therefore, LSD procedure may be applied to **Phase-III clinical trials** to compare two doses of an active treatment against placebo in the confirmatory sense.

Global test:

- The global test is meant for data sets in which **many features (or covariates) have been measured for the same subjects**, together with a response variable, for example a **class label, a survival time or a continuous measurement**.
- The global test can be used on a **group (or subset) of the covariates**, testing whether that group is associated with the response variable.

- The null hypothesis of the global test is that **none of the covariates in the tested group is associated** with the response.
- The alternative is that, **at least one of the covariates has such an association**.

Uses of LSDs:

- LSDs allow data to be observed at closely, without having sufficient prior statistical knowledge.**It is a simple calculation that allows the means of two or more pre-determined varieties to be compared.**
- It helps to identify the populations whose means are statistically different. LSD used to evaluate the result of chance and gain confidence that the inferences drawn from the data are correct.

Limitations to using LSDs:

1. LSDs must not be applied until the F-test or t-test shows significant differences between means.
2. LSDs are only valid for testing mean comparisons that were pre-set in the goal of the experiment.
3. LSD is reasonably acceptable for comparing each category separately with a standard control.
4. LSDs are at their most robust when more than two categories are involved.
5. LSD is used only when category means are arranged in order of their magnitudes.