Unit 4

Random Variable

A rule that assigns a real number to each outcome of a random experiment is called a random variable (r.v.). It can take any one of the various possible values each with definite probability. For example, in a throw of a die if X denotes the number obtained then X is a random variable which can take any one of the values 1, 2, 3, 4, 5, 6, each with equal probability 1/6. A random variable is usually denoted by any of the Capital Latin letters X, Y, Z, U, V... etc and particular values which the random variable takes are denoted by the corresponding small letters.

Let us consider a random experiment of three tosses of a coin then the sample space S consist of 2^3 =8 sample points given as

S ={HHH, HHT,HTH,THH, HTT,THT, TTH, TTT}.

Let us consider the variable X, which is the number of heads obtained, then X is a random variable which can take any one of the values 0, 1, 2, 3.

Outcomes	HHH	HHT	HTH	THH	HTT	THT	TTH	TTT
Values of X	3	2	2	2	1	1	1	0

Types of random Variables

There are two types of random variable namely

(a) Discrete random variable

(b) Continuous random variable

Discrete random variable

A random variable, say X, which can take a finite or a countably infinite numbers of values in an interval of the real line is called discrete random variable. For example, if we toss a coin, the variable can take only two values 0 and 1 assigned to tail and head respectively.

$$\text{i.e. } X = \begin{cases} 0 & \text{if x is Tail} \\ 1 & \text{if x is head} \end{cases}$$

In a rolling of a die, only six values of the variable x, i.e. 1,2,3,4,5&6 are possible outcomes, hence the variable x is discrete , here the variable X = { x: x is 1,2,3,4,5&6}

Continuous random Variable

A random variable X, which can take infinite and uncountable set of values in an interval of real line is said to be a continuous random variable , the probability of a point x is zero i.e. P(X=x) = 0. But the probability is ascribable in an interval. For example, the weight of middle aged people in Kathmandu lying between 40 kg and 150 kg is a continuous random variable .Symbolically X ={ x: 40 $\leq$ x $\leq$ 150}

Properties of random variable:

The important properties of random variable are as follows,

i. If X is a random variable and a and b are any two constant, then aX+b is also a random variable.

ii. If X is a random variable, then X^2 is also a random variable.

iii. If X is a random variable then 1/ X is also a random variable.

iv. If X and Y are random variables defined on the same sample space S then, X+Y, X-Y, aX, bY and aX+bY are also random variables with a and b are non negative constants.

Probability mass function (pmf)

If X is a one –dimensional discrete random variable taking at most a countable numbers x_1, x_2............... x_n with each value of the variable X , we associate a number $P(x_i)$ = $P(X = x_i)$; i = 1, 2,……….n. Which is known as the probability of x_i and satisfying the following two conditions :

(i) All P (x_i) are non-negative

(ii) $\Sigma P(x_i) = 1$ i.e. the total probability is unity.

Then P (x_i) = P(X = x_i) is called probability mass function of random variable X. The set of ordered pairs $\{x_i, P(x_i)\}$ i = 1, 2,……….n specifies the probability distribution of the random variable X.

Probability density function (pdf)

In case of continuous random variable, we do not talk of probability at a particular point, (which is always zero), but we always talk of probability in an interval, If X is continuous random variable and $f_X(x)$ is a continuous function of X. Then $f_X(x)$ dx, gives the probability of the event that X lies in the interval

$$\left(x - \frac{dx}{2}\ \&\ x + \frac{dx}{2}\right) \text{ i.e. } f_x\,(x)\,dx = P\left(x - \frac{dx}{2} \pounds X \pounds x + \frac{dx}{2}\right)$$

Then f_x (x) or simply f(x) is called probability density function (pdf). It is also known as frequency function because it also gives the proportion of units lying in the interval $x - \frac{dx}{2}\ \&\ x + \frac{dx}{2}$.

If X has the range [a, b], then

$\int_a^b f(x)\,dx = 1$, it implies that the total area under the frequency curve is always unity.

Example

If probability function of a variable X is defined as

$f(x) = \frac{3}{4}x(2 - x)$, $0 < x < 2$, , test whether f(x) is pdf or not.

Solution

If given probability function is probability density function then we must have

$$\int_a^b f(x)\,dx = 1.$$

$$\text{L.H.S.} = \int_0^2 \frac{3}{4}x(2 - x)\,dx = \frac{3}{4}\left[\frac{2x^2}{2} - \frac{x^3}{3}\right]_0^2 = \frac{3}{4}\times\frac{4}{3} = 1$$

Hence, f(x) is pdf.

Distribution Function

A function $F_X(x)$ of a random variable X for a real value x giving the probability of an event $(X \leq x)$ is called a cumulative distribution function (cdf) or simply distribution function, symbolically, $F_X(x) = P(X \leq x)$, obviously X lies in the interval $[-\infty, x]$

It is to be noted that

- If a and b are two constant values such that a < b and F is the distribution function then $P(a < X \leq b) = F(b) – F(a)$
- If F(x) is the distribution function of a mono-variate X , then $0 \leq F(x) \leq 1$

Discrete distribution Function

If discrete random variable X takes sample points $x_1, x_2 \ldots\ldots\ldots\ldots x_n$ and the number p_i is the probability function satisfies the property $p_i \geq 0$ and $\Sigma\ p = 1$ such that $F_X(x) = P(X \leq x)$ is called the discrete distribution function of random variable X.

Continuous Distribution Function

If X is a continuous random variable with the pdf f(x) then the function F(x) defined as $F(x) = P\ (X \leq x) = \int_{-¥}^{x} f(x)\,dx$ is called the distribution function (df) or sometimes the cumulative distribution function (cdf) of the random variable X.

Example: A random variable X has the following probability function

x	0	1	2	3	4	5	6	7	8
P(x)	k	3k	5k	7k	9k	11k	13k	15k	17k

(i) Determine the values of k.

(ii) Find $P(X < 3)$, $P(X \geq 3)$, $P(0 < X < 5)$.

(iii) Find the distribution function of X.

Solution

i. If given p(x) is pmf

then total probability is unity $\sum_{i=0}^{¥} p(x) = 1$

$$\text{or, } 81k = 1, \therefore\ k = \frac{1}{18}.$$

ii. $P(X<3) = P(X = 0) + P(X = 1) + P(X = 2) = \frac{1}{81} + \frac{3}{81} + \frac{5}{81} = \frac{1}{9}$

$P(X \geq 3) = 1 - P(X<3) = 1 - \frac{1}{9} = \frac{8}{9}$

] and $P(\ 0 < X < 5) = P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4)$

$$= \frac{3}{81} + \frac{5}{81} + \frac{7}{81} + \frac{9}{81} = \frac{24}{81}$$

iii. The distribution function F(x) of random variable X is given in the adjoining Table

X	P(x)	F(x)= P $(X \leq x)$
0	0.012	0.012
1	0.037	0.049
2	0.062	0.111
3	0.086	0.198
4	0.111	0.309
5	0.136	0.444
6	0.160	0.605
7	0.185	0.790
8	0.210	1.000

Mathematical Expectation

Once we have constructed the probability distribution for a random variable, we often want to compute the mean or expected value of the random, variable. The expected value of a discrete random variable is a weighted average of all possible values of the random variable, where the weights are the probabilities associated with the corresponding values.

If X is a random variable which can assume any one of the values x_1, x_2.............. x_n with respective probabilities P (x_1), P (x_2)…………P (x_n). Then the mathematical expectation of r v X usually called the expected value of X and denoted by E(X) and it is defined as

$$E(X) = x_1 P(x_1) + x_2 P(x_2) + \dots\dots + x_n P(x_n) = \square xP(x) \text{ where } \sum_{i=1}^{n} p(x) = 1$$

i.e. total probability is unity

Similarly, if X is continuous random variable with probability density function f(x) then mathematical expectation of r.v X is defined as

$$E(X) = \int_a^b xf(x)\,dx \text{ where } f(x) \geq 0 \text{ for every } x \in [a, b] \text{ and } \int_a^b f(x)\,dx = 1$$

Physical meaning of Mathematical Expectation

Let us consider the following frequency distribution of the random variable X as

X:	x_1	x_2	………………	x_n
f:	f_1	f_2	…………………	f_n

Then the mean of the distribution is given by

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + \dots\dots f_nx_n}{N}$$

$\Rightarrow \bar{x} = \frac{f_1}{N} x_1 + \frac{f_2}{N} x_2 + \dots\dots + \frac{f_n}{N} x_n$, we observed that, out of total of N cases f_i cases are Favourable to x_i. ∴ $P(X = x_i) = \frac{f_i}{N} = P(x_i)$, i= 1, 2…………n.

Which implies that, $\frac{f_1}{N} = P(x_1)$, $\frac{f_2}{N} = P(x_2)$ .... $\frac{f_n}{N} = P(x_n)$

Therefore $\bar{x} = x_1P(x_1) + x_2P(x_2) + \dots\dots + x_nP(x_n) = E(X)$.

Hence mathematical expectation of a random variable is nothing but its arithmetic mean.

Vaiance of a random variable

$$Var(X) = E(X^2) - [E(X)]^2$$

Where, $E(X^2) = \Sigma x^2P(x)$

$E(X) = \Sigma xP(x)$

Example In three tosses of a coin, Let X, be the number of heads. Tabulate the possible outcomes with the corresponding value of X. By simply counting & derive the probability distribution of X and hence calculate the expected value and variance of X.

Solution: let H represents a head, T a tail and X, the random variable denoting the number of heads. Then the following table shows the value of random variable with sample points.

S. No.	outcome	No. of heads
1	HHH	3
2	HHT	2
3	HTH	2
4	THH	2
5	HTT	1
6	THT	1
7	TTH	1
8	TTT	0

The random variable X takes the value 0, 1, 2, 3. since from the above table, we find that the number of cases favorable to the coming of 0, 1, 2 & 3 heads are 1, 3, 3 and 1 respectively . Therefore $P(x = 0) = \frac{1}{8}$, $P(X = 1) = \frac{3}{8}$, $P(X = 2) = \frac{3}{8}$, $P(X = 2) = \frac{1}{8}$ then probability distribution of X can be summarized as follows.

X	0	1	2	3
P(x)	1/8	3/8	3/8	1/8

Then $E(X) = \Sigma xP(x) = 0 \times \frac{1}{8} + 1 \times \frac{3}{8} + 2 \times \frac{3}{8} + 3 \times \frac{1}{8} = \frac{12}{8} = 1.5$

And $E(X^2) = \Sigma x^2P(x) = 0^2 \times \frac{1}{8} + 1^2 \times \frac{3}{8} + 2^2 \times \frac{3}{8} + 3^2 \times \frac{1}{8} = 3$

Hence $Var(X) = E(X^2) - [E(X)]^2 = 3 - (1.5)^2 = 0.75$.

Thus expected value of random variable 1.5 & it's variance is 0.75

5.6 t -distribution

Let $X_1, X_2 \ldots\ldots\ldots\ldots X_n$ be a random sample of size n drawn from a normal population with mean μ and variance σ^2 . Then student's 't' statistics is defined by

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$ follows student t-distribution with (n-1) degrees of freedom where, $\bar{x} = \Sigma X/n$ and

$$S^2 = \frac{1}{n-1}\Sigma(X - \bar{X})^2 \text{ and } s^2 = \frac{1}{n}\Sigma(X - \bar{X})^2$$

$$\Rightarrow ns^2 = (n - 1)\, S^2$$

S = Unbiased sample standard deviation
= Unbiased estimate of population standard deviation
s = Biased sample standard deviation
= Biased estimate of population standard deviation

And the probability density function of t is defined as

$$f(t) = \frac{1}{\sqrt{\nu}} \frac{1}{\beta\left(\frac{1}{\nu}, \frac{\nu}{2}\right)(1 + t^2/\nu)^{\frac{\nu+1}{2}}} \qquad -\infty < t < +\infty$$

and $\nu = (n - 1)$ degrees of freedom.

When sample size n is less than 30, we consider sample is small and use t-test

Properties of t-distribution

The important properties of t-distribution are given as follows

i. Like Z- distribution, t-distribution is a continuous distribution having symmetrical and bell shaped curve. The value of t ranges from. $-\infty$ to ∞ i.e. $-\infty < -t < \infty$

ii. As $n \to \infty$, the t-distribution tends to standard normal distribution.

iii. Mean of the t-distribution is zero and variance is equal to $\frac{\nu}{\nu - 2}$; $\nu > 2$

iv. The t-distribution is symmetrical i.e. $\beta_1 = 0$.

v. The t-distribution curve is unimodal with mean = mode = median

vi. The t-axis is the asymptote of the t-distribution.

vii. The t-distribution is flatter than the normal distribution and there is a different t-distribution for every possible sample size. For example a t-distribution for a sample size of 15 is different than the t-distribution for a sample size of 2 and so on.

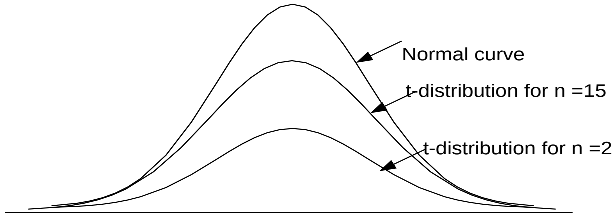


Fig: of t-distribution as compared with normal curve

viii. The t-distribution can be used even in case of large sample but the large sample theory can't be used for small sample.

ix. The very interesting property of the sampling distribution of t is that it does not depend on the population parameter and depends only on $\nu = n - 1$ i.e. on the sample size.

7.2.5 Degrees of Freedom

The number of independent observations in a set is called degrees of freedom (d.f.)

For example, we select two samples values x_1 and x_2 such that their mean is 30 then $\frac{x_1 + x_2}{2} = 30$. How can we find what value of x_1 and x_2 can take in this situation?

The answer is that x_1 and x_2 can be any two values whose sum is 60. Suppose x_1 has value of 20 then x_2 is no longer free to take on any value but must have value of 40 i.e. if $x_1 = 20$ then $\frac{x_1 + x_2}{2} = 30$ or $20 + x_2 = 60$ then $x_1 = 40$.

This example shows that when there are two elements in a sample and we know the sample mean of these two elements; we are free to specify only one of the elements to find the other element. So, we can say that there is one degree of freedom.

So with two sample values, number of degree of freedom = 2 -1 = 1.
If we have three sample values, number of degree of freedom = 3 -1 = 2.
If we have four sample values, number of degree of freedom = 4 -1 = 3.
In general for n sample values, number of degree of freedom = n -1 .

For t-distribution in estimating population mean, the number of degree of freedom (d.f.) = n-1, where n is the sample size.

Assumptions about t-distribution

For derivation of student's t-distribution following basic assumption are made

i. The parent population from which the sample is drawn is normal with mean μ and variance σ^2.

ii. All observations in the sample are independent i.e. one item selected in the sample doesn't effect the other items included in the sample.

iii. The sample size is small i.e. less than 30 as a usual practice. Also the sample should not be less tan 5 observations.

iv. The hypothetical value μ_0 of μ is a correct value of population mean of parent population form which the sample are drawn.

v. The sample values are correctly measured and recorded.

vi. The value of population standard deviation σ is unknown.

7.3.2 Application of t-distribution:

The student's t-distribution has wide number of applications in statistics; some of them are given below

(i) To test whether the sample mean ($\bar{x}$) differs significantly from the hypothetical value of the population mean (μ) or not. Population variance being unknown.

(ii) To test the significance of the difference between two independent means. The population variance being equal but unknown.

(iii) To perform Paired t-test for difference of two means.

(iv) To test the significance of an observed sample correlation coefficient

(vi) To test the significance about the regression coefficient s

t-test for single mean

We can use the t-test for single mean if random sample $x_1, x_2 ....x_n$ of size n has been drawn from a normal population with a specified mean say μ_0

Following are the steps for test of significance of a single mean in small sample (t-test)

i. **Set up the null hypothesis H_0: $\mu = \mu_0$** i.e. the sample has been drawn from the population with mean μ_0 or there is no significant difference between the sample mean $\bar{x}$ & the population mean μ_0.

ii. **Set up Alternative hypothesis** $H_1 : \mu \neq \mu_0$ (Two tailed)

or, $H_1 : \mu > \mu_0$ (right tailed)

or, $H_1 : \mu < \mu_0$ (left tailed)

i.e. there is significance difference between the sample means

iii. Fix the level of significance α

iv. Under the null hypothesis $H_0 : \mu = \mu_0$ the test statistic

$$t = \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n-1}}}$$, follows student's t-distribution with (n-1) d.f.

where, $\bar{x} = \frac{\sum x}{n}$ = sample mean &

$$S^2 = \frac{1}{n-1}\Sigma(x - \bar{x})^2 \ \& \ s^2 = \frac{1}{n}\Sigma(x - \bar{x})^2$$

v. For α level of significance & (n-1) d.f. the critical value of t from table is t(α).

vi. conclusion: If $|t| < t_\alpha$ we accept H_0 at α level

If $|t| \geq t_\alpha$ we reject H_0 & accept H_1

Confidence limit for estimating population means μ for small sample

(1 - α)% confidence limit to estimate the population mean μ for small sample is given by

$$\bar{x} \pm t_v(\alpha)\frac{S}{\sqrt{n}}$$ or,

Thus 95% fiducial limits

$$\bar{x} \pm t_v\,(0.05)\frac{S}{\sqrt{n}}$$

and 99% confidence or fiducial limits $\bar{x} \pm t_v(0.01)\frac{S}{\sqrt{n}}$

If biased sample standard deviation is given then,

$$\bar{x} \pm t_v\,(\alpha)\frac{s}{\sqrt{n-1}}$$

Thus 95% fiducial limits

$$\bar{x} \pm t_v\,(0.05)\frac{s}{\sqrt{n-1}}$$

and 99% confidence or fiducial limits $\bar{x} \pm t_v(0.01)\frac{s}{\sqrt{n-1}}$

t- test for difference of two independent sample means

If we want to test if two independent samples have been drawn from two normal populations having the same means, the population variances being equal but unknown, we use t-test for difference of means.

Let $X_1, X_2 ....X_{n_1}$ & Y_1, Y_2 .... Y_{n_2} be two independent random samples from the given normal population when sample size of each sample is less than 30. Then we apply the following steps for test of significance for different of mean in small sample.

i. Null hypothesis $H_0 : \mu_1 = \mu_2$ i.e. sample has been drawn from the normal population with same mean

ii. Alternative hypothesis $H_1: \mu_1 \neq \mu_2$ (Two tailed)

or, $H_1: \mu_1 > \mu_2$ (right tailed test)

or, $H_1: \mu_1 < \mu_2$ (left tailed test)

iii. Fix the level of significance α.

iv. Under the null hypothesis H_0: $\mu_1 = \mu_2$ & the assumption that $\sigma_1^2 = \sigma^2_2 = \sigma^2$ two population variance are equal but unknown the test statistics is

$$t = \frac{\overline{x} - \overline{y}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Where, S =

n_1 = First sample size
n_2 = Second sample size

$\overline{x}$ = Mean of first sample

$\overline{y}$ = Mean of second sample

when unbiased sample variance S_1^2 & S_2^2 are known then

S^2 =

Also when biased sample variance s_1^2 & s_2^2 are known then

$$S^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$$

v. For α level of significance & $(n_1 + n_2 - 2)$ d.f. find the critical value of t from table.

vi. Conclusion: If $t_{cal} < t_{tab}$ we accepts H_0 at α level of significant
& If $t_{cal} \geq t_{tab}$ we reject H_0 & accept H_1

Confidence limit for difference of means

(1-α)% confidence limit for difference of population mean from small sample is given as

$$(\overline{x} - \overline{y}) \pm t_{tab}\, S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

95% confidence limit for different of population mean is

$$(\overline{x} - \overline{y}) \pm t_{\nu}(0.05)\, S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

and 99% confidence limit for different of population mean is

$$(\overline{x} - \overline{y}) \pm t_{\nu}(0.01)\, S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Paired t-test for difference of two dependent means

(Paired t-test)

In the t-test for difference of independent means, the two samples were independent of each other. However, there are many situations where the samples are pair wise dependent to each other. For example, if we are testing sales of goods after & before advertisement; a testing the productivity level of workers before & after a training program or checking memory capacity of person before & after training and so on. Then their values are related to each other. In such situations, we are concerned with the difference between the pair of related observations instead of the value of the individual observation.

The following are the steps in testing paired t test for difference of mean as:

1. Set up the null hypothesis H_0: $\mu_1 = \mu_2$ or i.e. there is no significance difference in the observation before & after treatment.

2. Set up the Alternative Hypothesis H_1: $\mu_1 \neq \mu_2$ i.e. there is significance difference in the observation before & after treatment

or

H_1: $\mu_1 > \mu_2$ i.e.

or

H_1: $\mu_1 < \mu_2$

3. Fix the level of significant α

4. Test statistic

test statistic is, $t = \dfrac{\overline{d}}{\frac{S}{\sqrt{n}}}$ follow student t=distribution with (n-1) d.f.

where, d = x - y difference between two set of observations.

$$\overline{d} = \frac{\Sigma d}{n} \;\&\; S^2 = \frac{1}{n-1}\,\Sigma\,(d - \overline{d})^2 = \frac{1}{n-1}\left[\Sigma\, d^2 - \frac{(\Sigma d)^2}{n}\right]$$

5. For α level of significant & (n -1) d.f. find critical value t as $t_{\sqcap}(\alpha)$
if $t_{cal} < t_{tab}$ (We accept null hypothesis)
$t_{cal} \geq t_{tab}$ (We reject null hypothesis)

8.1.5 t-test for significant of an observed sample correlation coefficient (Test of correlation coefficient)

In order to test whether the sample correlation coefficient (r) is significant of any correlation between the variable in the population t-test for significant of an observed sample correlation coefficient is applied. The steps for testing of significance of an observed sample correlation coefficient are as follows.

i. Set up null hypothesis H_0: $\rho = 0$ i.e. the variables are uncorrelated in the population or population correlation coefficient is zero

ii. Set up alternative hypothesis H_1: $\rho \neq 0$ i.e. the variables are correlated in the population or population correlation coefficient is not zero

or H_1: $\rho > 0$, (Right tailed test) i.e. there is the positive correlation in the population

or H_1: $\rho < 0$, (Left tailed test) i.e. there is the negative correlation in the population.

iii. Fix the level of significance α

iv. Under the null hypothesis H_0 the test statistics is

$$t = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2} \text{ follows students t-distribution with (n-2) d.f.}$$

Where, r = Sample correlation coefficient

n = Number of pair of observations

v. For α level of significance at (n-2) d.f. find the critical value of t from table as t_{tab}.

vi. Conclusion: If $t_{cal} < t_{tab}$ we accept H_0 at α level.

& If $t_{cal} \geq t_{tab}$ we reject H_0 & accept H_1.

Confidence limits for estimating population correlation coefficient (ρ)

The $(1-\alpha)$% Confidence limits for estimating population correlations coefficient (ρ) are given by.

$r \pm t_\alpha(n-2)$ S.E.(r)

$r \pm t_\alpha(n-2)\frac{1-r^2}{\sqrt{n}}$ where, $S.E(r) = \frac{1-r^2}{\sqrt{n}}$

8.2.5 Test of significance of regression coefficient β_{YX} (β)

We have the regression equation of Y on X is Y =a + b_{YX} X and the population regression coefficient of Y on X as $Y = \beta_0 + \beta_{YX}$ X.

We proceed in the following stapes.

1. **Set up null hypothesis H_0:** β_{YX} =0 i.e. regression coefficient in population is zero.
2. **Set up alternative hypothesis H_1:** $\beta_{YX} \neq 0$ (two tailed) i.e. regression coefficient in population is not equal to zero.

Or H_1: $\beta_{YX} > 0$ (right tailed) i.e. regression coefficient in population is greater than zero.

Or H_1: $\beta_{YX} > 0$ (left tailed) i.e. regression coefficient in population is less than zero.

3. Fix the level of significance α.
4. **Test statistics;** Under the null hypothesis **H_0:** β_{YX} =0, the test statistics for regression coefficient β_{YX} is $t = \frac{b - \beta_{YX}}{S_b}$ follows student t-distribution with (n-2) degrees of freedom , where S_b is the standard error of an coefficient b_{YX}

5. For α level of significance and the (n-2) d.f. find the critical value of t as t_α from table.
6. **Conclusion:** - If $|t| < t\alpha$ we accept our null hypothesis at α level of significance and If $|t| \geq t_\alpha$ we reject our null hypothesis and accept alternative hypothesis.

Confidence limit for regression coefficient β_{YX}

The $(1 - \alpha)$ confidence limit for regression coefficient β_{YX} is given as

$b_{YX} \pm t_\alpha S_b$

7.3.3 F-test or variance ratio test

(Test based on population variance)

If x is a chi-square variate with $\nu_1 = (n_1-1)$ d.f. & y is also a chi-square variate with $\nu_2 = n_2 - 1$ d.f. where x & y are independent then F- statistic is defined as

$F = \frac{\frac{x}{n_1}}{\frac{y}{n_2}}$ in other words 'F' is defined as the ratio of two independent chi-square variates divided by their respective d.f. and it follow F-distribution with (ν_1, ν_2) d.f. with pdf given as

$$P(F) = \frac{\left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}}}{\beta\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)} \frac{F^{\frac{\nu_1}{2}-1}}{\left(1 + \frac{\nu_1}{\nu_2}F\right)^{\frac{\nu_1+\nu_2}{2}}}, \ 0 \leq F < \infty.$$

Remarks:

(i) The sampling distribution of F-statistics doesn't involve any population parameter and depends only on the degrees of freedom ν_1 & ν_2

(ii) A statistic F following F-distribution with (ν_1, ν_2) d.f. will be denoted by $F \approx F(\nu_1, \nu_2)$

(iii) It is a continuous r.v. which assumes only non-negative value.

Assumptions of F-test

Following are the important theoretical assumptions based on F-test

(1) The population from which the samples are drawn are normally distribution with mean μ & variance σ^2.

(2) The random samples drawn from the normal population are independent.

(3) The ratio of σ_1^2 to σ_2^2 should be greater than or equal to unity. That is the reason that larger variance is divided by the smaller variance in F-test.

(4) Value of F-distribution can never be negative, since it is always formed by a ratio of squared values.

Applications of F-distribution

F-distribution has a number of application is statistics, some of important application are

(1) F-test for equality of population variance.

(2) F-test for testing the equality of several population means.

(3) F-test for testing the significance of an observed sample multiple correlation

(4) F-test for testing the linearity of regression

Test procedure based on equality of population variance

We use the F-test for the following cases.

(a) To test whether the two estimates of population variances are significantly different or not.

(b) To test whether they establish the fact that both the sample have came from the some universes and have a common variance

(c) To test whether a given population follows a uniform distribution or how well one population compares with the other in terms of the uniformity or consistency of the distribution.

Following are the steps for f-test based on equality of two population variance

1. Set up null hypothesis H_0: $\sigma_1^2 = \sigma_2^2 = \sigma^2$(say) i.e. two population variances are equal

2. Set up alternative hypothesis H_1: $\sigma_1^2 \neq \sigma_2^2$ (Two tailed test) i.e. the population variances are not equal.

3. Fix the level of significance α

4 Test statistics under H_0, the test statistics is

$$F = \frac{S_1^2}{S_2^2} \approx F(\nu_1,\nu_2) \text{ d.f. if } S_1^2 > S_2^2$$

$$\text{or, } F = \frac{S_2^2}{S_1^2} \approx F(\nu_2, \nu_1) \text{ d.f. if } S_2^2 > S_1^2$$

Where S_1^2 & S_2^2 are unbiased estimate of the common population variance σ^2 & are give by

$$S_1^2 = \frac{1}{n_1-1}\,\Sigma\,(x_1 - \overline{x}_1)^2 = \frac{1}{n_1-1}\left[\Sigma\,(x_1^2) - \frac{(\Sigma x_1)^2}{n}\right]$$

$$S_2^2 = \frac{1}{n_2 - 1}\,\Sigma\,(x_2 - \overline{x}_2)^2 = \frac{1}{n_2 - 1}\left[\Sigma(x_2^2) - \frac{(\Sigma x_2)^2}{n}\right]$$

If biased sample variance s_1^2 & s_2^2 are known then

$$S_1^2 = \frac{n_1 s_1^2}{n_1 - 1} \quad \& \quad S_2^2 = \frac{n_2 s_2^2}{n_2 - 1}$$

$$\text{Thus } F = \frac{\text{greater variance}}{\text{smaller variance}} \approx F(\nu_1, \nu_2)$$

5. For α level of significance & (ν_1, ν_2)_ d.f. find the critical value of F from table

6. Conclusion

If $F_{cal} < F_{tab}$ we accept H_0 at α level &

If $f_{cal} \geq f_{tab}$ we reject H_0 & accept H_1.

Unit 7.1

Estimation

Parameter The statistical constants of whole population are called parameters. Population mean (μ), population variance (σ^2), population proportion (P). Population correlation coefficient (ρ), population regression coefficient (β_{YX}) etc. are parameters. The parameters mean and variance are largely used. In a theoretical probability distribution by parameters one means those population constant which appear in probability density (mass) function.

Statistic According to Ronald A. Fisher the statistical constants of the sample selected from the population are called statistics. In general, statistics are the function of observable random variable and do not involve any unknown parameters. For example sample mean $\bar{x}$, sample variance (S^2 or s^2), sample correlation coefficient (r) etc**.**

Name	Parameter	Statistics
Size	N	n
Mean	μ	$\bar{x}$
Proportion	P	p
Variance	σ2	S^2 or s^2
Correlation coefficient	ρ	r
Regression coefficient	β_{YX}	b_{yx}

In practice parameter values are not known and thus they are estimated on the basis sample values. Thus statistic can be regarded as the estimate of population parameter.

Estimation The theory of estimation was founded by Prof. R.A. Fisher. Estimation is the process by which population characteristics are estimated from characteristics of the sample studies with desired degree of precision. The main objective of estimation is to obtain a guess or estimate of the unknown, true value from the sample data or past experience. There are many situations in our daily life where we make estimation. When crossing a road, we estimate the speed of taxi that is approaching, the distance between us and that taxi and our own speed. Having made these quick estimates, we decide whether to wait, walk or run. We must make quick estimates to save the accidents. Sometimes, a decision maker makes rational decision without complete information and with a great deal of uncertainty about what the future will bring.

In estimation, the sample statistics are used to estimate the population parameters. For example, a campus administration makes the estimates of

enrollments for next year from current enrollments in the campus. A Forester estimates the production Kurilo tuber on the basis of observed sample of the plants. A businessman estimates his future sales of computer and electronic goods from the past records. In each case, some one is trying to draw conclusion about population from sample information.

The sample distributions of a statistic and its standard error play a vital role in both the estimation of parameters and the testing of statistical hypothesis.

Types of estimation

A random sample of a given size is selected from a given population and then computes a statistic which is a characteristic of the sample and this becomes an estimate of the similar characteristic of the population. There are two types of estimation about a population parameter namely point estimation and interval estimation.

i. **Point estimation:** A point estimate of a parameter is a single numerical value compute from sample information. For example a sample mean ($\bar{x}$) and a sample variance s^2 is the point estimate of population mean μ and population variance σ^2 .

We write as point estimator of μ is $\bar{x}$ as $\hat{\mu} = \bar{x}$ and $\hat{\sigma}^2 = s^2$

ii. Interval estimation: A range of values used to estimate a population parameter is called an interval estimate. There are two ways for indicating error by the extent of its range and by the probability of the true population parameter lying within that range.

For example,

P (a < μ < b) = 0.95.

UNIT 7.2

Testing of Hypothesis

The theory of testing of parametric statistical hypothesis was originally set forth by J. Neyman in 1928 and Karl Pearson in 1933.

When parametric values are unknown, we estimate them through sample values. If the sample value is exactly the same as per our hypothesis, there is no hitch in accepting it. And it is far from our contention; there is not reason to accept it. But the problem arises when the sample provides a value which is neither exactly equal to the parametric value, nor too far. In that situation one has to develop some procedures which enables one to decide whether to accept a hypothetical value or not on the basis of sample values. Such a procedure is known as testing of hypothesis.

A testing of hypothesis is a rule or procedure which makes one to decide about the acceptance or rejection of the hypothesis. This is simply a procedure of drawing the conclusion about the population based on sample information.

Statistical Hypothesis: Null and alternative hypothesis

A quantitative statement about the population parameter is called a hypothesis. Hypotheses are of two types. Null and alternative hypothesis,

A hypothesis which is usually a hypothesis of no difference is called null hypothesis and is usually denoted by H_0. For example, in case of a single statistic, H_0 will be that the sample statistics doesn't differ significantly from the hypothetical parametric value, and in the case of two statistics, H_0 will be that the sample statistics do not differ significantly.

Having setting up the null hypothesis we compute the probability 'p' that the deviation between the observed sample statistic and hypothetical parameter value might have occurred due to fluctuation of sampling. If the deviation comes out to be significant, null hypothesis is rejected at the particular level of significance adopted and if the deviation is not significant null hypothesis may be accepted at that level.

Any hypothesis which is complementary to the null hypothesis is called an alternative hypothesis It is denoted by H_1. For example, in case of a single statistic, H_0 will be that the sample statistic differs significantly from the hypothetical parametric value, and in the case of two statistics, H_0 will be that the sample statistics differ significantly. If H_0 is accepted, H_1 is rejected and vice versa.

If we want to test the null hypothesis that the population has a specified mean μ_0 (say) i.e. $H_0 : \mu = \mu_0$ then the alternative hypothesis could be,

i. H_0: $\mu \neq \mu_0$ (i.e., $\mu > \mu_0$ or $\mu < \mu_0$)

ii. $H_0 : \mu < \mu_0$

iii. H_0: $\mu > \mu_0$.

The alternative hypothesis (i) is known as a two tailed alternative and the alternative hypotheses (ii) and (iii) are known as left tailed and right tailed alternatives respectively.

The setting of alternative hypothesis is very important since it enables us to decide whether we have to use a single tailed (right or left) or two tailed test.

Types of Error

The main objective in sampling theory is to draw valid conclusions about the population parameters on the basis of the sample results.

There are the four possibilities in the testing of statistical hypothesis,

i. Accepting the null hypothesis when the null hypothesis is true.

ii. Rejecting the null hypothesis when the null hypothesis is true.

iii. **Accepting the null hypothesis when the null hypothesis is false**.

iv. Rejecting the null hypothesis when the null hypothesis is false.

The above decisions can be presented in the following table

Real situation (States of nature)	**Decision from sample**	
	Accept H_0	**Reject H_0**
When H_0 is True	correct decision (no error)	Wrong decision (Type I error)
When H_0 is False	Wrong decision (type II error)	correct decision (no error)

In the testing of hypothesis, we may commit two types of errors.

Type I error

The error committed in rejecting null hypothesis H_0 when it is true is called type I error. A type I error can also be referred to as an error of the first kind and its probability is denoted by α (alpha).

Thus, Type I error = Reject null hypothesis when it is true.

α = Probability (Reject null hypothesis when it is true)

= P (reject H_0 when H_0 is true)

For example, in a clinical trial of a new drug, the null hypothesis might be that the new drug is no better, on average, than the current drug; that is H0: there is no difference between the two drugs on average. A type I error would occur if we concluded that the two drugs produced different effects when in fact there was no difference between them.

Type II error

The error committed in accepting null hypothesis H_0 when it is false is called type II error or the error of second kind and its probability is denoted by β (beta).

Thus, Type II error = Accept null hypothesis when it is wrong

β = Probability (Accept null hypothesis when it is wrong)

=P (accept H_0 when H_1 is true)

For example, in a clinical trial of a new drug, the null hypothesis might be that the new drug is no better, on average, than the current drug; that is H0: there is no difference between the two drugs on average. A type II error would occur if it was concluded that the two drugs produced the same effect, that is, there is no difference between the two drugs on average, when in fact they produced different ones.

A type II error is frequently due to sample sizes being too small

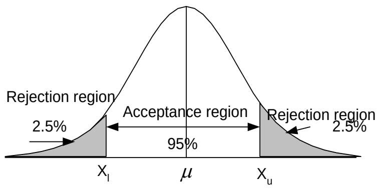
Level of significance

The level of significance may be defined as the probability of type of I error which we are ready to tolerate in making a decision about H_0.

i.e. P (reject H_0 when H_0 is true) = α.

It is our endeavour to carry out a test which minimizes both type of error; unfortunately for given set of observations, both the errors can't be controlled simultaneously. Hence it is a general practice to assign a bound to type I error and to minimize type II error, Thus one choose a value of α lying between 0 and 1 which is known as the level of significance . Hypotheses in general are tested at 1% or 5% level of significance. But the most commonly used level of significance in practice is 5%.

Critical region or region of rejection

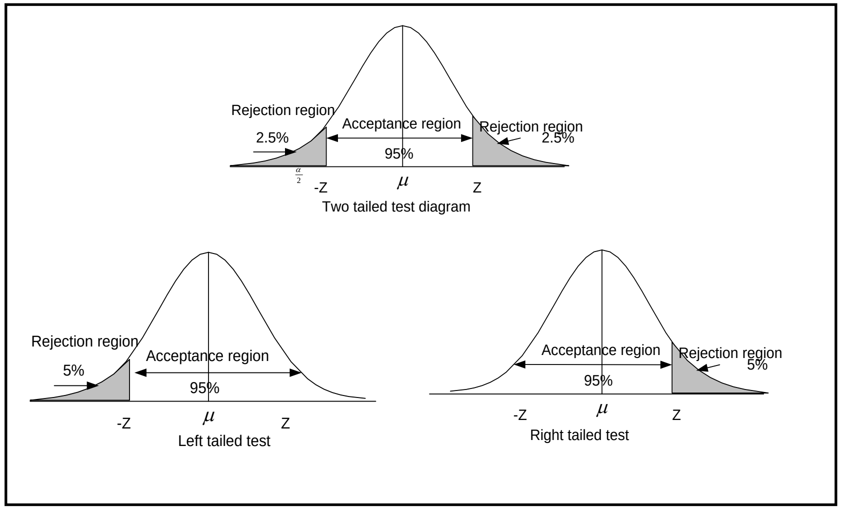


The critical region or rejection region is a set of values of the test statistic for which the null hypothesis is rejected in a hypothesis test; that is, the sample space for the test statistic is partitioned into two regions; one region (the critical region) will lead us to reject the null hypothesis H_0', the other not. So, if the observed value of the test statistic is a member of the critical region, we conclude 'reject H_0'; if it is not a member of the critical region then we conclude 'do not reject H_0.

The region of rejection may be situated on both the tails or on only one tail depending on the alternative hypothesis. If the value of the test statistics lies in the acceptance region, the null hypothesis H_0 is accepted and if it is lies in the critical region H_0 is rejected.

One Tailed and Two Tailed Tests

If an alternative hypothesis is such that it leads to two alternatives to the null hypothesis, it is said to be a two tailed test. For example, Test H_0: μ =20 verses H_1: $\mu \neq$ 20 leads to two sided test as μ can be greater than 20 or less than20. In this situation half of the area of critical region lies on the left tail and the half on the right. If α is the area of the critical region $\frac{\alpha}{2}$ is the area one both the tails.



Two tailed test diagram

Left tailed test

Right tailed test

Again, if the alternative hypothesis provides one sided alternative to H_0 e.g. H_0: μ =20 verses H_1: μ >20 or H_1: μ < 20, the critical region or size α lies only on one tail. Specifically an area equal to α lies on the right tail when H_1 is μ >20 and on the left tail when H_1 is μ< 20.

The choice between a one-sided and a two-sided test is determined by the purpose of the investigation.

Critical Value

The critical value for a hypothesis test is a value of the test statistic that defines the region of acceptance and rejection. The Thus critical value for a hypothesis test is a threshold to which the value of the test statistic in a sample is compared to determine whether or not the null hypothesis is rejected. The critical value for any hypothesis test depends on

i.the significance level at which the test is carried out

ii.nature of sampling distribution

iii.types of alternative hypothesis (i.e. whether the test is one-sided or two-sided.)

Critical values of Z for large samples (n>30) at commonly used levels of significance of both two tailed and single tailed test, have been obtained from the normal probability table as given below.

Critical value (Z_α)	Level of significance (Z_α)				
	1%	2%	4%	5%	10%
Two tailed test	$\|Z_\alpha\|$= 2.58	$\|Z_\alpha\|$= 2.33	$\|Z_\alpha\|$= 2.05	$\|Z_\alpha\|$= 1.96	$\|Z_\alpha\|$= 1.645
Right tailed test	Z_α =2.33	Z_α = 2.05	Z_α = 1.75	Z_α =1.645	Z_α =1.28
Left tailed test	Z_α = -2.33	Z_α = -2.05	Z_α = -1.75	Z_α =-1.645	Z_α = -1.28

It is to be noted that critical value of Z for a single tailed test (left or right) at level of significance α is same as the critical value of Z for a two –tailed test at level of significance ‘2α’.

Remarks: we can’t use the above critical values for small samples ($n \le 30$) as sampling distribution with this size ($n \le 30$) does not follow normal distribution.

Procedure for testing of hypothesis or test of significance

We now summaries below the various steps on testing of a statistical hypothesis in a systematic manner,

1 **Set up null hypothesis H_o:** there is no significance difference between sample statistic and population parameter or there is no significance difference between two sample statistics.

2 **Set up alternative hypothesis H_1**: there is significance difference between sample statistic and population parameter or there is significance difference between two sample statistics.

3 **Select the level of significance**: Select the level of significance α at which the hypothesis is to be tested.

4 **Apply suitable test statistic:** Choose an appropriate test statistic under null hypothesis H_0 on the basis of nature of sampling distribution such as t-test, Z-test, F-test, χ^2 -test etc.

5 **Obtain critical value:** Get the critical value for α level of significance from table.

6 **Draw conclusion:** If calculated value of test statistic under H_0 is less than tabulated value of the test we accept null hypothesis. If calculated value of test statistic under H_0 is greater than or equal to the tabulated value of the test we reject null hypothesis and accept our alternative hypothesis at level of significance α .

7.3.1Test of significance for large Sample (Z-test)

Large sample (normal) test is used when the sample size ‘n’ is greater than 30. Therefore the test is based on normal distribution (Z-test) in which the area under the normal curve is applied using standard normal variate Z.

The Z-tests of large sample are given by

1. Test of significance of a single mean.
2. Test of significance of a difference between two means.
3. Test of significance of a sample proportion.
4. Test of significance difference between two sample proportions.

Test of significance for single mean

We have for random sample X_1 , X_2.............X_n of size n from a normal population with mean μ and variance σ^2, then the sample mean $\bar{x}$ is distributed normally with mean μ and variance $\frac{\sigma^2}{n}$ i.e. $\bar{x} \approx N(\mu, \sigma^2)$.

Then Z statistic

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \approx N(0, 1)$$

Following are the steps for test of significance of single mean in large sample

1 **Set up null hypothesis H_o**: $\mu = \mu_0$ i.e. population mean has specified value say μ_0 or there is no significance difference between population mean μ and sample mean $\bar{x}$

2 **Set up alternative hypothesis H_1** : $\mu \ne \mu_0$ (two tailed test) i.e.there is no significance difference between population mean μ and sample mean

Or H_1: $\mu > \mu_0$ (right tailed test) population mean μ is greater than sample mean

Or H$_1$: $\mu < \mu_0$ (left tailed test) population mean μ is less than sample mean.

3 **Choose level of significance**: Determine the level of significance α at which the hypothesis is to be tested.

4 **Compute test statistics:** under the null hypothesis $H_0 : \mu = \mu_0$ the test statistics is given by $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \approx N(0, 1)$

If the population standard deviation is unknown then we use its estimate provided by the sample variance given by $\hat{\sigma}^2 = s^2 \Rightarrow \hat{\sigma} = s$ (for large samples size) i.e. $Z = \frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}} \approx N(0, 1)$

5 **Critical value:-** for α level of significance, find the critical value of Z form standard normal table as Z_α

6 **Conclusion:-**If calculated value of Z i.e. |Z| is less than Z_α we accept our null hypothesis H_0 at α level of significance and otherwise rejected. And at last draw valid conclusion on the basis of hypothesis testing.

Confidence limits or fiducial limits

The (1 - α) % confidence limits for estimation population mean μ from large sample at α level of significance is given by

$\bar{X} \pm Z\alpha$ S.E.($\bar{X}$)

$\Rightarrow \bar{X} \pm Z\alpha \frac{\sigma}{\sqrt{n}}$ i.e. $P\left(\bar{X} - Z\alpha \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + Z_\alpha \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$

$\therefore$ Upper limit for population mean $\mu = \bar{X} + Z\alpha \frac{\sigma}{\sqrt{n}}$ and

Lower limit for population mean $\mu = \bar{X} - Z\alpha \frac{\sigma}{\sqrt{n}}$

Thus 95%

$\bar{X} \pm 1.\ 96 \frac{\sigma}{\sqrt{n}}$

and 99% confidence limits for population mean μ are

$\bar{X} \pm 2.\ 58 \frac{\sigma}{\sqrt{n}}$.

Test of significance for single proportion

In a sample of size n, let X be the number of persons possessing the given attribute. Then observed proportion of success = $\frac{X}{n}$ = p (say)

$E(p) = E\left(\frac{X}{n}\right) = \frac{1}{n} E(X) = \frac{nP}{n} = P$, which showed that sample proportion 'p' is an unbiased estimate of the population proportion "P".

Also, $V(p) = V\left(\frac{X}{n}\right) = \frac{1}{n^2} V(X) = \frac{nPQ}{n^2} = \frac{PQ}{n}$

Now standard error of p = S.E.(p) = $\sqrt{V(p)} = \sqrt{\frac{PQ}{n}}$

Thus for large n, x and consequently X/n is asymptotically normal, the normal test for proportion of success being $Z = \frac{p - E(p)}{\sqrt{V(p)}} = \frac{p - P}{\sqrt{\frac{PQ}{n}}} \approx N(0, 1)$

In order to test whether there is significance difference between the sample proportion 'p' and the population proportion 'P' we proceed as follows

1. **Set up null hypothesis H$_0$:** $P = P_0$ (say) i.e. population proportion has specified value say P_0.
2. **Set up alternative hypothesis H$_1$**: $P \neq P_0$ (two tailed) i.e. population proportion is not equal to specified value say P_0.

Or H$_1$: $P > P_0$ (right tail tailed) i.e. population proportion is greater than specified value say P_0.

Or H$_1$: $P < P_0$ (left tailed) i.e. population proportion is less than specified value say P_0

3. Fix the level of significance α .
4. **Test statistics;** Under the null hypothesis $H_0 : P = P_0$, the test statistics for proportion of success(p) is

$Z = \frac{p - E(p)}{\sqrt{V(p)}} = \frac{p - P}{\sqrt{\frac{PQ}{n}}} \approx N(0, 1)$, For α level of significance find the critical value of Z as Z_α from normal tale.

5. Conclusion:- If $|Z| < Z_\alpha$ we accept our null hypothesis at α level of significance and If $|Z| \geq Z_\alpha$ we reject our null hypothesis and accept alternative hypothesis.

Confidence limits

95% confidence limits for 'P' are given by p $\pm$ 1.96 $\sqrt{\frac{pq}{n}}$ and 99% confidence limits for 'P' are given by p $\pm$ 2.58 $\sqrt{\frac{pq}{n}}$

Test of significance for difference of mean for large sample

Let two samples of sizes n_1 and n_2 are drawn from two normal population with mean μ_1 & μ_2 and variances σ_1^2 & σ_2^2 Let $\bar{x}_1$ & $\bar{x}_2$ be sample means of two samples , then for large sample $\bar{x}_1$ & $\bar{x}_2$ is normally distributed with means μ_1 &

μ_2 and variances σ_1^2/n_1 & σ_2^2/n_2 Also($\bar{x}_1 - \bar{x}_2$) , being the difference of two independence normal variate also a normal variate the value of Z(SNV) corresponding to ($\bar{x}_1 - \bar{x}_2$) is given by

$$Z= \frac{(\bar{x}_1 - \bar{x}_2) - E(\bar{x}_1 - \bar{x}_2)}{S.E.(\bar{x}_1 - \bar{x}_2)} \approx N(0, 1)$$

Then the stapes for test of significance of difference between two means are as follows.

1. **Null hypothesis H_o:** $\mu_1 = \mu_2$ i.e. two population means are equal or samples are drawn form two population having equal population mean.
2. **Alternative hypothesis H_1:** $\mu_1 \neq \mu_2$ (two tailed test) i.e. two population means are not equal or samples are drawn form two population having different population mean.

   Or H_1: $\mu_1 > \mu_2$ (right tailed test) i.e. population mean of first is greater than second

   Or H_1: $\mu_1 < \mu_2$ (left tailed test) i.e. Population mean of first is smaller than second
3. Fix the level of significance α.
4. **Test statistics:-** Under the null hypothesis **H_o:** $\mu_1 = \mu_2$ the test statistics is $Z = \frac{(\bar{x}_1 - \bar{x}_2) - E(\bar{x}_1 - \bar{x}_2)}{S.E.(\bar{x}_1 - \bar{x}_2)} \approx N(0, 1)$

   $$\text{Or, } Z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \approx N(0, 1)$$

Remarks 1: If $\sigma_1^2 = \sigma_2^2 = \sigma^2$ i.e. if the samples have been drawn from the population with common variance then under null hypothesis

H_o: $\mu_1 = \mu_2$ the test statistics is $Z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sigma\sqrt{1/n_1 + 1/n_2}}$

Remarks 2: If σ is not known then its estimate is

$$\hat{\sigma}^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2}$$

5. For α level of significance find the critical value of Z as Z_α from normal table.
6. **Conclusion:** if |Z| less than Z_α, we accept our null hypothesis **H_o:** $\mu_1 = \mu_2$ at α level of significance, reject otherwise.

Confidence limits

95% confidence limits for difference of mean is ($\bar{x}_1 - \bar{x}_2$)±1.96 S.E ($\bar{x}_1 - \bar{x}_2$)

99% confidence limits for 'difference of mean is ($\bar{x}_1 - \bar{x}_2$) ±2.58 S.E ($\bar{x}_1 - \bar{x}_2$)

Test of significance difference between two proportions

Let two sample proportions be $p_1 = \frac{x_1}{n_1}$ & $p_2 = \frac{x_2}{n_2}$

$$Z = \frac{(p_1 - p_2) - E(p_1 - p_2)}{\sqrt{var\ (p_1 - p_2)}} \approx N(0,1)$$

Following are the steps in test of significance for difference of two proportions are as follow.

1. Null hypothesis $H_0 = P_1 = P_2$ = P(say) i.e. two population proportion are same in other words there is no significance difference between two sample proportion p_1 and p_2

2. Alternative hypothesis $H_1 = P_1 \neq P_2$ (Two tailed test i.e. Two population proportion are not same or two samples are drawn from difference population or **H_1: $P_1 > P_2$** (right tailed test)i.e. population proportion of 1st population is greater than that of second or

H_1: $P_1 < P_2$ (Left tailed test)

That is population proportion of first is less then that of second

(3) Fix level of significance α

(4) Under H_0 : $P_1 = P_2$ the test statistics is

$$Z = \frac{(p_1 - p_2) - E(p_1 - p_2)}{\sqrt{var\ (p_1 - p_2)}} \approx N(0,1)$$

$$= \frac{(p_1 - p_2)}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad [\because P_1 = P_2 = P]$$

In general if common population proportion of two population is unknown, then an unbiased estimate of the population proportion P, based on both the sample is given by

$$\hat{P} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{x_1 + x_2}{n_1 + n_2}$$

(5) For α level of significance we find the critical value of Z from Table as $Z\alpha$.

(6) If $|Z| < Z\alpha$ we accept H_0 at α % level.

& If $|Z| \geq Z\alpha$ we reject H_0 & accept H_1.

Conclusion: The valid conclusion is made on the basis of hypothesis testing.

Confidence interval for difference of population proportion

99% confidence limits are:($p_1 - p_2$) ± 2.58 S.E ($p_1 - p_2$**)**

And 95**%**& ($p_1 - p_2$) ± 1.96 S.E.($p_1 - p_2$)