

Sampling Methods :

Population: The complete set of group of individuals, whether that group comprises a nation or a group of people with a common characteristic.

[A group of individuals under study is called population]

Sample : A part selected from population is called as sample.

Estimation : To study about some of the feature of the population on the basis of the random sample drawn from the population is called estimation.

Testing of hypothesis : A statistical hypothesis is a method of statistical inference used to decide whether the data sufficiently support a particular hypothesis.

Statistical Inference : A branch of statistics using probability concept to deal with uncertainty in decision making.

→ Statistical inference treats estimation & testing of hypothesis.

Procedure of testing hypothesis:

Hypothesis : A hypothesis is an assumption, an idea that is proposed for the sake of some statement so that it can be tested to check its validity.

① Set up a hypothesis :

Set up a hypothesis about a population parameter then collection of sample data.
Test the validity of assumption.

There are two hypothesis in a statistical test are —

- (a) Null hypothesis
- (b) Alternative hypothesis.

~~Ex:~~ Null hypothesis: A neutral attitude of the decision maker before the sample observations are taken, is the basic idea of null hypothesis

Notation : H_0

Alternative hypothesis: The acceptance or rejection of null hypothesis is meaningful only when it is being tested against rival hypothesis. is called alternative hypothesis

Notation : H_A or H_1 or H_α

EX: Consider a light bulb problem.

An enquirer wants to compare two manufacturing process of bulbs

— ① New process

② Standard process.

Let the light bulb has an average life of μ_1 hours under standard manufacturing process & μ_2 hours under new manufacturing process.

In this process, problem, decision maker have three hypothesis as —

- (i) New process is better than standard process
- (ii) Standard process is better than new process
- (iii) There is no significant difference between two processes.

Note: First two statements reflect a preferential attitude towards one or the other of the two processes. Whereas third statement is neutral which is null hypothesis.

Here $H_0: \mu_1 = \mu_2$

In above example, alternative hypothesis could be —

$H_A: (i) \mu_1 > \mu_2$ (ii) $\mu_1 < \mu_2$ (iii) $\mu_1 \neq \mu_2$

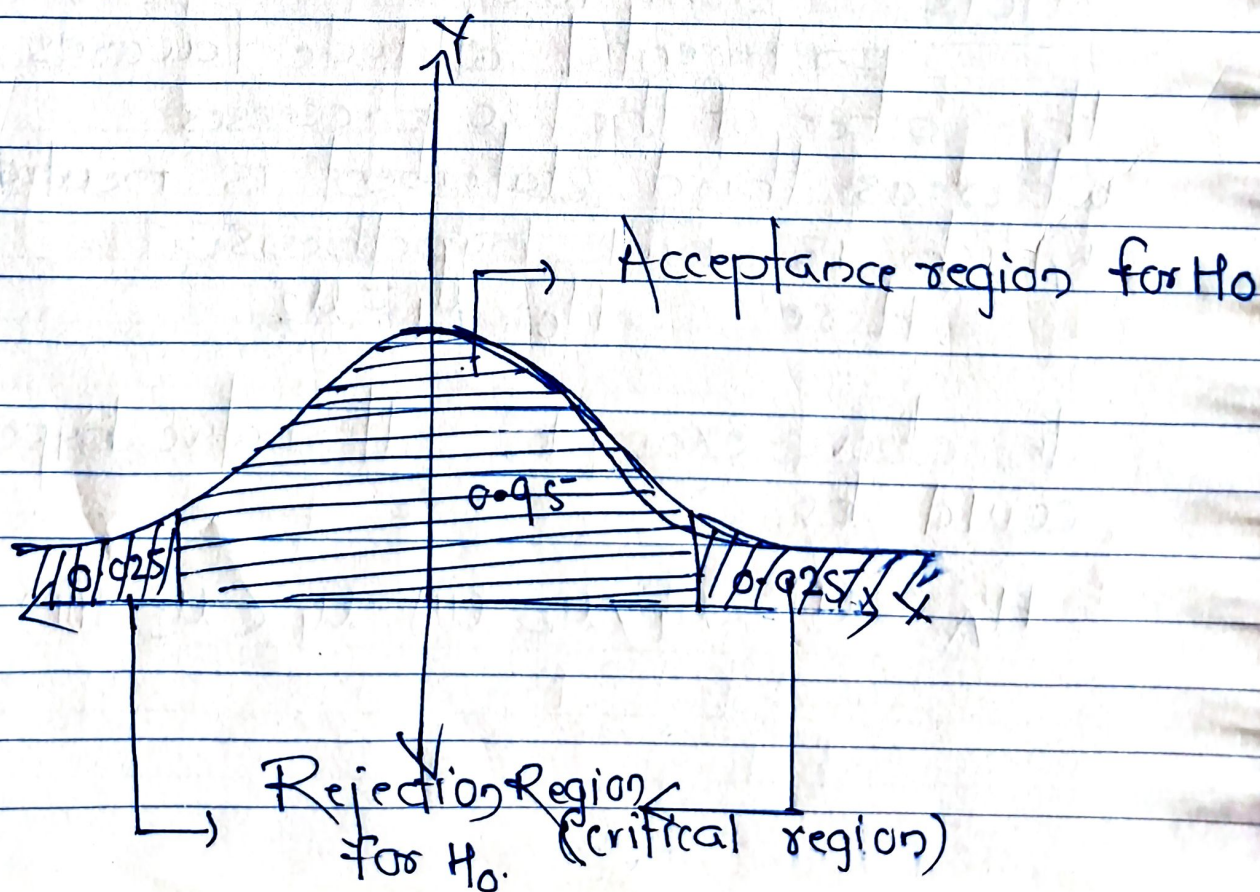
②. Set up a suitable significant level :

To test the validity of H_0 against that of H_1 at a certain level of significance.

Note : ① The confidence with which an experimenter or enquirer rejects or accepts a null hypothesis depends upon the significance level adopted.

It is expressed as 5% or 1%.

② When a hypothesis is accepted in 5% level, the statistician is running the risk that in long run, he will be making wrong decision about 5% of time.
Similarly for 1%.



③ Setting a test criterion :

The third step in hypothesis testing procedure is to construct a test criterion.

Note : Some common probability distributions that are commonly used in testing procedures are t , F & χ^2 .

Decision from sample	True statement	
	H_0 True	H_0 False
Reject H_0	Type-I error	No error
Accept H_0	No error	Type-II error

$P_x(\text{Type I error}) = \alpha$, this α is called the size of the test or size of critical region.

$$P_x(\text{Type II error}) = \beta$$

$1 - \beta = P_x(\text{Accept } H_0 \text{ when it is true})$
 = the capacity of the test to detect the falsity of H_0 .

// $H_0 \Rightarrow \text{True}$ — $\begin{cases} \rightarrow \text{Accept (No error)} \\ \rightarrow \text{Reject (Type-I error)} \end{cases}$

$H_0 \Rightarrow \text{False}$ — $\begin{cases} \rightarrow \text{Accept (Type-II error)} \\ \rightarrow \text{Reject (No error)} \end{cases}$

Note : ① In order to get low β , we will have to put up with a high α .

② The level of significance is also known as the size of rejection region or size of critical region or simply size of the test.

Notation : significance level = α

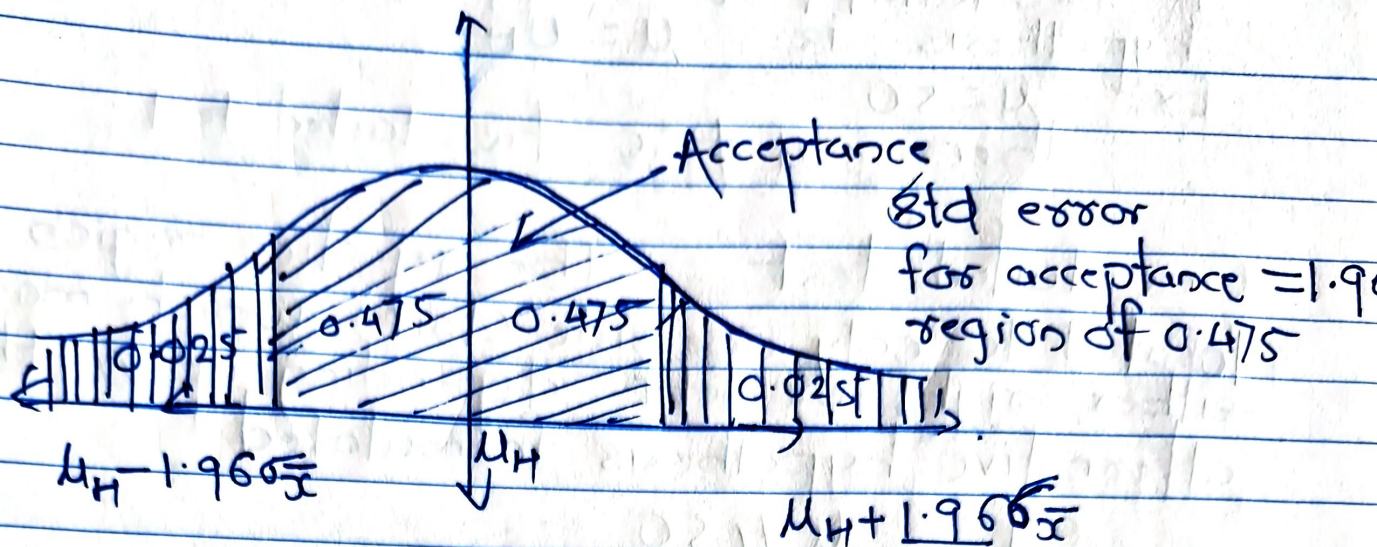
③ In most of the statistical tests, the significance level is generally fixed at 5%.
 $\alpha = 5\%$.

Hence, prob. of accepting true hypothesis is 95%.

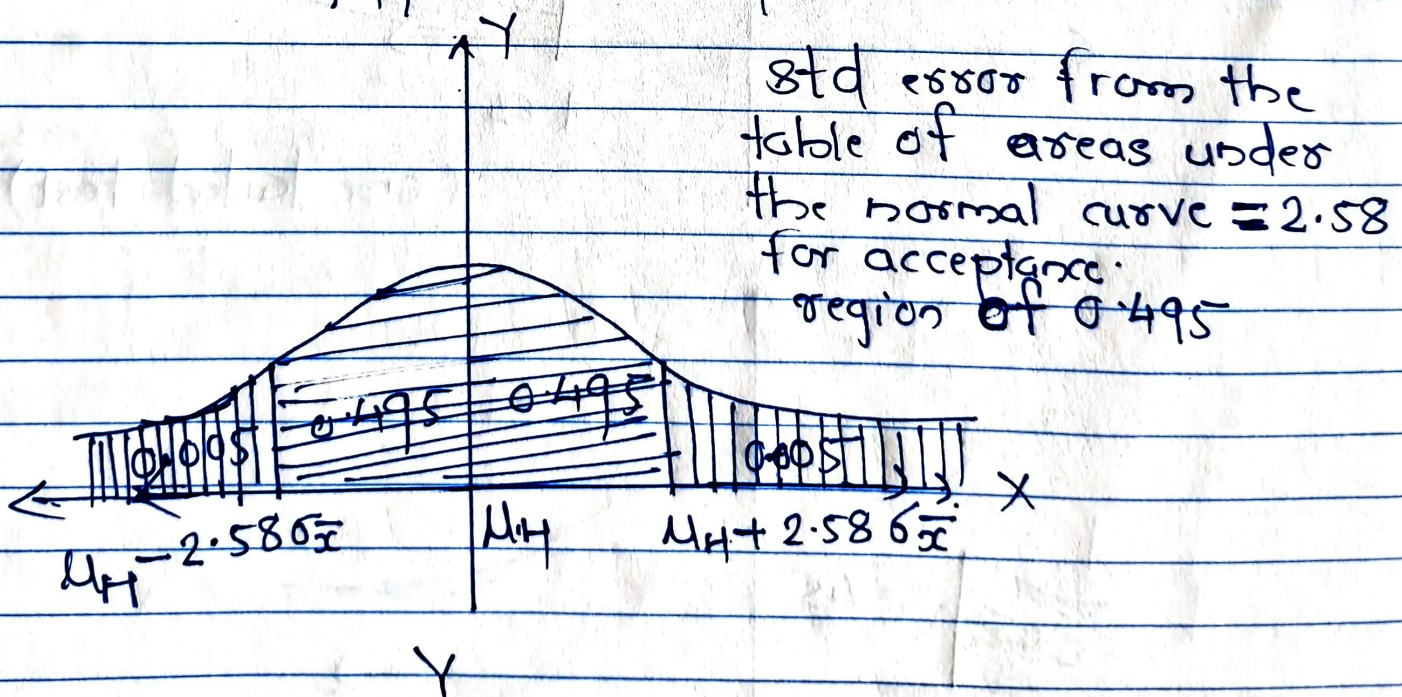
Two tailed & one tailed hypothesis :

A two tailed hypothesis will reject the null hypothesis if the sample statistics is significantly higher/lower than the hypothesized population parameter.

Note : In two tailed test, the rejection region is located on (both sides).



Note : To reduce the risk of committing Type-I error. This is done by reducing the size of rejection region.
For this, hypothesis may be treated at 1% level of significance
i.e. 1% is rejection
& 99% is acceptance.



A two tailed test is appropriate when null hypothesis is $\mu = \mu_H$

Ex: $\mu = 50$

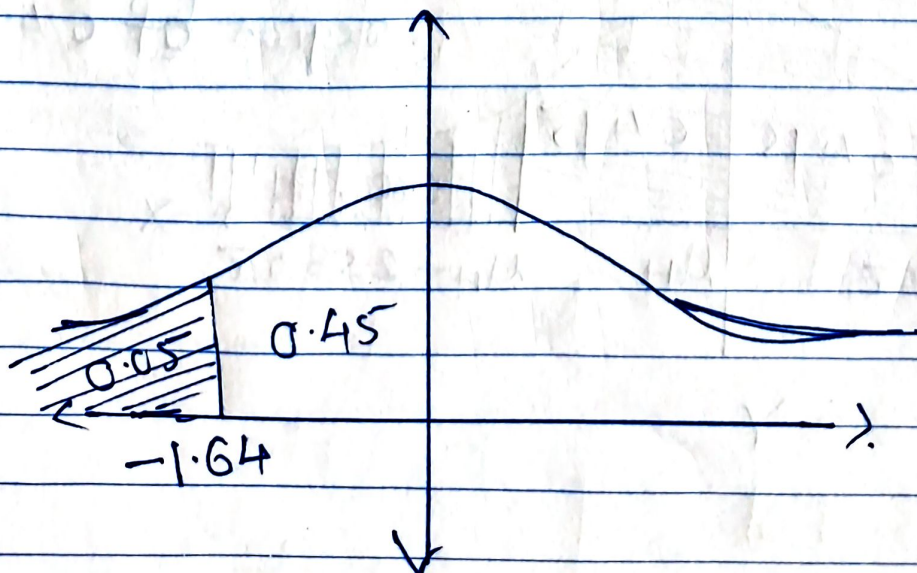
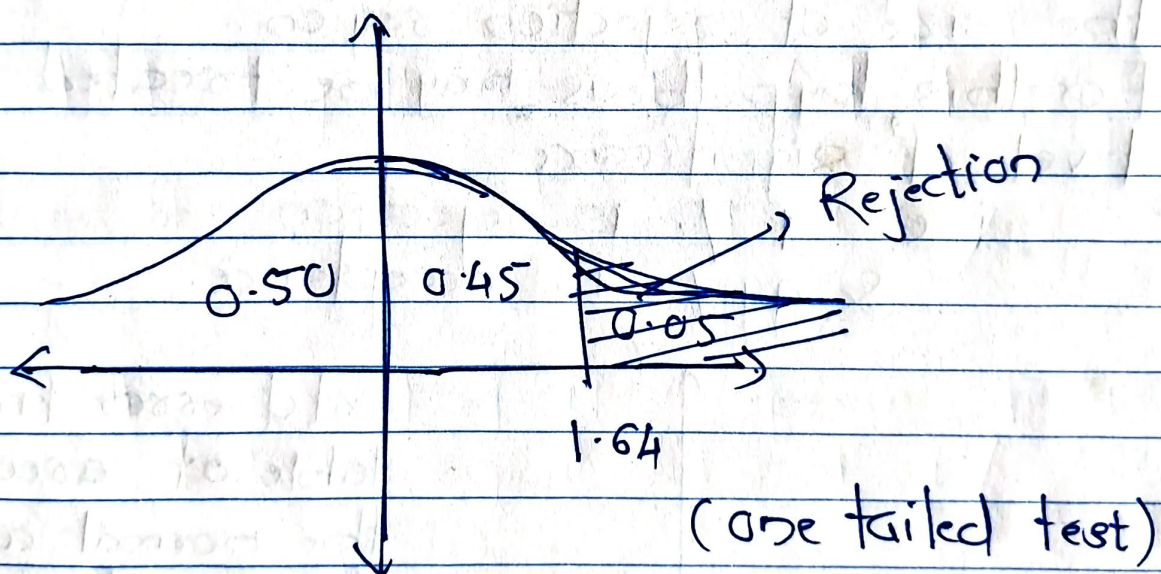
then $\mu \neq 50$ two tailed test.

One tailed test: The rejection region is located to only one side (tail) which may be either right or left depending upon the alternative hypothesis formulated.

e.g. $\mu > 50$

$H_1: \mu < 50$ alpha risk is on left side.

This test is left tailed test.



Ex: $H_0 \Rightarrow \mu = \mu_0$
alternative hypothesis be

- ① $H_1: \mu \neq \mu_0 \rightarrow$ two tailed
- ② $H_1: \mu > \mu_0 \rightarrow$ Right tailed
- ③ $H_1: \mu < \mu_0 \rightarrow$ left tailed

Measuring the power of Hypothesis test:

Note: ① The measure of how well the test is working is called the power of the test.

In hypothesis testing α & β should be small

② Rejecting a null hypothesis when it is false is exactly a good test outcome, a high value of $1 - \beta$ (something) near to 1 means the test is working quite well.

Low value of $1 - \beta$ means test is working very poorly.

Thus, $1 - \beta$ is the measure of how well the test is working & called power of the test.

Standard Error & Sampling distribution:

The standard deviation of the sampling distribution is called standard error (S.E)

Note: ① S.E is used as an instrument in testing a given hypothesis, generally tested at 5% level of significance.

$$\textcircled{2} \text{ If, } \frac{\text{Observed mean} - \text{expected mean}}{\text{S.E}} > 1.96$$

then the result of the experiment does not support the hypothesis at 5% level.

$$\text{If, } \frac{\text{Observed mean} - \text{expected mean}}{\text{S.E}} > 2.58$$

then the result of the experiment is considered to be significant at 1% level.

③ In practice, quite often a hypothesis is accepted if the difference is less than 3 S.E because the probability of a difference greater than 3 S.E, arising by chance, is only about 3 in thousand (0.27%) as 99.73% items are covered between $\text{mean} \pm 3\text{S.E}$
i.e. $\mu \pm 3\sigma$

Note that the use of 3 σ rule is justified only if n is large.

④ S.E provides an idea about the unreliability of a sample.

⑤ If S.E is greater, then greater unreliability of a sample.

⑥ $\frac{1}{S.E}$ is a measure of reliability (or precision) of the sample.

⑦ With the help of S.E, we can determine the limits within which the parameter values are expected to lie.

Test of significance for large samples ($n > 30$)
(problems of sampling of variables such as height, weight etc, which may take any value)

① The assumptions while dealing with problems related to large samples.

② Values given by samples are sufficiently close to the population value and can be used in its place for calculating the S.E of the estimate.

List of formulae for obtaining S.E for different statistics:

(1) S.E of mean :

When the S.D of population is known

$$S.E \bar{X} = \frac{\sigma_p}{\sqrt{n}}$$

Where $S.E \bar{X} \Rightarrow$ S.E of the mean

$\sigma_p \Rightarrow$ S.D of population

$n \Rightarrow$ no. of observations in the sample.

(2) When the S.D of population is not known, we use S.D of sample in calculating S.E of mean.

$$S.E \bar{X} = \frac{\sigma(\text{sample})}{\sqrt{n}}$$

$$(3) \quad S.E \text{ of median} = 1.25331 \cdot \frac{\sigma}{\sqrt{n}}$$

$$(4) \quad S.E \text{ of Quartiles} = S.E_a \\ = 1.36263 \frac{\sigma}{\sqrt{n}}$$

$$(5) \quad S.E \text{ of Quartile deviation} = 0.78672 \frac{\sigma}{\sqrt{n}}$$

⑥ S.E of mean deviation = $0.6028 \frac{\sigma}{\sqrt{n}}$

⑦ S.E of standard deviation = $\frac{\sigma}{\sqrt{2n}}$

⑧ S.E of variance = $\sigma^2 \sqrt{\frac{2}{n}}$

⑨ S.E of coefficient of correlation = $\frac{1-r^2}{\sqrt{n}}$

⑩ S.E of regression coefficient y on x

$$\text{S.E } b_{yx} = \frac{\sigma_y \sqrt{1-r^2}}{\sigma_x \sqrt{n}}$$

⑪ S.E of regression coefficient x on y

$$\text{S.E } b_{xy} = \frac{\sigma_x \sqrt{1-r^2}}{\sigma_y \sqrt{n}}$$

⑫ S.E of regression estimate x on y
= $\sigma_x \sqrt{1-r^2}$

⑬ S.E of regression estimate y on x
= $\sigma_y \sqrt{1-r^2}$

EX: 1. 5 students in a college were selected at random and their ages were found to be 18, 21, 19, 20 & 26

(a) Calculate the standard deviation of the ages in the sample.

(b) Calculate the standard error.

Sol : $n = 5$

$$\bar{x} = \frac{18 + 21 + 19 + 20 + 26}{5} = \frac{104}{5} = 20.8$$

$$S.D = \sigma = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

x	$x - \bar{x}$	$(x - \bar{x})^2$
18	-2.8	7.84
21	0.2	0.04
19	-1.8	3.24
20	-0.8	0.64
26	5.2	27.04

$$\sum (x - \bar{x})^2 = 38.8$$

$$\sigma = \sqrt{\frac{38.8}{4}} = 3.11448$$

$$SE = \frac{\sigma}{\sqrt{n}} = \frac{3.114}{\sqrt{5}} = 1.393$$

Ex: 2. In a certain university, the mean age of students is 20.5 with a standard deviation of 0.8

(a) Calculate the standard error of the mean if a sample of 25 students were selected.

(b) What would the standard error of the mean be if a sample of 100 students were selected?

Soln:

(a) $\sigma = 0.8$

$$S.E.\bar{x} = \frac{\sigma}{\sqrt{n}} = \frac{0.8}{\sqrt{25}} = \frac{0.8}{5} = 0.16$$

(b) $S.E.\bar{x} = \frac{\sigma}{\sqrt{n}} = \frac{0.8}{\sqrt{100}} = \frac{0.8}{10} = 0.08$

Note: If we increase 'n' (sample size) standard error will decrease
↓
of the mean.

$$\begin{array}{ccc} n & & n \\ 25 & \xrightarrow{\times 4} & 100 \end{array}$$

$$\begin{array}{ccc} S.E. & \xrightarrow{\div 2} & \\ 0.16 & & 0.08 \end{array}$$

As we increase 'n', we will get more accurate answer. (i.e. of $\bar{x}$)

Ex: 3. Calculate standard error of mean from the following data showing the amount paid by 100 firms in Kolkata on the occasion of Durga Puja.

Amount paid ('000 Rs)	39	49	59	69	79	89	99
No. of firms.	2	3	11	20	32	25	7

Soln : $S.E \bar{X} = \frac{\sigma}{\sqrt{n}}$

Here evaluate $\sigma = \sqrt{\left(\frac{1}{n} \sum fx^2\right) - (\bar{x})^2}$
 $= 13.11$

$$S.E \bar{X} = \frac{13.11}{\sqrt{100}} = 1.311$$

Ex: 4. A manufacturer intends that his electric bulbs have a life of 1000 hours. He tests a sample of 20 bulbs, drawn at random from a batch and discovers that the mean life of the sample bulbs is 990 hours with S.D of 22 hours.

Does this signify that the batch is not upto the standard?

(Given: The table value of t is 1% level of significance with 19 degrees of freedom is 2.539)

[The t distⁿ is used when sample size $n \leq 30$ & popⁿ S.D is unknown]
The t-statistic is defined as

$$t = \frac{\bar{x} - \mu}{s} \times \sqrt{n} \quad \text{where } s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

$$\bar{x} = 990, \text{ popⁿ mean} = 1000, \text{ S.D} = 22$$

$$n = 20,$$

$$H_0 : \text{mean} = 1000 \text{ hrs}$$

$$\text{vs } H_1 : \text{mean} \neq 1000 \text{ hrs}$$

$$S.D = 22 = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$$

$$(22)^2 = \frac{\sum (x_i - \bar{x})^2}{n}$$

$$(22)^2 \times 20 = \sum (x_i - \bar{x})^2$$

$$9680 = \sum (x_i - \bar{x})^2$$

Hence,

$$S = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

$$= \sqrt{\frac{9680}{19}}$$

$$= 22.5715$$

$$t \text{ statistic} = \frac{\bar{x} - \mu}{S} \times \sqrt{n}$$

$$= \frac{990 - 1000}{22.5715} \times \sqrt{20}$$

$$= -1.98$$

Now, $|t| = 1.98 < 2.539$ (t-table value given as 2.539)

// If $|t| < t\text{-table value} \Rightarrow \text{accept } H_0$ //

So accept H_0 .

Hence, 1% level of significance, we

accept the life of a bulb equal to 1000 hrs.

Ex: 5: In above Example,
take S.D = 25, $\bar{x} = 900$
& solve again.

Problem 4:

A sample of 6 persons in an office revealed an average daily smoking of 10, 12, 8, 9, 16, 5 cigarettes. The average level of smoking in the whole office has to be estimated at 90% level of confidence.

(Given: $t=2.015$ for 5 degree of freedom.)

Solution

Sample mean= 10, $s= 3.7417$

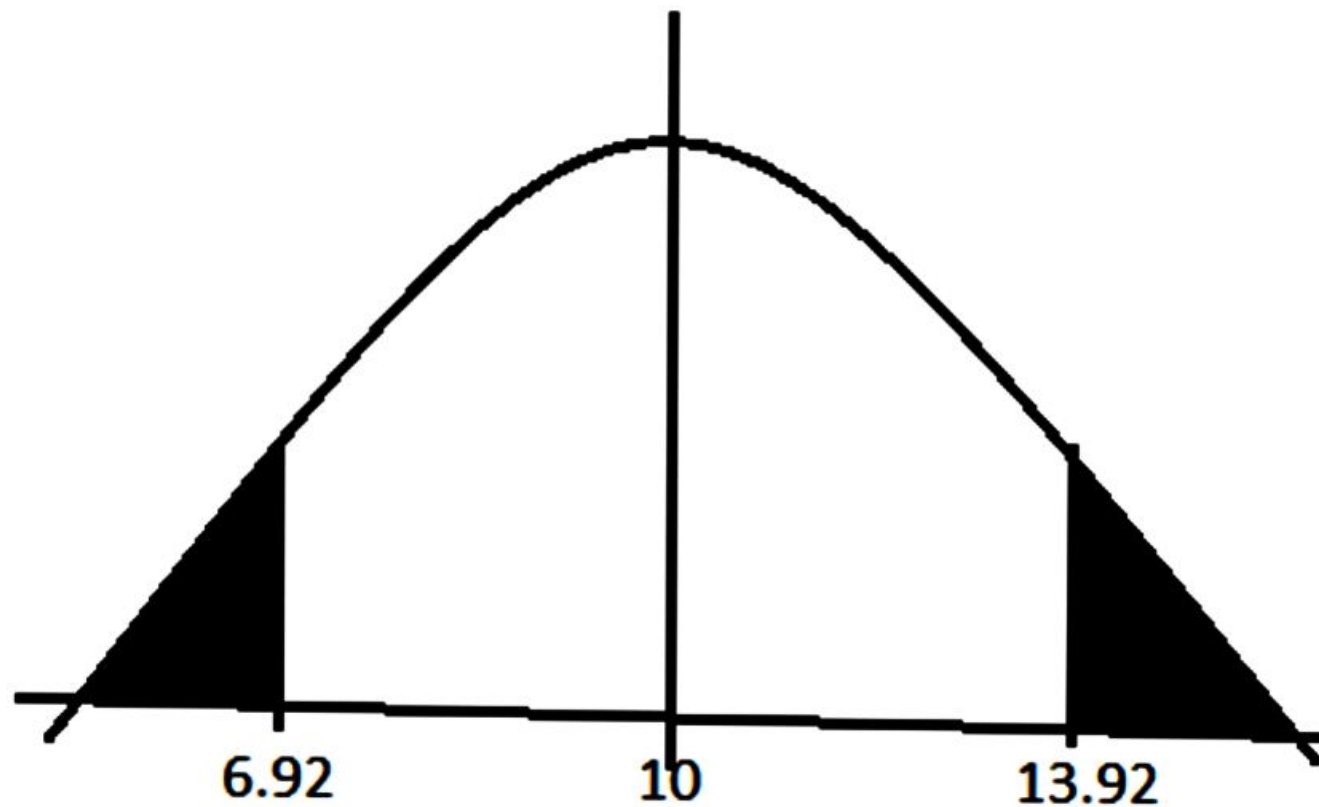
	X	(x-mean) ²
1	10	0
2	12	4
3	8	4
4	9	1
5	16	36
6	5	25
sum	60	70
x bar	10	14
sample sd	S	3.741657

$$T = \frac{(\bar{x} - \mu)}{\frac{s}{\sqrt{n}}} = \frac{(10 - \mu)}{\frac{3.7417}{\sqrt{6}}} = 2.015$$

$$t=2.015$$

$$\mu = 10 - 2.015 \cdot (3.7417/46) = 6.92$$

$$\mu = 10 + 2.015 \cdot (3.7417/46) = 13.08$$



Click on the format to make changes, make a copy of the file in the new format.

Exercise

The following data give the number of accidents occurred during the various days of a week.

Days	Mon	Tue	Wed	Thu	Fri	Sat	Total
No of accidents	15	19	13	12	16	15	90

Test whether the accidents are uniformly distributed over the week.

(Given: $\chi^2_{(v=5)} = 11.07$)

Solution:

H0: The number of accidents follows Uniform distribution

H1: The number of accidents do not follow Uniform distribution

Level of significance= 0.05

Chi square table =11.07

$$\chi^2 = \frac{\sum (O_i - E_i)^2}{E_i} \text{ degrees of freedom} = n - 1$$

Days	Mon	Tue	Wed	Thu	Fri	Sat	Total
O _i	15	19	13	12	16	15	90
E _i	15	15	15	15	15	15	90
(O _i -E _i) ² /E _i	0	1.06666667	0.266667	0.6	0.066667	0	2



Chi square calculated = 2 < 11.07

So, accept H_0

Conclusion

At 5% level of significance we accept the null hypothesis that is we could accept the number accidents are uniformly distributed on all days of a week.

Ex: 6. A sample of 100 tyres is taken from a lot. The mean life of tyres is found to be 39,350 kms. with S.D of 3,260.

Could the sample come from a population with mean life of 40,000 kms,

Establish 99% confidence limits within which the mean life of tyres is expected to lie.

Soln : Hypothesis : there is no significant difference between the sample mean & hypothetical population mean.

$$S.E \bar{x} = \frac{s}{\sqrt{n}} = \frac{3260}{\sqrt{100}} = 326.$$

$$\frac{\text{Difference}}{S.E} = \frac{40,000 - 39,350}{326} = 1.994$$

The difference is less than 2.58 S.E at 1% L.O.S & hence hypothesis is accepted.

& the difference could have arisen due to fluctuations in sampling

99.1% confidence limits $\bar{x} \pm 2.58(s.e)$

$$39350 \pm 2.58(326)$$

$$\therefore \text{f.e. } 38508.92 \text{ to } 40191.08$$

Hence the mean life of tyres is expected to lie between 38,509 to 40,191

Students t-distribution :

The t-distribution is used when sample size is 30 or less & the population std. deviation is unknown.

The t-statistic is defined as

$$t = \frac{\bar{x} - \mu}{s} \times \sqrt{n} \quad \text{where } s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

The t-distribution has been derived mathematically under the assumption

Note : The t-distribution has been derived mathematically under the assumption of a normally distributed population.

where,

$$f(t) = c \left(1 + \frac{t^2}{v} \right)^{-(v+1)/2}$$

$$t = \frac{\bar{x} - \mu}{s} \sqrt{n}$$

c = constant required to make the area under the curve unity
 $v = n - 1$, the no. of degree of freedom.

Properties :

① The t -distⁿ ranges from $-\infty$ to ∞
② Like std. normal distribution, t -distⁿ is symmetric & mean = 0

③ If $|t| < t_{0.05}$ (5% 1.d.s)
then accept H_0
otherwise reject it

11% for 1% 2.d.s.

Ex. 7. The life time of electric bulb for a random sample of 10 from a large consignment gave the following data.

Item	1	2	3	4	5	6	7	8	9	10
Life in '000 hrs	4.2	4.6	3.9	4.1	5.2	3.8	3.9	4.3	4.4	5.6

Can we accept the hypothesis that the avg life time of bulbs is 4,000 hrs.

Soln: Hypothesis: There is no significant difference in the sample mean & the hypothetical population mean.

$$t\text{-test} \Rightarrow t = \frac{|\bar{x} - \mu|}{s} \times \sqrt{n}$$

$$\bar{x} = \frac{44}{10} = 4.4$$

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{3.12}{9}} = 0.589$$

x	x	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
	4.2	-0.2	0.04
	4.6	0.2	0.04
	3.9	-0.5	0.25
	4.1	-0.3	0.09
	5.2	-0.8	0.64
	3.8	-0.6	0.36
	3.9	-0.5	0.25
	4.3	-0.1	0.01
	4.4	0	0
	5.6	-1.2	1.44

$$\sum (x_i - \bar{x})^2 = 3.12$$

$$t = \frac{|4.4 - 4|}{0.589} \sqrt{10} = 2.147$$

tabular value $t_{0.05} = 2.262$ for

$$n-1=9$$

d.o.f

$$2.147 < 2.262$$

Hence accept the hypothesis

Ex: 8 The manufacturer of a certain make of electric bulbs claims that his bulbs have a mean life of 25 months with std deviation of 5 months.

A random sample of 6 such bulbs given as follows.

Life in months : 24, 26, 30, 20, 20, 18

Can you regard the producer's claim to be valid at 1% I.O.S?

(Given that : $t_{5;0.01} = 4.032$, $t_{6;0.01} = 5.707$
 $t_{7;0.01} = 3.499$)

Soln : Hypothesis : There is no significant difference in the mean life of bulbs in the sample and that of the population.

$$t\text{-test: } t = \frac{\bar{x} - \mu}{s} \sqrt{n}$$

$$\begin{aligned}\bar{x} &= \frac{\sum x_i}{n} = \frac{24+26+30+20+20+18}{6} \\ &= 23\end{aligned}$$

x	$x - \bar{x}$	$(x - \bar{x})^2$
24	1	1
26	3	9
30	7	49
20	-3	9
20	-3	9
18	-5	25

$$\sum (x - \bar{x})^2 = 102$$

$$S = \sqrt{\frac{102}{5}} = 4.517$$

$$t = \frac{|23 - 25|}{4.517} \times \sqrt{6}$$

$$t = 1.084 < t_{5:0.01} = 4.03$$

Hence, Hypothesis is accepted.

III] Testing difference between means of two samples: (dependent samples)

In previous case, it was assumed that two samples are independent.

There are many situations where we get dependent samples (i.e. element of one sample is related to those of the other in any significant or meaningful manner)

i.e. there can be a case, two samples may consist of pairs of observations made on the same object

$$t\text{-test: } t = \frac{\bar{d} - 0}{\frac{s}{\sqrt{n}}}$$

where $\bar{d}$ = the mean of the differences
 s = S.D of differences

$$s = \sqrt{\frac{\sum (d - \bar{d})^2}{n-1}} \quad \text{OR} \quad \sqrt{\frac{\sum d^2 - n(\bar{d})^2}{n-1}}$$

Ex: 9 The following data relates to weights of 10 persons before joining a health club, & after 6 months ~~of~~ from joining it.

Test whether there is a significant change in the weights after 6 months from joining the club.

(Given: $t_9: 3.25$, $t_{10} = 3.16$)

Participant No.	1	2	3	4	5	6	7	8	9	10
wt before	120	125	115	130	123	124	122	127	128	118
wt. after	111	114	107	120	115	112	112	120	119	112

Hypothesis: There is no significant change in the weight after (6 months) joining the club.

Applying t-test: $t = \frac{\bar{d}\sqrt{n}}{s}$

Participant No	wt before x_1	wt. after x_2	$d = x_2 - x_1$	d^2
1	120	111	-9	81
2	125	114	-11	121
3	115	107	-8	64
4	130	120	-10	100
5	123	115	-8	64
6	119	112	-7	49
7	122	112	-10	100
8	127	120	-7	49
9	128	119	-9	81
10	118	112	-6	36

$$\sum d = -85 \quad \sum d^2 = 745$$

$$\bar{d} = \frac{\sum d}{n} = \frac{-85}{10} = -8.5$$

$$S = \sqrt{\frac{\sum d^2 - n(\bar{d})^2}{n-1}} = \sqrt{\frac{745 - 10(-8.5)^2}{9}} = 1.58$$

$$t = \frac{\bar{d}\sqrt{n}}{S} = \frac{(-8.5)\sqrt{10}}{1.58} = -17$$

$$\therefore |t| = 17 > t_9 = 2.26$$

$\therefore$ Hypothesis is rejected.

Hence there seems to be a significant difference in weights after joining club.

Ex: 10. A drug is given to 10 patients & the increments in their blood pressure were recorded to be
 3, 6, -2, 4, -3, 4, 6, 0, 0, 2

Is it reasonable to believe that the drug has no effect on change of blood pressure?

(5% value $t_9 = 2.26$)

→ H_0 : the drug has no effect on change of blood pressure.

$$t = \frac{\bar{d}\sqrt{n}}{s}$$

d	d^2
3	9
6	36
-2	4
4	16
-3	9
4	16
6	36
0	0
0	0
2	4

$$\sum d = 20 \quad \sum d^2 = 130$$

$$\bar{d} = \frac{\sum d}{n} = \frac{20}{10} = 2$$

$$s = \sqrt{\frac{\sum d^2 - n(\bar{d})^2}{n-1}}$$

$$= \sqrt{\frac{130 - 10 \times 4}{9}}$$

$$= 3.162$$

$$t = \frac{\bar{d}\sqrt{n}}{s} = \frac{2\sqrt{10}}{3.162}$$

$$= 2$$

Given: $t_g = 2.26$ at $5\% \cdot 1.0.5$

$$t = 2 < 2.26$$

Hence accept the hypothesis.