Mathematical statistics - UNIT-5

Small sample tests

Definition:- Consider a random sample $\{x_1, x_2, \ldots x_n\}$ of size n drawn from a Normal population with mean μ and variance σ^2.

The sample mean is $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$, and the unbiased estimate of the population variance σ^2 is denoted as S^2.

we define $S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$ Then the student's t - statistic is defined as $t = \frac{|\bar{x} - \mu|}{S/\sqrt{n}}$, where n = Sample size

The degrees of freedom of this statistic $\nu = n-1$.

t - test for Single mean

we proceed as follows.

(i) Formulate the null and alternative hypotheses.

$H_0 : \mu =$ a specified value

$H_1 : \mu \neq$ a specified value.

2. Select the level of significane.

3. The test statistic

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n-1}} \text{ (or) } t = \frac{\bar{x} - \mu}{S/\sqrt{n}}.$$

where s = sample S.D $S^2 = \frac{1}{n-1} \sum (x - \bar{x})^2$ (or) $S = \sqrt{\frac{ns^2}{n-1}}$.

4. Calculate the number of degrees of freedom $= v = n-1$

5. If the computed value of t is greater than the critical value t_α, H_0 is rejected (or) if $|t| < t_\alpha$, the null hypothesis is accepted at α level.

Applications of t-test :-

1. To test the significance of a single mean.
2. To test the significance of the difference between two sample means.
3. To test the significance of the coefficient of correlation.

Assumptions for students t-Test :-

1. The parent population from which the sample is drawn is normal.
(ii) The sample observations are independent i.e. the sample is random.
(iii) The population standard deviation σ is unknown.

Note: The test statistic

$$t = \frac{\bar{x} - \mu_0}{\sqrt{s^2/n}} = \frac{\bar{x} - \mu_0}{\sqrt{s^2/(n-1)}} \sim t_{n-1}.$$

Example - 1

A machinist is making engine parts with axle diameters of 0.700 inch. A random sample of 10 parts shows a mean diameter of 0.742 inch with a standard deviation of 0.040 inch. Compute the statistic you would use to test whether the work is meeting the specifications. Also state how you would proceed further.

Soln:

$\mu = 0.700$ inche $\bar{x} = 0.742$ inche.

$s = 0.040$ inche and $n = 10$.

Null Hypothesis $H_0 : \mu = 0.700$

Alternative Hypothesis $H_1 : \mu \neq 0.700$.

Test statistic $t = \dfrac{\bar{x} - \mu}{\sqrt{\frac{s^2}{n}}} = \dfrac{0.742 - 0.700}{\sqrt{(0.040)^2/10}}$

$= 3.15$

How to proceed further.

Degrees of freedom (d.f) $= n-1$

$= 10-1 = 9$.

Level of significance say 5%

Table value of t at 5% level $= 2.26 = t_0$

$\therefore t = 3.15, t_0 = 2.26$.

$3.15 > 2.26$ i.e. $|t| > t_0$

$\therefore H_0$ is rejected at this level of significance.

—— * ——

Example 2: The mean weekly sales of soap bars in departmental stores was 146.3 bars per store. After an advertising campaign the mean weekly sales in 22 stores for a typical week increased to 153.7 and showed a standard deviation of 17.2. was the advertising campaign successful?

Soln: $n = 22$, $\bar{x} = 153.7$, $s = 17.2$.

$\mu = 146.3$.

Null Hypothesis $H_0 : \mu = 146.3$

Alternative Hypothesis $H_1 : \mu > 146.3$ (Right tai

Test statistic: $t = \dfrac{\bar{x} - \mu}{\sqrt{s^2/(n-1)}} \sim t_{22-1} = t_{21}$.

$\therefore t = \dfrac{153.7 - 146.3}{\sqrt{\dfrac{(17.2)^2}{(22-1)}}} = 9.03$.

Tabulated value of t for 21 d.f. at 5% level of significance $= 1.72$

$d.f = n-1 = 22-1 = 21$.

$t = 9.03$. $t_0 = 1.72$

$|t| > t_0$ $\therefore$ H_0 is rejected at this level of significance. and conclude the advertising compaign was definitely successful in promoting sales.

———— x ————

Example - 3

A random sample of 10 boys had the following I.Q's : 70, 120, 110, 101, 88, 83, 95, 98, 107, 100. Do these data support the assumption of a population mean I.Q. of 100 ? Find a reasonable range in which of the mean I.Q. lie. Values of samples of 10 boys

Soln: Null Hypothesis $H_0 : \mu = 100$

Alternate Hypothesis $H_1 : \mu \neq 100$

Test statistic $t = \dfrac{\bar{x} - \mu}{\sqrt{s^2/n}} \sim t_{(n-1)}$.

calculation for sample mean and S.D:

x	$x - \bar{x}$	$(x - \bar{x})^2$
70	-27.2	739.84
120	22.8	519.84
110	12.8	163.84
101	3.8	14.44
88	-9.2	84.64
83	-14.2	201.64
95	-2.2	4.84
98	0.8	0.64
107	9.8	96.04
100	2.8	7.84
972		1833.60

Here $n = 10$, $\bar{x} = \dfrac{972}{10} = 97.2$

$s^2 = \dfrac{1833.60}{9} = 203.73$

$$\therefore |t| = \frac{|97.2 - 100|}{\sqrt{203.73/10}} = 0.62.$$

Tabulated $t_{0.05}$ for $(10-1) = 2.262 = t_0$

$d.f = n-1 = 10-1 = 9.$

$\because |t| < t_0$, H_0 is accepted at 5% level of significance and we may conclude that the data are consistent with the assumption of mean I.Q. of 100 in the population.

95% Confidence limits :-

$$\bar{x} \pm t_{0.05}\, s/\sqrt{n} = 97.2 \pm 2.262 \times 4.514$$

$$= 97.2 \pm 10.21$$

$$= 107.41 \text{ and } 86.99.$$

Hence the required 95% confidence interval is [86.99, 107.41].

———— * ————

t - Tests for Difference of means.

Let $\bar{x}_1$ and $\bar{x}_2$ be the means of the two sample and s_1^2 and s_2^2 be the samples variances. To test the difference between $\bar{x}_1$ and $\bar{x}_2$ is significant or not, we set up the null hyp: $H_0 : \mu_1 = \mu_2$

The alt. hyp: $H_1 : \mu_1 \neq \mu_2$

The test statistic is $t = \dfrac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$

where $s^2 = \dfrac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$

(or) $s^2 = \dfrac{\sum(x_1 - \bar{x}_1)^2 + \sum(x_2 - \bar{x}_2)^2}{n_1 + n_2 - 2}$

The no. of degrees of freedom $= v = n_1 + n_2 - 2$.

If $|t| < t_0$, H_0 is accepted.

If $|t| > t_0$, H_0 is rejected.

Example - 1

Samples of two types of Electric light bulbs were tested for length of life and following data were obtained:

	Type I	type II
Sample No	$n_1 = 8$	$n_2 = 7$
Sample means	$\bar{x}_1 = 1,234$ hrs	$\bar{x}_2 = 1,036$ hrs
Sample S.D's	$s_1 = 36$ hrs	$s_2 = 40$ hrs.

is the difference in the means sufficient to warrant that type I is superior to type II regarding length of life?

Soln:

Null Hypothesis $H_0 : \mu_x = \mu_y$

Alternative Hypothesis $H_1 : \mu_x \neq \mu_y$

i.e. H_0 : type I and type II are identical

H_1 : type I is superior to type II.

Test statistic $t = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2}$

where $s^2 = \dfrac{1}{n_1+n_2-2}\left[\sum(x_1-\bar{x}_1)^2 + \sum(x_2-\bar{x}_2)^2\right]$

$= \dfrac{1}{n_1+n_2-2}\left(n_1 s_1^2 + n_2 s_2^2\right)$

$= \dfrac{1}{13}\left[8(36)^2 + 7(40)^2\right]$

$= 1659.08.$

$\therefore t = \dfrac{1234 - 1036}{\sqrt{1659.08\left(\frac{1}{8}+\frac{1}{7}\right)}} = 9.39.$

degrees of freedom $= n_1+n_2-2 = 13.$

The tabulated value of t for 13 d.f at 5% level of significance for right tailed test is 1.77.

$\therefore t = 9.39, \quad t_0 = 1.77.$

$\therefore |t| > t_0$. H_0 is rejected

$\therefore$ The two types of electric bulbs differ significantly also type I is definitely superior to type II.

*

Example - 2:

Below are given the gain in weights (in kgs) of pigs fed on two diets A and B.

Diet A : 25, 32, 30, 34, 24, 14, 32, 24, 30, 31, 35, 25

Diet B: 44, 34, 22, 10, 47, 31, 40, 30, 32, 35, 18, 21, 35, 29, 22.

Test if two diets differ significantly as regards.

Soln: We set up. $H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$

L.O.S : $\alpha = 0.05$.

Test statistic $t = \dfrac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$, $s^2 = \dfrac{\Sigma(x_1-\bar{x}_1)^2 + \Sigma(x_2-\bar{x}_2)^2}{n_1+n_2-2}$

To find t:

Diet A			Diet B		
x_1	$x_1 - \bar{x}_1$	$(x_1-\bar{x}_1)^2$	x_2	$x_2-\bar{x}_2$	$(x_2-\bar{x}_2)^2$
25	-3	9	44	14	196
32	4	16	34	4	16
30	2	4	22	-8	64
34	6	36	10	-20	400
24	-4	16	47	17	289
14	-14	196	31	1	1
32	4	16	40	10	100
24	-4	16	30	0	0
30	2	4	32	2	4
31	3	9	35	5	25
35	7	49	18	-12	144
25	-3	9	21	-9	81
			35	5	25
			29	-1	1
			22	-8	64
$\Sigma x_1 = 336$	$\Sigma(x_1-\bar{x}_1) = 0$	$\Sigma(x_1-\bar{x}_1)^2 = 380$	$\Sigma x_2 = 450$	$\Sigma(x_2-\bar{x}_2) = 0$	$\Sigma(x_2-\bar{x}_2)^2 = 1410$

$\bar{x}_1 = \dfrac{\Sigma x_1}{n_1} = \dfrac{336}{12} = 28$ $\quad s^2 = \dfrac{\Sigma(x_1-\bar{x}_1)^2 + \Sigma(x_2-\bar{x}_2)^2}{n_1+n_2-2} = 71.6$

$\bar{x}_2 = \dfrac{\Sigma x_2}{n_2} = \dfrac{450}{15} = 30$ $\quad n_1 = 12, \ n_2 = 15.$

Test statistic :

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{28-30}{\sqrt{71.6\left(\frac{1}{12}+\frac{1}{15}\right)}}$$

$$= -0.609.$$

The number of d.o.f $= n_1+n_2-2$

$= 12+15-2$

$= 25$

The table value t for 25 d.o.f at 5% L.O.S. = 2.06.

$\therefore |t| = 0.609 \quad t_0 = 2.06.$

$\therefore |t| < t_0$. $\therefore$ H_0 is accepted at 5% level of significance.

The two diets do not differ significantly as regards their effect on increase in weight.

---*---

Chi-Square Test

Chi-square Distribution:

Definition:

If O_i $(i = 1, 2, \ldots n)$ are set of observed frequencies and E_i $(i = 1, 2, \ldots n)$ are the corresponding set of expected frequencies, then the statistic χ^2 is defined as $\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$

and the degrees of freedom this statistic is $v = n-1$.

Note: The density function of χ^2 is given by

$$f(\chi^2) = \frac{1}{2^{n/2}\Gamma(n/2)} \exp(-\chi^2/2)(\chi^2)^{n/2-1} d\chi^2, \quad 0 \le \chi^2 < \infty$$

and the degrees of freedom is n.

Degrees of freedom:

While comparing the calculated value of χ^2 with the table value, we have to determine the degrees of freedom.

The number of degrees of freedom is the total number of observations less the number of independent constraints imposed on the observations.

$\therefore$ d.f $= v = n-k$.

where k is the number of independent constraints in a set of data of n observations.

chi - Square Test of Goodness of fit

It is very powerful test for testing the significance of the discrepancy between theory and experiment. It is enables us to find if the deviation of the experiment from theory is just by chance or is it really due to the inadequacy of the theory to fit the observed data.

χ^2 test enable us to ascertain how well the theoretical distributions such as Binomial, poisson Normal etc. fit empirical distributions. i.e. distribution obtained from sample data.

If the calculated value of χ^2 is less than the table value at a specified level of significance, The fit is considered to be good.

If the calculated value of χ^2 is greater than the value, the fit is considered to be poor.

Conditions for Applying χ^2 Test:-

1. The sample observations should be independent.
2. Constraints on the cell frequencies, if any should be linear.
3. N. the total frequency should be reasona-bly large, say greater than 50.
4. No theoretical frequency should be less than 5.

If any theoretical cell frequency is less than 5. Then for application of χ^2 test, it is pooled with the preceeding or succeeding frequency so that the pooled frequency is greater than 5 and finally adjust for the d.f. lost in pooling.

Example-1

The following table gives the number of aircraft accidents that occurred during the various days of the week. Test whether the accidents are uniformly distributed over the week.

Days	Mon	Tue	Wed	Thu	Fri	Sat	Total
No. of accidents	14	18	12	11	15	14	84

Soln:

We set up H_0: The accidents are uniformly distributed over the week.

L.O.S : $\alpha = 0.05$

Test statistic : $\chi^2 = \sum_{i=1} \frac{(O_i - E_i)^2}{E_i}$

Under the null Hypothesis,

The expected frequency of the accidents on each day $= \frac{84}{6} = 14. = E_i$

O_i :	14	18	12	11	15	14
E_i :	14	14	14	14	14	14
$(O_i - E_i)^2$:	0	16	4	9	1	0

Scanned with CamScanner

$$\chi^2 = \frac{1}{14}[0+16+4+9+1+0]$$

$= 2.14.$

Number of degrees of freedom $v = n-1$
$= 7-1$
$= 6.$

Critical value:

The tabulated value of χ^2 at 5% level for 6 d.f. is $\chi^2_{0.05} = 12.59.$

Conclusion:

Since the calculated value of χ^2 is less than the tabulated value, (i.e) $2.14 < 12.59$

$\therefore$ we accept the null hypothesis H_0.

Hence we conclude that the accidents are uniformly distributed over the week.

Example 2:

A die is thrown 264 times with the following results

No. appeared on the die :	1	2	3	4	5	6
Frequency :	40	32	28	58	54	60.

Show that the die is biased.

Soln: We set up the null Hypothesis H_0: the die is unbiased and the distribution is uniform

L.O.S : $\alpha = 0.05$

Test statistic $\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$

Under the null hypothesis,

The Expected frequency of each of the number 1, 2, ... 6 is $= \frac{264}{6} = 44$.

O_i :	40	32	28	58	54	60
E_i :	44	44	44	44	44	44
$(O_i - E_i)^2$:	16	144	256	196	100	256

$$\therefore \chi^2 = \frac{1}{44}\left[16 + 144 + 256 + 196 + 100 + 256\right]$$

$$= 22.$$

No. of degrees of freedom $= \nu = n-1$
$= 6-1 = 5$.

Critical Value :

The table value of χ^2 at 5% level for 5 d.f $= 11.07$.

Conclusion :

The calculated value of χ^2 is greater than the table value

i.e. $22 > 11.07$. $\therefore H_0$ is rejected at 5% level.

The results of the experiment do not support the hypothesis.

Hence the die is biased.

Tests for Independence of Attributes

Let $s \times t$ contingency table

	A_1	A_2	......	A_t	Total
B_1	O_{11}	O_{12}	......	O_{1t}	(B_1)
B_2	O_{21}	O_{22}	......	O_{2t}	(B_2)
⋮	⋮	⋮		⋮	⋮
B_s	O_{s1}	O_{s2}	......	O_{st}	(B_s)
Total	(A_1)	(A_2)	..	(A_t)	N

Here N = Total frequency

O_{ij} = observed frequency

<u>Expected Frequency</u> $e_{ij} = \dfrac{(\text{row total } B_i)(\text{column total } A_j)}{N}$

where $i = 1, 2, \ldots s$ and $j = 1, 2, \ldots t$.

<u>Degrees of freedom:</u>

No. of degrees of freedom associated with a ~~set~~ $s \times t$ contingency table

$= (s-1) \times (t-1)$

<u>Chi - Square value for 2x2 Contingency table</u>

In 2x2 contingency table wherein the frequencies are

a	b
c	d

. The value of χ^2 is

$$\chi^2 = \frac{(a+b+c+d)(ad-bc)^2}{(a+b)(a+c)(c+d)(b+d)}.$$

Example - 1 An opinion poll was conducted to find the reaction to a proposed civic reform in 100 members of each of the two political parties. The information is tabulated below.

	Favourable	Unfavourable	Indifferent
Party A	40	30	30
Party B	42	28	30

Test for independence of reaction with the party affiliations.

Soln:

H_0: Reactions and the party affiliations are independent.

	1	2	3	Total
A	40	30	30	100
B	42	28	30	100
Total	82	58	60	200

To find Expected frequencies:

$$E(40) = \frac{82 \times 100}{200} = 41$$

$$E(30) = \frac{58 \times 100}{200} = 29$$

$$E(30) = \frac{60 \times 100}{200} = 30$$

$$E(42) = \frac{82 \times 100}{200} = 41$$

$$E(28) = \frac{58 \times 100}{200} = 29$$

$$E(30) = \frac{60 \times 100}{200} = 30.$$

$\therefore O_i$: 40 30 30 42 28 30

E_i : 41 29 30 41 29 30

$(O_i - E_i)^2$: 1 1 0 1 1 0

Test statistic $\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$

$\chi^2 = \frac{1}{41} + \frac{1}{29} + \frac{0}{30} + \frac{1}{41} + \frac{1}{29} + \frac{0}{30}$

$= \frac{1}{41} + \frac{1}{29} + \frac{1}{41} + \frac{1}{29}$

$= 0.12.$

No. of d.f $= (2-1)(3-1) = 2.$

The ~~cal~~ table value of χ^2 at 5%.

level for 2 d.f $= 5.99.$

Conclusion:

The calculated $\chi^2 <$ table value.

i.e, $0.12 < 5.99.$

$\therefore H_0$ is accepted at 5% level.

$\therefore$ The hypothesis of independent of reactions with the party affiliations may be correct.

Example - 2

From the following data, test whether there is any association between intelligency and economic conditions.

		Intelligency				
		Excellent	good	Medium	Dull	Total
Economic Conditions	Good	48	200	150	80	478
	Not good	52	180	190	100	522
	Total	100	380	340	180	N= 1000

Soln:

Ho: There is no association between intelli-gency and economic conditions.

To find Expected frequency:

$E(48) = \frac{478 \times 100}{1000} = 47.8$

$E(200) = \frac{478 \times 380}{1000} = 181.64$

$E(150) = \frac{478 \times 340}{1000} = 162.52$

$E(80) = \frac{478 \times 180}{1000} = 86.04.$

$E(52) = \frac{522 \times 100}{1000} = 52.2$

$E(180) = \frac{522 \times 380}{1000} = 198.36$

$E(190) = \frac{522 \times 340}{1000} = 177.48$

$E(100) = \frac{522 \times 180}{1000} = 93.96.$

Test statistic $= \chi^2 = \sum_{i=1} \frac{(O_i - E_i)^2}{E_i}$

$$\chi^2 = \frac{(48-47.8)^2}{47.8} + \frac{(200-181.64)^2}{181.64} + \frac{(150-162.52)^2}{162.52}$$

$$+ \frac{(80-86.04)^2}{86.04} + \frac{(52-52.2)^2}{52.2} + \frac{(180-198.36)^2}{198.36}$$

$$+ \frac{(190-177.48)^2}{177.48} + \frac{(100-93.96)^2}{93.96}$$

$= 6.2168$. [Check it].

No of degrees of freedom $= (2-1)(4-1)$
$= 3$

The tabulated value of χ^2 at 5% level for 3 d.f $= 7.815$

Conclusion:

Since the calculated χ^2 < table value.

i.e) $6.21 < 7.81$

$\therefore$ H_0 is accepted at 5% level.

Hence we conclude there is no association between intelligency and economic conditions.