CHAPTER-8

Sampling Theory (Smal & Large)

The group of individuals under study is called populationor universe. It may be finite or infinite and a part selected from the population is called a sample and process of selection of a sample is called sampling.

A random sample is one in which each member of population has an equal chance of being selected. There are ${}^{N}C_{n}$ different samples of size N. Mean, standerd deviation are para meters. The logic of sampling theory is the logic of induction. In Induction, we pass from a particular (sample) to general (population) which is known as Statistical Inference.

On the basis of sample information, we make decisions about the population. In taking such decisions, we make certain assumptions and these assumptions are known as statistical hyprothesis. These hypotheses are tested. Assuming the hypothesis correct, we calculate the probalility of getting the observed sample. If this probability is less than a certain assigned value, the hypothesis is to be rejected.

Null Hypothesps is the hyprothesis of no difference i.e. No differences in observed value and expected value and then we test whether this hypothesis is satisfied by the data or not. Null hypothesis is denoted by Ho.

Test of Significance:

The tests which enables us to decide whether to accept or to reject the null hyprothesis is called the test of significance. If the difference as the sample values and the population values are so high, it is to be rejected.

Level of Significane:

There are two critical regions which cover 5% and 1% areas of the normal curve. The probability of the value of the variable

falling in the critical region is the level of significance. If the variate falls in the critical region, the hypothesis is to be rejected.

Test of significance (limit)

Normal distribution is the limiting case of binomial distribution when x is large enough For normal distribution 5% of the item lie outside $\mu \pm 1.96\sigma$ while only 1% of the items lie outside $\mu \pm 2.586\sigma$

$$z = \frac{x - \mu}{\sigma}$$

Where z is the standard normal variate and x is observed number of success.

Test of significance depends on value of z.

(a) If $|z| < 1.96$, difference between observed and expected values is not significant at 5% level of significance.

(b) If $|z| > 1.96$, difference is significant at 5% level of significance.

(c) If $|z| < 2.58$, difference between between observed and expected values is not significant at 1% level of significance.

(d) If $|z| > 2.58$, difference is significant at 1% level of significance.

Tests:

t-test.

The student t-distribution is used to test the significance of the mean of small sample t is difined as

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

Where $\bar{x}$ = mean of sample

M = mean of population

S' = S.D. of sample

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

Working Rule :

To calculate significance of sample mean at 5% level, calculate $t = \frac{\bar{x} - \mu}{s} \times \sqrt{n}$ and campare it with *(n-1)* degree of freedom at 5% level obtained from table. Let tabulated value be t_1 then

a) If $t<t_1$, then we accept the hypothesis ie. We say that the value of t is significant.
b) If $t>t_1$, we cmpare it with the tabulated value of t at 1% level of significant for *(n-1)* degree of freedom and denot it by t_2.
c) If $t_1<t<t_2$, then value of t is significant (accept).
d) If $t>t_2$, we reject the hypothesis and sample is not drawn from the population.

Problem: Ten individuals are chosen at random from a population and their heights are found to be in inches 63, 63, 64, 65, 66, 69, 70, 70, 71. Discuss the suggestin that the mean height of universe is 65. For 9 degree of freedom at 5% level of significance = 2.262

Solution:

x	***x-67***	***(x-67)²***
63	-4	16
63	-4	16
64	-3	9
65	-2	4
66	-2	4
69	-1	1
69	+2	4
70	+3	9
70	+3	9
71	4	16
$\sum x = 670$		$\sum (x-\bar{x})^2 = 88$

$$\bar{x} = \frac{\sum x}{x} = \frac{670}{67} = 67$$

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}} = \sqrt{\frac{88}{9}} = 3.13$$

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{67 - 65}{\frac{3.13}{\sqrt{10}}} = \frac{2\sqrt{10}}{3.13} = 2.02$$

$2.02 < 2.262$

Calculated value of t (2.02), is less then the tabulated value of t (2.262), hence the hypothesis is accepted so mean height of universe is 65 inches.

Problem: The mean life time of sample of 100 fluorescent light bulbs produced bty a company is computed to be 1570 hours with a standard deviation of 120 hours. The company claims that the average life of the bulbs produced by it is 1600 hours using the level of significance 5%. Is claim acceptable?

Solution:

$$\bar{x} = 1570,\ \ s = 120,\ \ n = 100,\ \ \mu = 1600$$

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{1570 - 1600}{\frac{120}{\sqrt{100}}} = \frac{1570 - 1600}{12} = 2.5$$

At 0.05 level of significance $t=1.96$

Calculated value of t> tabulated value of t

$$2.5 > 1.96$$

Hence the claim is to be rejected.

Chi-Square Test:

The square of a standard normal variate z is known as chi-square variate (x^2) with one degree function.

$z = \frac{x - \mu}{\sigma}$ is a normal variate, hence $\left(\frac{x - \mu}{\sigma}\right)^2$ is a chi-square variate.

Condition for Chi-square Test :

1. The sample under study must be large (not less then 50)
2. Member of cell should be independent
3. Cell frequencies of each cell should be greater than 5.

Working Rule to Calculate χ^2

Step-I : Calculate the expected frequencies E_i

Step-II. Calculate $\frac{(0i-Ei)^2}{Ei}$.

Step-III. Add all these then $\chi^2 = \frac{\sum_{i=1}^{n}(0i-Ei)^2}{Ei}$

Step-IV : The value of x^2 lies between 0 and ∞

Degree of Freedom

Cast-I: If the data is given in the form of series of variables in a row or column then the degree of freedom = (Number of item in series) - 1

Case-II: Degree of freedom = (R-1) (C-1)

R is number of rows and C is number of columns If frequencies are given in the form of table.

Remark: If calucated value < Tabulated value then the null hypothesis is accepted otherwise not accepted.

Problem: A survey of 320 families with 5 children is given below

No. of boys	5	4	5	2	1	0	**Total**
No. of girls	0	1	2	3	4	5	
No. of families	14	56	110	88	40	12	**320**

Is this result consistent with hypothesis i.e. male and female birth are equally possible.

Solution: Null hypothesis Ho:

Male and female births are equally probable.

Calculation of Expected frequencies $(q+p)^n$

Probability of female birth p = ½

probability of male birth q = ½

$$(q+p)^n = q^n + {}^nC_1 pq^{n-1} + {}^nC_2 p^2 q^{n-2} + {}^nC_3 p^3 q^{n-3} + \ldots\ldots\ldots\ldots + p^n$$

$$\left(\frac{1}{2}+\frac{1}{2}\right)^5 = \left(\frac{1}{2}\right)^5 + 5\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)^4 + 10\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^3 + 10\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^2 + 5\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right) + \left(\frac{1}{2}\right)^5$$

So, Number of girls =

$$= 320\left[\frac{1}{32}+\frac{5}{32}+\frac{10}{32}+\frac{40}{32}+\frac{5}{32}+\frac{1}{32}\right]$$

$$= 320\times\frac{1}{32}+320\times\frac{5}{32}+320\times\frac{10}{32}+320\times\frac{40}{32}+320\times\frac{5}{32}+320\times\frac{1}{32}$$

$$= 10+50+100+100+50+10$$

These are expected frequencies of the female birth.

O	**E**	**O - E**	**(O-E)^2**	$\frac{(O-E)^2}{E}$
14	10	4	16	1.60
56	50	6	36	0.72
110	100	10	100	1.00
88	100	-12	144	1.44
40	50	-10	100	2.00
12	10	2	4	0.40
				$\sum = 7.16$

Level of significance = 0.05

The tabulated value of χ^2 at α =0.05 for (6 - 1) (2 - 1) = 5 degree of freedom is 11.07.

Since the calculated value of χ^2 (7.16) < tabulated value of χ^2 at level of significance 0.05 for *5 degree of freedom = 11.07*

So null hypothesis is accepted i.e. male and female birth is equally probable.

Problem: The table below gives the number of air craft accidents that occurred during the various days of the week. Test wether the accidents are uniformally distributed over the week.

Days	M	T	W	T	F	S	Sun	Total
Number of Accidents	14	18	12	11	15	14	14	98

Solution: Ho: Null hypothesis: accidents are uniformally distributed.

The expected frequencies of the accidents on each day $= \frac{98}{7} = 14$

O	**E**	**O - E**	**(O-E)^2**	$\frac{(O-E)^2}{E}$
14		0	0	0
18		4	16	1.14

12	14	-2	4	0.29
11		-3	9	0.64
15		1	1	0.07
14		0	0	0
14		0	0	0
98				$\sum = 2.14$

The tabulated value of χ^2 at α=0.05 is for (7-1) (2-1) i.e. 6 degree is $\chi^2 = 12.592$

Since the calculated value of χ^2 (2.14) < tabulated value of χ^2 at level of significance 0.05 for six degree of freedom = 12.592

Hence null hypothesis is accepted i.e. air craft accidents are uniformally distributed over the week.

Chi-square test as a test of independence

$$\text{Expected frequency} = \frac{\text{Row total} \times \text{Column total}}{\text{Grand total}}$$

Problem: The following table gives a classification of a sample of 160 plants of their flower colour and flatness of leaf

	Flat leafs	**Coloured leafs**	**Total**
White flower	99	36	135
Red Flower	20	5	35
Total	**119**	**41**	**160**

Test whether the flower colour is independent of the flat ness of leafs.

Solution: Null hypothesis : There is no association between colour flowers and flatness of leafs.

$$\text{Expected frequency} = \frac{\text{Row total} \times \text{Column total}}{\text{Grand total}}$$

	Flat leafs	**Colour leafs**	**Total**
White flower	$\frac{135\times119}{160} = 100.41$	$\frac{135\times41}{160} = 34.59$	135
Red Flower	$\frac{25\times119}{160} = 18.59$	$\frac{25\times41}{160} = 6.41$	25
Total	**119**	**41**	**160**

Observed value (0)	Expected value (E)	O - E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
99	100.41	-1.41	1.99	0.019
20	18.59	1.41	1.99	0.107
36	34.59	1.41	1.99	0.056
5	6.41	-1.41	1.99	0.310
			Total	0.492

Degree of freedom = (R-1) (C-1) = (2-1) (2-1) = 1

The tabulated value of χ^2 at α =0.05 for 1 degree of freedom is 3.841. Since the tabulated value of χ^2 (0.492) is less than the tabulated value (3.841) hence null hypothesis should be rejected, i.e. they are independent flatness of leafs with the colour of flower i.e. there is no association between them.

Tests of Significance for Large Samples: Suppose a large number of sample is classified according to the frequencies of an attribute,

If the number of successes in a large sample of size *n* differs from the expected value np, by more than $3\sqrt{npq}$ we call the difference highly significant and the truth of the hypothesis is very improbable.

Definition: The testing of a statistical hypothesis is meant a procedure for deciding whether to accept or reject the hypothesis.

Generally we accept the hypothesis as correct and then we calculate np, npq , and apply the above test.

Example: A coin is tossed 400 times and it turns up head 216 times. Discuss whether the coin may be unbiased one.Suppose that the coin is unbiased.Expected number of heads in 400 tosses = np = 400 x ½ = 200

The deviation of the actual number of head from expected = 216 – 200 = 16.

The standard deviation = $\sqrt{npq}$ = $\sqrt{400\,\frac{1}{2}\,\frac{1/2}{}}$ =10

The deviation is only 1.6 times the standard deviation and hence it is likely to appear as a result of fluctuations of simple sampling. We conclude that the coin may be taken as unbiased one.

Assignment

1. The following is the distribution of 106 nine pig-litters according to the numbers of males in the litters-

No.of Males	0	1	2	3	4	5	6	7	8	Total
No.of Litters	6	5	8	22	23	25	12	1	4	106

Fit a Binomial distribution under the hypothesis that the sex ratio is 1:1.Test the goodness of fit. Given that χ^2 for 4 degrees of freedom at 5% level of significance = 9.488.

2. The following table gives the number of aircraft accident that occurred during the various days of the week. Find whether the accident is uniformly distributed over the week.

Days	**Mon.**	**Tue.**	**Wed.**	**Thur.**	**Fri.**	**Sat.**	**Total**
No. of Accident	**14**	**16**	**12**	**19**	**9**	**14**	**84**

[Given that χ^2 for 5 degrees of freedom at 5% level of significance=11.07].

3. Records taken of the number of male and female birth in 800 families having four children are as follows-

No.of Male Births	No.of Female Births	No.of Families
0	4	32
1	3	178
2	2	290

3	1	236
4	0	64

Test whether the data are consistent with the hypothesis that the binomial law holds and that the chance of a male birth is equal to that of a female birth, that is q=p=1/2.You may use the table given below-

Degrees of Freedom	1	2	3	4	5
5%Value of χ^2	3.84	5.99	7.82	9.49	11.07

4. The following data shows the suicides of women in eight German states during fourteen years-

No. of suicides in a state per year	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Total**
ObservedFrequency	**9**	**19**	**7**	**20**	**15**	**11**	**8**	**2**	**3**	**5**	**3**	**112**

Fit a Poisson distribution and test its goodness of fit. The value of χ^2 for 6d.f. at 5% level of significance=12.592

5. Genetic theory states that children having one parent of blood type M and the other of blood type N will always be one of the three types M,MN,N and that the proportions of three type s will be on average as 1:2:1.A report states that oout of 300 childern having one M parent and N parent 30% were found to be type M,45%type MN and remainder type N Test the hypothesis by χ^2 for 2 degrees of freedom at 5% level 5.991.

6. The following table occurs in a memoir of karl pearson:

Eye colour in father	Eye colour in sons	
	Not light	Light
Not light	230	148
Light	300	471

Test whether the colour of the son's eyes is associated with that of the father's (χ^2=3.84,v=1)

7. In an experiment on the immunization of goats from anthrax the following results were obtained. Derive your inference on the efficiency of the vaccine-

	Died of Anthrax	Survived	Total
Inoculated withVaccine	2	10	12
Not Inoculated	6	6	12
Total	8	16	24

The value of χ^2 for 1 degree of freedom at 5% is 3.481.

8. The following table gives the series of controlled experiment. Discuss whether the treatment may be considered to have any positive effect-

	Positive	No effect	Negative	Total
Treatment	9	2	1	12
Control	3	6	3	12
Total	12	8	4	24

The value of χ^2for2 degrees of freedom at 5% level of significance is 5.99.

9. In an experiment on immunization of cattle from tuberculosis, the following results were obtained-

	Affected	Unaffected
Inoculated	12	26
Not inoculated	16	6

Examine the effect of vaccine in controlled susceptibility to tuberculosis.

10. Find the value of χ^2 for the following table-

Diet	Males	Females	Total
A	123	153	276
B	145	150	295
Total	268	303	571

11. Five's computed for four fold tables is independent replications of an experiment are 0.50,4.10,1.20,2.79 and 5.41. Does the aggregate of these tests yield a significant? Given-

$\chi_{0.05}^2$	**9.488**	**11.070**	**12.592**	**14.067**
Degrees of freedom	**4**	**5**	**6**	**7**

12. Write a short note on application of χ^2 in tests of significance.

13. What is the use of chi-square distribution in tests of goodness of fit?

14. Explain briefly the uses and limitations of χ^2–test.

15.. Write a short note on tests based on χ^2–distribution.

16. Explain any two applications of χ^2–test.

17. Discuss conditions for applications of χ^2–test.

18. Write down an expression for testing the independence of two attributes.

ANSWERS :

1.f_e: 0.4,3.3,11.6,23.2,29,23.2,11.6,3.3,0.4; $\chi 2$=2.52,v=4.fit is good

2. Accidents appears to be uniformly distributed over all days of the week.

3. χ_2^2 =19.63>$\chi_{0.0}^2$(4).

4.theoretical frequencies are:4,12,21,24,21,15,8,4,2,1,0, .

$\chi 2$ =18.35,v=6.

5. f_e:75,150,75; . χ^2 =4.5. 6. Appears to be associated.

7. Survival is not associated with in occulation of vaccine.

9.Vaccine is effective.

Tutorial

1. Describe how to test the significance of an observed correlation coefficient when the corresponding population value is 0.
2. A random sample of 11 observations from a bivariate population gave a correlation coefficient 0.239. Could the observed value have arisen from an uncorrelated population

Ans. T= 0.74, $t_{0.05}$ (9) =2.262, H_0 is accepted

3. A random sample of size 15 from a bivariate normal population gave a correlation coefficient of – 0.5. Is this an indication of the existence of correlation in the population?

Ans. T = -2.082 is not significant

4. Show that in the random sample of size 25 from an uncorrelated normal population the chance is 1 in 100 that *r* is greater than about 0.43.

5. Find the least value of r in a sample of 27 paired observations from a bivariate normal population that is significant at *5%* level of significance.

Ans. | r | > 0.38

6. Discuss any two tests of significance based on t-distribution

7. Calculate the value of *t* in the case of two characters *A* and *B* whose corresponding value are given below:

A	16	10	8	9	9	8
B	8	4	5	9	12	4

Ans. t = 1.66.

8. The figures below are for protein tests of the same variety of wheat grown in two districts. The average in District I is 12.74 and in District II is 13.03. Calculate r for testing the significance between the means of the two districts:

Protein results								
District	I	12.6	13.4	11.9	12.8	13		
District	II	13.1	13.4	12.8	13.5	13.3	12.7	12.4

Ans. t = 0.85

9. In a Test Examination given to two groups of students the marks obtained were as follows:

First group	18	20	36	50	49	36	34	49	41
Second group	29	28	26	35	30	44	46		

Examine the significance of difference between the arithmetic averages of the marks secured by the students of the above two groups.
(The value of *t* for 14 degrees of freedom at 5% level of significance = 2 .14.

Ans. Not significant

10. For a random sample of 12 boys fed on diet *A,* the increases in weight in pounds in a certain period were

25, 32, 30, 34, 24, 25,14, 32,24, 30, 31, 35.

For another random sample of 15 boys fed on diet *B,* the increase weight in pounds in the same period were

44, 34, 22,10,47, 31,40, 30, 32, 35,18, 21, 35, 29, 22. Find whether diet *B* is superior to diet *A*.

Given that the value of *t* for 25 degrees of freedom at 5% level of significance is 2 06.

Ans. Diet B is superior to dit A

11. The means of two random samples of sizes 9 and 7 respectively, are 196.4 and 198.82 respectively. The sum of the squares of the deviations from the means are 26.94 and 18.73 respectively. Can the samples be considered to have been drawn from the same normal population?

It being given that the value of t for 14 d.f. at 5% level of significance is 2.145 and at 1% level of significance is 2.977.

Ans. t = 2.65

12. Two types of batteries-A and *B* are tested for their length of life and following results are obtained:

No of Samples		Mean	Variance
A	10	500 hours	100
B	10	560 hours	121

Is their a significant difference in the two means?

The value of t for 18 degrees of freedom at 5% level of significance is 2.1.

Ans. Not significant