Year	Subject Title	Sem	Sub Code
2018–19 Onwards	II B.Sc Psychology - STATISTICS - II	IV	

Objective: To impart the basic knowledge of Statistical tools and their applications in Psychology.

UNIT I

Probability Distribution – Binomial, Poisson and Normal Distributions – Properties and Applications (without Proof) – Simple Problems.

UNIT II

Sampling – Advantages and Disadvantages – Simple Random Sampling – Stratified Random Sampling – Systematic Sampling – (Concept Only) – Sampling Distribution – Standard Error – Tests of Significance – Type I and Type II Errors – Large Sample Tests for Single Mean and Two Means. Tests for single proportion and difference of two proportions.

UNIT III

Small Sample Tests – Test for Single Mean and Two Means – Paired 't' Test Chi-Square Test for Independence of Attributes. Association of Attributes – Contingency Tables – Methods of Studying Association – Yule's Coefficient of Association

UNIT IV

Measurement and scaling techniques- Categorical variables-Data types-Metric, Interval and Ratio data. Non-Metric data- Nominal, ordinal data. Scales of measurement -Comparative scale, paired Comparison scale, rank order scale, constant sum scale, Non-comparative scale-continuous rating scale, Itemized rating scale- Likert scale, Guttmann scale

UNIT V

Non – Parametric Tests– Introduction advantages and disadvantages. Run test, Sign test, Median test, Mann-Whitney U test(one sample only) Kolmogrov Smirnov test(two samples).

Text Books:

1. R.S.N. Pillai and V. Bagavathi - Statistics – Theory and Practice, S.Chand & Sons Company Ltd, New Delhi.
2. S.C.Gupta and V.K.Kapoor - Fundamentals of Applied Statistics, Sultan Chand & Sons, New Delhi, 11th revised Edition, June 2012.
3. J.P Verma and Mohammed Ghufran- Statistics for Psychology, Tata Mcgraw Hill Education (P)Ltd. New Delhi.

UNIT – III

1

6. TESTS OF SIGNIFICANCE
(Small Samples)

6.0 Introduction:

In the previous chapter we have discussed problems relating to large samples. The large sampling theory is based upon two important assumptions such as

(a) The random sampling distribution of a statistic is approximately normal and

(b) The values given by the sample data are sufficiently close to the population values and can be used in their place for the calculation of the standard error of the estimate.

The above assumptions do not hold good in the theory of small samples. Thus, a new technique is needed to deal with the theory of small samples. A sample is small when it consists of less than 30 items. ($n < 30$)

Since in many of the problems it becomes necessary to take a small size sample, considerable attention has been paid in developing suitable tests for dealing with problems of small samples. The greatest contribution to the theory of small samples is that of Sir William Gosset and Prof. R.A. Fisher. Sir William Gosset published his discovery in 1905 under the pen name 'Student' and later on developed and extended by Prof. R.A.Fisher. He gave a test popularly known as 't-test'.

6.1 t - statistic definition:

If $x_1, x_2, \ldots\ldots x_n$ is a random sample of size n from a normal population with mean μ and variance σ^2, then Student's t-statistic is defined as $t = \dfrac{\bar{x} - \mu}{\dfrac{S}{\sqrt{n}}}$

where $\bar{x} = \dfrac{\sum x}{n}$ is the sample mea and $S^2 = \frac{1}{n-1}\sum(x-\overline{x})^2$ 2

is an unbiased estimate of the population variance σ^2 It follows student's t-distribution with $\nu = n-1$ d.f

6.1.1 Assumptions for students t-test:

1. The parent population from which the sample drawn is normal.
2. The sample observations are random and independent.
3. The population standard deviation σ is not known.

6.1.2 Properties of t- distribution:

1. t-distribution ranges from $-\infty$ to ∞ just as does a normal distribution.
2. Like the normal distribution, t-distribution also symmetrical and has a mean zero.
3. t-distribution has a greater dispersion than the standard normal distribution.
4. As the sample size approaches 30, the t-distribution. approaches the Normal distribution

$-\infty \quad -t_\alpha \quad t=0 \quad t_\alpha$

Applications of t-distribution: 3

The t-distribution has a number of applications in statistics. of which we shall discuss the following in the coming sections:

(i) t-test for significance of single mean, population variance being unknown.

(ii) t-test for significance of the difference between two sample means, the population variances being equal but unknown.

(a) Independent samples

(b) Related samples: paired t-test

6.2 Test of significance for Mean:

We set up the corresponding null and alternative hypotheses as follows:

4

$H_0: \mu = \mu_0$; There is no significant difference between the sample mean and population Mean.

$H_1: \mu \neq \mu_0$ ($\mu < \mu_0$ (or) $\mu > \mu_0$)

Level of significance:

5% or 1%

Calculation of statistic:

Under H_0 the test statistic is

$$t_0 = \left|\frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}}\right| \quad \text{or} \quad \left|\frac{\bar{x} - \mu}{s / \sqrt{n-1}}\right|$$

$$\text{where } \bar{x} = \frac{\sum x}{n} \text{ is the sample mean}$$

$$\text{and} \quad S^2 = \frac{1}{n-1}\sum(x - \bar{x})^2 \quad \text{(or)} \quad s^2 = \frac{1}{n}\sum(x - \bar{x})^2$$

Expected value :

$$t_e = \left|\frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}}\right| \sim \text{student's t-distribution with (n-1) d.f}$$

1.4 **Inference :**

If $t_0 \leq t_e$ it falls in the acceptance region and the null hypothesis is accepted and if $t_0 > t_e$ the null hypothesis H_0 may be rejected at the given level of significance.

Example 1:

Certain pesticide is packed into bags by a machine. A random sample of 10 bags is drawn and their contents are found to weigh (in kg) as follows: 48 46 45 49 45 50 49 52 44 45

Test if the average packing can be taken to be 50 kg.

Solution:

Null hypothesis:

$H_0 : \mu = 50$ kgs in the average packing is 50 kgs.

mandatory

Alternative Hypothesis:

$H_1 : \mu \neq 50$kgs (Two -tailed)

Level of Significance:

Let $\alpha = 0.05$

Calculation of sample mean and S.D

X	d = x –48	d^2
50	2	4
49	1	1
52	4	16
44	–4	16
45	–3	9
48	0	0
46	–2	4
45	–3	9
49	+1	1
45	–3	9
Total	–7	69

$$\bar{x} = A + \frac{\Sigma d}{n}$$

$$= 48 + \frac{-7}{10}$$

$$= 48 - 0.7 = 47.3$$

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

$$= \frac{1}{9}\left[69 - \frac{(7^2)}{10}\right]$$

$$= \frac{64.1}{9} = 7.12$$

Calculation of Statistic:

Under H_0 the test statistic is :

$$t_0 = \left|\frac{\bar{x} - \mu}{\sqrt{S^2/n}}\right|$$

$$= \left|\frac{47.3-50.0}{\sqrt{7.12/10}}\right|$$

6

$$= \frac{2.7}{\sqrt{0.712}} = 3.2$$

Expected value: ?

$$t_e = \left|\frac{\bar{x}-\mu}{\sqrt{S^2/n}}\right| \text{ follows t distribution with } (10-1) \text{ d.f}$$

$= 2.262$ (3.2 - 1)

Inference:

Since $t_0 > t_e$, H_0 is rejected at 5% level of significance and we conclude that the average packing cannot be taken to be 50 kgs.

Example 2:

A soap manufacturing company was distributing a particular brand of soap through a large number of retail shops. Before a heavy advertisement campaign, the mean sales per week per shop was 140 dozens. After the campaign, a sample of 26 shops was taken and the mean sales was found to be 147 dozens with standard deviation 16. Can you consider the advertisement effective?

Solution:

We are given

$n = 26$; $\bar{x} = 147$dozens; $s = 16$

Null hypothesis:

H_0: $\mu = 140$ dozens i.e. Advertisement is not effective.

Alternative Hypothesis:

H_1: $\mu > 140$kgs (Right -tailed) upto 50 kg two tailed (Pg 146)

Calculation of statistic:

Under the null hypothesis H_0, the test statistic is

$$t_0 = \left|\frac{\bar{x}-\mu}{S/\sqrt{n-1}}\right|$$

$$= \left|\frac{147-140}{16/\sqrt{25}}\right| = \frac{7\times 5}{16} = 2.19$$

7

Expected value:

$$t_e = \left| \frac{\bar{x} - \mu}{s/\sqrt{n-1}} \right|$$ follows t-distribution with (26-1) = 25d.f

$= 1.708$

Inference:

Since $t_0 > t_e$, H_0 is rejected at 5% level of significance. Hence we conclude that advertisement is certainly effective in increasing the sales.

6.3 Test of significance for difference between two means:

6.3.1 Independent samples:

Suppose we want to test if two independent samples have been drawn from two normal populations having the same means, the population variances being equal. Let $x_1, x_2, \ldots x_{n_1}$ and $y_1, y_2, \ldots y_{n_2}$ be two independent random samples from the given normal populations.

Null hypothesis:

$H_0 : \mu_1 = \mu_2$ i.e. the samples have been drawn from the normal populations with same means.

Alternative Hypothesis:

$H_1 : \mu_1 \neq \mu_2$ ($\mu_1 < \mu_2$ or $\mu_1 > \mu_2$)

Test statistic:

Under the H_0, the test statistic is

$$t_0 = \left| \frac{\bar{x} - \bar{y}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right|$$

where $\bar{x} = \frac{\Sigma x}{n_1}$; $\bar{y} = \frac{\Sigma y}{n_2}$

and $S^2 = \frac{1}{n_1 + n_2 - 2} [\Sigma(x-\bar{x})^2 + \Sigma(y-\bar{y})^2] = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$

Expected value:

8

$$t = \frac{\bar{x} - \bar{y}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$ follows t-distribution with $n_1 + n_2 - 2$ d.f

Inference:

If the $t_0 < t_e$ we accept the null hypothesis. If $t_0 > t_e$ we reject the null hypothesis.

Example 3:

A group of 5 patients treated with medicine 'A' weigh 42, 39, 38, 60 and 41 kgs: Second group of 7 patients from the same hospital treated with medicine 'B' weigh 38, 42 , 56, 64, 68, 69 and 62 kgs. Do you agree with the claim that medicine 'B' increases the weight significantly?

Solution:

Let the weights (in kgs) of the patients treated with medicines A and B be denoted by variables X and Y respectively.

Null hypothesis:

$H_0 : \mu_1 = \mu_2$

i.e. There is no significant difference between the medicines A and B as regards their effect on increase in weight.

Alternative Hypothesis:

$H_1 : \mu_1 < \mu_2$ (left-tail) i.e. medicine B increases the weight significantly.

Level of significance : Let $\alpha = 0.05$

Computation of sample means and S.Ds

Medicine A

X	$x - \bar{x}$ ($\bar{x} = 46$)	$(x - \bar{x})^2$
42	−4	16
39	−7	49
48	2	4
60	14	196
41	−5	25
230	**0**	**290**

9

$$\bar{x} = \frac{\sum x}{n_1} = \frac{230}{5} = 46$$

Medicine B

Y	$y - \bar{y}$ $(\bar{y} = 57)$	$(y - \bar{y})^2$
38	–19	361
42	–15	225
56	–1	1
64	7	49
68	11	121
69	12	144
62	5	25
399	**0**	**926**

$$\bar{y} = \frac{\sum y}{n_2} = \frac{399}{7} = 57$$

$$S^2 = \frac{1}{n_1 + n_2 - 2} [\sum(x - \bar{x})^2 + \sum(y - \bar{y})^2]$$

$$= \frac{1}{10} [290 + 926] = 121.6$$

5 + 7 - 2

Calculation of statistic:

Under H_0 the test statistic is

$$t_0 = \left| \frac{\bar{x} - \bar{y}}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \right|$$

$$= \left| \frac{46 - 57}{\sqrt{121.6\left(\frac{1}{5} + \frac{1}{7}\right)}} \right|$$

$$= \frac{11}{\sqrt{121.6 \times \frac{12}{35}}}$$

10

$$= \frac{11}{6.57} = 1.7$$

Expected value:

$$t_e = \left| \frac{\bar{x} - \bar{y}}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \right|$$ follows t-distribution with (5+7–2) =10 d.f

= 1.812

Inference:

Since $t_0 < t_e$ it is not significant. Hence H_0 is accepted and we conclude that the medicines A and B do not differ significantly as regards their effect on increase in weight.

Example 4:

Two types of batteries are tested for their length of life and the following data are obtained:

	No of samples	Mean life (in hrs)	Variance
Type A	9	600	121
Type B	8	640	144

Is there a significant difference in the two means?

Solution:

We are given

$n_1=9$; $\bar{x}_1=600$hrs; $s_1^2=121$; $n_2=8$; $\bar{x}_2=640$hrs; $s_2^2=144$

Null hypothesis:

$H_0 : \mu_1 = \mu_2$ i.e. Two types of batteries A and B are identical i.e. there is no significant difference between two types of batteries.

Alternative Hypothesis: 11

$H_1 : \mu_1 \neq \mu_2$ (Two- tailed)

Level of Significance:

Let $\alpha = 5\%$

Calculation of statistics:

Under H_0, the test statistic is

$$t_0 = \left| \frac{\bar{x} - \bar{y}}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \right|$$

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

$$= \frac{9 \times 121 + 8 \times 144}{9 + 8 - 2}$$

$$= \frac{2241}{15} = 149.4$$

$$\therefore \quad t_0 = \left| \frac{600 - 640}{\sqrt{149.4\left(\frac{1}{9} + \frac{1}{8}\right)}} \right|$$

$$= \frac{40}{\sqrt{149.4\left(\frac{17}{72}\right)}} = \frac{40}{5.9391} = 6.735$$

Expected value:

$$t_e = \left| \frac{\bar{x} - \bar{y}}{\sqrt{S^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \right|$$ follows t-distribution with $9+8-2=15$ d.f

$= 2.131$

Inference:

12

Since $t_0 \geq t_e$ it is highly significant. Hence H_0 is rejected and we conclude that the two types of batteries differ significantly as regards their length of life.

6.3.2 Related samples –Paired t-test:

In the t-test for difference of means, the two samples were independent of each other. Let us now take a particular situations where

(i) The sample sizes are equal; i.e., $n_1 = n_2 = n$(say), and

(ii) The sample observations ($x_1, x_2, \ldots\ldots x_n$) and ($y_1, y_2, \ldots\ldots y_n$) are not completely independent but they are dependent in pairs.

That is we are making two observations one before treatment and another after the treatment on the same individual. For example a business concern wants to find if a particular media of promoting sales of a product, say door to door canvassing or advertisement in papers or through T.V. is really effective. Similarly a pharmaceutical company wants to test the efficiency of a particular drug, say for inducing sleep after the drug is given. For testing of such claims gives rise to situations in (i) and (ii) above, we apply paired t-test.

Paired – t –test:

Let $d_i = X_i - Y_i$ ($i = 1, 2, \ldots\ldots n$) denote the difference in the observations for the i^{th} unit.

Null hypothesis:

$H_0 : \mu_1 = \mu_2$ ie the increments are just by chance

Alternative Hypothesis:

$H_1 : \mu_1 \neq \mu_2$ ($\mu_1 > \mu_2$ (or) $\mu_1 < \mu_2$)

Calculation of test statistic:

$$t_0 = \left| \frac{\bar{d}}{S/\sqrt{n}} \right|$$

$$\text{where } \bar{d} = \frac{\sum d}{n} \text{ and } S^2 = \frac{1}{n-1}\sum(d-\bar{d})^2 = \frac{1}{n-1}\left[\sum d^2 - \frac{(\sum d)^2}{n}\right]$$

Expected value:

$$t_e = \left| \frac{\bar{d}}{S/\sqrt{n}} \right| \text{ follows t-distribution with } n-1 \text{ d.f}$$

13

Inference:

By comparing t_0 and t_e at the desired level of significan usually 5% or 1%, we reject or accept the null hypothesis.

Example 5:

To test the desirability of a certain modification in typis desks, 9 typists were given two tests of as nearly as possible t same nature, one on the desk in use and the other on the new typ The following difference in the number of words typed per minut were recorded:

Typists	A	B	C	D	E	F	G	H	I
Increase in number of words	2	4	0	3	−1	4	−3	2	5

Do the data indicate the modification in desk promotes speed in typing?

Solution:

Null hypothesis:

$H_0 : \mu_1 = \mu_2$ i.e. the modification in desk does not promote speed in typing.

Alternative Hypothesis:

$H_1 : \mu_1 < \mu_2$ (Left tailed test)

Level of significance: Let $\alpha = 0.05$

Typist	d	d^2
A	2	4
B	4	16
C	0	0
D	3	9
E	1	1
F	4	16
G	3	9
H	2	4
I	5	25
	Σd 16	Σd^2 [illegible]

14

$$\bar{d} = \frac{\sum d}{n} = \frac{16}{9} = 1.778$$

$$S = \sqrt{\frac{1}{n-1}\left[\sum d^2 - \frac{(\sum d)^2}{n}\right]}$$

$$= \sqrt{\frac{1}{8}\left[84 - \frac{(16)^2}{9}\right]} = \sqrt{6.9} = 2.635$$

Calculation of statistic:

Under H_0 the test statistic is

$$t_0 = \left|\frac{\bar{d}.\sqrt{n}}{S}\right| = \frac{1.778 \times 3}{2.635} = 2.024$$

$\sqrt{9}$

Expected value:

$t_e = \left|\frac{\bar{d}.\sqrt{n}}{S}\right|$ follows t- distribution with $9 - 1 = 8$ d.f

$= 1.860$

Inference:

When $t_0 < t_e$ the null hypothesis is accepted. The data does not indicate that the modification in desk promotes speed in typing.

Example 6:

An IQ test was administered to 5 persons before and after they were trained. The results are given below:

Candidates	I	II	III	IV	V
IQ before training	110	120	123	132	125
IQ after training	120	118	125	136	121

Test whether there is any change in IQ after the training programme (test at 1% level of significance)

Solution:

Null hypothesis:

$H_0: \mu_1 = \mu_2$ i.e. there is no significant change in IQ after the training programme.

Alternative Hypothesis:

$H_1 : \mu_1 \neq \mu_2$ (two tailed test)

15

Level of significance :

$\alpha = 0.01$

x	110	120	123	132	125	Total
y	120	118	125	136	121	-
$d = x - y$	-10	2	-2	-4	4	-10
d^2	100	4	4	16	16	140

$$\bar{d} = \frac{\sum d}{n} = \frac{-10}{5} = -2$$

$$S^2 = \frac{1}{n-1}\left[\sum d^2 - \frac{(\sum d)^2}{n}\right]$$

$$= \frac{1}{4}\left[140 - \frac{100}{5}\right] = 30$$

Calculation of Statistic:

Under H_0 the test statistic is

$$t_0 = \left|\frac{\bar{d}}{S/\sqrt{n}}\right|$$

$$= \left|\frac{-2}{\sqrt{30/5}}\right|$$

$$= \frac{2}{2.45}$$

$$= 0.816$$

Expected value:

$$t_e = \left|\frac{\bar{d}}{\sqrt{S^2/n}}\right|$$ follows t-distribution with $5 - 1 = 4$ d.f

$= 4.604$

Inference:

Since $t_0 < t_e$ at 1% level of significance we accept the null hypothesis. We therefore, conclude that there is no change in IQ after the training programme.

between the observed and expected frequencies the greater is the value of χ^2. chi-square Test: 1

UNIT-III Continuation.

Chi square - Distribution:

The square of a standard normal variate is a Chi-square variate with 1 degree of freedom i.e., If X is normally distributed with mean μ and standard deviation σ, then $\left(\frac{x-\mu}{\sigma}\right)^2$ is a Chi-square variate (χ^2) with 1 d.f. The distribution of Chi-square depends on the degrees of freedom. There is a different distribution for each number of degrees of freedom.

distribution.

chi-square 2

6.6 Test of independence Attributes Attributes

Let us suppose that the given population consisting of N items is divided into r mutually disjoint (exclusive) and exhaustive classes $A_1, A_2, \ldots, A_r$ with respect to the attribute A so that randomly selected item belongs to one and only one of the attributes $A_1, A_2, \ldots, A_r$ Similarly let us suppose that the same population is divided into c mutually disjoint and exhaustive classes $B_1, B_2, \ldots, B_c$ w.r.t another attribute B so that an item selected at random possess one and only one of the attributes $B_1, B_2, \ldots, B_c$. The frequency distribution of the items belonging to

3

the classes $A_1, A_2, \ldots, A_r$ and $B_1, B_2, \ldots, B_c$ can be represented in the following $r \times c$ manifold contingency table.

$r \times c$ manifold contingency table

B / A	B_1	B_2	…	B_j	…	B_c	Total
A_1	(A_1B_1)	(A_1B_2)	…	(A_1B_j)	…	(A_1B_c)	(A_1)
A_2	(A_2B_1)	(A_2B_2)	…	(A_2B_j)	…	(A_2B_c)	(A_2)
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
A_i	(A_iB_1)	(A_iB_2)	…	(A_iB_j)	…	(A_iB_c)	(A_i)
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
A_r	(A_rB_1)	(A_rB_2)	…	(A_rB_j)	…	(A_rB_c)	(A_r)
Total	(B_1)	(B_2)	…	(B_j)	…	(B_c)	$\Sigma Ai = \Sigma Bj = N$

(A_i) is the number of persons possessing the attribute A_i, $(i=1,2,\ldots r)$, (Bj) is the number of persons possing the attribute $B_j,(j=1,2,3,\ldots c)$ and $(A_i\ B_j)$ is the number of persons possessing both the attributes A_i and B_j $(i=1,2,\ldots r\ ,j=1,2,\ldots c)$.

Also $\Sigma A_i = \Sigma B_j = N$

Under the null hypothesis that the two attributes A and B are independent, the expected frequency for (A_iB_j) is given by

$$= \frac{(Ai)(Bj)}{N}$$

Calculation of statistic:

Thus the under null hypothesis of the independence of attributes,the expected frequencies for each of the cell frequencies of the above table can be obtained on using the formula

$$\chi_0^2 = \Sigma\left(\frac{(O_i - E_i)^2}{E_i}\right)$$

4

Expected value:

$$\chi_e^2 = \Sigma\left(\frac{(O_i - E_i)^2}{E_i}\right)$$ follows χ^2-distribution with $(r-1)(c-1)$ d.f

Inference:

Now comparing χ_0^2 with χ_e^2 at certain level of significance ,we reject or accept the null hypothesis accordingly at that level of significance.

6.6.1 2×2 contingency table :

Under the null hypothesis of independence of attributes, the value of χ^2 for the 2×2 contingency table

			Total
	a	b	a+b
	c	d	c+d
Total	a+c	b+d	N

is given by

$$\chi_0^2 = \frac{N(ad - bc)^2}{(a+c)(b+d)(a+b)(c+d)}$$

5

Example 9:

1000 students at college level were graded according to their I.Q. and the economic conditions of their homes. Use χ^2 test to find out whether there is any association between economic condition at home and I.Q.

Economic Conditions	IQ		Total
	High	Low	
Rich	460	140	600
Poor	240	160	400
Total	700	300	1000

Solution:

Null Hypothesis:

There is no association between economic condition at home and I.Q. i.e. they are independent.

$$E_{11} = \frac{(A)(B)}{N} = \frac{600 \times 700}{1000} = 420$$

The table of expected frequencies shall be as follows.

			Total
	420	180	600
	280	120	400
Total	700	300	1000

Observed Frequency O	Expected Frequency E	$(O - E)^2$	$\left(\frac{(O-E)^2}{E}\right)$
460	420	1600	3.81
240	280	1600	5.714
140	180	1600	8.889
160	120	1600	13.333
			31.746

$$\chi_0^2 = \Sigma \left(\frac{(O-E)^2}{E}\right) = 31.746$$

6

Expected Value:

$$\chi_e^2 = \Sigma\left(\frac{(O - E)^2}{E}\right) \text{ follow } \chi^2 \text{ distribution with } (2-1)(2-1) = 1 \text{ d.f}$$

$$= 3.84$$

Inference :

$\chi_o^2 > \chi_e^2$, hence the hypothesis is rejected at 5 % level of significance. ∴ there is association between economic condition at home and I.Q.

Methods of studying Association

9. THEORY OF ATTRIBUTES

UNIT-III Continuations... 1

9.0 Introduction:

Generally statistics deal with quantitative data only. But in behavioural sciences, one often deals with the variable which are not quantitatively measurable. Literally an attribute means a quality on characteristic which are not related to quantitative measurements. Examples of attributes are health, honesty, blindness etc. They cannot be measured directly. The observer may find the presence or absence of these attributes. Statistics of attributes based on descriptive character.

9.1 Notations:

(Association of attribute is studied by the presence or absence of a particular attribute. If only one attribute is studied, the population is divided into two classes according to its presence or absence and such classification is termed as division by dichotomy. If a class is divided into more than two scale-classes, such classification is called manifold classification.

Positive class which denotes the presence of attribute is generally denoted by Roman letters generally A,B,….etc and the negative class denoting the absence of the attribute and it is denoted by the Greek letters α, β….etc For example, A represents the attribute 'Literacy' and B represents 'Criminal'. α and β represents the 'Illiteracy' and 'Not Criminal' respectively.)

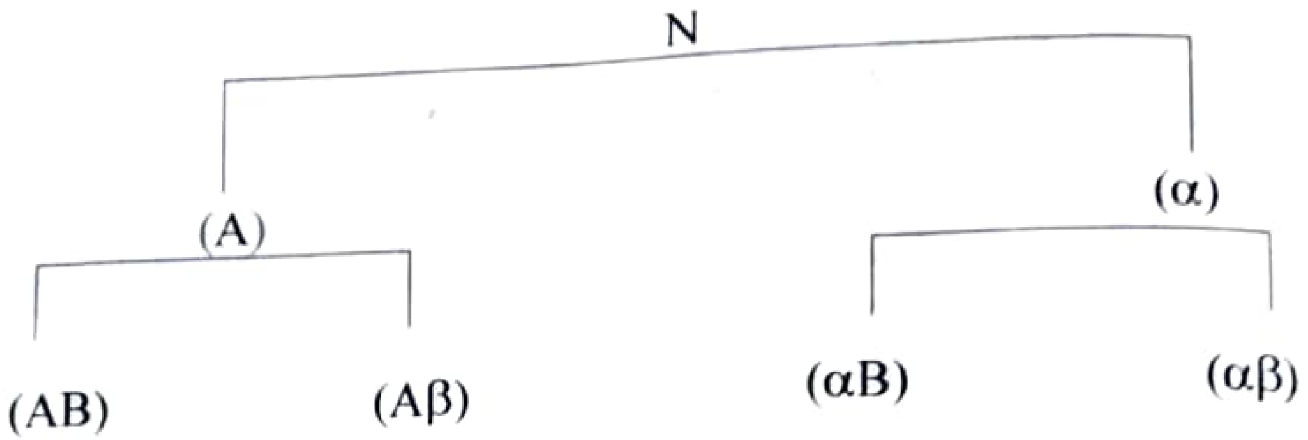
9.2 Classes and Class frequencies:

Different attributes, their sub-groups and combinations are called different classes and the number of observations assigned to them are called their class frequencies.

If two attributes are studied the number of classes will be 9. (i.e.,) (A) , (α), (B), (β), (A β) (α β), (α B), (α β) and N.

2

The chart given below illustrate it clearly.



The number of observations or units belonging to class is known as its frequency are denoted within bracket. Thus (A) stands for the frequency of A and (AB) stands for the number objects possessing the attribute both A and B. The contingency table of order (2×2) for two attributes A and B can be displayed as given below

	A	α	Total
B	(AB)	(αB)	(B)
β	(Aβ)	(αβ)	(β)
Total	(A)	(α)	N

Relationship between the class frequencies:

The frequency of a lower order class can always be expressed in terms of the higher order class frequencies.

i.e., $N = (A) + (\alpha) = (B) + (\beta)$

$(A) = (AB) + (A\beta)$

$(\alpha) = (\alpha B) + (\alpha\beta)$

$(B) = (AB) + (\alpha B)$

$(\beta) = (A\beta) + (\alpha \beta)$

If the number of attributes is n, then there will be 3^n classes and we have 2^n cell frequencies.

3

9.3 Consistency of the data:

13 (In order to find out whether the given data are consistent or not we have to apply a very simple test. The test is to find out whether any or more of the ultimate class-frequencies is negative or not. If none of the class frequencies is negative we can safely calculate that the given data are consistent (i.e the frequencies do not conflict in any way each other). On the other hand, if any of the ultimate class frequencies comes to be negative the given data are inconsistent.) 13

Example 1:

Given N = 2500, (A) = 420, (AB) = 85 and (B) = 670. Find the missing values.

Solution:

We know N = (A) +(α) = (B) + (β) — 1

(A) = (AB) + (Aβ) — 2

(α) = (αB) + (αβ) — 3

(B) = (AB) + (α B) — 4

(β) = (Aβ) + (α β) — 5

	A	α	Total
B	AB	Bα	B
β	βA	βα	β
Total	A	α	N

85	585	670
335	2062	1830
420	2080	2500

From (2) 420 = 85 + (Aβ)

∴ (Aβ) = 420 – 85

(A β) = 335

From (4) 670 = 85 + (αB)

∴ (αB) = 670 – 85

(αB) = 585

From (1) 2500 = 420 + (α)

∴ (α) = 2500 – 420

(α) = 2080

From (1) (β) = 2500 – 670

(β) = 1830

From (3) = 2080 = 585 + (αβ)

∴ (αβ) = 1495

Example 2:

4

Test the consistency of the following data with the symbols having their usual meaning.

N = 1000 (A) = 600 (B) = 500 (AB) = 50

Solution:

	A	α	Total
B	50	450	500
β	550	-50	500
Total	600	400	1000

Since $(\alpha\beta)) = -50$, the given data is inconsistent.

Example 3:

Examine the consistency of the given data. N = 60 (A) = 51 (B) = 32 (AB) = 25

Solution:

	A	α	Total
B	25	7	32
β	26	2	28
Total	51	9	60

Since all the frequencies are positive, it can be concluded that the given data are consistent.

9.4 Independence of Attributes:

If the attributes are said to be independent the presence or absence of one attribute does not affect the presence or absence of the other. For example, the attributes skin colour and intelligence of persons are independent.

5

If two attributes A and B are independent then the actual frequency is equal to the expected frequency

$$(AB) = \frac{(A).(B)}{N}$$

Similarly $$(\alpha\ \beta) = \frac{(\alpha).(\beta)}{N}$$

9.4.1 Association of attributes:

14 Two attributes A and B are said to be associated if they are not independent but they are related with each other in some way or other.

The attributes A and B are said to be positively associated if

$$(AB) > \frac{(A).(B)}{N}$$

If $(AB) < \dfrac{(A).(B)}{N}$,. then they are said to be negatively associated 14

Example 4:

Show that whether A and B are independent, positively associated or negatively associated.

$(AB) = 128$, $(\alpha B) = 384$, $(A\beta) = 24$ and $(\alpha\beta) = 72$

Solution:

$$(A) = (AB) + (A\beta)$$
$$= 128 + 24$$
$$(A) = 152$$
$$(B) = (AB) + (\alpha B)$$
$$= 128 + 384$$
$$(B) = 512$$
$$(\alpha) = (\alpha B) + (\alpha\beta)$$
$$= 384 + 72$$
$$\therefore (\alpha) = 456$$
$$(N) = (A) + (\alpha)$$
$$= 152 + 456$$
$$= 608$$

6

$$\frac{(A)\times(B)}{N} = \frac{152\times 512}{608}$$
$$= 128$$
$$(AB) = 128$$
$$\therefore (AB) = \frac{(A)\times(B)}{N}$$

Hence A and B are independent

Example 5:

From the following data, find out the types of association of A and B.

1) N = 200	(A) = 30	(B) = 100	(AB) = 15
2) N = 400	(A) = 50	(B) = 160	(AB) = 20
3) N = 800	(A) = 160	(B) = 300	(AB) = 50

Solution:

1. Expected frequency of $(AB) = \frac{(A).(B)}{N}$

$$= \frac{(30)(100)}{200} = 15$$

Since the actual frequency is equal to the expected frequency, ie 15 = 15, therefore A and B are independent.

2. Expected frequency of $(AB) = \frac{(A).(B)}{N}$

$$= \frac{(50)(160)}{400} = 20$$

Since the actual frequency is greater than expected frequency, i.e., 25 > 20, therefore A and B are positively associated.

3. Expected frequency of $(AB) = \frac{(A).(B)}{N} = \frac{(160)(300)}{800} = 60$

Since Actual frequency is less than expected frequency i.e., 50 < 60 therefore A and B are negatively associated.

0.5 Yules' co-efficient of association:

15 The above example gives a rough idea about association but not the degree of association. For this Prof. G. Undy Yule has suggested a formula to measure the degree of association. It is a relative measure of association between two attributes A and B.

If (AB), (αB), (Aβ) and ($\alpha\beta$) are the four distinct combination of A, B, α and β then Yules' co-efficient of association is

$$Q = \frac{(AB)(\alpha\beta) - (A\beta).(\alpha B)}{(AB)(\alpha\beta) + (A\beta).(\alpha B)}$$

Note:

1. If Q = +1 there is perfect positive association
   If Q = -1 there is perfect negative association
   If Q = 0 there is no association (ie) A and B are independent
1. For rememberance of the above formula , we use the table below

	A	α
B	AB	αB
β	Aβ	$\alpha\beta$

115

Example 6:

Investigate the association between darkness of eye colour in father and son from the following data.

Fathers' with dark eyes and sons' with dark eyes = 50
Fathers' with dark eyes an sons' with no dark eyes = 79
Fathers' with no dark eyes and sons with dark eyes = 89
Neither son nor father having dark eyes = 782

Solution:

Let A denote the dark eye colour of father and B denote dark eye colour of son.

	A	α	Total
B	50	89	139
β	79	782	861
Total	129	871	1000

Yules' co-efficient of association is

8

$$Q = \frac{(AB)(\alpha\beta)-(A\beta).(\alpha B)}{(AB)(\alpha\beta)+(A\beta).(\alpha B)}$$

$$= \frac{50\times 782 - 79\times 89}{50\times 782 + 79\times 89}$$

$$= \frac{32069}{46131} = 0.69$$

$\therefore$ there is a positive association between the eye colour of fathers' and sons'.

Example 7 :

Can vaccination be regarded as a preventive measure of small pox from the data given below.

Of 1482 persons in a locality, exposed to small pox, 368 in all were attacked, among the 1482 persons 343 had been vaccinated among these only 35 were attacked.

Solution:

Let A denote the attribute of vaccination and B denote that of attacked.

	A	α	Total
B	35	333	368
β	308	806	1114
Total	343	1139	1482

Yules' co-efficient of association is

$$Q = \frac{(AB)(\alpha\beta)-(A\beta).(\alpha B)}{(AB)(\alpha\beta)+(A\beta).(\alpha B)}$$

$$= \frac{35\times 806 - 308\times 333}{35\times 806 + 308\times 333}$$

$$= \frac{-74354}{130774} = -0.57$$

i.e., there is a negative association between attacked and vaccinated. In other words there is a positive association between not attacked and vaccinated. Hence vaccination can be regarded as a preventive measure for small pox.

9

Example 8:

In a co-educational institution, out of 200 students, 150 were boys. They took an examination and it was found that 120 passed. 10 girls failed. Is there any association between sex and success in the examination.

Solution:

Let A denote boys and α denote girls. Let B denote those who passed the examination and β denote those who failed.

We have given $N = 200$ $(A) = 150$ $(AB) = 120$ $(\alpha\beta) = 10$

Other frequencies can be obtained from the following table

	A	α	Total
B	120	40	160
β	30	10	40
Total	150	50	200

Yule's co-efficient of association is

$$Q = \frac{(AB)(\alpha\beta) - (A\beta).(\alpha B)}{(AB)(\alpha\beta) + (A\beta).(\alpha B)}$$

$$= \frac{120 \times 10 - 30 \times 40}{120 \times 10 + 30 \times 40} = 0$$

Therefore, there is no association between sex and success in the examination.

Recall

(A) (B) denote positive attributes

(α) (β) denote negative attributes

2 ×2 contingency table.

X	A	α	Total
B	(AB)	(αB)	(B)
β	(Aβ)	($\alpha\beta$)	(β)
Total	(A)	(α)	N

Vertical Total

$$(AB) + (A\beta) = (A)$$
$$(\alpha B) + (\alpha\beta) = (\alpha)$$
$$(A) + (\alpha) = N$$

Horizontal Total

$$(AB) + (\alpha B) = B$$
$$(A\beta) + (\alpha\beta) = \beta$$
$$(B) + (\beta) = N$$

Types of Association

Positive Association if $(AB) > \dfrac{(A).(B)}{N}$

Negative Association if $(AB) < \dfrac{(A).(B)}{N}$

Independent if $(AB) = \dfrac{(A).(B)}{N}$

Yule's co-efficient of Association

$$Q = \frac{(AB)(\alpha\beta) - (A\beta).(\alpha B)}{(AB)(\alpha\beta) + (A\beta).(\alpha B)}$$