UNIT V – CHI-SQUARE TEST (Part 2)

2. CHI-SQUARE TEST FOR INDEPENDENCE OF ATTRIBUTES

Under this test, we can find out whether two or more attributes are associated or not. Let us consider two attributes A and B, A is divided into 'r' classes A_1, A_2... A_r, and B is divided into 's'classes B_1 , B_2 ,..., B_s . Such a classification in which attributes are divided into more than two classes is known as manifold classification. The various cell frequencies can be expressed in the following table known as **r×s manifold contingency table** where (A_i) denotes the number of persons possessing the attribute A_i, (i=1,2,…,r), (B_j) denote the number of persons possessing the attributes B_j, (j=1,2,…,s) and (A_iB_j) denote the number of persons possessing both the attributes (A_i) and (B_i). Also $\sum_{i=1}^{r} A_i$ = $\sum_{i=1}^{s} B_i$ = N, is the total frequency.

	A_1	A_2	…….	A_i	…….	A_r	Total
B_1	(A_1B_1)	(A_2B_1)	…….	(A_iB_1)	…….	(A_rB_1)	*(B_1)*
B_2	(A_1B_2)	(A_2B_2)	…….	(A_iB_2)	…….	(A_rB_2)	*(B_2)*
⋮			…….		…….		⋮
B_j	(A_1B_j)	(A_2B_j)	…….	(A_iB_3)	…….	(A_rB_j)	*(B_j)*
⋮			…….		…….		⋮
B_s	(A_1B_s)	(A_2B_j)	…….	(A_iB_4)	…….	(A_rB_s)	*(B_s)*
	(A_1)	*(A_2)*	…….	*(A_i)*	…….	*(A_r)*	N

Under the null hypothesis that the two attributes A and B are independent, the expected frequencies are calculated as follows

P(A_i) = probability that a person possessing the attribute A*i*

$$\frac{(A_i\ )}{N};i=1,2,\dots,r$$

$$\frac{(j\ )}{N};j=1,2,\dots,s$$

P(A_iB_j) = P(A_i) P(B_j) (attributes Ai and Bj are independent under the null hypothesis)

$$\mathrm{P(AiBj)}\ =\frac{(A_i\ )}{N}\times\frac{(B_j\ )}{N}$$

If $(A_iB_j)_o$ denote the expected frequency of (A_iB_j), then

$$\mathrm{(AiBj)o}=\mathrm{N}*\mathrm{P(AiBj)}=\frac{(A_i\ B_{j\ )}}{N},(i=1,2,\dots,r;\ j=1,2,\dots,s)$$

By using this formula expected frequencies for each of the cell frequencies
$\mathrm{(AiBj)},(i=1,2,\dots,r;\ j=1,2,\dots,s)$ can be worked out.

The exact test for independence of attributes is very complicated but a fair degree of approximation is given, for large samples by the $\chi\ 2$ -test of goodness of fit i.e

$$\chi^2=\ \sum\sum[\frac{\left((A_i\ B_J\ )-(A_i\ B_J\ )o\ \right)^2}{(A_i\ B_J\ )o}]$$

Expected Frequency = (RT×CT)/GT

follows $\chi\ 2$ distribution with **(r-1)(s-1)** degrees of freedom.

Now comparing this calculated value with the tabulated value for (r-1)(s-1) d.f at certain level of significance, we reject or retain the null hypothesis of independence of attributes at that level of significance.

Illustration: A certain drug was administered to 456 males out of total 720 in a certain locality to test its efficacy against typhoid. The **incidence of typhoid** is shown below. Find out the effectiveness of the **drug** against the disease

	Infection	No infection	Total
Administering the drug:	144	312	456
Without administering the drug:	192	72	264
Total	336	384	720

Solution: We set up the null hypothesis that the two attributes “incidence of typhoid” and the administration of the drug are independent.

Step1: H_0: There is no significnt association between use of drugs and infection of disease.(it is hypothesised that “incidence of typhoid” and the administration of the drug are independent)

H_1: There is significnt association between use of drugs and infection of disease.(it is hypothesised that “incidence of typhoid” and the administration of the drug are dependent)

Step 2: $\alpha = 0.05$ (Let us assume that significance level is 5%)

Step 3: Expected Frequency = (RT×CT)/GT

Under the hypothesis of independence,

E(144)= $\frac{336\times456}{720} = 212.8$; E(312)= $\frac{384\times456}{720} = 243.2$; E(192)= $\frac{336\times264}{720} = 123.2$; E(72)= $\frac{264\times384}{720} = 140.8$

O	E	$(O - E)^2$	$\frac{(O-E)^2}{E}$
144	212.8	4733.44	22.244
192	123.2	4733.44	38.420
312	243.2	4733.44	19.46
72	140.8	4733.44	33.62
			$\sum\frac{(O-E)^2}{E} = 113.74$

Degrees of freedom=(r-1)(c-1)=(2-1)(2-1)=1 d.f

For 1 dof at 5% level of significance the table value of $\chi 2$ =3.84. Since calculated value is very much greater than the table value. It is highly significant. Hence the null hypothesis is rejected at 5% level of significance and we conclude that the drug is certainly effective in controlling typhoid.

CHI-SQUARE AS A TEST FOR COMPARING VARIANCE

The chi-square value is often used to judge the significance of population variance i.e., we can use the test to judge if a random sample has been drawn from a normal population with mean (μ) and with a specified variance (σ_p^2). The test is based on $\chi 2$ -distribution. Such a distribution we encounter when we deal with collections of values that involve adding up squares. Variances of samples require us to add a collection of squared quantities and, thus, have distributions that are related to $\chi 2$ -distribution. If we take each one of a collection of sample variances, divided them by the known population variance and multiply these quotients by (n – 1), where n means the number of items in the sample, we shall obtain a $\chi 2$ -distribution. Thus, $\frac{\sigma_s^2}{\sigma_p^2}$ (n-1) = $\frac{\sigma_s^2}{\sigma_p^2}$ × (dof) would have the same distribution as $\chi 2$ -distribution with (*n* – 1) degrees of freedom.

In brief, when we have to use chi-square as a test of population variance, we have to work out the value of $\chi 2$ to test the null hypothesis (viz., H_0: $\sigma_p^2 = \sigma_s^2$) as under:

$$\chi^2 = \frac{\sigma_s^2}{\sigma_p^2}(n-1)$$

Where, σ_s^2 = variance of the sample;

σ_p^2 = variance of the population;

(*n* – 1) = degrees of freedom, *n* being the number of items in the sample.

Then by comparing the calculated value with the table value of $\chi 2$ for (n – 1) degrees of freedom at a given level of significance, we may either accept or reject the null hypothesis. If the calculated value of $\chi 2$ is less than the table value, the null hypothesis is accepted, but if the calculated value is equal or greater than the table value, the hypothesis is rejected. All this can be made clear by an example.

Example:

S.No:	1	2	3	4	5	6	7	8	9	10
Weight (kg)	38	40	45	53	47	43	55	48	52	49

Can we say that the variance of the distribution of weight of all students from which the above sample of 10 students was drawn is equal to **20** kgs? Test this at 5 per cent and 1 per cent level of significance.

Solution: First of all we should work out the variance of the sample data or σ_s^2 and the same has been worked out as under:

s.no	*X_i (weight in kgs)*	*$(X_i – X)^2$*
1	38	81
2	40	49
3	45	4
4	53	36
5	47	0
6	43	16
7	55	64
8	48	1
9	52	25
10	49	4
n= 10	$\sum X_i = 470$	$\sum(X_i - \overline{X})^2$*=280*

$$\bar{X} = \frac{\sum X_i}{n} = \frac{470}{10} = 47$$

$$\sigma_s \quad = \sqrt{\frac{\sum(X_i - \overline{X})^2}{n-1}} = \sqrt{\frac{280}{10-1}}$$

$$\sigma_S{}^2 = 31.11$$

Let the null hypothesis be H$_o$: $\sigma_p^2 = \sigma_s^2$. In order to test this hypothesis we work out the c2 value as

$\chi^2 = \frac{\sigma_s^2}{\sigma_p^2}(n-1) = \frac{31.11}{20} \times (10-1) = 14$

Degrees of freedom in the given case is (n – 1) = (10 – 1) = 9. At 5 per cent level of significance the table value of 9 dof = 16.92 and at 1 per cent level of significance, it is 21.67 for 9 d.f. and both these values are greater than the calculated value of chi-square which is 13.999. Hence we accept the null hypothesis and conclude that the variance of the given distribution can be taken as 20 kgs at 5 per cent as also at 1 per cent level of significance. In other words, the sample can be said to have been taken from a population with variance 20 kgs.

Illustration 2

A sample of 10 is drawn randomly from a certain population. The sum of the squared deviations from the mean of the given sample is 50. Test the hypothesis that the variance of the population is 5 at 5 per cent level of significance.

Solution: Given information is n = 10; $\sum(X_i - \overline{X})^2 = 50$; α =.05

$$\sigma_s^2 = \frac{\sum(X_i - \overline{X})^2}{n-1} = \frac{50}{9}$$

Let the null hypothesis be H$_o$: $\sigma_p^2 = \sigma_s^2$. In order to test this hypothesis we work out the $\chi2$ value as

$$\chi^2 = \frac{\sigma_s^2}{\sigma_p^2}(n-1) = \frac{\frac{50}{9}}{5}\;(10-1) = \;10$$

d.f = (10-1) = 9

9.

The table value of $\chi2$ at 5 per cent level for 9 d.f. is 16.92. The calculated value of $\chi2$ is less than this table value, so we accept the null hypothesis and conclude that the variance of the population is 5 as given in the question.

CONDITIONS FOR THE APPLICATION OF $\chi2$ TEST

The following conditions should be satisfied before $\chi2$ test can be applied:

(i) Observations recorded and used are collected on a random basis.

(ii) All the itmes in the sample must be independent.

(iii) No group should contain very few items, say less than 10. In case where the frequencies are less than 10, regrouping is done by combining the frequencies of adjoining groups so that the new frequencies become greater than 10. Some statisticians take this number as 5, but 10 is regarded as better by most of the statisticians.

(iv) The overall number of items must also be reasonably large. It should normally be at least 50, howsoever small the number of groups may be.

(v) The constraints must be linear. Constraints which involve linear equations in the cell frequencies of a contingency table (i.e., equations containing no squares or higher powers of the frequencies) are known as linear constraints.

IMPORTANT CHARACTERISTICS OF $\chi2$ TEST

(i) This test (as a non-parametric test) is based on frequencies and not on the parameters like mean and standard deviation.

(ii) The test is used for testing the hypothesis and is not useful for estimation.

(iii) This test possesses the additive property as has already been explained.

(iv) This test can also be applied to a complex contingency table with several classes and as such is a very useful test in research work.

(v) This test is an important non-parametric test as no rigid assumptions are necessary in regard to the type of population, no need of parameter values and relatively less mathematical details are involved.

CAUTION IN USING $\chi2$ TEST

The chi-square test is no doubt a most frequently used test, but its correct application is equally an uphill task. It should be borne in mind that the test is to be applied only when the individual observations of sample are independent which means that the occurrence of one individual observation (event) has no effect upon the occurrence of any other observation (event) in the sample under consideration.

Small theoretical frequencies, if these occur in certain groups, should be dealt with under special care. The other possible reasons concerning the improper application or misuse of this test can be (i) neglect of frequencies of non-occurrence; (ii) failure to equalize the sum of observed and the sum of the expected frequencies; (iii) wrong determination of the degrees of freedom; (iv) wrong computations, and the like. The researcher while applying this test must remain careful about all these things and must thoroughly understand the rationale of this important test before using it and drawing inferences in respect of his hypothesis.

Questions

1. What is Chi-square text? Explain its significance in statistical analysis.

2. Write short notes on the following:

(i) Property of Chi-square;
(ii) Chi-square as a test of 'goodness of fit';
(iii) Precautions in applying Chi-square test;
(iv) Conditions for applying Chi-square test.

Q1. A die is thrown 132 times with following results:

Number turned up	1	2	3	4	5	6
Frequency	16	20	25	14	29	28

Is the die unbiased?

Q2. Find the value of $\chi2$ for the following information:

Class	A	B	C	D	E
Observed Frequency	8	29	44	15	4
Expected Frequency	7	24	38	24	7

Q3. Genetic theory states that children having one parent of blood type *A* and the other of blood type *B* will always be of one of three types, *A*, *AB*, *B* and that the proportion of three types will on an average be as 1 : 2 : 1. A report states that out of 300 children having one *A* parent and *B* parent, 30 per cent were found to be types *A*, 45 per cent per cent type *AB* and remainder type *B*. Test the hypothesis by $\chi2$ test.

Q4. The table given below shows the data obtained during outbreak of smallpox:

	Attacked	Not Attacked	Total
Vaccinated	31	469	500
Not Vaccinated	185	1315	1500
Total	216	1784	2000

Test the effectiveness of vaccination in preventing the attack from smallpox. Test your result with the help of $\chi2$ at 5 per cent level of significance

Q5. Two research workers classified some people in income groups on the basis of sampling studies. Their results are as follows:

investigatiors	Income group			
	Poor	Middle	Rich	Total
A	160	30	10	200
B	140	120	40	300
Total	300	150	50	500

Q6. Eight coins were tossed 256 times and the following results were obtained:

Number of heads	0	1	2	3	4	5	6	7	8
Frequency	2	6	30	52	67	56	32	10	1

Are the coins biased? Use $\chi2$ test.

Q7. The following information is obtained concerning an investigation of 50 ordinary shops of small size:

	Shops		
	In town	In villages	Total
Run by men	17	18	35
Run by women	3	12	15
Total	20	30	50

Can it be inferred that shops run by women are relatively more in villages than in towns? Use $\chi2$ test.

Q8. An experiment was conducted to test the efficacy of chloromycetin in checking typhoid. In a certain hospital chloromycetin was given to 285 out of the 392 patients suffering from typhoid. The number of typhoid cases were as follows:

	Typhoid	No Typhoid	Total
Chloromycetin	35	250	285
No Chloromycetin	50	57	107
Total	85	307	392

With the help of $\chi 2$, test the effectiveness of chloromycetin in checking typhoid. (The $\chi 2$ value at 5 per cent level of significance for one degree of freedom is 3.841).

Q9. On the basis of information given below about the treatment of 200 patients suffering from a disease, state whether the new treatment is comparatively superior to the conventional treatment.

Treatments	No. of Paitent	
	Favorable Response	Unfavorable Response
New	60	20
Conventional	70	50

For drawing your inference, use the value of $\chi 2$ for one degree of freedom at the 5 per cent level of significance, viz., 3.84.

Q10. You are given a sample of 150 observations classified by two attributes *A* and *B* as follows:

	A1	A2	A3	Total
B1	40	25	15	80
B2	11	26	8	45
B3	9	9	7	25
Total	60	60	30	150

Use the $\chi 2$ test to examine whether *A* and *B* are associated.

Q11. A brand manager is concerned that her brand's share may be unevenly distributed throughout the country. In a survey in which the country was divided into four geographical regions, a random sampling of 100 consumers in each region was surveyed, with the following result:

	REGION				
	NE	NW	SE	SW	TOTAL
Purchase the brand	40	55	45	50	190
Do not Purchase the brand	60	45	55	50	210
Total	100	100	100	100	400

(a) Develop a table of observed and expected frequencies for this problem.
(b) Calculate the sample χ^2 value
(c) State the null and alternative hypotheses.
(d) At α = 0.05, test whether brand share is the same across the four regions.

Q12. At the 0.01 level of significance, can we conclude that the following 400 observations follow a Poisson distribution with λ = 3?

Number of arrivals per hour	0	1	2	3	4	5 or more
Number of hours	20	57	98	85	78	60

Q13. Mr. George, president of NIC, is opposed to National health insurance. He argues that it would be too costly to implement, particularly, since the existence of such a system would, among other effects, tend to encourage people to spend more time in hospitals. **George believes that lengths of stays in hospitals are dependent of the types of health insurance** that people have. He asked Donna, his staff statisticians, to check the matter. Donna collected data on random sample of 660 hospitals and summarized them in Table.

Table: Hospital stay data classified by type of insurance coverage & length of stay

		Days in Hospital			
		< 5	**5 - 1O**	**> 10**	**TOTAL**
Fraction of costs covered by insurance	**< 25%**	40	75	65	180
	25 - 50%	30	45	75	150
	> 50%	40	100	190	330
	Total	110	220	330	660

Q14. An advertising firm is trying to determine the demographics for a new product. They have randomly selected 75 people in each of 5 different age groups and introduced the product to them. The results of the surveys are given below:

	Age Group				
	18-29	30-39	40-49	50-59	60-69
Purchase frequently	12	18	17	22	32
Seldom purchase	18	25	29	24	30
Never purchase	45	32	29	29	13

(a) Develop a table of observed and expected frequencies for this problem.
(b) Calculate the sample χ^2 value
(c) State the null and alternative hypotheses.
(d) At α = 0.01, test whether hypothesis is rejected?