UNIT 14: CHI-SQUARE TEST

UNIT STRUCTURE

14.1 Learning Objectives

14.2 Introduction

14.3 The Chi-Square (χ^2) Distribution

14.4 Applications of χ^2 Test

- 14.4.1 χ^2 Test of Goodness of Fit
- 14.4.2 χ^2 Test for Independence of Attributes
- 14.4.3 χ^2 Test for a Specified Population Variance
- 14.4.4 χ^2 Test of Homogeneity

14.5 Let Us Sum Up

14.6 Further Readings

14.7 Answer To Check Your Progress

14.8 Model Questions

14.1 LEARNING OBJECTIVES

After going through this unit you will be able to

- introduce the Chi-square distribution for the comparison of more than two means or proportions.
- use the Chi-square distribution to test whether two classifications of the same data are independent of each other.
- use the Chi-square distribution for testing the hypothesis about a single population variance.

14.2 INTRODUCTION

In the previous unit we have learnt some statistical tests (using data from either one or two samples) that are used to test the hypothesis about the population parameters, with a number of business related examples. These tests are based on the assumption that the samples are drawn from a normally distributed population or, more appropriately, that the sample means are normally distributed. Such tests are known as parametric tests as the testing procedure involves the assumption about the type of population of parameters. On the other hand, in some cases, the researcher may not

be sure of whether the population distribution is normal. The statistical tests that do not require prior knowledge about the population are termed as non-parametric tests. Chi-square test is one of the important non-parametric methods.

A non-parametric method, also known as distribution free test satisfies at least one of the following criteria:

(i) The method does not take into consideration any population parameter.

(ii) It is applied only on categorical data that are non-numerical and frequency counts of categories for one or more variables.

(iii) It does not depend on the form of the underlying population distribution, in particular, the requirement of normality.

In this unit we will focus on χ^2 (Chi-square) test. This test enables us to test whether more than two population proportions are equal. In fact, it allows to do a lot more than just test for the equality of several proportions. If we classify a population into several categories with respect to two attributes (such as age and job performance), we can then employ a chi-square test to test whether the two attributes are independent of each other.

$$f(\chi^2) = C(\chi^2)^{\frac{\nu}{2}-1} e^{-\frac{\chi^2}{2}}$$

14.3 THE CHI-SQUARE (χ^2) DISTRIBUTION

The symbol χ stands for the Greek letter 'chi'. This test was first developed by Karl Pearson in testing of hypothesis in 1900. The sampling distribution of is called χ^2 distribution. This distribution has only one parameter called the degrees of freedom (df) and the shape of the curve depends on the degrees of freedom. The distribution is skewed to the right. Being a square quantity, χ^2 distribution can never be negative. In fact, χ^2 is a continuous probability distribution with range 0 to ∞ (Figure 9.1)

The probability density function of distribution is given by

where ν is is the degrees of freedom, is a constant depending upon the

degrees of freedom, and $e = 2.71828.$

χ^2 test statistic can be defined as below:

$$\chi^2 = \sum \frac{(O - E)^2}{E} \text{ with } df = k - 1 - c$$

where O is the observed frequency, E the expected frequency, k the number of categories, and c the number of parameters being estimated from the sample data.

The computed value of is compared with the critical value of χ^2 at a specified level of significance and degrees of freedom. The decision to accept or reject the null hypothesis is given by

If $\chi^2_{cal} < \chi^2_{critical}$ then we may accept the null hypothesis, otherwise we reject the null hypothesis.

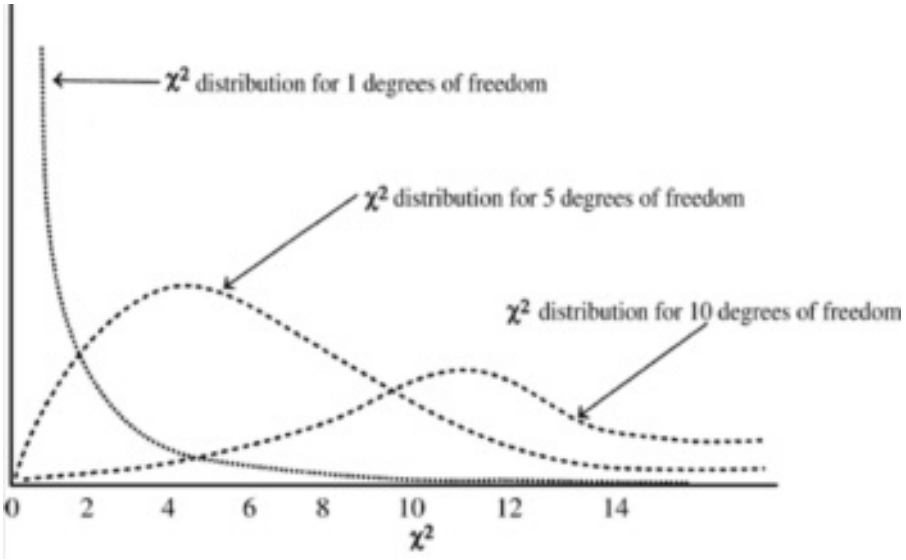


Fig. 14.1 χ^2 distribution with 1,5 and 10 degrees of freedom

Properties of Chi-square test:

The following properties are useful while using χ^2 test statistic to analyze its sampling distribution:

(i) The density function of χ^2 does not involve any parameter and the shape

of the curve depends upon the degrees of freedom only. The distribution becomes more and more symmetrical as the degrees of freedom increase and tends to normal distribution for high value of degrees of freedom ($\nu \geq 30$).

(ii) Mean = Degrees of freedom = , mode= and variance= 2ν

(iii) The values of χ^2 lie between 0 and ∞.

(iv) If be a random sample drawn from a normal population with mean μ and variance , then

(a) $\sum_{i=1}^{n}(x_i-\mu)^2/\sigma^2$ follows χ^2 -distribution with n degrees of freedom.

(b) follows χ^2 -distribution with degrees of freedom $n-1$, $\bar{x}$ being the sample mean.

(v) For large degrees of freedom (ν), the standard variate

$$\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2/\sigma^2-n}{\text{S.E.}(\chi^2\sigma^2)}=\frac{\chi^2-\nu}{\sqrt{2\nu}}$$

may approximately be considered as a standard normal variate.

Conditions for applications of Chi-square test:

The following conditions need to be satisfied before applying χ^2 as a test statistic for hypothesis testing:

(i) In a contingency table, an expected frequency of less than 5 in a cell is less than the frequency required to apply the test. In such cases we apply the χ^2 technique of 'pooling' which consists in adding the

may approximately be considered as a standard normal variate.

Conditions for applications of Chi-square test:

The following conditions need to be satisfied before applying χ^2 as a test statistic for hypothesis testing:

(i) In a contingency table, an expected frequency of less than 5 in a cell is less than the frequency required to apply the test. In such cases we apply the technique of 'pooling' which consists in adding the frequencies which are less than 5 with the preceding or succeeding frequency, so that the resulting sum is 5 or more and adjust for the degrees of freedom accordingly.

(ii) The sample should consist of at least 50 observations and should be drawn randomly from the population.

(iii) The data should be expressed in original units, rather than in percentage or ratio form.

(iv) The constraints on the cell frequencies, if any, should be linear such as $\sum O = \sum E = N$, the total frequency.

14.4 APPLICATIONS OF χ^2 TEST

The Chi-square test is not based on assumption of normality and can be used to both large and small samples. Chi-square distribution is a widely used since it can be employed in situations in which the normal distribution cannot be applied. Chi-square distribution has a number of applications, some of which are enumerated as below:

(i) to test goodness of fit,

(ii) to test the independence of attributes,

(iii) to test if the population has a specified value of standard deviation or variance.

(iv) to test if two or more populations are homogenous with respect to some characteristics of interest.

Critical Values of χ^2 :

From the χ^2 table (given in the Appendix) it may be observed that that the tabulated (critical) values of χ^2 increase as the degrees of freedom increases and the level of significance decreases.

Let us define the symbol $\chi_n^2(\alpha)$ as the value of χ^2 for $n\ d.f.$ such that the area to the right of this point is α, i.e.,

is the rejection region. This is shown in the following diagram

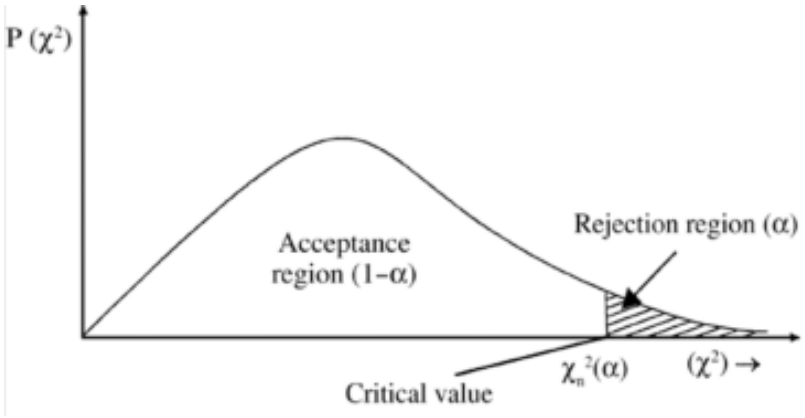


Fig. 14.2 χ^2 Critical Value and rejection region of χ^2 statistic



CHECK YOUR PROGRESS

Q 1. State whether the following statements are true or false:

(i) The critical value of a statistic is the value which cuts off the region for the rejection of null hypothesis.

(ii) The χ^2 -test requires that the sampling distribution should be normally distributed.

(iii) If null hypothesis is accepted, then it is concluded that data analysis did not produce significant results.

(iv) In order to reject the null hypothesis $\chi^2_{cal} > \chi^2_{table}$

(v) The sampling distribution is a family of curves, the distribution of which varies with the degrees of freedom.

Now we discuss various applications of χ^2 as mentioned in 14.4.

14.4.1 χ^2 TEST OF GOODNESS OF FIT

χ^2 test enables us to ascertain whether the known probability distributions such as Binomial, Poisson and Normal distributions fit with an observed sample observation. Thus χ^2 test for goodness of fit, a statistical test, provides a platform that can be used to ascertain

whether theoretical probability distributions coincide with empirical sample distributions. This test compares the theoretical (expected) frequencies with the observed (actual) frequencies to obtain the deviation between theoretical and observed frequencies. For the application of χ^2 test, first a theoretical distribution is hypothesized for a given population and then the test is carried out to make sure whether or not the sample data could come from the population of interest with the hypothesized theoretical distribution. The steps used for hypothesis testing can also be performed using the χ^2 goodness of fit test.

Steps to conduct χ^2 test for goodness of fit: The general steps to conduct a goodness of fit test for any hypothesized population distribution are summarized as below:

Step 1: Set up the null and alternative hypotheses as

Null hypothesis H_0 : No difference between the observed and expected sets of frequencies.
Alternative hypothesis H_1 : There is a difference.

Step 2: Calculate expected frequencies (E values) in each category by multiplying the category probability by the sample size.

Step 3: Compute the test statistic

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

where O is the observed frequency, E the expected frequency.

Step 4: Using a specified level of significance α and provided that the number of expected frequencies are 5 or more for all categories, determine the critical value of χ^2.

Step 5: Compare the computed value of χ^2 with the critical value of χ^2 at a specified level of significance and degrees of freedom. The decision to accept or reject the null hypothesis is given by

If χ^2_{cal} is less than the critical value $\chi^2_{\alpha,n-1}$ then we may accept the null hypothesis, otherwise we reject the null hypothesis.

Example 14.1: A company is concerned about the increasing violent altercations between its employees. The number of violent incidents recorded by the management during six randomly selected months is given by

Months	Number of violent incidents
Jan	55
Feb	65
Mar	68
Apr	72
May	80
Jun	85

Test whether the data fits a uniform distribution. (Use $\alpha = 0.05$)

Solution: Let us set up null and alternative hypotheses as

Null hypothesis H_0: Number of violent altercations are uniformly distributed over the months.

Alternative hypothesis H_1: Number of violent altercations are not uniformly distributed over the months.

In order to test the null hypothesis, the appropriate test statistic is given by

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

where O is the observed frequency, E the expected frequency.

Now the expected frequencies will be calculated by dividing total observed frequencies by number of

months. Hence, the expected frequency $\frac{\sum O}{6} = \frac{420}{6} = 70.$

Now to compute the test statistic we prepare the following table:

Months	O	E	$\frac{(O-E)^2}{E}$
Jan	55	70	3.2142
Feb	65	70	0.3571
Mar	68	70	0.0571
Apr	72	70	0.0571
May	80	70	0.9142
Jun	85	70	2.0571
Total	420		6.65

Thus

$\therefore$ The calculated value of is 6.65

The degrees of freedom= $n-1=6-1=5$

The critical value of χ^2 at 5% level of significance for 5 degrees of freedom i.e., $\chi^2_{0.05,5}=11.07$

Since the calculated value of is less than the critical value , therefore we may accept the null hypothesis and conclude that the number of violent altercations is uniformly distributed over the months.

Example 14.2: 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 information:

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

Develop a table of observed and expected frequencies for this problem.

(i) Calculate the sample Chi-square value.
(ii) State the null and alternative hypothesis.

(iii) At $\alpha = 0.05$, test whether brand share is the same across the four regions.

Solution: To calculate the expected frequencies for the corresponding observed frequencies, we apply the following formula

(Row total × Column total)/Grand total.

The calculations of expected frequencies for the corresponding observed frequencies are given below:

		O	E	$O-E$	$(O-E)^2$	$(O-E)^2/E$
Row 1	NE	40	47.5	-7.5	56.25	1.18
	NW	55	47.5	7.5	56.25	1.18
	SE	45	47.5	-2.5	6.25	0.13
	SW	50	47.5	2.5	6.25	0.13
Row 2	NE	60	52.5	7.5	56.25	1.07
	NW	45	52.5	-7.5	56.25	1.07
	SE	55	52.5	2.5	6.25	0.12
	SW	50	52.5	-2.5	6.25	0.12
Total						5.00

(i) The sample Chi-square statistic is given by

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

$\chi^2 = 5.00$

where O is the observed frequency, E the expected frequency.

$\therefore$ From the above table we have,

(ii) We state the null and alternative hypotheses as:

Null hypothesis H_0 : The brand share is evenly distributed.

Alternative hypothesis H_1 : The brand share is not evenly distributed.

(iii) The calculated value of χ^2 is 5.00

The degrees of freedom = $(r-1)(c-1) = (2-1)(4-1) = 3$

The critical value of χ^2 at 5% level of significance for 3 degrees of freedom i.e., $\chi^2_{0.05,3} = 7.815$

Since the calculated value of is less than the critical value, therefore we may accept the null hypothesis and conclude that the brand share is evenly distributed in all the four regions of the country.

Example 14.3: A survey of 800 families with 4 children each revealed the following distribution:

Number of births		Number of families
Male	Female	
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 the chance of a male birth is equal to that of a female birth.

Solution: Here we set up the null hypothesis that the data are consistent with the binomial law of equal probability for male and female births, i.e.,

$p = q = \frac{1}{2}$, (where p denotes the probability of a male birth.)

Using the binomial probability law, the expected frequency of x male births is given by:

$^{n}C_x\, p^x q^{n-x}$, where

$N = 800,\ n = 4,\ \ x = 0, 1, 2, 3, 4.$

$= 800 \times {}^{4}C_x \left(\frac{1}{2}\right)^4$

$= 50 \times {}^{4}C_x$ .....(*)

Thus putting $x = 0, 1, 2, 3, 4$ successively in (*), we can determine the expected binomial frequencies as given below:

Number of Male births (x)	Expected frequency $f(x)$
0	$50\times{}^{4}C_0=50\times1=50$
1	$50\times{}^{4}C_1=50\times4=200$
2	$50\times{}^{4}C_2=50\times6=300$
3	$50\times{}^{4}C_3=50\times4=200$
4	$50\times{}^{4}C_4=50\times1=50$
Total	800

Let us now apply χ^2 test to examine the goodness of fit of the given data to the above binomial distribution. Calculations are given below:

	Observed frequency(O)	Expected frequency (E)	$O-E$	$(O-E)^2$	$\frac{(O-E)^2}{E}$
0	32	50	-18	324	6.48
1	178	200	-22	484	2.42
2	290	300	-10	100	0.33
3	236	200	36	1296	6.48
4	64	50	14	196	3.92
Total	800	800	0		19.63

Thus $\chi^2 = 19.63$

$\therefore$ The calculated value of is 19.63

The degrees of freedom= $5-1=4$

The critical value of at 5% level of significance for 4 degrees of freedom i.e., $\chi^2_{0.05,4} = 9.488$

Since the calculated value of χ^2 is greater than the critical value $\chi^2_{0.05,4}$, therefore we reject the null hypothesis and conclude that the hypothesis of equal male and female births is not valid.

χ^2

Example 14.4: The figures given below are (a) the observed frequencies of a distribution, and (b) the frequencies of the normal distribution having the same mean, standard deviation and total frequency as in (a). Apply the χ^2 test of goodness of fit.

(a) 1 12 66 220 495 792 924 792 495 220 66 12 1

(b) 2 15 66 210 484 799 944 799 484 210 66 15 2

Solution: Since we are given the observed frequencies and expected frequencies of the normal distribution, therefore we apply the χ^2 test of goodness of fit.

Now we set up the null hypothesis that normal distribution is a good fit to the given frequency distribution, i.e., the given observed frequencies and the expected normal frequencies do not differ significantly.

In order to test the null hypothesis, the appropriate test statistic is given by

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

where O is the observed frequency, E the expected frequency.

Now to compute the χ^2 test we prepare the following table:

	O	E	$O-E$	$(O-E)^2$	$(O-E)^2/E$
	1, 12	2, 15	-4	16	0.94
	66	66	0	0	0.00
	220	210	10	100	0.48
	495	484	11	121	0.25
	792	799	-7	49	0.06
	924	944	-20	400	0.42
	792	799	-7	49	0.06
	495	484	11	121	0.25
	220	210	10	100	0.48
	66	66	0	0	0.00
	12, 1	15, 2	-4	16	0.94
Total	4096	4096	0		3.88

Thus $\chi^2 = 3.88$

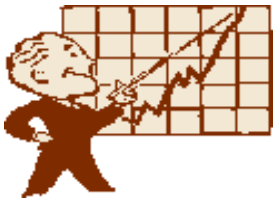
$\therefore$ The calculated value of is 3.88

The degrees of freedom=13-5=8

(The 13 original classes have been reduced to 11 by grouping, thus 2 degrees of freedom are lost in pooling. In addition, the mean, the standard deviation and total frequency of the original distribution have been used in calculating the theoretical frequencies, thus introducing three restraints. The number of degrees of freedom is accordingly 8.)

The critical value of χ^2 at 5% level of significance for 8 degrees of freedom i.e., $\chi^2_{0.05,8} = 15.507$

Since the calculated value of χ^2 is much less than the critical value $\chi^2_{0.05,8}$, it is highly non-significant. Hence, we may accept the null hypothesis and conclude that the normal distribution is a good fit to the given distribution.



CHECK YOUR PROGRESS

Q 2. In a particular chi-square goodness-of-fit test there are four categories and 200 observations. Use the 5% significance level.

(a) How many degrees of freedom are there?

(b) What is the critical value of chi-square?

Q 3. The null hypothesis and alternative hypothesis are

H_0 : The frequencies are equal and H_1 : The frequencies are not equal

Category	f
A	10
B	20
C	30

(a) State the decision rule ($\alpha = 0.05$), (b) Compute the value of Chi-square.

(c) What is your decision regarding

χ^2

14.4.2 TEST FOR INDEPENDENCE OF ATTRIBUTES

In many business situations, a market researcher might be interested to study the relationship between two variables or to check whether they are independent of each other. For example, an edible oil company may be interested in knowing whether the purchase of oil is independent of the customer's age or whether it is dependent on the customer's age. These are two different situations and consequently the company needs to frame a production and selling strategy. Similarly following are the situations where the Chi-square test of independence can be employed to analyse attributes with multiple categories.

- Whether university students classified by gender are independent of courses of study.
- Whether wage-earners classified by education level are independent of income.
- Whether a product produced is acceptable or not is

independent of the shifts in which it was produced.

Thus when observations are classified on the basis of two qualitative variables or attributes and arranged in a table, the resulting table is known as a contingency table (Table 9.1). χ^2 test of independence uses this contingency table for determining independence of two qualitative variables.

Table 9.1: Contingency table

Variable B	Variable A				Total
	A_1	A_2	.....	A_k	
B_1	O_{11}	O_{12}	.....	O_{1k}	R_1
B_2	O_{21}	O_{22}		O_{2k}	R_2
.	.	.		.	.
.	.	.		.	.
.	.	.		.	.
B_r	O_{r1}	O_{r2}	.....	O_{rk}	R_r
Total	C_1	C_2		C_k	N

It is to be noted that the variables A and B have been classified into mutually exclusive categories. The value O_{ij} is the observed frequency for the cell in ith row and jth column. R_i is the total of the i^{th} row and C_j is the total of the j^{th} column. The row and column totals are added up to get a grand total (N), which represents the sample size. It is to be noted that to apply the Chi-square test we need to calculate the expected frequency.

The calculation of expected frequency, E_{ij} , corresponding to an observed frequency O_{ij} in row i and column is based on the concept of multiplicative law of probability. Using this concept of probability the expected frequency, , is calculated as

$$\frac{\text{i}^{\text{th}}\text{ Row total}}{\text{sample size}} \times \frac{\text{j}^{\text{th}}\text{ column total}}{\text{sample size}} \times \text{Grand total}$$

$$= \frac{\text{R}_\text{i}}{\text{N}} \times \frac{\text{C}_\text{j}}{\text{N}} \times \text{N} = \frac{\text{R}_\text{i} \times \text{C}_\text{j}}{\text{N}}$$

Thus expected frequency for any cell can be computed by using the formula

$$E_{ij} = \frac{\text{Row total} \times \text{Column total}}{\text{Grand total}}$$

The analysis of a two-way contingency table helps to determine whether the two variables are unrelated or independent of each other. Thus the null hypothesis for a Chi-square test of independence is that the two variables are independent.

Under the null hypothesis, the χ^2 test statistic is defined as

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

The degrees of freedom for the Chi-square test of independence are given by

Degrees of freedom= (Number of rows-1) × (Number of columns-1)

$$= (r-1) \times (c-1)$$

Example 14.5: A certain drug is claimed to be effective in curing colds. In an experiment on 500 persons with cold, half of them were given the drug and half of them were given sugar pills. The patients' reactions to the treatment are recorded in the following table:

	Consequence			
Treatment	Helped	Reaction	No effect	Total
Drug	150	30	70	250
Sugar pills	130	40	80	250
Total	280	70	150	500

On the basis of the data, can it be concluded that there is a significant difference in the effect of the drug and sugar pills?

Solution: We state the null hypothesis H_0 that there is no significant difference in the effect of the drug and sugar pills. On the basis of this hypothesis expected frequencies are calculated as:

Treatment	Consequence Helped	Reaction	No effect	Total
Drug	$\frac{250\times280}{500}=140$	$\frac{250\times70}{500}=35$	250-(140+O5)=75	250
Sugar pills	140	35	75	250
Total	280	70	150	500

Now we arrange the observed and expected frequencies in the following table in order to calculate the χ^2 test statistic.

Observed (O)	Expected (E)	$(O-E)$	$(O-E)^2$	$(O-E)^2/E$
150	140	10	100	0.714
130	140	-10	100	0.714
30	35	-5	25	0.714
40	35	5	25	0.714
70	75	-5	25	0.333
80	75	5	25	0.333
Total				3.522

Thus the χ^2 test statistic is

$$\chi^2=\sum\frac{(O-E)^2}{E}=3.522$$

Number of degrees of freedom= (r-1) (c-1) = (2-1) (3-1) =2

The critical value of the χ^2 at 5% level of significance for 2 degrees of freedom = 5.99

Since the calculated value of the is less than the critical value, therefore we may accept the null hypothesis. Hence conclude that drug is no better than sugar pills in curing colds.

Example 14.6: Two researchers adopted different sampling techniques while investigating the same group of students to find the number of students falling in different intelligence levels. The results are as follows:

Researcher	Number of students in each level				Total
	Below average	Average	Above average	Genius	
X	86	60	44	10	200
Y	40	33	25	2	100
Total	126	93	69	12	300

Would you say that the sampling techniques adopted by the two researchers are significantly different?

Solution: We state he null hypothesis that the data obtained are independent of the sampling techniques adopted by the two researchers.

Under the null hypothesis of independence, the expected frequencies are calculated as:

$$E_{11} = \frac{126 \times 200}{300} = 84; \qquad E_{12} = \frac{93 \times 200}{300} = 62; \qquad E_{13} = \frac{69 \times 200}{300} = 46$$

Now we arrange the observed and expected frequencies in the following table in order to calculate the χ^2 test statistic.

Observed (O)	Expected (E)	$(O-E)$	$(O-E)^2$	$(O-E)^2/E$
86	84	2	4	0.048
60	62	-2	4	0.064
44	46	-2	4	0.087
10	8	2	4	0.500
40	42	-2	4	0.095
33	31	2	4	0.129
25	23	2	4	0.174
2	4	-2	4	1.000
Total	300	0		2.097

Thus the χ^2 test statistic is

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 2.097$$

Number of degrees of freedom= (r-1)(c-1)=(4-1)(2-1)=3

The critical value of the χ^2 at 5% level of significance for 3 degrees of freedom = 7.82

Since the calculated value of χ^2 the is less than the critical value, therefore we may accept the null hypothesis. Hence conclude that the sampling techniques adopted by the two investigators do not differ significantly.



CHECK YOUR PROGRESS

Q 4. A social scientist sampled 140 people and classified them according to income level and whether or not they played a state lottery in the last month. The sample information is reported below. Is it reasonable to conclude that playing the lottery is related to income level? Use the 5% significance level.

	Income			
	Low	Middle	High	Total
Played	46	28	21	95
Did not play	14	12	19	45
Total	60	40	40	140

(a) What is this table called?

(b) State the null and alternative hypotheses.

(c) What is the degrees of freedom?

(d) What is the decision rule?

(e) Determine the value of χ^2.

(f) Make a decision about H_0.

14.4.3 χ^2 TEST FOR A SPECIFIED POPULATION VARIANCE

Let $x_1, x_2, ..., x_n$ be a random sample of size n drawn from a normal population with mean and variance . We want to test the hypothesis that the population variance has a specified value.

We set up the null hypothesis and the alternative hypothesis as

$H_0 : \sigma^2 = \sigma_0^2$ and $H_1 : \sigma^2 > \sigma_0^2$

Then under the null hypothesis, the test statistic

$$\chi^2 = \sum \frac{(x - \bar{x})^2}{\sigma^2} = \frac{ns^2}{\sigma^2}$$

follows χ^2 distribution with $(n-1)\ d.f.$, where

, denotes the sample variance.

By comparing the calculated value of χ^2 with the tabulated value of for at the specified level of significance, we may accept or reject the null hypothesis at that level of significance.

Remarks: 1. test given by

$$\chi^2 = \frac{ns^2}{\sigma^2}$$

can be applied only if the population is normal.

2. If the sample size is large, say, greater than 30, then we can use Fisher's approximation, viz.,

$$\sqrt{2\chi^2} \sim N\left(\sqrt{2n-1}, 1\right)$$

And then apply the normal test by considering the statistic

$$\chi^2 = \frac{\sum(x-\bar{x})^2}{\sigma(0.9)^2} = 17.78 \qquad Z = \sqrt{2\chi^2} - \sqrt{2n-1} \sim N(0,1)$$

Example 14.7: A battery manufacturer claims that the lives of his batteries have a standard deviation of 0.9 year. Can we accept manufacture's claim if the lives of a random sample of 10 batteries have a standard deviation of 1.2 years? (Use a 5% level of significance)

Solution: We set up the null and alternative hypotheses as

Null hypothesis $H_0 : \sigma = 0.9\ years$

Alternative hypothesis $H_1 : \sigma > 0.9\ years$

Under the null hypothesis, the test statistic

~Chi-square distribution with $(n-1)\ d.f.$

Now

Degrees of freedom= n-1 =10-1=9

The critical value of χ^2 at 5% level for 9 d.f. i.e., $\chi^2_{0.05}(9) = 16.92.$

Since the critical value of χ^2 is greater than the tabulated value of value of χ^2 at 5% level for 9 d.f., we reject the null hypothesis and conclude that the standard deviation of life of the batteries is more than 0.9 years.

Example 14.8: The weights in kg. of 10 students are given below:

38, 40, 45, 53, 47, 43, 55, 48, 52, 49

Can we say that the variance of the distribution of weights of all students from which the above sample of 10 students was drawn, is equal to 20 square kg.?

Solution: We set up the null and alternative hypotheses as:

Null hypothesis $H_0 : \sigma^2 = 20\ sq.\ kg.$

Alternative hypothesis $H_1 : \sigma^2 > 20\ sq.\ kg.$

Under the null hypothesis, the test statistic

$\chi^2 = \frac{ns^2}{\sigma^2}$ ~Chi-square distribution with $(n-1)\ d.f.$

Now we prepare the following table to calculate the sample variance

Weight in kg. (x)	$(x-\bar{x})$ $x-47$	$(x-\bar{x})^2$
38	-9	81
40	-7	49
45	-2	4
53	6	36
47	0	0
43	-4	16
55	8	64
48	1	1
52	5	25
49	2	4
Total=470		280

Now, $\chi^2 = \frac{ns^2}{\sigma^2} = \frac{10 \times 28}{20} = 14$

Degrees of freedom= n-1=10-1=9

The critical value of χ^2 i.e., $\chi^2_{0.05}(9) = 16.92$

Since the calculated value of χ^2 is less than the critical value of $\chi^2_{0.05}(9)$, therefore we may accept our null hypothesis and conclude that the variance of the distribution of weights of all students in the population is 20 sq. kg.

14.4.4 χ^2 TEST OF HOMOGENEITY

χ^2 test of homogeneity is used to determine whether two or more independent variables are drawn from the same population or from different populations. In other words, χ^2 test of homogeneity is used to know whether two or more populations are homogeneous with respect to some characteristics of interest. For example, a researcher may be interested in knowing whether the employees from three departments, production, finance, and personnel, feel the same about the requirements of the top management in terms of hard work expected from employees. The amount of hard work can also be classified into three groups, namely, very hard, hard, and easy going. In this situation, we can state a null hypothesis that the opinion of all the groups is the same about the requirement of hard work. In other words, the null hypothesis assumes that the three classifications are homogeneous in terms of their opinion about the amount of hard work required by the top management.

It may be noted that, in both χ^2 test of homogeneity and χ^2 test of independence, one is concerned with the cross tabulation of the data and hypotheses testing procedure is also same. The differences that exist between the χ^2 test of homogeneity and χ^2 test of independence are: (i) In χ^2 test of independence, a researcher is interested in testing whether two attributes are independent. On the

other hand in, χ^2 intest of homogeneity, a researcher investigates whether two or more populations are homogenous with respect to some characteristic of interest. (ii) in χ^2 test of homogeneity, two or more independent samples are drawn from each population whereas in case of χ^2 test of independence, a single sample is drawn from a population.

Example 14.9: A television company has launched a new product with some advanced features. The company wants to know the opinion of consumers about this product with respect to four characteristics: preferred brand with new features did not prefer brand with new features, preferred only a few new features, and indifferent. The company has divided consumers into three groups—executives/ officers; businessmen, and private consultants. It has taken a random sample of size 459 and obtained the following results:

Consumers / Opinion	Executives/Officers	Businessmen	Private consultants	Row total
Preferred brand with new features	35	25	40	100
Did not prefer brand with new features	30	45	34	109
Preferred only a few new features	45	50	25	120
Indifferent	25	55	50	130
Column total	135	175	149	459

Use χ^2 test of homogeneity and draw inference from the data.

Solution: Let us set up null and alternative hypotheses as

Null hypothesis H_0 : Opinion of all the groups is the same about the product with new features.

And alternative hypothesis H_1 : Opinion of all the groups is not the same about the product with new features.

In order to test the null hypothesis, the appropriate test statistic is given by

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

where O is the observed frequency, E the expected frequency.

To calculate the expected frequencies for the corresponding observed frequencies, we apply the following formula

(Row total × Column total)/Grand total.

Now for the calculation of the expected frequencies against the corresponding observed frequencies and to compute the test statistic we prepare the following table:

Observed (O)	Expected (E)	$(O-E)$	$(O-E)^2/E$
35	29.41	5.59	1.0625
30	32.06	-2.06	0.1324
45	35.29	9.71	2.6717
25	38.24	-13.24	4.5841
25	38.13	-13.13	4.5213
45	41.56	3.44	0.2847
50	45.75	4.25	0.3948
55	49.56	5.44	0.5971
40	32.46	7.54	1.7514
34	35.38	-1.38	0.0538
25	38.95	-13.95	4.9962
50	42.20	7.8	1.4417
420			22.4917

Thus the χ^2 test statistic is

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 22.4917$$

The number of degrees of freedom $= (r-1)(c-1) = (4-1)(3-1) = 6$

The critical value of the χ^2 at 5% level of significance for 6 degrees of freedom = 12.59

Since the calculated value of the χ^2 is greater than the critical value, therefore we reject the null hypothesis and conclude that the opinion of all the groups is not the same about the product with new features. Hence, the company has to consider different groups and their needs separately.

Precautions about using the Chi-square test: In order to apply Chi-square hypothesis test properly, one has to be extremely careful and keep in mind certain precautions. First, a sample should be large enough. If the expected frequencies are too small, the value of Chi-square gets over estimated. This will result in the rejection of the null hypothesis in several situations. To overcome this problem, we must ensure that the expected frequency in any cell of the contingency table should not be less than 5. In case the expected frequency less than 5 in more than one cell, we can combine these to obtain the expected frequency of at least 5. Another point to note is that the calculations must be made with actual numbers in each cell and not with proportions or percentages. If the proportions or percentages were used, then the theoretical distribution would not be applicable.

When the calculated value of Chi-square turns out to be more than the critical value at a specified level of significance, we reject our null hypothesis. On the other hand, when Chi-square value is less than the critical value, we may accept our null hypothesis. However, when the Chi-square value turns out to be zero, we have to be extremely careful to confirm that there is no difference between the observed and the expected frequencies. Such a situation may sometimes arise on account of a faulty manner used in the data collection.

In most of the cases, the problems of Chi-square involve simple calculations. However, for large sets of data the Chi-square test involves very comprehensive calculations. In all such situations, computer should be used. Several computer Statistics packages contain routines for carrying out Chi-square tests.



CHECK YOUR PROGRESS

Q 5: When chi-square test of homogeneity is useful?

……



14.5 LET US SUM UP

- Statistical tests that do not require any prior information about the population are termed as non-parametric tests. This unit focuses on only χ^2 (Chi-square) distribution and the related χ^2 test. χ^2 distribution is the family of curves with each distribution being defined by the degree of freedom associated to it.

- χ^2 test can be used for a variety of purposes. χ^2 test provides a platform that can be used to ascertain whether the theoretical probability distribution coincides with the empirical sample distribution. This is commonly known as χ^2 goodness of fit test. It can also be used to test the independence of independence of two variables.

- χ^2 test of independence makes use of the contingency table to determine the independence of two variables. This test can also be used for estimating the population variance. Again the test of homogeneity is applied to determine whether two or more populations are homogeneous with respect to some characteristics of interest.



14.6 FURTHER READINGS

1) Sharma, J.K. (2007). Business Statistics. New Delhi. Pearson Education Ltd.

2) Hazarika, P.L. (2016). Essential Statistics For Economics And Business Studies. New Delhi. Akansha Publishing House.

3) Lind, D.A., Marshal, W.G., Wathen, S.A. (2009) Statistical Techniques in Business and Economics. New Delhi. Tata McGraw Hill Education Private Limited.



14.7 ANSWERS TO CHECK YOUR PROGRESS

Ans. to Q No 1: (i) T, (ii) F, (iii) T, (iv) T, (v) T.

Ans. to Q No 2: (a) , (b) The critical value of at 5% level of significance for 3 degrees of freedom i.e., $\chi^2_{0.05,3} = 7.8147$

Ans. to Q No 3: $d.f. = 3-1 = 2$

The critical value of at 5% level of significance for 2 degrees of freedom i.e., $\chi^2_{0.05,2} = 5.9915$

(a) The decision rule is to reject if

(b) $\chi^2 = \frac{(10-20)^2}{20} + \frac{(20-20)^2}{20} + \frac{(30-20)^2}{20} = 10.0$

(c)Reject H_0.

Ans. to Q No 4: (a) Contingency table.

(b) Null hypothesis is H_0 : There is no relationship between income and whether the person played the lottery.

Alternative hypothesis is H_1 : There is a relationship between income and whether the person played the lottery.

(c) The degrees of freedom, $d.f. = (2-1)\times(3-1) = 2$

(d) Reject H_0 if $\chi^2 > 5.9915$,The critical value of χ^2 at 5% level of significance for 2 degrees of freedom i.e., $\chi^2_{0.05,2} = 5.9915$

(e)

$$\chi^2 = \frac{(46-40.71)^2}{40.71} + \frac{(28-27.14)^2}{27.14} + \frac{(21-27.14)^2}{27.14} + \frac{(14-19.29)^2}{19.29} + \frac{(12-12.86)^2}{12.86} + \frac{(19-12.86)^2}{12.86}$$

= 6.544

(f) Reject H_0.

Ans. to Q No 5: Chi-square test of homogeneity is useful in a situation when we intend to verify whether several populations are homogeneous with respect to some characteristic of interest.



14.8 MODEL QUESTIONS

1. Define Chi-square distribution. Discuss the important properties of the distribution.

2. Under what situations Chi-square test can be applied?

3. Write down the applications of Chi-square test.

4. The figures given below are (a) the theoretical frequencies of a distribution, and (b) the frequencies of the normal distribution having the same mean, standard deviation and the total frequency as in (a).

(a)	1	5	20	28	42	22	15	5	2
(b)	1	6	18	25	40	25	18	6	1

Apply χ^2 -test of goodness of fit.

[Hints: Since the observed and the expected frequencies are less than 5 at the beginning and at end of the series, these frequencies are grouped

Degrees of freedom = 4, Tabulated value of $\chi^2_{0.05}(4) = 9.49$

H_0 may be accepted implying that the fit is good.]

5. A Mumbai film director claims that his films are liked equally by males and females. An opinion survey of a random sample of 1000 film-goers revealed the following results:

	Liked	Disliked
Male	402	193
Female	245	160

Is the film director's claim supported by data?

[Given : $\chi^2_{0.05} =$ 3.84 5.99 7.81

d.f.= 1 2 3]

6. A certain oil corporation conducts a seismic test on land to assess the possibility of extraction of oil from the land. From the past 200 records, the following test results and oil yields were obtained:

Test results	Oil yield frequencies		
	High	Medium	Low
Good	44	22	13
Fair	15	19	26
Bad	5	8	47

Can we say, at 1% level of significance that oil yields and test results are associated?

7. The following table gives the classification of 100 workers according to sex and the nature of work. Test whether the nature of work is independent of the sex of the work.

	Skilled	Unskilled
Male	40	20
Female	10	30

8. A certain drug was administered to 456 males out of a total of 720 in a certain locality to test its efficiency against typhoid. The incidence of typhoid is shown below. Find out the effectiveness of the drug against the disease. (The table value of χ^2 for 1 degree of freedom at 5% level of significance is 3.84)

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

9. A sample of 20 observations gave a standard deviation of 3.72. Is this compatible with the hypothesis that the sample is drawn from a normal population with variance 4.35?

[Hints: H_0 : The sample is drawn from a normal population.

Under H_0, the test statistic $\chi^2 = \frac{ns^2}{\sigma^2}$ follows χ^2 distribution with $n-1 = 20-1 = 19\ d.f.$ Table value of is rejected at 5% level.]

*** ***** ***