
UNIT 12 CATEGORICAL DATA ANALYSIS

Structure

- 12.0 Introduction
- 12.1 Objectives
- 12.2 Chi-Square Test for Goodness of Fit
- 12.3 Chi-Square Test for Independence of Attributes
- 12.4 Kolmogorov–Smirnov Goodness of Fit Test
- 12.5 Summary
- 12.6 Solutions / Answers
- References

12.0 INTRODUCTION

In many real-world situations, like in business and other areas, the data are collected in the form of counts. In some cases, the collected data are classified into different categories or groups according to one or more attributes. Such type of data is known as categorical data. For example, the number of people of a colony can be classified into different categories according to age, gender, income, job, etc. or the books in a library can be classified according to their subjects such as books of Science, Commerce, Art, etc. Now, the question arises “how do we tackle the inference problems arising out of categorical data?”. For testing various hypotheses involving population parameter(s) such as mean(s), proportion(s), variance(s), etc., we perform the following two points:

- We make an assumption regarding the form of the distribution function of the parent population from which the sample has been drawn, or in other words, the form of the distribution is known to us, and
- We test a statistical hypothesis regarding population parameters under study, such as mean(s), variance(s), proportion(s), etc.

Those types of tests where we deal with the above two points are known as “**parametric tests**”.

But in many real-life problems particularly in social and behavioural sciences, the requirements of the parametric tests do not meet. Therefore, in such situations, the parametric tests are not applicable. Thus, statisticians discovered various tests and methods which are independent of the form of population distribution and also applicable when the observations are not measured in interval scale that is observations are measured in ordinal scale or nominal scale. These tests are known as “**Non-parametric tests**” or “**Distribution Free Tests**”. The name distribution-free test is suggested since the distribution of the test statistic is free (independent) from the form of population distribution. In addition to it, the non-parametric test is suggested since these tests are developed for such hypothesis testing which is not involving the population parameter(s).

These tests are based on the “**order statistics**” theory which has the property that “The distribution of the area under the density function between any two ordered observations is independent of the form of the density function”. Therefore, in these tests, we use median, range, quartile, etc., in our hypothesis.

Assumptions Associated with Non-parametric Tests:

Although the non-parametric tests are not based on the assumption that the form of parent population is known, but these tests require some basic assumptions which are:

When the data are classified into different categories or groups according to one or more attributes, such type of data is known as categorical data.

A parametric test is a test which is based on the fundamental assumption that the form of parent population is known and involved parameter(s).

A non-parametric test is a test that does not make any assumption regarding the form of the parent population from which the sample is drawn.

- (i) Sample observations are independent.
- (ii) The variable under study is continuous.
- (iii) Lower order moments exist.

Obviously, these assumptions are fewer and weaker than those associated with the parametric tests.

This unit discusses non-parametric tests for categorical data. The tests discussed in this unit include the chi-square test for goodness of fit, the chi-square test for independence of attributes and the Kolmogorov-Smirnov test for goodness of fit.

12.1 OBJECTIVES

After studying this unit, you should be able to:

- describe the need for tests that are applied for categorical data;
- apply the chi-square test for goodness of fit in different situations;
- define the contingency table;
- apply the chi-square test for independence of two attributes; and
- apply the Kolmogorov-Smirnov test.

12.2 CHI-SQUARE TEST FOR GOODNESS OF FIT

The parametric tests are the ones in which one first assumes the form of the parent population and then performs a test about some parameter(s) of the population, such as mean, variance, proportion, etc. Generally, the parametric tests are based on the assumption of a normal population. The suitability of a normal distribution or some other distribution may itself be verified by means of the goodness of fit test. The chi-square (χ^2) test for goodness of fit was given by Karl Pearson in 1900. It is the oldest non-parametric test. With the help of this test, we test whether the random variable under study follows a specified distribution, such as binomial, Poisson, normal or any other distribution, when the data are in categorical form. Here, we compare the actual or observed frequencies of each category with theoretically expected frequencies that would have occurred if the data followed a specified or assumed or hypothesised probability distribution. This test is known as “**goodness of fit test**” because we test how well an observed frequency distribution (distribution from which the sample is drawn) fit to the theoretical distribution such as normal, uniform, binomial, etc.

Assumptions

This test works under the following assumptions:

- (i) The sample observations are random and independent.
- (ii) The sample size is large.
- (iii) The observations may be classified into non-overlapping categories.
- (iv) The expected frequency of each class is greater than five.
- (v) The sum of observed frequencies is equal to the sum of expected frequencies, i.e., $\sum O = \sum E$.

As usual, the first step in testing of hypothesis is to set up null and alternative hypotheses, so our null and alternative hypotheses are setup, below in Step 1.

Step 1: Generally, we are interested to test whether data follow a specified or assumed or hypothesised distribution $F_0(x)$ or a sample has come from a specified distribution or not. So here we consider only a two-tailed case. Thus, we can take the null and alternative hypotheses as

H_0 : Data follow a specified distribution

H_1 : Data does not follow a specified distribution

In symbolical form

$H_0 : F(x) = F_0(x)$ for all values of x

$H_1 : F(x) \neq F_0(x)$ for at least one value of x

Step 2: After setting the null and alternative hypotheses, our next step is to draw a random sample. So, let a random sample of size N be drawn from a population with unknown distribution function $F(x)$ and the data are categorised into k groups or classes. Also, let $O_1, O_2, \dots, O_k$ are the observed frequencies and $E_1, E_2, \dots, E_k$ are the corresponding expected frequencies. If the parameter(s) of assumed distribution is (are) unknown then in this step, we estimate the value of each parameter of the assumed distribution with the help of sample data by calculating the sample mean, variance, proportion, etc as may be the case.

Step 3: After that, we find the probability of each category or group in which an observation falls with the help of the assumed probability distribution.

Step 4: If p_i ($i = 1, 2, \dots, k$) is the probability that an observation falls in i^{th} category then we find the expected frequency by the formula given below

$$E_i = Np_i; \quad \text{for all } i = 1, 2, \dots, k$$

Note 1: Sometimes, (generally, when the expected frequencies come in the form of decimal) it is observed that the sum of expected frequencies not come equal to the sum of observed frequencies yet it is a necessary condition for this test so in such cases, last expected frequency is obtained by subtracting the sum of all expected frequencies except last expected frequency from the sum of observed frequencies, that is,

$$E_k = N - (E_1 + E_2 + \dots + E_{k-1})$$

Step 5: Test statistic:

Since this test compares observed frequencies with the corresponding expected frequencies, therefore, we are interested in the magnitudes of the differences between the observed and expected frequencies. Specifically, we wish to know whether the differences are small enough to be attributed to chance or they are large due to some other factors. With the help of observed and expected frequencies, we may compute a test statistic that reflects the magnitudes of differences between these two quantities when H_0 is true. The test statistic is given by

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \sim \chi_{(k-1)}^2 \text{ under } H_0$$

Here, we take the notation for sample size is 'N' instead of 'n' since in this test generally we deal with frequencies of the observations and to represent the sum of frequencies we take notation 'N'.

where k represents the number of classes. If any expected frequency is less than 5 then it is pooled or combined with the preceding or succeeding class then k represents the number of classes that remain after the combining classes.

The χ^2 -statistic follows approximately the chi-square distribution with $(k - 1)$ degrees of freedom.

If the parameter (s) of the distribution to be fitted is (are) unknown, that is, not specified in null hypothesis then test statistic χ^2 follows approximately the chi-square distribution with $(k - r - 1)$ degrees of freedom, that is,

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{(k-r-1)} \text{ under } H_0$$

where r is the number of unknown parameters which are estimated from the sample.

Step 6: Obtain the critical value of the test statistic at a given level of significance under the condition that the null hypothesis is true. **Table VI** in the Appendix at the end of this course provides critical values of the test statistic χ^2 for various degrees of freedom and different level of significance.

Step 7: Take the decision about the null hypothesis as:

To take the decision about the null hypothesis, the test statistic (calculated in Step 5) is compared with the chi-square critical (tabulated) value (observed in Step 6) for a given level of significance (α) under the condition that the null hypothesis is true.

If the calculated value of the test statistic is greater than or equal to the critical value with $(k - r - 1)$ degrees of freedom at α level of significance then we reject the null hypothesis H_0 at α level of significance, otherwise, we do not reject H_0 .

Let us do some examples based on the above test.

Example 1: The following data are collected during a test to determine consumer preference among five leading brands of bath soaps:

Brand Preferred	A	B	C	D	E	Total
Number of Customers	194	205	204	196	201	1000

Test that the preference is uniform over the five brands at 5% level of significance.

Solution: Here, we want to test that the preference of the customers over five brands is uniform. So our claim is “the preference of the customers over the five brands is uniform” and its complement is “the preference of the customers over the five brands is not uniform”. So we can take the claim as the null hypothesis and the complement as the alternative hypothesis. Thus,

H_0 : The preference of the customers over the five brands of bath soap is uniform

H_1 : The preference of the customers over the five brands of bath soap is not uniform

In other words, we can say that

H_0 : The probability distribution is uniform

H_1 : The probability distribution is not uniform

Since the data are given in the categorical form and we are interested to fit a distribution, so we can go for the chi-square goodness of fit test.

If X denotes the preference of the customers over the five brands of bath soap that follows uniform distribution then the probability mass function of uniform distribution is given by

$$P[X = x] = \frac{1}{N}; \quad x = 0, 1, 2, \dots, N$$

The uniform distribution has a parameter N which is given so for testing the null hypothesis, the test statistic is given by

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{(k-1)} \text{ under } H_0$$

where O_i and E_i are the observed and expected frequencies of i^{th} brand of bath soap respectively.

Here, we want to test the null hypothesis that the preference of the customers is uniform i.e. follows a uniform distribution. Uniform distribution is one in which all outcomes have an equal (or uniform) probability. Therefore, the probability that the customers prefer one of any five brands is the same. Thus,

$$p_1 = p_2 = p_3 = p_4 = p_5 = p = \frac{1}{5}$$

The theoretical or expected number of customers or frequency for each brand is obtained by multiplying the appropriate probability by the total number of customers, that is, sample size N . Therefore,

$$E_1 = E_2 = E_3 = E_4 = E_5 = Np = 1000 \times \frac{1}{5} = 200$$

Calculations for $\frac{(O - E)^2}{E}$:

Soap Brand	Observed Frequency (O)	Expected Frequency (E)	(O-E)	(O-E) ²	$\frac{(O-E)^2}{E}$
A	194	200	-6	36	0.18
B	205	200	5	25	0.13
C	204	200	4	16	0.08
D	196	200	-4	16	0.08
E	201	200	1	1	0.01
Total	1000	1000			0.48

From the above calculations, we have

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} = 0.48$$

The critical value of the chi-square with $k - 1 = 5 - 1 = 4$ degrees of freedom at 5% level of significance is 9.49.

Since the calculated value of the test statistic ($= 0.48$) is smaller than the critical value ($= 9.49$) so we do not reject the null hypothesis i.e. we support the claim at 5% level of significance.

Thus, we conclude that the sample fails to provide us sufficient evidence against the claim so we may assume that the preference of the customers over the five brands of bath soap is uniform.

Example 2: The following data give the number of weekly accidents occurring on a mile stretch of a particular road:

Number of Accidents	0	1	2	3	4	5	6 or more
Frequency	10	12	12	9	5	3	1

A highway engineer wants to know whether the data follow a Poisson distribution or not at 5% level of significance.

Solution: Here, the highway engineer wants to test that the number of accidents follows a Poisson distribution. So our claim is “the number of accidents follows the Poisson distribution” and its complement is “the number of accidents does not follow the Poisson distribution”. So we can take the claim as the null hypothesis and the complement as the alternative hypothesis. Thus,

H_0 : The number of accidents follows the Poisson distribution

H_1 : The number of accidents does not follow the Poisson distribution

Since the data are given in the categorical form and we are interested to fit a distribution, so we can go for the chi-square goodness of fit test.

Here, the parameter of the Poisson distribution is unknown so testing the null hypothesis, the test statistic is given by

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \sim \chi_{k-r-1}^2 \text{ under } H_0$$

where, k = the number of classes

r = the number of parameters estimated from the sample data

If X denotes the number of accidents per week that follows the Poisson distribution then its probability mass function is given by

$$P[X = x] = \frac{e^{-\lambda} \lambda^x}{x!}; \quad x = 0, 1, 2, \dots$$

where λ is the mean number of accidents per week.

The difficulty here is that the parameter λ is unknown (it is not specified in the null hypothesis), therefore, it is estimated from sample data. Since λ represents the mean of the population so it can be estimated by the value of the sample mean. Thus, first, we find the sample mean and the value of the sample mean would be taken as the estimate of the mean of the Poisson distribution.

S. No.	Number of Accidents(X)	Frequency(f)	fX
1	0	10	0
2	1	12	12
3	2	12	24
4	3	9	27
5	4	5	20
6	5	3	15
7	6	1	6
		$N = 52$	$\sum fX = 104$

Here, we can not use the Kolmogorov-Smirnov test for goodness of fit because this test is applied only when all the parameters of the fitting distribution are known. This test is discussed in Section 12.4

The formula for calculating mean is

$$\bar{X} = \frac{1}{N} \sum fX = \frac{1}{52} \times 104 = 2$$

Therefore, $\hat{\lambda} = \bar{X} = 2$ and the pmf of the Poisson distribution becomes as

$$P[X = x] = \frac{e^{-\lambda} \lambda^x}{x!}; \quad x = 0, 1, 2, \dots \quad \dots (1)$$

Now, to find the expected or theoretical number of accidents, we first find the probability of each class ($X = 0, 1, 2, \dots, 6$) by putting $X = 0, 1, \dots, 6$ in equation (1) respectively. **Table of Poisson probabilities** (refer to references) at a specified value of X and at various values of λ . Therefore, with the help of this table, we can also find these probabilities by taking $X = 0, 1, \dots, 6$ and $\lambda = 2$ as:

$$p_1 = P[X = 0] = 0.1353, p_2 = P[X = 1] = 0.2707, p_3 = P[X = 2] = 0.2707,$$

$$p_4 = P[X = 3] = 0.1804, p_5 = P[X = 4] = 0.0902, p_6 = P[X = 5] = 0.0361,$$

$$p_7 = P[X = 6 \text{ or more}] = 1 - P[X < 6] = 1 - [P[X = 0] + \dots + P[X = 5]]$$

$$= 1 - (0.1353 + 0.2707 + 0.2707 + 0.1804 + 0.0902 + 0.0361) = 0.0166$$

Thus, the expected number of accidents is obtained by multiplying the appropriate probability by the total number of accidents, that is, N . Therefore,

$$E_1 = Np_1 = 52 \times 0.1353 = 7.0356$$

Similarly,

$$E_2 = Np_2 = 14.0764, E_3 = Np_3 = 14.0764, E_4 = Np_4 = 9.3808,$$

$$E_5 = Np_5 = 4.6904, E_6 = Np_6 = 1.8772,$$

$$E_7 = N - (E_1 + E_2 + \dots + E_6)$$

$$= 52 - (7.0356 + 14.0764 + \dots + 1.8772) = 0.8632$$

Since the expected frequencies in the last three classes are < 5 , therefore, we combine the last three classes in order to realise a cell total of at least 5.

Calculations for $\frac{(O - E)^2}{E}$:

S.No.	Observed Frequency	Expected Frequency	(O - E)	(O - E) ²	$\frac{(O - E)^2}{E}$
1	10	7.0356	2.9644	8.7877	1.2490
2	12	14.0764	-2.0764	4.3114	0.3063
3	12	14.0764	-2.0764	4.3114	0.3063
4	9	9.3808	-0.3808	0.1450	0.0155
5	9	7.4308	1.5692	2.4624	0.3314
Total	52	52			2.2085

From the above calculations, we have

$$\chi^2 = \sum_{i=1}^5 \frac{(O_i - E_i)^2}{E_i} = 2.2085$$

Here, we combined the classes so

k = the number of classes that remain after the combining classes = 5

r = the number of parameters estimated from the sample data = 1
The critical value of χ^2 with $k - r - 1 = 5 - 1 - 1 = 3$ degrees of freedom at 5% level of significance is 7.81.

Since the calculated value of the test statistic (= 2.2085) is less than the critical value (= 7.81) so we do not reject the null hypothesis i.e. we support the claim at 5% level of significance.

Thus, we conclude that the sample fails to provide us sufficient evidence against the claim so the number of accidents follows the Poisson distribution.

Example 3: A random sample of 400 car batteries revealed the following distribution of battery life (in years):

Life	0-1	1-2	2-3	3-4	4-5	5-6
Frequency	16	90	145	112	27	10

Do these data follow a normal distribution at 1% level of significance?

Solution: Here, we want to test that the life of the car batteries follows the normal distribution. So our claim is “the life of the car batteries follows the normal distribution” and its complement is “the life of the car batteries does not follow the normal distribution”. So we can take the claim as the null hypothesis and complement as the alternative hypothesis. Thus,

H_0 : The life of the car batteries follows the normal distribution

H_1 : The life of the car batteries does not follow the normal distribution

Since the data are given in the categorical form and we are interested to fit a distribution, so we can go for the chi-square goodness of fit test.

Here, the parameters of the normal distribution are not given so for testing the null hypothesis, the test statistic is given by

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \sim \chi_{k-r-1}^2 \text{ under } H_0$$

where k = the number of classes

r = the number of parameters estimated from the sample data

If X represents the life of the car battery that follows normal distribution then the probability density function of the normal distribution is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}; -\infty < x < \infty, -\infty < \mu < \infty, \sigma > 0$$

There are two parameters (mean μ and standard deviation σ) in the normal distribution that can be estimated from the sample data as:

Life	Mid Value (X)	Frequency (f)	fX	fX ²
0-1	0.5	16	8.0	4.00
1-2	1.5	90	135.0	202.50
2-3	2.5	145	362.5	906.25
3-4	3.5	112	392.0	1372.00
4-5	4.5	27	121.5	546.75
5-6	5.5	10	55.0	302.50
Total		N = 400	1074	3334.00

The formula for calculating mean is

$$\hat{\mu} = \bar{X} = \frac{1}{N} \sum fX = \frac{1}{400} \times 1074 = 2.68$$

The formula for calculating standard deviation is

$$\begin{aligned} \hat{\sigma} = S &= \sqrt{\frac{1}{N-1} \left\{ \sum fX^2 - \frac{(\sum fX)^2}{N} \right\}} \\ &= \sqrt{\frac{1}{399} \left\{ 3334 - \frac{(1074)^2}{400} \right\}} = \sqrt{1.199} = 1.06 \end{aligned}$$

The next step is to find the expected frequency for each class under the assumption of a normally distributed population. The expected frequencies of the normal distribution can be obtained as follows:

First, we find the standard normal variate $Z = \frac{X - \hat{\mu}}{\hat{\sigma}}$ corresponding to each lower

limit of the class and then find $F(x) = P[X \leq x] = P[Z \leq z]$ for each value with the help of the **Table of Normal distribution** (Table II of the Appendix).

Thus, when $x = 0$, $z = \frac{x - \hat{\mu}}{\hat{\sigma}} = \frac{0 - 2.68}{1.06} = -2.53$ and

$$\begin{aligned} F(0) &= P[Z \leq z] = P[Z \leq -2.53] = 0.5 - P[-2.53 < Z < 0] \\ &= 0.5 - P[0 < Z < 2.53] = 0.5 - 0.4943 = 0.0057 \end{aligned}$$

Similarly, when $x = 3$, $z = \frac{3 - 2.68}{1.06} = 0.30$ and

$$\begin{aligned} F(3) &= P[Z \leq z] = P[Z \leq 0.30] = 0.5 + P[0 < Z < 0.30] \\ &= 0.5 + 0.1179 = 0.6179 \end{aligned}$$

Therefore, the expected frequencies are calculated as follows:

Life	Low er Limi t	Standard Normal Variate $Z = \frac{X - \hat{\mu}}{\hat{\sigma}}$	Area Under Normal Curve to the Left of Z i.e. $P[Z \leq z]$	Difference between Successive Areas	Expected Frequency = $400 \times \text{col.5}$
Below 0	$-\infty$	$-\infty$	0	$0.0057 - 0$ $= 0.0057$	2.28
0-1	0	-2.53	0.0057	$0.0571 -$ $0.0057 =$ 0.0514	20.56
1-2	1	-1.58	0.0571	$0.2611 -$ $0.0571 =$ 0.2040	81.60
2-3	2	-0.64	0.2611	$0.6179 -$ $0.2611 =$ 0.3568	142.72
3-4	3	0.30	0.6179	$0.8944 -$ $0.6179 =$ 0.2765	110.6

4-5	4	1.25	0.8944	$0.9857 - 0.8944 = 0.0913$	36.52
5-6	5	2.19	0.9706	$0.9991 - 0.9706 = 0.0134$	5.36
6 and above	6	3.13	0.9991	--	--

Here, the last cell expected frequency is obtained by the formula given by

$$E_7 = N - (E_1 + E_2 + \dots + E_6) = 11.76$$

Since the expected frequency in the first cell is < 5 , therefore, we combine the first two classes in order to realize a cell total of at least 5.

Calculations for $\frac{(O-E)^2}{E}$:

Life	Observed Frequency	Expected Frequency	(O-E)	(O-E) ²	$\frac{(O-E)^2}{E}$
0-1	16	22.84	-6.84	46.7856	2.0484
1-2	90	81.60	8.40	70.5600	0.8647
2-3	145	142.72	2.28	5.1984	0.0364
3-4	112	110.60	1.40	1.9600	0.0177
4-5	27	36.52	-9.52	90.6304	2.4817
5-6	10	5.36	4.64	21.5296	4.0167
Total	400	400			9.4656

Therefore, from the above calculations, we have

$$\chi^2 = \sum_{i=1}^6 \frac{(O_i - E_i)^2}{E_i} = 9.4656$$

Here, we combined the classes so

k = the number of classes that remain after the combining classes = 6

r = the number of parameters estimated from the sample data = 2

The critical value of χ^2 with $k - r - 1 = 6 - 2 - 1 = 3$ degrees of freedom at 1% level of significance is 11.34.

Since the calculated value of the test statistic (= 9.4656) is less than the critical value (= 11.34) so we do not reject the null hypothesis i.e. we support the claim at 1% level of significance.

Thus, we conclude that the sample fails to provide us sufficient evidence against the claim so we may assume that the life of car batteries follows the normal distribution.

Example 4: The following data represent the results of an investigation of the gender distribution of the children of 30 families containing 4 children each:

Number of Sons	0	1	2	3	4
Number of Families	4	8	8	7	3

Apply the chi-square test to see whether the number of sons in a family follows a binomial distribution with the probability that a child to be a son is 0.5 at 5% level of significance.

Solution: Here, we want to test whether the number of sons in a family follows a binomial distribution with $p = 0.5$. So our claim is “the number of sons in a

family follows the binomial distribution with $p = 0.5$ ” and its complement is “the number of sons in a family does not follow the binomial distribution with $p = 0.5$ ”. So we can take the claim as the null hypothesis and complement as the alternative hypothesis. Thus,

H_0 : The number of sons in a family follows the binomial distribution with $p = 0.5$

H_1 : The number of sons in a family does not follow the binomial distribution with $p = 0.5$

Since the data are given in the categorical form and we are interested to fit a distribution, so we can go for the chi-square goodness of fit test.

Here, the parameter p of the binomial distribution is given so for testing the null hypothesis, the test statistic is given by

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \sim \chi_{k-1}^2$$

where k = the number of classes

If X denotes the number of sons in a family that follows binomial distribution then the binomial probability mass function is given by

$$P[X = x] = {}^nC_x p^x (1-p)^{n-x}; \quad x = 0, 1, 2, \dots, n$$

Here, the total number of children is 4 and $p = 0.5$. Therefore, we have

$$\begin{aligned} P[X = x] &= {}^4C_x (0.5)^x (1-0.5)^{4-x}; \quad x = 0, 1, 2, 3, 4 \\ &= {}^4C_x (0.5)^x (0.5)^{4-x} = {}^4C_x (0.5)^4; \quad x = 0, 1, 2, 3, 4 \quad \dots(2) \end{aligned}$$

Now, to find the expected or theoretical number of families, we first find the probability of each class by putting $X = 0, 1, 2, 3, 4$ in equation (2) respectively. Thus,

$$\begin{aligned} p_1 &= P[X = 0] = {}^4C_0 (0.5)^4 = 0.0625 \\ p_2 &= P[X = 1] = {}^4C_1 (0.5)^4 = 4 \times 0.0625 = 0.2500 \\ p_3 &= P[X = 2] = {}^4C_2 (0.5)^4 = 6 \times 0.0625 = 0.3750 \\ p_4 &= P[X = 3] = {}^4C_3 (0.5)^4 = 4 \times 0.0625 = 0.2500 \\ p_5 &= P[X = 4] = {}^4C_4 (0.5)^4 = 1 \times 0.0625 = 0.0625 \end{aligned}$$

The expected number of families is obtained by multiplying the appropriate probability by the total number of families, that is, the sample size N . Therefore,

$$E_1 = Np_1 = 30 \times 0.0625 = 1.875$$

Similarly,

$$E_2 = Np_2 = 7.5, \quad E_3 = Np_3 = 11.25, \quad E_4 = Np_4 = 7.5,$$

$$E_5 = N - (E_1 + E_2 + E_3 + E_4) = 30 - (1.875 + 7.5 + 11.25 + 7.5) = 1.875$$

Since the expected frequency in the last cell is less than 5 therefore, we combine the last two classes in order to realize a cell total of at least 5.

Calculation for $\frac{(O - E)^2}{E}$:

S. No.	Observed Frequency	Expected Frequency	(O-E)	(O-E) ²	$\frac{(O-E)^2}{E}$
1	4	1.875	2.125	4.5156	2.4083
2	8	7.5	0.500	0.2500	0.0333
3	8	11.25	-3.250	10.5625	0.9389
4	10	9.375	0.625	0.3906	0.0417
Total	30	30			3.4222

From the above calculations, we have

$$\chi^2 = \sum_{i=1}^5 \frac{(O_i - E_i)^2}{E_i} = 3.4222$$

Here, we combined the classes so

k = the number of classes that remain after the combining classes = 4

The tabulated value of χ^2 with $k-1=4-1=3$ degrees of freedom and at 5% level of significance is 7.81.

Since the calculated value of the test statistic (= 3.4222) is less than the critical value (= 7.81) so we do not reject the null hypothesis i.e. we support the claim at 5% level of significance.

Thus, we conclude that the sample fails to provide us sufficient evidence against the claim so we may assume that the number of sons in a family follows binomial distribution with probability that a child to be a son is 0.5.

Check Your Progress 1

- 1) Write one difference between the chi-square test and Kolmogorov-Smirnov test for goodness of fit.

.....

.....

- 2) The following table gives the numbers of road accidents that occurred during the various days of the week:

Days	Mon	Tue	Wed	Thu	Fri	Sat	Sun
Number of Accidents	14	15	8	20	11	9	14

Test whether the accidents are uniformly distributed over the week by chi-square test at 1% level of significance.

.....

.....

.....

.....

.....

.....

.....

- 3) The number of customers waiting for service on the checkout counter line of a big supermarket is examined at random on 64 occasions during a period. The results are as follows:

Waiting Time (in minutes)	0 or 1	2	3	4	5	6	7	8 or more
Frequency	5	8	10	11	10	9	7	4

Does the number of customers waiting for service per occasion follows a Poisson distribution with mean 8 minutes at 5% level of significance?

.....

.....

.....

.....

12.3 CHI-SQUARE TEST FOR INDEPENDENCE OF ATTRIBUTES

There are many situations where we need to test the independence of two characteristics or attributes of categorical data. For example, a sociologist may wish to know whether the level of formal education is independent of income, whether the height of sons depending on the height of their fathers or not, etc. If there is no association between two variables, we say that they are independent. In other words, we can say that two variables are independent if the distribution of one is not depending on the distribution of another. To test the independence of two variables when observations in a population are classified according to some attributes we may use the chi-square test for independence. This test will indicate only whether or not any association exists between the attributes.

To conduct the test, a sample is drawn from the population and the observed frequencies are cross-classified according to the two characteristics so that each observation belongs to one and only one level of each characteristic. The cross-classification can be conveniently displayed by mean of a table called a contingency table. Therefore, a contingency table is an arrangement of data into a two-way classification. One of the classifications is entered in rows and the other in columns.

The chi-square test for independence of attributes can be used in the situation in which the data classified according to two attributes or characteristics.

A contingency table is an arrangement of data into a two-way classification. One of the classifications is entered in rows and the other in columns.

Assumptions

This test works under the following assumptions:

- The sample observations are random and independent.
- The observations may be classified into non-overlapping categories.
- The observed and expected frequencies of each class are greater than 5.
- The sum of observed frequencies is equal to the sum of expected frequencies, i.e., $\sum O = \sum E$.
- Each observation in the sample may be classified according to two characteristics so that each observation belongs to one and only one level of each characteristic.

Let us discuss the procedure of this test:

Step 1: As usual, the first step is to set up null and alternative hypotheses. Generally, we are interested to test whether two characteristics or attributes say, A and B are independent or not. So here we consider two-tailed case only. Thus, we can take the null and alternative hypotheses as

H_0 : The two characteristics of classification, say, A and B are independent

H_1 : They are not independent

Step 2: After setting the null and alternative hypotheses, the next step is to draw a random sample. So, let a sample of N observations is drawn from a population and they are cross-classified according to two characteristics, say, A and B. Also let characteristic A be assumed to have 'r' categories $A_1, A_2, \dots, A_r$ and characteristic B be assumed to have 'c' categories $B_1, B_2, \dots, B_c$. The various observed frequencies in different classes can be expressed in the form of a table known as a contingency table.

B A	B₁	B₂	... B_j ...	B_c	Total
A₁	O ₁₁	O ₁₂	... O _{1j} ...	O _{1c}	R₁
A₂	O ₂₁	O ₂₂	... O _{2j} ...	O _{2c}	R₂
·	·	·	·	·	·
·	·	·	·	·	·
·	·	·	·	·	·
A_i	O _{i1}	O _{i2}	... O _{ij} ...	O _{ic}	R_i
·	·	·	·	·	·
·	·	·	·	·	·
·	·	·	·	·	·
A_r	O _{r1}	O _{r2}	... O _{rj} ...	O _{rc}	R_r
Total	C₁	C₂	... C_j ...	C_c	N

Here, O_{ij} represents the number of observations corresponding to i^{th} level of characteristic A and j^{th} level of characteristic B. R_i and C_j represent the sum of the number of observations corresponding to i^{th} level of characteristic A, i.e. i^{th} row and j^{th} level of characteristic B, i.e. j^{th} column respectively.

Step 3: The chi-square test of independence of attributes compares observed frequencies with frequencies that are expected when H_0 is true. To calculate the expected frequencies, we first find the probabilities that observation lies in a particular cell for all $i=1, 2, \dots, r$ and $j=1, 2, \dots, c$. These probabilities are calculated according to the multiplication law of probability. According to this law, if two events are independent then the probability of their joint occurrence is equal to the product of their individual probabilities. Therefore, the probability that an observation falls in cell (i, j) is equal to the probability that observation falls in the i^{th} row multiplied by the probability of falling in the j^{th} column. We estimate these probabilities from the sample data by R_i/N and C_j/N , respectively. Thus,

$$P[A_i B_j] = \frac{R_i}{N} \cdot \frac{C_j}{N}$$

Step 4: To obtain the expected frequency E_{ij} for (i, j) cells, we multiply this estimated probability by the total sample size. Thus,

$$E_{ij} = N \cdot \frac{R_i}{N} \cdot \frac{C_j}{N}$$

This equation reduces to

$$E_{ij} = \frac{R_i \times C_j}{N} = \frac{\text{Sum of } i^{\text{th}} \text{ row} \times \text{Sum of } j^{\text{th}} \text{ column}}{\text{Total sample size}}$$

This form of the equation indicates that we can easily compute an expected cell frequency for each cell by multiplying together the appropriate row and column totals and dividing the product by the total sample size.

Step 5: Test statistic:

Since this test also compares the observed cell frequencies with the corresponding expected cell frequencies, therefore, we are interested in the magnitudes of the differences between the observed and the expected cell frequencies. Specifically, we wish to know whether the differences are small enough to be attributed to chance or they are large due to some other factors. With the help of observed and expected frequencies, we may compute a test statistic that reflects the magnitudes of differences between these two quantities. When H_0 is true, the test statistic is

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi_{(r-1)(c-1)}^2 \text{ under } H_0$$

The test statistic follows as a chi-square distribution with $(r-1)(c-1)$ degrees of freedom.

Step 6: Obtain the critical value of the test statistic corresponding to $df = (r-1)(c-1)$ at a given level of significance under the condition that the null hypothesis is true. **Table of chi-square** (refer to Table VI in Appendix) provides critical values of the test statistic χ^2 for various df and different level of significance.

Step 7: Take the decision about the null hypothesis as:

To take the decision about the null hypothesis, the test statistic (calculated in Step 5) is compared with the chi-square critical (tabulated) value (observed in Step 6) for a given level of significance (α) under the condition that the null hypothesis is true.

If the calculated value of the test statistic is greater than or equal to the tabulated value with $(r-1)(c-1)$ degrees of freedom at α level of significance, we reject the null hypothesis at α level of significance otherwise we do not reject it.

Let us do some examples to become more user friendly with this test.

Example 5: 1000 students at the college level were graded according to their IQ level and the economic condition of their parents.

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

Test that the IQ level of students is independent of the economic condition of their parents at 5% level of significance.

Solution: Here, we want to test that the IQ level of students is independent of the economic condition of their parents. So our claim is “the IQ level and the

economic condition are independent” and its complement is “the IQ level and the economic condition are not independent”. So we can take the claim as the null hypothesis and complement as the alternative hypothesis. Thus,

H_0 : The IQ level and economic condition are independent

H_1 : The IQ level and economic condition are not independent

Here, we are interested to test the independence of two characteristics IQ and economic condition, so we go for the chi-square test of independence.

For testing the null hypothesis, the test statistic is

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi_{(r-1)(c-1)}^2 \text{ under } H_0$$

Now, under H_0 , the expected frequencies can be obtained as:

$$\begin{aligned} E_{ij} &= \text{Expected frequency of the } i^{\text{th}} \text{ row and } j^{\text{th}} \text{ column} \\ &= \frac{R_i \times C_j}{N} = \frac{\text{Sum of } i^{\text{th}} \text{ row} \times \text{Sum of } j^{\text{th}} \text{ column}}{\text{Total sample size}} \end{aligned}$$

Therefore,

$$\begin{aligned} E_{11} &= \frac{R_1 \times C_1}{N} = \frac{400 \times 700}{1000} = 280, E_{12} = \frac{R_1 \times C_2}{N} = \frac{400 \times 300}{1000} = 120 \\ E_{21} &= \frac{R_2 \times C_1}{N} = \frac{600 \times 700}{1000} = 420, E_{22} = \frac{R_2 \times C_2}{N} = \frac{600 \times 300}{1000} = 180 \end{aligned}$$

Calculations for $\frac{(O - E)^2}{E}$:

Observed Frequency (O)	Expected Frequency (E)	(O - E)	(O - E) ²	$\frac{(O - E)^2}{E}$
240	280	-40	1600	5.71
160	120	40	1600	13.33
460	420	40	1600	3.81
140	180	-40	1600	8.89
Total = 1000	1000			31.74

Therefore, from the above calculations, we have

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 31.74$$

The degrees of freedom will be $(r - 1)(c - 1) = (2 - 1)(2 - 1) = 1$.

The critical value of χ^2 with 1 degree of freedom at 5% level of significance is 3.84.

Since the calculated value of the test statistic (= 31.74) is greater than the critical value (= 3.84) so we reject the null hypothesis i.e. we reject the claim at 5% level of significance.

Thus, we conclude that the sample provides us sufficient evidence against the claim so the IQ level of students is not independent of the economic condition of their parents.

Example 6: Calculate the expected frequencies for the following data presuming the two attributes and check that condition of home and condition of the child are independent at 5% level of significance.

Condition of Child	Condition of Home	
	Clean	Dirty
Clear	70	50
Fairly Clean	80	20
Dirty	35	45

Solution: Here, we want to test that the condition of home and condition of child are independent. So our claim is “the condition of home and condition of child are independent” and its complement is “condition of home and condition of child are not independent”. So we can take the claim as the null hypothesis and complement as the alternative hypothesis. Thus,

H_0 : The condition of home and condition of child are independent

H_1 : The condition of home and condition of child are not independent

Here, we are interested to test the independence of two characteristics, condition of home and condition of the child, so we go for the chi-square test of independence.

For testing the null hypothesis, the test statistic is

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi_{(r-1)(c-1)}^2 \text{ under } H_0$$

Now, under H_0 , the expected frequencies can be obtained as:

Condition of Child	Condition of Home		Total
	Clean	Dirty	
Clear	70	50	120
Fairly Clean	80	20	100
Dirty	35	45	80
Total	185	115	300

E_{ij} = Expected frequency of the i^{th} row and j^{th} column

$$= \frac{R_i \times C_j}{N} = \frac{\text{Sum of } i^{\text{th}} \text{ row} \times \text{Sum of } j^{\text{th}} \text{ column}}{\text{Total sample size}}$$

Therefore,

$$E_{11} = \frac{R_1 \times C_1}{N} = \frac{120 \times 185}{300} = 74; \quad E_{12} = \frac{R_1 \times C_2}{N} = \frac{120 \times 115}{300} = 46;$$

$$E_{21} = \frac{R_2 \times C_1}{N} = \frac{100 \times 185}{300} = 61.67; \quad E_{22} = \frac{R_2 \times C_2}{N} = \frac{100 \times 115}{300} = 38.33;$$

$$E_{31} = \frac{R_3 \times C_1}{N} = \frac{80 \times 185}{300} = 49.33; \quad E_{32} = \frac{R_3 \times C_2}{N} = \frac{80 \times 115}{300} = 30.67$$

Calculations for $\frac{(O - E)^2}{E}$:

Observed Frequency (O)	Expected Frequency (E)	(O - E)	(O - E) ²	$\frac{(O - E)^2}{E}$
70	74.00	-4.00	16.00	0.22
50	46.00	4.00	16.00	0.35

80	61.67	18.33	335.99	5.45
20	38.33	-18.33	335.99	8.77
35	49.33	-14.33	205.35	4.16
45	30.67	14.33	205.35	6.70
Total = 300	300			25.65

Therefore, from the above calculations, we have

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 25.65$$

The degrees of freedom will be $(r-1)(c-1) = (3-1)(2-1) = 2$.

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

Since the calculated value of the test statistic (= 25.65) is greater than the critical value (= 5.99) so we reject the null hypothesis i.e. we reject the claim at 5% level of significance.

Thus, we conclude that the sample provides us sufficient evidence against the claim so the condition of home and the condition of the child are not independent.

Check Your Progress 2

- 1) 1500 families were selected at random in a city to test the belief that high-income families usually send their children to public schools and low-income families often send their children to government schools. The following results were obtained in the study conducted.

Income	Public School	Government School	Total
Low	300	600	900
High	435	165	600
Total	735	765	1500

Use the chi-square test at 1% level of significance to test whether the two attributes are independent.

.....

.....

.....

- 2) The following contingency table presents the analysis of 300 persons according to hair colour and eye colour:

Hair Colour	Eye Colour			Total
	Blue	Grey	Brown	
Fair	30	10	40	80
Brown	40	20	40	100
Black	50	30	40	120
Total	120	60	120	300

Test the hypothesis that there is an association between hair colour and eye colour at 1% level of significance.

.....

.....

.....

12.4 KOLMOGOROV-SMIRNOV GOODNESS OF FIT TEST

This test has its name on the names of its discoverers A. N. Kolmogorov and N.V. Smirnov. It is a simple non-parametric test for testing whether data follow a specified or assumed distribution or sample has come from a specified or assumed distribution or there is a significant difference between an observed distribution and a theoretical distribution. Therefore, it is a measure of goodness of fit to a theoretical distribution. The main difference between the chi-square test and Kolmogorov-Smirnov (K-S) test is that the chi-square test is designed for categorical data whereas the K-S test is designed for continuous data.

Assumptions

This test works under the following assumptions:

- (i) The sample is randomly selected from some unknown distribution.
- (ii) The observations are independent.
- (iii) The variable under study is continuous.
- (iv) The variable under study is measured on at least ordinal scale.

Let $X_1, X_2, \dots, X_n$ be a random sample from a population with unknown continuous distribution function $F(x)$. Generally, we are interested to test whether data follow a specified distribution $F_0(x)$ or a sample has come from a specified or assumed distribution or not. So here we consider the only two-tailed case. Thus, we can take the null and alternative hypotheses as

H_0 : Data follow a specified distribution

H_1 : Data do not follow a specified distribution

In symbolical form

$$H_0 : F(x) = F_0(x) \quad \text{for all values of } x$$

$$H_1 : F(x) \neq F_0(x) \quad \text{for at least one value of } x \quad [\text{two-tailed test}]$$

We summarise the Kolmogorov-Smirnov goodness of fit test in the following steps:

Step 1: The K-S goodness of fit test is based on the comparison of the empirical (observed or sample) cumulative distribution function with the theoretical (specified or population) cumulative distribution function. Therefore, first of all, we compute the empirical cumulative distribution function, say, $S(x)$ from the sample data which is defined as the proportion of sample observations less than or equal to some value x , that is,

$S(x)$ = proportion of sample observations less than or equal to some value x

$$S(x) = \frac{\text{The number of sample observation less than or equal to } x}{\text{Total number of observations}}$$

Step 2: In this step, we find the theoretical cumulative distribution function $F_0(x)$ for all possible values of x on the basis of specified distribution.

Step 3: After finding the empirical and theoretical cumulative distribution functions for all possible values of x , we find the deviation between the empirical and theoretical cumulative distribution functions for all x . That is,

$$S(x) - F_0(x) \quad \text{for all } x$$

Step 4: Since $S(x)$ is the statistical image of the population distribution function $F(x)$ so if the null hypothesis is true then the difference between $S(x)$ and $F(x)$ should be small for all values of x . Thus, in all deviations obtained in Step 3, we find the point(s) at which the two functions show the maximum deviation.

The test statistic is denoted by D_n and it is the greatest vertical deviation between $S(x)$ and $F_0(x)$, that is,

$$D_n = \sup_x |S(x) - F_0(x)|$$

which is read as “ D_n equals the supreme overall x , of the absolute value of the difference $S(x) - F_0(x)$ ”

Step 5: Obtain the critical value of the test statistic at α % level of significance under the condition that the null hypothesis is true. You may refer to the Appendix provided in the references for the critical values at the α level of significance.

Step 6: Decision Rule:

To take the decision about the null hypothesis, the test statistic (calculated in Step 4) is compared with the critical (tabulated) value (obtained in Step 5) for a given level of significance (α).

If the computed value of D_n is greater than or equal to the critical value $D_{n,\alpha}$ at α level of significance, that is, if $D_n \geq D_{n,\alpha}$ then we reject the null hypothesis at α level of significance, otherwise we do not reject H_0 .

For large sample ($n > 40$):

For a sample size n greater than 40, the critical value of the test statistic at a given α level of significance is approximated by the formula given in the Appendix given in the references. As the critical value for the two-tailed test at 5% level of significance is calculated by the formula given by

$$D_{n,0.05} = \frac{1.36}{\sqrt{n}}$$

where n is the sample size.

Let us do some examples to become more user friendly with the calculations explained above:

Example 7: A coin is tossed 400 times in sets of four. The frequencies of getting 0 to 4 tails are:

Number of Tails	0	1	2	3	4	Total
Frequency	23	95	164	92	26	400

Examine the fairness of coin using the K-S test at 5% level of significance.

Solution: Here, we want to test that the coin is fair. We know that if a coin is fair then the number of tails follows a binomial distribution with parameter n and $p = 1/2$. So we can test that the number of tails follows the binomial distribution. If $F_0(x)$ denotes the cumulative distribution function of binomial distribution with parameter $n = 4$ and $p = 1/2$ then our claim is $F(x) = F_0(x)$ and its complement is $F(x) \neq F_0(x)$. Since the claim contains the equality sign so we can take the claim as the null hypothesis and the complement as the alternative hypothesis. Thus

$$H_0 : F(x) = F_0(x) \text{ for all values of } x$$

$$H_1 : F(x) \neq F_0(x) \text{ for at least one value of } x$$

For testing the null hypothesis, the test statistic is given by

$$D_n = \sup_x |S(x) - F_0(x)|$$

where $S(x)$ is the empirical distribution function based on the observed sample.

For obtaining the empirical cumulative distribution function $S(x)$, first, we obtain the cumulative frequencies and then by the definition of empirical distribution function we obtain $S(x)$ as

$$S(x) = \frac{\text{The number of sample observations } \leq x}{\text{Total number of observations}}$$

Therefore, at $x = 0, 1, 2, 3, 4$, we have

$$S(0) = \frac{23}{400} = 0.0575, S(1) = \frac{118}{400} = 0.2950, S(2) = \frac{282}{400} = 0.705,$$

$$S(3) = \frac{374}{400} = 0.935, S(4) = \frac{400}{400} = 1$$

The theoretical distribution can be obtained by the definition of cumulative distribution function as

$$F_0(x) = P[X \leq x]$$

In this case, our theoretical distribution is the binomial distribution with parameter $n = 4$ and $p = 1/2$. Therefore,

$$\begin{aligned} \text{i.e. } F_0(x) &= P[X \leq x] = {}^4C_x \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{4-x} \\ &= {}^4C_x \left(\frac{1}{2}\right)^4, x = 0, 1, 2, 3, 4 \end{aligned}$$

With the help of **table in the reference**, we can find the values of $F_0(x)$ at different values of x .

Thus, at $x = 0$, we have

$$F_0(0) = P[X \leq 0] = 0.0625$$

Similarly,

$$\text{At } x = 1, F_0(1) = P[X \leq 1] = 0.3125$$

$$\text{At } x = 2, F_0(2) = P[X \leq 2] = 0.6875$$

$$\text{At } x = 3, F_0(3) = P[X \leq 3] = 0.9375$$

$$\text{At } x = 4, F_0(4) = P[X \leq 4] = 1$$

Calculation for $|S(x) - F_0(x)|$:

Number of Tails (x)	Observed Frequency	Cumulative Observed Frequency	Observed Distribution Function $S(x) = (3)/n$	Theoretical Distribution Function	Absolute Difference $ S(x) - F_0(x) $
(1)	(2)	(3)	(4)	(5)	(6)
0	23	23	0.0575	0.0625	0.0050
1	95	118	0.2950	0.3125	0.0175
2	164	282	0.7050	0.6875	0.0175
3	92	374	0.9350	0.9375	0.0025
4	26	400	1	1	0

From the above calculation, we have

$$D_n = \sup_{0 \leq x \leq 4} |S(x) - F_0(x)| = 0.0175$$

Here, $n = 400 > 40$, therefore, the critical value of the test statistic can be obtained by the formula given by

$$\begin{aligned} D_{n,0.05} &= \frac{1.36}{\sqrt{n}} \\ &= \frac{1.36}{\sqrt{400}} = 0.068 \end{aligned}$$

Since the calculated value of the test statistic $D_n (= 0.0175)$ is less than the critical value $D_{n,0.05} (= 0.068)$ so we do not reject the null hypothesis i.e. we support the claim at 5% level of significance.

Thus, we conclude that the sample fails to provide us sufficient evidence against the claim so we may assume that number of tails follows the binomial distribution or the coin is fair or unbiased.

Example 8: The following data were obtained from a table of random numbers of a normal distribution with mean 3 and standard deviation 1:

2.1, 1.9, 3.2, 2.8, 1.0, 5.1, 4.2, 3.6, 3.9, 2.7

Test by the K-S test that the data have come to the same normal distribution at 5% level of significance.

Solution: Here, we want to test that the data have come from the normal distribution with mean 3 and variance 1. If $F_0(x)$ is the cumulative distribution function of $N(3,1)$ then our claim is $F(x) = F_0(x)$ and its complement is $F(x) \neq F_0(x)$. Since the claim contains the equality sign so we can take the claim as the null hypothesis and the complement as the alternative hypothesis. Thus

$$H_0 : F(x) = F_0(x) \text{ for all values of } x$$

$$H_1 : F(x) \neq F_0(x) \text{ for at least one value of } x$$

The Kolmogorov-Smirnov test for goodness of fit is applied only when all the parameters of the fitting distribution are known. If all parameter (s) of the fitted distribution is (are) not known then we use chi-square goodness of fit test.

For testing the null hypothesis, the test statistic is given by

$$D_n = \sup_x |S(x) - F_0(x)|$$

where $S(x)$ is the empirical distribution function based on the observed sample.

For obtaining the empirical cumulative distribution function $S(x)$, first, we obtain the cumulative frequencies and then by the definition of empirical distribution function we obtain $S(x)$ as

$$S(x) = \frac{\text{The number of sample observations} \leq x}{\text{Total number of observations}}$$

Here, the theoretical cumulative distribution function $F_0(x)$ can be obtained with the help of the method as described in Unit 4 of MST-003. Therefore, by the definition of the distribution function, we have

$$F_0(x) = P[X \leq x] = P\left[\frac{X - \mu}{\sigma} \leq \frac{x - \mu}{\sigma}\right] = P[Z \leq z]$$

where $P[Z \leq z]$ it can be obtained with the help of the Table II in the Appendix.

Thus, when $x = 1.0$, $z = \frac{x - \mu}{\sigma} = \frac{1.0 - 3}{1} = -2.0$ and

$$\begin{aligned} F_0(1.0) &= P[Z \leq z] = P[Z \leq -2.0] = 0.5 - P[-2.0 < Z < 0] \\ &= 0.5 - P[0 < Z < -2.0] \\ &= 0.5 - 0.4772 = 0.0228 \end{aligned}$$

Similarly, when $x = 3.2$, $z = \frac{3.2 - 3}{1} = 0.2$ and

$$\begin{aligned} F_0(3.2) &= P[Z \leq z] = P[Z \leq 0.2] \\ &= 0.5 + P[0 < Z \leq 0.2] \\ &= 0.5 + 0.0793 = 0.5793 \text{ and so on.} \end{aligned}$$

Calculation for $|S(x) - F_0(x)|$:

X	Observed Frequency	Cumulative Observed Frequency	Observed Distribution Function $S(x)$	$z = \frac{X - \mu}{\sigma} = \frac{X - 3}{1}$	Theoretical Distribution Function $F_0(x)$	Absolute Difference $ S(x) - F_0(x) $
1.0	1	1	$1/10 = 0.1$	-2.0	0.0228	0.0772
1.9	1	2	$2/10 = 0.2$	-1.1	0.1357	0.0643
2.1	1	3	$3/10 = 0.3$	-0.9	0.1841	0.1159
2.7	1	4	$4/10 = 0.4$	-0.3	0.3821	0.0179
2.8	1	5	$5/10 = 0.5$	-0.2	0.4207	0.0793

3.2	1	6	6/10 = 0.6	0.2	0.5793	0.0207
3.6	1	7	7/10 = 0.7	0.6	0.7257	0.0257
3.9	1	8	8/10 = 0.8	0.9	0.8159	0.0159
4.2	1	9	9/10 = 0.9	1.2	0.8849	0.0151
5.1	1	10	10/10 = 1.0	2.1	0.9821	0.0179

From the above calculations, we have

$$D_n = \sup_x |S(x) - F_0(x)| = 0.1159$$

The critical value of the test statistic at 5% level of significance is 0.409.

Since the calculated value of the test statistic $D_n (= 0.1159)$ is less than the critical value ($= 0.409$) so we do not reject the null hypothesis i.e. we support the claim at 5% level of significance.

Thus, we conclude that the sample fails to provide us sufficient evidence against the claim so we may assume that the data have come from the normal distribution with mean 3 and standard deviation 1.

Check Your Progress 3

- 1) Write the main difference between the chi-square test and Kolmogorov-Smirnov (K-S) test for goodness of fit.

.....

.....

.....

- 2) The following data were obtained from a table of random numbers of a normal distribution with mean 5.6 and standard deviation 1.2:

6.0	6.6	5.9	5.0	6.2	4.8
5.5	6.3	5.9	5.8	6.6	6.2

Test the hypothesis by the K-S test that the data have come from the same normal distribution at 5% level of significance.

.....

.....

.....

12.5 SUMMARY

In this unit, we have discussed the following points:

1. When the data are classified into different categories or groups according to one or more attributes such type of data is known as categorical data.
2. Needs of tests which are applied for categorical data.
3. The chi-square test for goodness of fit.
4. A contingency table is an arrangement of data into a two-way classification. One of the classifications is entered in rows and the other in columns.
5. The chi-square test for independence of two attributes.
6. The Kolmogorov-Smirnov test for goodness of fit.

12.6 SOLUTIONS / ANSWERS

Check Your Progress 1

- 1) The main difference between the chi-square test and Kolmogorov-Smirnov (K-S) test for goodness of fit is that the chi-square test is designed for categorical data whereas the K-S test is designed for continuous data.
- 2) Here, we want to test that the road accidents are uniformly distributed over the week. So our claim is “the accidents are uniformly distributed over the week” and its complement is “the accidents are not uniformly distributed over the week”. So we can take the claim as the null hypothesis and the complement as the alternative hypothesis. Thus,

H_0 : The accidents are uniformly distributed over the week

H_1 : The accidents are not uniformly distributed over the week

Since the data are given in the categorical form and we are interested to fit a distribution, so we can go for the chi-square goodness of fit test.

If X denotes the number of accidents per day that follows uniform distribution then the probability mass function of the uniform distribution is given by

$$P[X = x] = \frac{1}{N}; \quad x = 0, 1, 2, \dots, N$$

The uniform distribution has a parameter N which is given so for testing the null hypothesis, the test statistic is given by

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{(k-1)}$$

Since the uniform distribution is one in which all outcomes considered have equal or uniform probability. Therefore, the probability that the accident occurs in any day is same. Thus,

$$p_1 = p_2 = p_3 = p_4 = p_5 = p = \frac{1}{7}$$

The theoretical or expected frequency for each day is obtained by multiplying the appropriate probability by the total number of accidents, that is, sample size N . Therefore,

$$E_1 = E_2 = E_3 = E_4 = E_5 = Np = 91 \times \frac{1}{7} = 13$$

Calculations for $\frac{(O - E)^2}{E}$:

Days	Observed Frequency (O)	Expected Frequency (E)	(O-E)	(O-E) ²	$\frac{(O-E)^2}{E}$
Mon	14	13	1	1	0.0769
Tue	15	13	2	4	0.3077
Wed	8	13	-5	25	1.9231
Thu	20	13	7	49	3.7692
Fri	11	13	-2	4	0.3077

Sat	9	13	-4	16	1.2308
Sun	14	13	1	1	0.0769
Total	91	91			7.6923

From the above calculations, we have

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} = 7.6923$$

The critical value of the chi-square with $k-1=7-1=6$ degrees of freedom at 1% level of significance is 16.81.

Since the calculated value of the test statistic ($= 7.6923$) is less than the critical value ($= 16.81$) so we do not reject the null hypothesis i.e. we support the claim at 1% level of significance.

Thus, we conclude that the sample fails to provide us sufficient evidence against the claim so we may assume that the accidents are uniformly distributed over the week.

- 3) Here, we want to test that the number of costumers waiting for servicing follows a Poisson distribution with mean waiting time 8 minutes. So our claim is “the number of costumers waiting for servicing follow a Poisson distribution with mean waiting time 8 minutes” and its complement is “the number of costumers waiting for servicing does not follow a Poisson distribution with mean waiting time 8 minutes”. So we can take the claim as the null hypothesis and complement as the alternative hypothesis. Thus,

H_0 : The number of customers waiting for servicing follows the Poisson distribution with mean waiting time 8 minutes

H_1 : The number of customers waiting for servicing does not follow the Poisson distribution with mean waiting time 8 minutes

Since the data are given in the categorical form and we are interested to fit a distribution, so we can go for the chi-square goodness of fit test.

Here, the parameter λ of Poisson distribution is given so for testing the null hypothesis, the test statistic is given by

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{(k-1)}$$

where k = the number of classes

If X denotes the number of customers waiting for servicing per occasion that follows the Poisson distribution then the Poisson probability mass function is given by

$$P[X = x] = \frac{e^{-\lambda} \lambda^x}{x!}; \quad x = 0, 1, 2, \dots$$

where λ is the mean number of customers waiting for servicing per occasion. Since it is specified 8 minutes in the null hypothesis, therefore, the Poisson probability mass function is,

$$P[X = x] = \frac{e^{-8} 8^x}{x!}; \quad x = 0, 1, 2, \dots \quad \dots (3)$$

Now, to find the expected number of customers waiting for servicing per occasion, we first find the probability of each class by putting $X = 0, 1, \dots$ in equation (3). With the help of **Table of Poisson distribution (refer the table from References)**, we can also find these probabilities by taking $X = 0, 1, \dots$ and $\lambda = 8$ as:

$$\begin{aligned} p_1 &= P[X = 0 \text{ or } 1] = P[X = 0] + P[X = 1] = 0.0003 + 0.0027 = 0.0030, \\ p_2 &= P[X = 2] = 0.0107, \quad p_3 = P[X = 3] = 0.0286, \quad p_4 = P[X = 4] = 0.0573, \\ p_5 &= P[X = 5] = 0.0916, \quad p_6 = P[X = 6] = 0.1221, \quad p_7 = P[X = 7] = 0.1396, \\ p_8 &= P[X = 8 \text{ or more}] = 1 - P[X < 8] = 1 - [P[X = 0] + \dots + P[X = 7]] \\ &= 1 - (0.0003 + 0.0027 + 0.0107 + 0.0286 + 0.0573 + 0.0916 + 0.1221 + 0.1396) = 0.5471 \end{aligned}$$

The expected number of costumers waiting for servicing is obtained by multiplying the appropriate probability by total number of accidents, that is, N . Therefore,

$$\begin{aligned} E_1 &= Np_1 = 64 \times 0.0030 = 0.1920 \\ E_2 &= Np_2 = 0.6848, \quad E_3 = Np_3 = 1.8304, \quad E_4 = Np_4 = 3.6672, \\ E_5 &= Np_5 = 5.8624, \quad E_6 = Np_6 = 7.8144, \quad E_7 = Np_7 = 8.9344, \\ E_8 &= N - (E_1 + E_2 + \dots + E_7) = 35.0144 \end{aligned}$$

Since the expected frequencies in the first four classes are less than 5 therefore, we combine the first four classes in order to realize a cell total of at least 5.

Calculations for $\frac{(O - E)^2}{E}$:

S. No.	Observed Frequency	Expected Frequency	(O-E)	(O-E) ²	$\frac{(O-E)^2}{E}$
1	34	6.3744	27.6256	763.1738	119.7248
2	10	5.8624	4.1376	17.1197	2.9203
3	9	7.8144	1.1856	1.4056	0.1799
4	7	8.9344	1.9344	3.7419	0.4188
5	4	35.0144	-31.0144	961.8930	27.4714
	64	64			150.7151

From the above calculations, we have

$$\chi^2 = \sum_{i=1}^5 \frac{(O_i - E_i)^2}{E_i} = 150.7151$$

Here, we combined the classes so

k = the number of classes that remain after the combining classes = 5

The critical value of the chi-square with $k-1 = 5-1 = 4$ degrees of freedom at 5% level of significance is 9.49.

Since the calculated value of the test statistic (= 150.7151) is greater than the critical value (= 9.49) so we reject the null hypothesis i.e. we reject the claim at 5% level of significance.

Here, the value of parameter λ is given so we take $r = 0$

Thus, we conclude that the sample provides us sufficient evidence against the claim so the number of customers waiting for servicing per occasion does not follow the Poisson distribution.

Check Your Progress 2

- 1) Here, we want to test that the family income and selection of school are independent. So our claim is “the family income and selection of school are independent” and its complement is “the family income and selection of school are not independent”. So we can take the claim as the null hypothesis and the complement as the alternative hypothesis. Thus,

H_0 : Family income and selection of school are independent

H_1 : Family income and selection of school are not independent

Here, we are interested to test the independence of two attributes family income and selection of school, so we go for the chi-square test of independence.

For testing the null hypothesis, the test statistic is

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi_{(r-1)(c-1)}^2 \text{ under } H_0$$

Now, under H_0 , the expected frequencies can be obtained as:

$$\begin{aligned} E_{ij} &= \text{Expected frequency of the } i^{\text{th}} \text{ row and } j^{\text{th}} \text{ column} \\ &= \frac{R_i \times C_j}{N} = \frac{\text{Sum of } i^{\text{th}} \text{ row} \times \text{Sum of } j^{\text{th}} \text{ column}}{\text{Total sample size}} \end{aligned}$$

Therefore,

$$E_{11} = \frac{R_1 \times C_1}{N} = \frac{900 \times 735}{1500} = 441, E_{12} = \frac{R_1 \times C_2}{N} = \frac{900 \times 765}{1500} = 459,$$

$$E_{21} = \frac{R_2 \times C_1}{N} = \frac{600 \times 735}{1500} = 294, E_{22} = \frac{R_2 \times C_2}{N} = \frac{600 \times 765}{1500} = 306$$

Calculations for $\frac{(O - E)^2}{E}$:

Observed Frequency (O)	Expected Frequency (E)	(O - E)	(O - E) ²	$\frac{(O - E)^2}{E}$
300	441	-141	19881	45.08
600	459	141	19881	43.31
435	294	141	19881	67.62
165	306	-141	19881	64.97
Total = 1500	1500			220.98

Therefore, from the above calculations, we have

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 220.98$$

The degrees of freedom will be $(r-1)(c-1) = (2-1)(2-1) = 1$.

The critical value of the chi-square with 1 df at 1% level of significance is 6.63.

Since the calculated value of the test statistic (= 220.98) is greater than the critical value (= 6.63) so we reject the null hypothesis i.e. we reject the claim at 1% level of significance.

Thus, we conclude that the sample provides us sufficient evidence against the claim so two attributes family income and selection of school are not independent.

- 2) Here, we want to test that there is an association between hair colour and eye colour that means we want to test that hair colour and eye colour are not independent. So our claim is “hair colour and eye colour are not independent” and its complement is “hair colour and eye colour are independent”. So we can take complement as the null hypothesis and the claim as the alternative hypothesis. Thus,

H_0 : Hair and eye colour are independent

H_1 : Hair and eye colour are associated

Here, we are interested to test the independence of two attributes hair colour and eye colour, so we go for the chi-square test of independence.

For testing the null hypothesis, the test statistic is

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{(r-1)(c-1)} \text{ under } H_0$$

Now, under H_0 , the expected frequencies can be obtained as:

$$E_{ij} = \text{Expected frequency of the } i^{\text{th}} \text{ row and } j^{\text{th}} \text{ column} \\ = \frac{R_i \times C_j}{N} = \frac{\text{Sum of } i^{\text{th}} \text{ row} \times \text{Sum of } j^{\text{th}} \text{ column}}{\text{Total sample size}}$$

Therefore,

$$E_{11} = \frac{R_1 \times C_1}{N} = \frac{80 \times 120}{300} = 32; E_{12} = \frac{R_1 \times C_2}{N} = \frac{80 \times 60}{300} = 16; \\ E_{13} = \frac{R_1 \times C_3}{N} = \frac{80 \times 120}{300} = 32; E_{21} = \frac{R_2 \times C_1}{N} = \frac{100 \times 120}{300} = 40; \\ E_{22} = \frac{R_2 \times C_2}{N} = \frac{100 \times 60}{300} = 20; E_{23} = \frac{R_2 \times C_3}{N} = \frac{100 \times 120}{300} = 40; \\ E_{31} = \frac{R_3 \times C_1}{N} = \frac{120 \times 120}{300} = 48; E_{32} = \frac{R_3 \times C_2}{N} = \frac{120 \times 60}{300} = 24; \\ E_{33} = \frac{R_3 \times C_3}{N} = \frac{120 \times 120}{300} = 48$$

Calculations for $\frac{(O - E)^2}{E}$:

Observed Frequency (O)	Expected Frequency (E)	(O - E)	(O - E) ²	$\frac{(O - E)^2}{E}$
30	32	-2	4	0.13
10	16	-6	36	2.25
40	32	8	64	2
40	40	0	0	0
20	20	0	0	0

40	40	0	0	0
50	48	2	4	0.08
30	24	6	36	1.50
40	48	-8	64	1.33
Total = 300	300			7.29

Therefore, from the above calculations, we have

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 7.29$$

The degrees of freedom will be $(r-1)(c-1) = (3-1)(3-1) = 4$.

The critical value of the chi-square with 4 df at 1% level of significance is 13.28.

Since the calculated value of the test statistic chi-square (= 7.29) is less than the critical value (= 13.28) so we do not reject the null hypothesis and reject the alternative hypothesis i.e. we reject the claim at 1% level of significance.

Thus, we conclude that the sample provides us sufficient evidence against the claim so hair colour is independent of eye colour.

Check Your Progress 3

- 1) As stated in check your progress 1, the main difference between the chi-square test and the Kolmogorov-Smirnov (K-S) test is that the chi-square test is designed for categorical data whereas the K-S test is designed for continuous data.
- 2) Here, we want to test that the data have come from the normal distribution with mean 5.6 and standard deviation 1.2. If $F_0(x)$ is the cumulative distribution function of this normal distribution then our claim is $F(x) = F_0(x)$ and its complement is $F(x) \neq F_0(x)$. Since the claim contains the equality sign so we can take the claim as the null hypothesis and the complement as the alternative hypothesis. Thus

$$H_0 : F(x) = F_0(x) \text{ for all values of } x$$

$$H_1 : F(x) \neq F_0(x) \text{ for at least one value of } x$$

For testing the null hypothesis, the test statistic is given by

$$D_n = \sup_x |S(x) - F_0(x)|$$

where $S(x)$ is the empirical cumulative distribution function based on the observed sample.

For obtaining the empirical cumulative distribution function $S(x)$, first, we obtain cumulative frequencies and then by the definition of empirical cumulative distribution function we obtain $S(x)$ as

$$S(x) = \frac{\text{The number of sample observations } \leq x}{\text{Total number of observations}}$$

Here, the theoretical cumulative distribution function $F_0(x)$ can be obtained with the help of the method as described in the unit. Therefore, by the definition of the distribution function, we have

$$F_0(x) = P[X \leq x] = P\left[\frac{X - \mu}{\sigma} \leq \frac{x - \mu}{\sigma}\right] = P[Z \leq z]$$

where $P[Z \leq z]$ can be obtained with the help of **Table II of Appendix**.

Thus, when $x = 4.8$, $z = \frac{x - \mu}{\sigma} = \frac{4.8 - 5.6}{1.2} = -0.67$ and

$$\begin{aligned} F_0(4.8) &= P[Z \leq z] = P[Z \leq -0.67] \\ &= 0.5 - P[0 < Z < 0.67] \\ &= 0.5 - 0.2486 \\ &= 0.2514 \end{aligned}$$

Similarly, when $x = 5.8$, $z = \frac{5.8 - 5.6}{1.2} = 0.17$ and

$$\begin{aligned} F_0(5.8) &= P[Z \leq z] = P[Z \leq 0.17] \\ &= 0.5 + P[0 < Z \leq 0.17] \\ &= 0.5 + 0.0675 = 0.5675 \text{ and so on.} \end{aligned}$$

Calculation for $|S(x) - F_0(x)|$:

X	Observed Frequency	Cumulative Observed Frequency	Observed Distribution Function $S(x)$	$z = \frac{X - \mu}{\sigma} = \frac{X - 5.6}{1.2}$	Theoretical Distribution Function $F_0(x)$	Absolute Difference $ S(x) - F_0(x) $
4.8	1	1	$1/12 = 0.0833$	-0.67	0.2514	0.1681
5.0	1	2	$2/12 = 0.1667$	-0.50	0.3085	0.1418
5.5	1	3	$3/12 = 0.2500$	-0.08	0.4681	0.2181
5.8	1	4	$4/12 = 0.3333$	0.17	0.5675	0.2342
5.9	2	6	$6/12 = 0.5000$	0.25	0.5987	0.0987
6.0	1	7	$7/12 = 0.5833$	0.33	0.6293	0.0460
6.2	2	9	$9/12 = 0.7500$	0.50	0.6915	0.0585
6.3	1	10	$10/12 = 0.8333$	0.58	0.7190	0.1143
6.6	2	12	$12/12 = 1.0000$	0.83	0.7967	0.2033

From the above calculations, we have

$$D_n = \sup_x |S(x) - F_0(x)| = 0.2342$$

The critical value of the test statistic D_n corresponding $n = 12$ at 5% level of significance is 0.375.

Since the calculated value of the test statistic $D_n (= 0.2342)$ is less than the critical value ($= 0.375$) so we do not reject the null hypothesis i.e. we support the claim at 5% level of significance.

Thus, we conclude that the sample fails to provide us sufficient evidence against the claim so we may assume that data have come from the normal distribution with mean 5.6 and standard deviation 1.2.

12.6 REFERENCES

IGNOU Course Material

1. Post-Graduate Diploma in Applied Statistics (PGDAST), MST 004: Statistical Inference, Block 4: Non-Parametric Tests UNIT 16: Analysis of Frequencies (for Sections 12.2, 12.3 and related portions in Sections 12.0, 12.1, 12.5, 12.6 and Check Your Progress questions)
2. Post-Graduate Diploma in Applied Statistics (PGDAST), MST 004: Statistical Inference, Block 4: Non-Parametric Tests UNIT 13: One Sample Tests (for Sections 12.4 and related portions in Sections 12.0, 12.1, 12.5, 12.6 and Check Your Progress questions)