UNIT 16 ANALYSIS OF FREQUENCIES

Structure

16.1 Introduction

Objectives

16.2 Chi-Square Test for Goodness of Fit

16.3 Chi-Square Test for Independence of Attributes

16.4 Summary

16.5 Solutions / Answers

16.1 INTRODUCTION

In testing of hypothesis and estimation of parameter(s), we generally assume that the random variable follows a distribution such as normal, binomial, Poisson distribution, etc., but often the need is felt to confirm whether our assumption is true or not. Therefore, on the basis of observational data for such testing we use Kolmogorov-Smirnov test (described in Unit 13). But 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, sex, income, job, etc. or the books of library can be classified according to their subjects such as books of Science, Commerce, Art, etc. Now, the question arises “how we tackle the inference problems arising out of categorical data?” The chi square test is usually used in such situations. In this unit, we shall discuss two most widely used applications of chi-square test as goodness of fit and independence of attributes.

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

This unit is divided into five sections. Section 16.1 is described the need of tests which are applied for categorical data. The chi-square test for goodness of fit is discussed in Section 16.2 in details whereas chi-square test for independence of attributes is discussed in Section 16.3. Unit ends by providing summary of what we have discussed in this unit in Section 16.4 and solution of exercises in Section 16.5.

Objectives

After studying this unit, you should be able to:

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

16.2 CHI-SQUARE TEST FOR GOODNESS OF FIT

In the previous block, we have discussed the parametric tests in which we first assume the form of the parent population and then perform a test about some parameter(s) of the population(s) 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 a 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 in each category with theoretically expected frequencies that would have occurred if the data followed a specified or assumed or hypothesized 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) Sum of observed frequencies is equal to sum of expected frequencies, i.e., $\sum O = \sum E$.

As usual first step in testing of hypothesis is to setup 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 hypothesized distribution $F_0(x)$ or a sample has come from a specified distribution or not. So here we consider only 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

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

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

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 sun of frequencies we take notation 'N'.

Step 2: After setting null and alternative hypotheses, our next set up 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 categorized into k groups or classes. Also let $O_1, O_2,\ldots,O_k$ are the observed frequencies and $E_1, E_2, \ldots, E_k$ 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 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, …, k) is the probability that an observation falls in ith

category then we find the expected frequency by the formula given below

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

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

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

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

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

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 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 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 critical value of test statistic at given level of significance under the condition that null hypothesis is true. **Table III** in the Appendix at the end of Block I 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 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 calculated value of test statistic is greater than or equal to 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 example based on 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 customers over five brands is uniform. So our claim is "the preference of customers over five brands is uniform" and its complement is "the preference of customers over five brands is not uniform". So we can take claim as the null hypothesis and complement as the alternative hypothesis. Thus,

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

H_1: The preference of 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 chi- square goodness of fit test.

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

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

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

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

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 uniform distribution. Uniform distribution is one in which all outcomes have equal (or uniform) probability. Therefore, the probability that the customers prefer one of any brand is 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 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 chi-square with $k - 1 = 5 - 1 = 4$ degrees of freedom at 5% level of significance is 9.49.

Since calculated value of test statistic (= 0.48) is smaller than 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 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 Poisson distribution or not at 5% level of significance.

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

H_0: The number of accidents follows Poisson distribution

H_1: The number of accidents does not follow Poisson distribution

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.

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

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

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

where, k = number of classes

r = number of parameters estimated from the sample data

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

$$P\left[X = x\right] = \frac{\bar{e}^{\lambda}\lambda^{x}}{x!}; \quad x = 0, 1, 2, ...$$

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 null hypothesis), therefore, it is estimated from sample data. Since λ represents the mean of the population so it can be estimated by value of sample mean. Thus, first we find the sample mean and the value of 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	$\sum fX = 104$

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 Poisson distribution is

$$P\left[X = x\right] = \frac{\bar{e}^{2}\,2^{x}}{x!}; \quad x = 0, 1, 2, ... \qquad \text{… (1)}$$

Now, to find the expected or theoretical number of accidents, we first find the probability of each class (X = 0, 1, 2,…,6) by putting X = 0, 1,…,6 in equation (1) respectively. **Table X** in the Appendix given at the end of this block provides the Poisson probabilities 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, …,6 and λ = 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 or more] = 1−P[X < 6] =1− [P[X = 0]+…+ P[X = 5]]

=1 − (0.1353 + 0.2707 + 0.2707 + 0.1804 + 0.0902 + 0.0361) = 0.0166

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

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

Similarly,

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

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

$$E_7 = N - \left(E_1 + E_2 + ... + E_6\right)$$

$$= 52 - \left(7.0356 + 14.0764 + ... + 1.8772\right) = 0.8632$$

Since the expected frequencies in the last three classes are < 5, therefore, we combine the last three 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	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 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 = number of classes that remain after the combining classes = 5

r = 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 calculated value of test statistic (= 2.2085) is less than 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 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 life of car batteries follows a normal distribution. So our claim is "the life of car batteries follows normal distribution" and its complement is "the life of car batteries does not follow normal distribution". So we can take claim as the null hypothesis and complement as the alternative hypothesis. Thus,

H_0: The life of car batteries follows normal distribution

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

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

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

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

where, k = number of classes

r = number of parameters estimated from the sample data

If X represents the life of car battery that follows normal distribution then the probability density function of 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**	**fX²**
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	3330.00

The formula for calculating mean is

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

The formula for calculating standard deviation is

$$\hat{\sigma} = S = \sqrt{\frac{1}{N-1}\left\{\sum fX^2 - \frac{\left(\sum fX\right)^2}{N}\right\}}$$

$$= \sqrt{\frac{1}{399}\left\{3330 - \frac{(1074)^2}{400}\right\}} = \sqrt{1.199} = 1.06$$

The next step is to find the expected frequency for each class under the assumption of normally distributed population. The expected frequencies of normal distribution can be obtained with the help of the method as described in Unit 4 of MST-003. Therefore, first we find the standard normal veriate $Z = \frac{X - \hat{\mu}}{\hat{\sigma}}$ corresponding to each lower limit of the class and then find $F(x) = P[X \le x] = P[Z \le z]$ for each value with the help of **Table I** given in the Appendix at the end of Block 1of this course.

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

$$F(0) = P[Z \le z] = P[Z \le -2.53] = 0.5 - P[-2.53 < Z < 0]$$

$$= 0.5 - P[0 < Z < 2.53] = 0.5 - 0.4943 = 0.0057$$

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

$$F(3) = P[Z \le z] = P[Z \le 0.30] = 0.5 + P[0 < Z < 0.30]$$

$$= 0.5 + 0.1179 = 0.6179$$

Therefore, expected frequencies are calculated as follows:

Life	Lower Limit	Standard Normal Variate $Z=\frac{X-\hat{\mu}}{\hat{\sigma}}$	Area Under Normal Curve to the Left of Z i.e. $P[Z \le z]$	Difference between Successive Areas	Expected Frequency $= 400 \times$ 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	$0.9706 - 0.8944$ $= 0.0762$	30.48
5-6	5	1.89	0.9706	$0.9991 - 0.9706$ $= 0.0285$	11.76
6 and above	6	3.13	0.9991	--	--

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

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

Since 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	30.48	−3.48	12.1104	0.3973
5-6	10	11.76	−1.76	3.0976	0.2634
Total	400	400			3.6279

Therefore, from above calculations, we have

$$\chi^2 = \sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i} = 3.6279$$

Here, we combined the classes so

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

r = 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 calculated value of test statistic (= 3.6279) is less than 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 normal distribution.

Example 4: The following data represent the results of an investigation of the sex 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 binomial distribution with 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 binomial distribution with p = 0.5. So our claim is "the number of sons in a family follows binomial distribution with p = 0.5" and its complement is "the number of sons in a family does not follow binomial distribution with p = 0.5". So we can take claim as the null hypothesis and complement as the alternative hypothesis. Thus,

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

H_1: The number of sons in a family does not follow 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 chi-square goodness of fit test.

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

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

where, k = 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

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

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

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

$$= {}^{4}C_x (0.5)^x (0.5)^{4-x} = {}^{4}C_x (0.5)^4; \quad x = 0,1,2,3,4 \qquad \dots(2)$$

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,

$$p_1 = P[X = 0] = {}^{4}C_0 (0.5)^4 = 0.0625$$

$$p_2 = P[X = 1] = {}^{4}C_1 (0.5)^4 = 4 \times 0.0625 = 0.2500$$

$$p_3 = P[X = 2] = {}^{4}C_2 (0.5)^4 = 6 \times 0.0625 = 0.3750$$

$$p_4 = P[X = 3] = {}^{4}C_3 (0.5)^4 = 4 \times 0.0625 = 0.2500$$

$$p_5 = P[X = 4] = {}^{4}C_4 (0.5)^4 = 1 \times 0.0625 = 0.0625$$

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

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

Similarly,

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

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

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 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 = 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 calculated value of test statistic (= 3.4222) is less than 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.

Now, you can try the following exercises.

E1) Write one difference between chi-square test and Kolmogorov-Smirnov test for goodness of fit.

E2) 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.

E3) The number of customers waiting for service on the checkout counter line of a large 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 Poisson distribution with mean 8 minutes at 5% level of significance?

16.3 CHI-SQUARE TEST FOR INDEPENDENCE OF ATTRIBUTES

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.

There are many situations where we need to test the independency of two characteristic or attributes of categorical data. For example, a sociologist may wish to know whether level of formal education is independent with income, whether height of sons depending on 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 other. To test the independence of two variables when observations in a population are classified according to some attributes we may use chi-square test for independence. This test will indicate only whether or not any association exists between attributes.

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.

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.

Assumptions

This test work under the following assumptions:

(i) The sample observations are random and independent.

(ii) The observations may be classified into non-overlapping categories.

(iii) The observed and expected frequencies of each class are greater than 5.

(iv) Sum of observed frequencies is equal to sum of expected frequencies, i.e., $\sum O = \sum E$.

(v) 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, first step is to setup 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 only two-tailed case. 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 null and alternative hypotheses, 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 the characteristic A be assumed to have ‘r’ categories $A_1, A_2, \ldots, A_r$ and characteristic B be assumed to have ‘c’ categories $B_1, B_2, \ldots, B_c$. The various observed frequencies in different classes can be expressed in the form of a table known as contingency table.

B / **A**	**B_1**	**B_2**	**… B_j …**	**B_c**	**Total**
A_1	O_{11}	O_{12}	… O_{1j} …	O_{1c}	**R_1**
A_2	O_{21}	O_{22}	… O_{2j} …	O_{2c}	**R_2**
.	.	.	.	.	**.**
.	.	.	.	.	**.**
.	.	.	.	.	**.**
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	**C_1**	**C_2**	**… 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 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 an observation lies in a particular cell for all i =1, 2, …, r and j =1, 2, …, c. These probabilities are calculated according to the multiplication law of probability (described in Section 3.4 of Unit 3 of the course MST-003). 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,

$$P\left[A_i B_j\right] = \frac{R_i}{N}.\frac{C_j}{N}$$

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

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

This equation reduces to

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

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 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

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

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

Step 6: Obtain critical value of test statistic corresponding $(r-1)(c-1)$ df at given level of significance under the condition that null hypothesis is true. **Table III** in the Appendix at the end of Block I of this course 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 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 calculated value of test statistic is greater than or equal to 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 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 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 IQ level of students is independent of the economic condition of their parents. So our claim is "IQ level and economic condition are independent" and its complement is "IQ level and economic condition are not independent". So we can take claim as the null hypothesis and complement as the alternative hypothesis. Thus,

H_0 : IQ level and economic condition are independent

H_1 : IQ level and economic condition are not independent

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

For testing the null hypothesis, the test statistic is

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

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

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

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

Therefore,

$$E_{11} = \frac{R_1 \times C_1}{N} = \frac{400 \times 700}{1000} = 280,\ E_{12} = \frac{R_1 \times C_2}{N} = \frac{400 \times 300}{1000} = 120$$

$$E_{21} = \frac{R_2 \times C_1}{N} = \frac{600 \times 700}{1000} = 420,\ E_{22} = \frac{R_2 \times C_2}{N} = \frac{600 \times 300}{1000} = 180$$

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 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 calculated value of test statistic (= 31.74) is greater than 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 sample provides us sufficient evidence against the claim so 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 condition of home and condition of the child are independent. So our claim is "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 claim as the null hypothesis and complement as the alternative hypothesis. Thus,

H_0 : Condition of home and condition of child are independent

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

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

For testing the null hypothesis, test statistic is

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

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

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

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

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			25.64

Therefore, from above calculations, we have

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(O_{ij}-E_{ij}\right)^2}{E_{ij}} = 25.64$$

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 calculated value of test statistic (= 25.64) is greater than 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 condition of home and condition of the child are not independent.

Now, you can try the following exercises.

E4) 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 chi-square test at 1% level of significance to test whether the two attributes are independent.

E5) The following contingency table presents the analysis of 300 persons according to hair colour and eye colour:

Hair Colour	Eye Colour			
	Blue	Grey	Brown	Total
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.

We now end this unit by giving a summary of what we have covered in it

16.4 SUMMARY

In this unit, we have discussed following points:

1. When the data are classified into different categories or groups according to one or more attributes such type of data 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.

16.5 SOLUTIONS / ANSWERS

E1) The main difference between 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 K-S test is designed for the continuous data.

E2) Here, we want to test that 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 claim as the null hypothesis and 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 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 uniform distribution is given by

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

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

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

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 $\frac{\left(O-E\right)^2}{E}$:

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 calculation, we have

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

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

Since calculated value of test statistic (= 7.6923) is less than 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.

E3) Here, we want to test that the number of costumers waiting for servicing follow 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 claim as the null hypothesis and complement as the alternative hypothesis. Thus,

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

H_1: The number of costumers waiting for servicing does not follow 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 chi- square goodness of fit test.

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

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

where, k = number of classes

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

$$P\left[X = x\right] = \frac{\overline{e}^{\lambda}\,\lambda^{x}}{x!}; \quad x = 0, 1, 2, \ldots$$

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

$$P\left[X = x\right] = \frac{\overline{e}^{8}\,8^{x}}{x!}; \quad x = 0, 1, 2, \ldots \qquad \ldots (3)$$

Now, to find the expected number of costumers waiting for servicing per occasion, we first find the probability of each class by putting X = 0, 1, … in equation (3). With the help of **Table X** in the Appendix at the end of this block, we can also find these probabilities by taking X = 0, 1, … and λ = 8 as:

p_1 = P[X = 0 or 1] = P[X = 0] + P[X =1] = 0.0003 + 0.0027 = 0.0030,

p_2 = P[X = 2] = 0.0107, p_3 =P[X =3] = 0.0286, p_4 = P[X = 4] = 0.0573,

p_5 = P[X = 5]= 0.0916, p_6 = P[X = 6] = 0.1221, p_7 = P[X = 7] = 0.1396,

p_8 = P[X = 8 or more] = 1−P[X < 8] =1− [P[X = 0]+…+ P[X = 7]]

=1 − (0.0003 + 0.0027 + 0.0107 + 0.0286 + 0.0573 + 0.0916 + 0.1221 + 0.1396) = 0.5471

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

$$E_1 = Np_1 = 64 \times 0.0030 = 0.1920$$

$$E_2 = Np_2 = 0.6848,\ E_3 = Np_3 = 1.8304,\ E_4 = Np_4 = 3.6672,$$

$$E_5 = Np_5 = 5.8624, E_6 = Np_6 = 7.8144,\ E_7 = Np_7 = 8.9344,$$

$$E_8 = N - \left(E_1 + E_2 + \ldots + E_7\right) = 35.0144$$

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

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

Here, we combined the classes so

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

Here,
param
we tak

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

Since calculated value of test statistic (= 150.7151) is greater than critical value (= 9.49) 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 number of costumers waiting for servicing per occasion does not follow Poisson distribution.

E4) 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 claim as the null hypothesis and 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 interesting to test the independence of two attributes family income and selection of school, so we go for chi-square test of independence.

For testing the null hypothesis, the test statistic is

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

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

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

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

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 above calculations, we have

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

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

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

Since calculated value of test statistic (= 220.98) is greater than 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.

E5) 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 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 interesting to test the independence of two attributes hair colour and eye colour, so we go for chi-square test of independence.

For testing the null hypothesis, the test statistic is

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

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

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

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

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 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 chi square with 4 df at 1% level of significance is 13.28.

Since calculated value of test statistic chi-square (= 7.29) is less than 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 provide us sufficient evidence against the claim so hair colour is independent of eye colour.