UNIT 16 ASSOCIATION OF ATTRIBUTES FOR $r \times s$ CONTINGENCY TABLE

Structure

16.1 INTRODUCTION

In Unit 13, we have discussed that the classification of the data can be dichotomous or manifold. If an attribute has only two classes it is said to be dichotomous and if it has many classes, it is called manifold classification. For example the criterion ‘location’ can be divided into big city and small town. The other characteristic ‘nature of occupancy’ can be divided into ‘owner occupied’, ‘rented to private parties’. This is dichotomous classification. Now suppose we have N observations classified according to both criteria. For example, we may have a random sample of 250 buildings classified according to ‘location’ and ‘nature of occupancy’ as indicated in the table below:

table
ells.

Table 1

Nature of occupancy	Location		Total
	Big town	Small Town	
Owner occupied	54	67	121
Rented to parties	107	22	129
Total	161	89	250

Here we have classification by two criteria - one location (two categories) and the other nature of occupancy (two categories). Such a two-way table is called contingency table. The table above is 2×2 contingency table where both the attributes have two categories each. The table has 2 rows and 2 columns and $2\times 2 = 4$ distinct cells. We also discussed in the previous unit that the purpose behind the construction of such table is to study the relation between two attributes i.e. the two attributes or characteristics appear to occur independently of each other or whether there is some association between the two. In the above case our interest lies in ascertaining whether both the attributes i.e. location and nature of occupancy are independent.

In practical situations, instead of two classes, an attribute can be classified into number of classes. Such type of classification is called manifold classification. For example stature can be classified as very tall, tall, medium, short and very short. In the present unit, we shall discuss manifold classification; related contingency table and methodology to test the intensity of association between two attributes, which are classified into number of classes. The main focus of this unit would be the computation of chi-square and the coefficient of contingency, which would be used to measure the degree of association between two attributes.

In this unit, Section 16.2 deals with the concept of contingency table in manifold classification, while the Section 16.3 illustrates the calculation of chi-square and coefficient of contingency.

Objectives

After reading this unit, you should be able to

- describe the concept of contingency table for manifold classification;
- compute the expected frequencies for different cells, which are necessary for the computation of chi-square;
- compute chi-square; and
- calculate coefficient of contingency and interpret the level of association with the help of it.

16.2 CONTINGENCY TABLE: MANIFOLD CLASSIFICATION

We have already learnt that if an attribute is divided into more than two parts or groups, we have manifold classification. For example, instead of dividing the universe into two parts-heavy and not heavy, we may sub-divide it in a large number of parts very heavy, heavy, normal, light and very light. This type of subdivision can be done for both the attributes of the universe. Thus, attribute A can be divided into a number of groups $A_1, A_2, \ldots, A_r$. Similarly, the attribute B can be subdivided into $B_1, B_2, \ldots, B_r$. When the observations are classified according to two attributes and arranged in a table, the display is called contingency table. This table can be 3×3, 4×4, etc. In 3×3 table both of the attributes A and B have three subdivisions. Similarly, in 4×4 table, each of the attributes A and B is divided into four parts, viz. A_1, A_2, A_3, A_4 and B_1, B_2, B_3, B_4.

The number of classes for both the attributes may be different also. If attribute A is divided into 3 parts and B into 4 parts, then we will have 3×4 contingency table. In the same way, we can have 3×5, 4×3, etc. contingency tables. It should be noted that if one of the attributes has two classes and another has more than two classes, even then the classification is manifold. Thus, we can have 2×3, 2×4, etc. contingency tables.

If or attri class has class the c man

We shall confine our attention to two attributes A and B, where A is subdivided into r classes, $A_1, A_2, \ldots, A_r$ and B is subdivided into s classes $B_1, B_2, \ldots, B_s$. The various cell frequencies can be expressed in the following table known as $r\times s$ contingency table where (A_i) is the number of person possessing the attribute A_i $(i = 1, 2, \ldots, r)$, (B_j) is the number of persons possessing the attribute B_j $(j=1, 2, \ldots, s)$ and (A_iB_j) is the number of person possessing both attributes A_i and B_j $(i = 1, 2, \ldots, r;\ j = 1, 2, \ldots, s)$. Also, we have $\sum_{i=1}^{r} A_i = \sum_{j=1}^{s} B_j = N$ where N is the total frequency.

Following is the layout of r × s contingency table:

Table 2: r × s Contingency Table

B \ A	A_1	A_2	…	A_i	…	A_r	Total
B_1	(A_1B_1)	(A_2B_1)	…	(A_iB_1)	…	(A_rB_1)	(B_1)
B_2	(A_1B_2)	(A_2B_2)	…	(A_iB_2)	…	(A_rB_2)	(B_2)
⋮	⋮	⋮		⋮		⋮	⋮
B_j	(A_1B_j)	(A_2B_j)	…	(A_iB_j)	…	(A_rB_j)	(B_j)
⋮	⋮	⋮		⋮		⋮	⋮
B_s	(A_1B_s)	(A_2B_s)	…	(A_iB_s)	…	(A_rB_s)	(B_s)
Total	(A_1)	(A_2)	…	(A_i)	…	(A_r)	N

In the above table sum of columns A_1, A_2, etc. and the sum of rows B_1, B_2, etc. would be first order frequencies and the frequencies of various cells would be second order frequencies. The total of either A_1, A_2, etc. or B_1, B_2, etc. would give grand total N.

In the table

$$(A_1) = (A_1B_1) + (A_1B_2) + \dots + (A_1B_s),$$

$$(A_2) = (A_2B_1) + (A_2B_2) + \dots + (A_2B_s),$$

etc. Similarly,

$$(B_1) = (A_1B_1) + (A_2B_1) + \dots + (A_rB_1),$$

$$(B_2) = (A_1B_2) + (A_2B_2) + \dots + (A_rB_2),$$

etc. And

$$N = (A_1) + (A_2) + \dots + (A_r) \text{ or}$$

$$N = (B_1) + (B_2) + \dots + (B_s)$$

In the following section you will learn how to find degree of association between attributes in r × s contingency table.

16.3 CHI - SQUARE AND COEFFICIENT OF CONTINGENCY

The computation of coefficient of contingency requires the knowledge of observed frequencies as well as knowledge of theoretical or expected frequencies. Therefore, before computing coefficient of contingency it becomes necessary to construct a theoretical or expected frequency table. The expected frequencies are calculated in the following manner and are entered into the table of expected frequency.

Expected frequency of $(A_1B_1) = \frac{(A_1)(B_1)}{N}$

Expected frequency of $(A_2B_1) = \frac{(A_2)(B_1)}{N}$

In general, $(A_iB_j) = \frac{(A_i)(B_j)}{N}$; $\quad i = 1, 2, ..., r$ & $j = 1, 2, ..., s$

Similarly, expected frequencies corresponding to other observed frequencies can be computed.

For convenience a 3×3 contingency table for computing expected frequencies is displayed in Table 3. Likewise construction can be done for $r \times s$ contingency table.

Table 3: Expected Frequency Table

A / B	A_1	A_2	A_3	Total
B_1	$\frac{(A_1)(B_1)}{N}$	$\frac{(A_2)(B_1)}{N}$	$\frac{(A_3)(B_1)}{N}$	(B_1)
B_2	$\frac{(A_1)(B_2)}{N}$	$\frac{(A_2)(B_2)}{N}$	$\frac{(A_3)(B_2)}{N}$	(B_2)
B_3	$\frac{(A_1)(B_3)}{N}$	$\frac{(A_2)(B_3)}{N}$	$\frac{(A_3)(B_3)}{N}$	(B_3)
Total	(A_1)	(A_2)	(A_3)	N

$$\chi^2 = \sum\sum \frac{O_{ij}^2}{E_{ij}} - N$$

Unequal values of observed and expected frequencies of any cell indicate Association between two attributes.

If A and B are completely independent of each other, then the actual values of (A_1B_1), (A_2B_2), etc. must be equal to their corresponding expected values i.e. $\frac{(A_1)(B_1)}{N}$, $\frac{(A_2)(B_2)}{N}$, etc. respectively. If observed frequency of each cell of a contingency table is equal to the expected frequency of the same cell then we can say A and B are completely independent.

If these values are not equal for any of the cells then it indicates association between two attributes A and B. In order to measure the level of association, the difference between the observed and the expected frequencies for various cells are calculated. With the help of such differences the value of Chi-square is calculated which is abbreviated as χ^2.

Thus,

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{[\text{Difference of observed and expected frequencies}]^2}{\text{Expected frequencies}}$$

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

The above expression can also be written as

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{O_{ij}^2}{E_{ij}} - N$$

where, O - is the observed frequency of a class, and

E - is the expected frequency of that class.

χ^2 is also called "Square contingency". If the mean of χ^2 is calculated it is called "Mean Square Contingency" which is denoted by ϕ^2 (pronounced as phi- square).

Therefore, Square contingency = χ^2

Mean square contingency $\phi^2 = \dfrac{\chi^2}{N}$

χ^2 and ϕ^2 are sum of squares and hence they cannot assume negative values.

As far as the limit of χ^2 and ϕ^2 are concerned we see that χ^2 and ϕ^2 are sum of squares and hence they cannot assume negative values. The minimum value of χ^2 and ϕ^2would be 0. This will happen when the numerator in the expression of χ^2 is 0, i.e. when the observed and expected frequencies are equal in all the cells of the contingency table. This is the case when the attributes A and B are completely independent. The limits of χ^2 and ϕ^2 vary in different cases and we cannot assign upper limits to χ^2 and ϕ^2 and thus they are not suitable for studying the association in contingency tables. Karl Pearson has given the following formula for the calculation of "Co-efficient of Mean Square Contingency."

The coefficient of mean square contingency is defined as

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}} \quad \text{or} \quad C = \sqrt{\frac{\phi^2}{1 + \phi^2}}$$

If we calculate χ^2 by the formula

$$\chi^2 = \sum\sum\left(\frac{O^2}{E}\right) - N \quad \text{and if } \sum\sum\left(\frac{O^2}{E}\right) \text{ is represented by S}$$

then, $C = \sqrt{\dfrac{S-N}{N + S - N}} = \sqrt{\dfrac{S-N}{S}}$

The above coefficient has a drawback, that it will never attain the upper limit of 1. The limit of 1 is reached only if the number of classes is infinite. Ordinarily its maximum value depends on the values of r and s, i.e. number of rows and columns.

In $r\times r$ table (i.e. 2×2, 3×3, 4×4, etc.) the maximum value of $C = \sqrt{\dfrac{r-1}{r}}$

Thus, in 2×2 table the maximum value of $C = \sqrt{\dfrac{2-1}{2}} = 0.707$

Thus 3×3 table it is 0.816 and in 4×4 it is 0.866.
It is clear the maximum value of C depends upon how the data are classified. Therefore, coefficients calculated from different types of classification are not comparable.

We now illustrate the computation of χ^2 and coefficient of contingency through some examples.

Example 1: From the data given below, study the association between temperament of brothers and sisters

Table 4

Temperament of Brothers	Temperament of Sisters			
	Quick	**Good Natured**	**Sullen**	**Total**
Quick	850	571	580	2001
Good Natured	618	593	455	1666
Sullen	540	456	457	1453
Total	2008	1620	1492	5120

Solution: The expected frequencies for different cells would be calculated in the following fashion. For example the expected frequency of class (A_1B_1)

$$= \frac{(A_1)(B_1)}{N} = \frac{2001 \times 2008}{5120} = 785$$

Similarly, the expected frequency of class

$$(A_3B_2) = \frac{(A_3)(B_2)}{N} = \frac{1453 \times 1620}{5120} = 460$$

[The figures are rounded off as frequencies in decimals are not possible. The rounding is done keeping in mind that marginal totals of observed frequencies are equal to marginal totals of expected frequencies.]

The other frequencies are calculated in the same manner. Now we have for each cell, two sets of values

(i) O- the observed frequency

(ii) E -the expected frequency.

For conceptual clarity, let us put them in form of table

Table 5

Class	Observed frequency (O)	Expected frequency (E)	$(O - E)^2$	$\frac{(O - E)^2}{E}$
(A_1B_1)	850	785	4225	5.38
(A_1B_2)	571	633	3844	6.07
(A_1B_3)	580	583	0009	0.02
(A_2B_1)	618	653	1225	1.88
(A_2B_2)	593	527	4356	8.27
(A_2B_3)	455	486	0961	1.98
(A_3B_1)	540	570	0900	1.58
(A_3B_2)	456	460	0016	0.03
(A_3B_3)	457	423	1156	2.73
	5120	5120		27.94

Thus, the value of chi-square is

$$\chi^2 = 27.94$$

and

Coefficient of contingency, $C = \sqrt{\frac{\chi^2}{\chi^2 + N}}$

$$= \sqrt{\frac{27.94}{5120 + 27.94}}$$

$$= 0.0736$$

The strength of association can be measured by comparing the calculated value of C with the value calculated theoretically. We have seen maximum value of C in 3×3 table (the one in the question) is $\sqrt{\frac{r-1}{r}}$ where r denotes the columns or rows.

Hence, $C_{max} = \sqrt{\frac{3-1}{3}} = 0.816$

If we compare C calculated (0.0736) with its maximum value i.e. 0.816, we find that there is very weak association between nature of brothers and sisters.

Example 2: The table that follows contains a set of data in which 141 individuals with brain tumour have been doubly classified with respect to type and site of the tumour. The three types were as follows: A, benignant tumour; B, malignant tumour; C, other cerebral tumour. The sites concerned were : I, frontal lobes; II, temporal lobes; III, other cerebral areas. Compute the coefficient of contingency and interpret the result.

Table 6: Incidence of Cerebral Tumour

		Type			Total
		A	**B**	**C**	
Site	**I**	23	9	6	38
	II	21	4	3	28
	III	34	24	17	75
		78	37	26	141

Solution: Firstly, we will compute expected frequencies for different cells.

The entry in the first cell viz. $E_{11} = \frac{38 \times 78}{141} = 21$

Similarly, expected frequency

$$E_{12} = \frac{38 \times 37}{141} = 10$$

Likewise we calculate all the expected frequencies for different cells and enter them in the table given below:

Table 7

Observed frequency (O)	**Expected frequency (E)**	$(O - E)^2$	$\frac{(O - E)^2}{E}$
23	21	4	0.19
9	10	1	0.10
6	7	1	0.14
21	16	25	1.56
4	7	9	1.29
3	5	4	0.80
34	41	49	1.20
24	20	16	0.8
17	14	9	0.64
141	141		6.72

Hence, the value of $\chi^2 = 6.72$

Coefficient of contingency $C = \sqrt{\frac{\chi^2}{\chi^2 + N}}$

$$= \sqrt{\frac{6.72}{6.72 + 141}}$$

$$= 0.21$$

Comparing the coefficient of contingency with theoretical value of C_{max} for 3×3 table, (Recall $C_{max} = 0.816$, as given in previous example) we see that the association between cell and type of tumour is weak.

Note: Same procedure will be followed for 4×4 or 5×5, etc. contingency tables except for comparison the theoretical C_{max} would be different. Hope you remember it is $\sqrt{\frac{r-1}{r}}$ where r denotes the number of columns or rows.

Now, let us solve the following exercise:

E1) 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
Rich	460	140	600
Poor	240	160	400
Total	700	300	1000

Use the coefficient of contingency to determine the amount of association between economic condition and IQ level.

E2) The following contingency table presents the analysis of 300 persons according to hair colour and eye colour. Study the association between hair colour and eye colour.

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

E3) A company is interested in determining the strength of association between the communicating time of their employees and the level of stress–related problem observed on job. A study of 116 assembly line workers reveals the following:

	Stress			
	High	Moderate	Low	Total
Under 20 min	9	5	18	32
20-50 min	17	8	28	53
Over 50 min	18	6	7	31
Total	44	19	53	116

16.4 SUMMARY

In this unit, we have discussed:

1. Contingency table is a table of joint frequencies of occurrence of two variables classified into categories. For example, a contingency table for a sample of right and left handed boys and girls would show the number of right handed boys, right handed girls, left handed boys and left handed girls together with the total of sex and handedness. Thus, the table could be displayed as:

	Boys	Girls	Total
Right handed			
Left handed			
Total			

This is an example of 2×2 contingency table, where each attribute is divided into two categories. Similarly, 3×2 table would have 3 categories of one attribute and two of other attribute. In general, we have seen that a $r\times s$ table has r categories of one attribute and s categories of other attribute. 2×2 table is an example of dichotomous classification whereas 3×3 or $r\times s$ contingency tables are examples of manifold classification;

2. χ^2 is used for finding association and relationship between attributes;
3. The calculation of χ^2 is based on observed frequencies and theoretically determined (expected) frequencies. Here, it should be kept in mind that $\sum\sum O_{ij} = \sum\sum E_{ij} = N$;
4. We have seen that if the observed frequency of each cell is equal to the expected frequency of the respective cell for whole contingency table, then the attributes A and B are completely independent and if they are not same for some of the cells then it means there exists some association between the attributes;
5. The degree or the extent of association between attributes in $r\times s$ contingency table could be found by computing coefficient of mean square contingency $C = \sqrt{\dfrac{\chi^2}{\chi^2 + N}}$. The value of C lies between 0 and 1 but it never attains the value unity. A value near to 1 shows great degree of association between two attributes and a value near 0 shows no association; and
6. The theoretical maximum value of c depends upon the value of rows and columns. In $r\times r$ table the maximum value of C is $\sqrt{\dfrac{r-1}{r}}$ where r = 2, 3, 4, etc.

16.5 SOLUTIONS /ANSWERS

E1) Calculations for expected frequencies are below:

$$E_{11} = \frac{700\times 600}{1000} = 420$$

$$E_{12} = \frac{300\times 600}{1000} = 180$$

$$E_{21} = \frac{700\times 400}{1000} = 280$$

$$E_{22} = \frac{300\times 400}{1000} = 120$$

Enter the expected frequencies in the table given below

Observed frequency (O)	**Expected frequency (E)**	**(O – E)**	**$(O - E)^2$**	**$\frac{(O - E)^2}{E}$**
460	420	40	1600	3.810
140	180	– 40	1600	8.889
240	280	– 40	1600	5.714
160	120	40	1600	13.333
1000	1000			31.746

Therefore, $\chi^2 = 31.746$

The coefficient of contingency, $C = \sqrt{\dfrac{\chi^2}{\chi^2 + N}}$

$$= \sqrt{\frac{31.746}{31.746 + 1000}}$$

$$= 0.175$$

The maximum value of C for 2×2 contingency table is

$$C_{max} = \sqrt{\frac{r-1}{r}} = \sqrt{\frac{2-1}{2}} = \sqrt{\frac{1}{2}} = 0.7071$$

If we compare C with C_{max} we infer that association between IQ level and economic condition of students is weak.

E2) The expected frequencies and corresponding observed frequencies are computed as follows

$$E_{11} = \frac{120 \times 80}{300} = 32, \quad E_{12} = \frac{60 \times 80}{300} = 16$$

Similarly, we can compute them for other cells. Thus, we have following table:

Observed frequency (O)	**Expected frequency (E)**	**(O – E)**	$\mathbf{(O - E)^2}$	$\mathbf{\frac{(O - E)^2}{E}}$
30	32	-2	4	0.125
10	16	-6	36	2.250
40	32	8	64	2.000
40	40	0	0	0
20	20	0	0	0
40	40	0	0	0
50	48	2	4	0.083
30	24	6	36	1.500
40	48	-8	64	1.333
300	300			7.291

Hence $\chi^2 = 7.291$

The coefficient of contingency, $C = \sqrt{\dfrac{\chi^2}{\chi^2 + N}}$

$$= \sqrt{\frac{7.291}{7.291 + 300}} = 0.154$$

The value of C_{max} for 3×3 contingency table is 0.816 which is appreciably higher than C calculated, i.e. 0.154.

Thus we infer that the association between hair colour and eye colour is very weak.

E3) The expected frequencies for different cells can be computed as

$$E_{11} = \frac{32 \times 44}{116} = 12,\quad E_{12} = \frac{32 \times 19}{116} = 5,\ E_{13} = \frac{32 \times 53}{116} = 17$$

Other expected frequencies can be calculated similarly. The expected frequencies are rounded off maintaining that marginal totals of observed frequencies are equal to the marginal totals of expected frequencies. Thus, we have following table:

Observed frequency (O)	**Expected frequency (E)**	**(O – E)**	$\mathbf{(O - E)^2}$	$\mathbf{\frac{(O - E)^2}{E}}$
9	12	–3	9	0.75
5	5	0	0	0
18	15	3	9	0.60
17	20	–3	9	0.45
8	9	–1	1	0.11
28	24	4	16	0.57
18	12	6	36	2.00
6	5	1	1	0.20
7	14	–7	49	3.50
116	116			8.18

Therefore, $\chi^2 = 8.18$

The coefficient of contingency, $C = \sqrt{\frac{\chi^2}{\chi^2 + N}}$

$$= \sqrt{\frac{8.18}{8.18 + 116}} = 0.2566$$

Comparing calculated C with C_{max} for 3×3 contingency table, we find that C calculated (0.2566) is considerably less than $C_{max} = 0.816$.

Thus, we conclude that the association between commuting time of employees is weakly associated with the stress on the job.

GLOSSARY

Contingency table : A two-way table, in which columns are classified according to one criterion or attribute and rows are classified according to the other criterion or attribute.

Expected frequency : Frequencies expected under certain assumptions.

Observed frequency : Actually recorded frequency.