7.26 Blood donation society of a college claimed that 5% of the students in our college make blood donation during a given year. If in a random sample 10 of 250 students have given blood during the past year. Test H_0: $p = 0.05$ against H_1: $p \neq 0.05$ with $\alpha = 0.05$.

7.27 An electric company claimed that at least 85% of the parts which it supplied confirmed to specifications. A sample of 400 parts was tested and 75 did not meet specifications. Can we accept the company's claim at 1% level of significance?

7.28 Describe the steps used in hypothesis testing about difference of two population proportions in case of large samples.

7.29 Test the hypothesis that proportions of men and women favoring a political candidate are different on the basis of a sample survey in which 225 of 500 men and 275 of 500 women favor the candidate. Test at 5% level of significance.

7.30 In a random table of 600 persons from a certain large city 450 are found to be smokers. In another sample of 900 persons from another large city 450 are smokers. Test at $\alpha = 0.01$ that the two cities are significantly different with respect to the prevalence of smoking.

Unit - 8

Association of Attributes

After studying this unit, the students will be able to

- Recall variable and attribute.
- Recognize the notation and terminology to represent the presence and absence of attributes.
- Describe class and class frequency.
- Recognize the categorical data of two attributes.
- Explain independence of two attributes.
- Know the criterion of independence of two attributes.
- Discuss the association of two attributes; positive association, negative association, complete association and complete disassociation.
- Define Yule's coefficient of association.
- Find the coefficient of association and interpret its result.
- Define contingency table.
- Test whether two attributes in a given contingency table are statistically independent or not.
- Describe Pearson's coefficient of mean square contingencies.
- Calculate Pearson's co efficient of mean square contingency for a given contingency table and find its maximum value.
- Describe and apply Yule's correction for continuity to test the statistical independence of two given attributes.

8.1 Introduction to attributes

In our daily life, we study the characteristics like gender, health, satisfaction, religion, colour etc. that cannot be measured and expressed quantitatively but instead of its qualitative or descriptive nature, only their presence or absence can be noticed. These descriptive or qualitative characteristics are called attributes. Attributes cannot be measured accurately but they can be divided into classes and their numbers in each class can be counted e.g. the above characteristics can be classified as male or female, healthy or unhealthy, satisfied or unsatisfied, Muslim or non-Muslim, white or black etc.

8.1.1 Notation and terminology for attributes

- **Symbols**

For the sake of convenience capital English letters A, B, C... are used to denote presence of attributes and the Greek letters α, β, γ... denote the absence of these attributes respectively. For example if A represents the class "Muslim", then α will represent the class "non-Muslim". Similarly, if we are studying two or more attributes, their combination can be represented by combining the letters representing the attributes. For example, if A represent "blindness" and B "deafness", then AB will represent the "blindness and deafness". Similarly, if A represents the "Muslim" and B "male", then the following four classes will be formed for the presence or absence of these attributes

i) Muslim and male = AB

ii) Muslim and female = Aβ

iii) Non-Muslim and male = αB

iv) Non-Muslim an female = $\alpha\beta$

- **Positive and negative classes**

The classes A, B, AB are called positive classes because they contain the presence of attributes. The classes $\alpha, \beta, \alpha\beta$ are called negative classes because they contain absence of the attributes. The classes Aβ, αB contain both presence and absence of the attributes, hence are called mixed classes.

- **Class frequency**

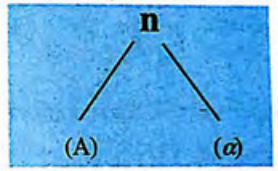
The number of observations falling in a class is called frequency of the class. For attributes class frequencies are denoted by enclosing the class symbols in parentheses.

Thus (A) denotes frequency of class/attribute A. (B) denotes frequency of attributes B. Similarly, (AB) denotes the number of individuals / objects possessing both attributes A and B. (ABC) is frequency for class ABC and so on.

8.1.2 Classification of attributes

- **Dichotomy or one-way classification**

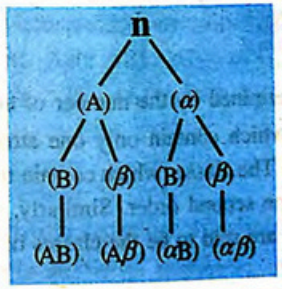
When a single attribute is under study, then simple process of classifying the whole data in two groups is called dichotomy which means classifying into two.



Since only one attribute is involved, so the division of data is called one-way classification.

- **Two-way classification**

When two attributes A and B are under study then whole data are classified into four classes or groups as follows:



Since two attributes are involved, the division of the sample is called two-way classification.

• 2 × 2 Contingency table

Classification of data about two attributes each having two classes or categories can be shown in tabular form as given below.

(2 × 2) Contingency table

Attributes	B	β	Row total
A	(AB)	(Aβ)	(A)
α	(αB)	($\alpha\beta$)	(α)
Column total	(B)	(β)	n

Since the table contains 2 rows and 2 columns is therefore called 2 × 2 contingency table.

Remember that:

$n = (A) + (\alpha)$

$n = (B) + (\beta)$

$(A) = (AB) + (A\beta)$

$(\alpha) = (\alpha B) + (\alpha\beta)$

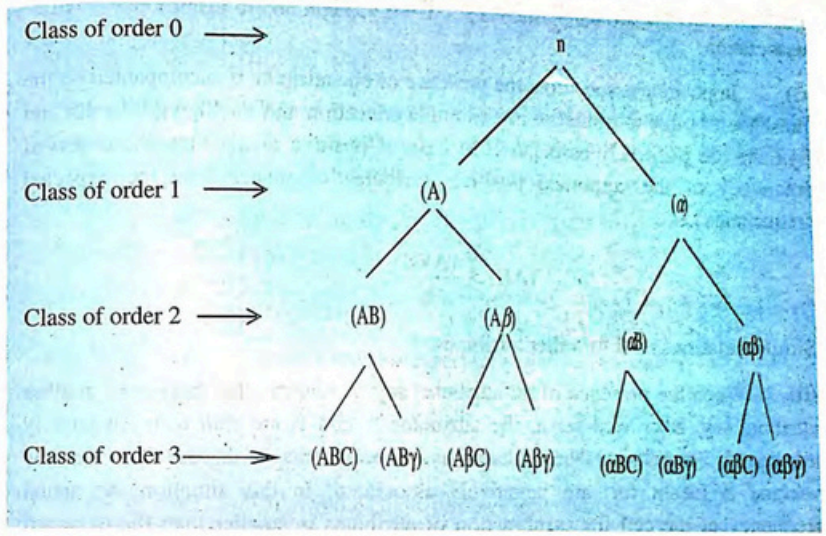
$(B) = (AB) + (\alpha B)$

$(\beta) = (A\beta) + (\alpha\beta)$

• Order of classes

The order of a class is determined by the number of attributes present in a class. For example, the classes which contain only one attribute say "A" or "B" are called classes of first order. The classes which contain two attributes say "AB" are said to be the classes of the second order. Similarly, the classes which contain three attributes say "ABC" are said to be the classes of the third order and so on.

The sample size n does not contain any attributes so is called class of order zero. Let us understand the above discussion through tree diagram as;



• Ultimate classes and ultimate class frequencies

The classes and class frequencies of the highest order are called the ultimate classes and ultimate class frequencies. For example, for one attribute ultimate classes are A, α and ultimate frequencies are (A), (α). For two attributes, ultimate classes are AB, Aβ, αB, $\alpha\beta$ and ultimate classes frequencies are (AB), (Aβ), (αB), ($\alpha\beta$). If we are considering three attributes A, B, C, then ultimate classes are ABC, ABγ, AβC, A$\beta\gamma$, αBC, αBγ, $\alpha\beta$C, $\alpha\beta\gamma$ and ultimate class frequencies are (ABC), (ABγ), (AβC), (A$\beta\gamma$), (αBC), (αBγ), ($\alpha\beta$C), ($\alpha\beta\gamma$)and so on.

8.1.3 Association of attributes

Association is a statistical technique which measures the strength and direction of relationship among qualitative variables.

• Kinds of association

There are three kinds of associations which ossibly occur between attributes namely (i) positive association, (ii) negative association and (iii) no association.

(i) In positive association, the presence of one attribute is accompanied by the presence of other attribute(s). For example education and intelligence, health and hygiene are positively associated. In case of positive association, the observed frequency of the combined positive attributes is greater than the expected frequencies i.e.

$$(AB) > \frac{(A)(B)}{n}.$$

Similar relations hold for other attributes.

(ii) When the presence of an attributes say, A ensures the absence of another attribute say, B or vice-versa, the attributes A and B are said to be negatively associated. For instance, the vaccination and occurrence of disease for which the vaccine is meant for, are negatively associated. In this situation the actual frequency of the cell for combination of attributes is smaller than the expected frequency i.e.

$$(AB) < \frac{(A)(B)}{n}.$$

(iii) If the two attributes are such that the presence or absence of one attribute has nothing to do with the presence or absence of the other, they are said to be independent. For instance, skin colour and intelligence of persons are independent attributes. If the two attributes A and B are independent, then $(AB) = \frac{(A)(B)}{n}$.

This is known as criterion of independence.

8.1.4 Methods of measures of association

There are mainly three methods of measures of association.

(i) Yule's coefficient of association.

(ii) Chi-square test for testing hypothesis about independence of attributes in contingency tables.

(iii) Coefficient of contingency for $(r \times c)$ contingency table.

• Yule's coefficient of association

Yule's coefficient of association is named after its inventor G. Undy Yule. It is a relative measure for the strength of association between two attributes say, A and B. It is defined by the formula given below.

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}, \qquad -1 < Q < +1$$

The result of Q is interpreted as:

If	$Q = -1$	Attributes have perfect negative association.
	$-1 < Q < 0$	Attributes have negative association.
	$Q = 0$	Attributes have no association means they are independent.
	$0 < Q < +1$	Attributes have positive association.
	$Q = +1$	Attributes have perfect positive association.

Example 8.1

Discuss the association between two attributes say A and B when:

(i) $(AB) = 110, (\alpha B) = 96, (A\beta) = 290, (\alpha\beta) = 510$

(ii) $(A) = 245, (AB) = 147, (\alpha) = 285, (\alpha B) = 190$

Solution:

(i) Given $(AB) = 110, (\alpha B) = 96, (A\beta) = 290, (\alpha\beta) = 510$

Coefficient of association is

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$$

$$Q=\frac{(110)(510)-(290)(96)}{(110)(510)+(290)(96)}$$

$$Q=\frac{28260}{83940}=0.34$$

It means that there is positive association between attributes A and B.

(ii) Given (A) = 245, (AB) = 147, (α) = 285, (αB) = 190

Since all values required for the coefficient of association formula are not known, so first put these values in a (2 × 2) contingency table to find the unknown values easily by addition or subtraction as.

(2 × 2) contingency table

Attributes	B	β	Row total
A	(AB) = 147	(Aβ) = 98	(A) = 245
α	(αB) = 190	($\alpha\beta$) = 95	(α) = 285
Column total	(B) = 337	(β) = 193	n = 530

Now

$$Q=\frac{(AB)(\alpha\beta)-(A\beta)(\alpha B)}{(AB)(\alpha\beta)+(A\beta)(\alpha B)}$$

$$Q=\frac{(147)(95)-(98)(190)}{(147)(95)+(98)(190)}$$

$$Q=\frac{13965-18620}{13965+18620}=\frac{-4655}{32585}=-0.14$$

Thus there is negative association between A and B.

Example 8.2

The following table shows the data obtained during an epidemic of cholera.

Attributes	attacked	not attacked
inoculated	31	469
not inoculated	185	1315

Test the effectiveness of inoculation in preventing attack of cholera.

Solution:

Let us denote inoculated by A and attack by B.

The given data can be written as under

(AB) = 31, (Aβ) = 469, (αB) = 185, ($\alpha\beta$) = 1315

Putting in the coefficient of association we get

$$Q=\frac{(AB)(\alpha\beta)-(A\beta)(\alpha B)}{(AB)(\alpha\beta)+(A\beta)(\alpha B)}$$

$$Q=\frac{(31)(1315)-(469)(185)}{(31)(1315)+(469)(185)}=\frac{40765-86765}{40765+86765}=\frac{-46000}{127530}=-0.36$$

There is negative association between inoculation and attack of cholera disease.

Example 8.3

For admission in a medical college 1660 candidates appeared in an entry test and 422 were successful. 256 attended a coaching class and of these 150 came out successful. Estimate the utility of the coaching classes.

Solution:

Putting the given information in (2 × 2) contingency table as;

Attributes	Successful	Not successful	Total
Coached A	AB = 150	$A\beta = 106$	256
Not coached α	$\alpha B = 272$	$\alpha\beta = 1132$	1404
Total	422	1238	1660

Yule's coefficient of association is given as

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$$

$$Q = \frac{(150)(1132) - (106)(272)}{(150)(1132) + (106)(272)} = \frac{140968}{198632} = 0.71$$

There is high positive association means that coaching helps in success.

8.1.5 Multi-way classification

When a population or sample is divided into many classes or categories according to an attribute is called multi-way or manifold classification. For example, a population according to "colour of eyes" can be divided as black eyes, grey eyes, green eyes etc. Population of students according to "performance" in an examination can be rated as excellent, good, average and poor. Heights of persons can be categorized as tall, medium and short. The classification of data about such attributes which have many classes can be shown by means of a contingency table.

8.2 Contingency table

A table consisting of r-rows and c-columns into which the data are classified according to two attributes is called (r × c) contingency table. For example, if an attribute "A" has $A_1, A_2, \ldots, A_r$ classes and attributes "B" has $B_1, B_2, \ldots, B_c$ classes, then the observed data by $(r \times c)$ contingency table is shown as:

$(r \times c)$ contingency table

Classes	B_1	B_2	...	B_j	...	B_c	Row total
A_1	(A_1B_1)	(A_1B_2)	...	(A_1B_j)	...	(A_1B_c)	(A_1)
A_2	(A_2B_1)	(A_2B_2)	...	(A_2B_j)	...	(A_2B_c)	(A_2)
⋮	⋮		...	⋮	...	⋮	...
A_i	(A_iB_1)	(A_iB_2)	...	(A_iB_j)	...	(A_iB_c)	(A_i)
⋮	⋮		...	⋮	...	⋮	...
A_r	(A_rB_1)	(A_rB_2)	...	(A_rB_j)	...	(A_rB_c)	(A_r)
Column total	(B_1)	(B_2)	...	(B_j)	...	(B_c)	n

This table is an extension of (2 × 2) contingency table. The values in the cells of contingency table shown in parentheses are called cell frequencies. For each observed frequency O_f, the expected frequency E_f is computed as $E_f = \frac{R \times C}{n}$, where R is the row total and C is the column total.

For example; for observed frequency (A_1B_1) given in above table, the corresponding expected frequency will be equal to $\frac{(A_1)(B_1)}{n}$ and similarly for (A_1B_2) the expected frequency = $\frac{(A_1)(B_2)}{n}$ and so on. Chi-square test statistic denoted by χ^2 (pronounced as kai square) is used to test the hypothesis about independence of attributes in the contingency tables.

8.2.1 General procedure for testing hypothesis about independence of attributes in contingency tables

i. H_0: The attributes are independent.

H_1: The attributes are associated.

ii. $\alpha = 0.01$ or 0.05 etc.

iii. Test statistic to be used here is

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-e_{ij})^2}{e_{ij}} \text{ with } \upsilon=(r-1)(c-1) \text{ d.f}$$

Whereas, O_{ij} denotes observed frequency

e_{ij} : expected frequency

r : number of rows

c : number of columns

iv. Calculation

v. Critical region

Reject H_0 if $\chi^2_{cal} \geq \chi^2_{tab}$, whereas χ^2 table value is obtained from χ^2 table 8.1

vi. Conclusion

Table 8.1 Critical values for the chi-square (χ^2) distribution.

df	α						
υ	0.10	0.05	0.025	0.02	0.01	0.005	0.001
1	2.706	3.841	5.024	5.412	6.635	7.879	10.827
2	4.605	5.991	7.378	7.824	9.210	10.597	13.815
3	6.251	7.815	9.348	9.837	11.345	12.838	16.268
4	7.779	9.488	11.143	11.668	13.277	14.860	18.465
5	9.236	11.070	12.832	13.388	15.086	16.750	20.517
6	10.645	12.592	14.449	15.033	16.812	18.548	22.457
7	12.017	14.067	16.013	16.622	18.475	20.278	24.322
8	13.362	15.507	17.535	18.168	20.090	21.955	26.125
9	14.684	16.919	19.023	19.679	21.666	23.589	27.877
10	15.987	18.307	20.483	21.161	23.209	25.188	29.588
11	17.275	189.675	21.920	22.618	24.725	26.757	31.264
12	18.549	21.026	23.337	24.054	26.217	28.300	32.909
13	19.812	22.362	24.736	25.472	27.688	29.819	34.528
14	21.064	23.685	26.119	26.873	29.141	31.319	36.123
15	22.307	24.996	27.488	28.259	30.578	32.801	37.697
16	23.542	26.296	28.845	29.633	32.000	34.267	39.252
17	24.769	27.587	30.191	30.995	33.409	35.718	40.790
18	25.989	28.869	31.526	32.346	34.805	37.156	42.312
19	27.204	30.144	32.852	33.687	36.191	38.582	43.820
20	28.412	31.410	34.170	35.020	37.566	39.997	45.315
21	29.615	32.671	35.479	36.343	38.932	41.401	46.797
22	30.813	33.924	36.781	37.659	40.289	42.796	48.268
23	32.007	35.172	38.076	38.968	41.638	44.181	49.728
24	33.196	36.415	39.364	40.270	42.980	45.558	51.179
25	34.382	37.652	40.646	41.566	44.314	46.928	52.620
26	35.563	38.885	41.923	42.856	45.642	48.290	54.052
27	36.741	40.113	43.194	44.140	46.963	49.645	55.476
28	37.916	41.337	44.461	45.419	48.278	50.993	56.893
29	39.087	42.557	45.722	46.693	49.588	52.336	58.302
30	40.256	43.773	46.979	47.962	50.892	53.672	59.703

Example 8.4

Consider the data given in the following contingency table

General ability \ Mathematical ability	Good B_1	Fair B_2	Poor B_3
Good	44	82	44
Fair	265	257	178
Poor	41	91	98

Discuss the association between the two attributes i.e. mathematical ability and general ability.

Solution:

i. H_0: There is no association between attributes.

H_1: There is association.

ii. We choose $\alpha = 0.05$

iii. Test statistic to be used here is

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-e_{ij})^2}{e_{ij}} \text{ with } v=(r-1)(c-1) \text{ d.f}$$

iv. Calculation

(O_{ij})

General ability \ Mathematical ability		Good B_1	Fair B_2	Poor B_3	Total
Good	A_1	44	82	44	170
Fair	A_2	265	257	178	700
Poor	A_3	41	91	98	230
Total		350	430	320	1100

The expected frequencies are obtained by multiplying respective rows and columns totals and are divided by total sample size n as given in the table.

(e_{ij})

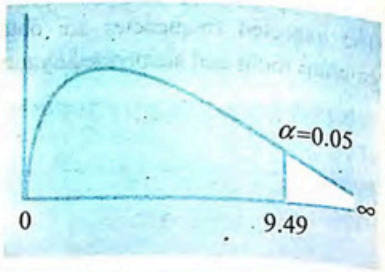
General ability \ Mathematical ability		Good B_1	Fair B_2	Poor B_3	Total
Good	A_1	$\frac{170\times350}{1100}=54.09$	$\frac{170\times430}{1100}=66.45$	$\frac{170\times320}{1100}=49.45$	169.99
Fair	A_2	$\frac{700\times350}{1100}=222.73$	$\frac{700\times430}{1100}=273.64$	$\frac{700\times320}{1100}=203.63$	700.01
Poor	A_3	$\frac{230\times350}{1100}=73.18$	$\frac{230\times430}{1100}=89.91$	$\frac{230\times320}{1100}=66.91$	230
Total		350	430	320	1100

Now χ^2 test statistic value is computed as;

O_{ij}	e_{ij}	$(O_{ij}-e_{ij})$	$(O_{ij}-e_{ij})^2$	$\frac{(O_{ij}-e_{ij})^2}{e_{ij}}$
44	54.09	−10.09	101.81	1.882
265	222.73	42.27	1786.75	8.022
41	73.18	−32.18	1035.55	14.151
82	66.45	15.55	241.80	3.639
257	273.64	−16.64	276.8896	1.012
91	89.91	1.09	1.188	0.013
44	49.45	−5.45	29.70	0.601
178	203.63	−25.64	657.41	3.228
98	66.91	31.09	966.59	14.446
1100	1100	-	-	$\chi^2 = 46.994$

$$\chi^2 = \sum_{i=1}^{3}\sum_{j=1}^{3}\frac{(O_{ij}-e_{ij})^2}{e_{ij}} = 46.994$$

v. Critical region



Reject H_0 if $\chi_c^2 \geq \chi_{\alpha,v}^2$ whereas from table 8.1 of χ^2 distribution, we have

$$\chi_{\alpha,v}^2 = \chi_{0.05,(3-1)(3-1)}^2 = \chi_{0.05,4}^2 = 9.49\ \chi^2$$

vi. Conclusion

Since our computed value of chi-square lies in the rejection region, therefore, we reject H_0 and conclude that there is association between mathematical ability and general ability.

8.2.2 Shortcut method of calculating χ^2 in case of (2 × 2) contingency table

When we have (2 × 2) contingency table having four cell frequencies as given below;

Attributes	B	β	Total
A	a	b	$a+b$
α	c	d	$c+d$
Total	$a+c$	$b+d$	$a+b+c+d=n$

The value of χ^2 can be calculated directly without computing the expected frequencies by using the following formula

$$\chi^2 = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} \quad \text{with 1 d.f}$$

Example 8.5

A random sample of 100 educated and 200 uneducated people were asked about liking and disliking of football game and the following data were recorded.

Attributes	Like football	Dislike football	Total
Educated	55	45	100
Uneducated	125	75	200
Total	180	120	300

Test the hypothesis about independence between education and liking of football at $\alpha = 0.05$

Solution:

i. H_0: There is no association between education and liking of football.

H_1: There is association

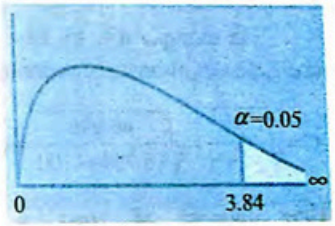
ii. $\alpha = 0.05$

iii. Test statistic to be used in this case by direct method is

$$\chi^2 = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} \quad \text{with 1 d.f}$$

iv. Calculation

$$\chi^2 = \frac{300(55\times75-45\times125)^2}{(100)(200)(180)(120)} = \frac{675}{432} = 1.5625$$



v. Critical region

Reject H_0 if $\chi_c^2 \geq \chi_{0.05,(1)}^2 = 3.84$

vi. Conclusion.

Since $\chi_c^2 = 1.5625$ falls in the acceptance χ^2 region, therefore, we accept H_0.

8.2.3 Coefficient of contingency for (r × c) contingency table

The chi-square test-statistic only decides about the independence or association of attributes in contingency tables but when H_0 is rejected it does not

tell anything about the strength of association, which we sometime need to measure. For this purpose, Karl Pearson has defined a formula for coefficient of contingency which is denoted by C and is known as Pearson's coefficient of mean-square contingency given by the relation.

$$C=\sqrt{\frac{\chi^2}{\chi^2+n}}, \quad 0\le C\le\sqrt{\frac{k-1}{k}}$$

Where χ^2 is the calculated value of chi-square test-statistic, n is the total sample size and k is the smaller one in rows and columns in number. A value of C near to "0" shows weak association and value of C near to $\sqrt{\frac{k-1}{k}}$ shows a strong association between the two attributes.

Example 8.6

Compute Pearson's coefficient of mean square contingency for the contingency table given in example 8.4.

Solution:

In example 8.4, we have $\chi^2 = 46.994$ and $n = 1100$, therefore Pearson's coefficient of mean square contingency is computed as:

$$C=\sqrt{\frac{\chi^2}{\chi^2+n}}=\sqrt{\frac{46.994}{46.994+1100}}=\sqrt{0.04097}=0.202$$

Here number of rows = number of columns = $k = 3$, so $\sqrt{\frac{k-1}{k}}=\sqrt{\frac{3-1}{3}}=\sqrt{0.6666}=0.82$. Hence the range of C is $0 \le C \le 0.82$.

Thus our computed value of C shows that the given attributes have a weak association.

8.2.4 Yate's correction for continuity

To get satisfactory results from the chi-square test in testing hypothesis about independence of attributes in contingency tables, it is necessary that expected frequency of each cell should at least be 5. If it is less, it should be added with the neighbour one to get 5 or more but in 2×2 contingency table it is not possible to combine the smaller frequency with the larger one otherwise the table will be finished. For this purpose, Frank Yate has suggested the following formula.

$$\chi^2=\sum_{i=1}^{2}\frac{\left(|O_i-e_i|-\frac{1}{2}\right)^2}{e_i}$$

This formula is known as Yate's correction for continuity and should be used only when one cell frequency is less than 5 in a 2×2 contingency table.

We also know that 2×2 contingency table is discrete frequency distribution and chi-square distribution is continuous distribution. This also needs correction for which the following formula has been suggested.

$$\chi^2=\frac{n\left(|ad-bc|-\frac{n}{2}\right)^2}{(a+b)(c+d)(a+c)(b+d)}$$

This formula should be used only when any cell expected frequency is less than 10 in a 2×2 contingency table.

8.2.5 Association versus correlation

i. Association measures the strength of relationship between two qualitative variables e.g. level of crime and education.

ii. Correlation measures the strength of relationship between two quantitative variables e.g. heights of fathers and their sons.

iii. Both are relative measures.

iv. Both measures are ranging from –1 to +1.

v. Measures of association are based on frequencies only, whereas, in correlation actual paired observations are used.

Key points

Descriptive or qualitative characteristics are called attributes.

- For the sake of convenience capital English letters A, B, C... are used to denote presence of attributes and the Greek letters $\alpha, \beta, \gamma, \ldots$ denote the absence of these attributes respectively.
- The classes A, B, AB are called positive classes because they contain the presence of attributes.
- The classes $\alpha, \beta, \alpha\beta$ are called negative classes because they contain absence of the attributes.
- For attributes class frequencies are denoted by enclosing the class symbols in parentheses.
- When a single attribute is under study then simple process of classifying the whole data in two groups is called dichotomy.
- The order of a class is determined by the number of attributes present in a class.
- The classes and class frequencies of the highest order are called the ultimate classes and ultimate class frequencies.
- Association is a statistical technique which measures the strength and direction of relationship among qualitative variables.
- $(AB) = \frac{(A)(B)}{n}$. This is known as criterion of independence for attributes.
- $Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$ is Yule's coefficient of association.
- When a population or sample is divided into many classes or categories according to an attribute is called multi-way or manifold classification.
- A table consisting of r-rows and c-columns into which the data are classified according to two attributes is called $(r \times c)$ contingency table.
- Chi-square test statistic is used to test the hypothesis about independence of attributes in the contingency tables.

Exercise

8.1 Mark the following statements as true and false.

i. Data recorded on attributes are called quantitative data.

ii. A, B, AB, ABC are called positive classes.

iii. $(B) + (\beta) = n$

iv. If $(AB) = \frac{(A)(B)}{n}$ means A and B are positively associated.

v. If $Q = 0$, then attributes are independent.

vi. Relationship between two qualitative variables is called correlation.

vii. χ^2 test has $v = (r-1)(c-1)$ df for an $(r \times c)$ contingency table.

viii. For a (4×5) contingency table the degrees of freedom for χ^2 test is 20.

ix. The range of χ^2 distribution is from 0 to ∞.

x. (2×2) contingency table contains 3-rows and 3 columns.

8.2 Fill in the suitable words in the blanks.

i. If an attribute has two classes, it is said to be _____.

ii. The number of letters representing a class determines the _____ of the class.

iii. The class represented by AB is of _____ order.

iv. The total of all frequencies n is of order _____.

v. If the attributes A and B are independent, frequency (AB) is equal to _____.

vi. If A and B are independent, Yule's coefficient of association Q is equal to _____.

vii. Association and correlation are _____.

viii. Yule's coefficient Q is ranging from _____.

ix. For an $(r \times c)$ contingency table, the sum of observed and expected frequencies must be _____.

x. For (5×6) contingency table the degree of freedom for χ^2 test is equal to _____.

8.3 Choose the correct answer.

i. Relationship between two categorical variables is called

(a) regression (b) correlation

(c) association (d) coefficient of variation

ii. The combination AB of attributes is known as the class of

(a) first order (b) second order

(c) third order (d) zero order

iii. For a 4×5 contingency table the degrees of freedom for the χ^2 is

(a) 20 (b) 16 (c) 15 (d) 12

iv. The range of Yule's coefficient of association is

(a) 0 to ∞ (b) $-\infty$ to 0

(c) 0 to 1 (d) -1 to $+1$

v. With two attributes, the total number of ultimate class frequency is

(a) two (b) four

(c) six (d) five

vi. If $(AB) < \frac{(A)(B)}{n}$, then the attributes are

(a) independent (b) positively associated

(c) negatively associated (d) no association

vii. χ^2 test statistic value varies form

(a) $-\infty$ to 0 (b) 0 to ∞ (c) $-\infty$ to ∞ (d) -1 to $+1$

viii. If A and B are independent attributes then the coefficient of association is equal to

(a) -1 (b) $+1$ (c) 0 (d) 0.5

ix. The degrees of freedom for a (2×2) contingency table will always be equal to

(a) 4 (b) 2 (c) 1 (d) 0

x. In case of consistent data, no class frequency can be

(a) positive (b) negative

(c) zero (d) one

8.4 What is meant by attributes and how they are classified?

8.5 Write short notes on the following:

i. Class symbol and class frequency

ii. Positive and negative classes

iii. Order of classes

iv. Ultimate class frequencies

8.6 Explain the following:

i. Two-way classification

ii. Association of attributes

iii. Positive and negative association

iv. Criterion of independence of attributes.

v. Yule's coefficient of association.

8.7 Distinguish between associations and correlation?

8.8 When are two attributes said to be

i. Independent

ii. Positively associated

iii. Negatively associated

8.9 Calculate the coefficient of association between extravagance in fathers and sons:

Attributes	extravagant sons	miserly sons
extravagant fathers	237	545
miserly fathers	741	235

8.10 Discuss the association when $(AB) = 256$, $(\alpha B) = 768$, $(A\beta) = 48$, $(\alpha\beta) = 144$.

8.11 Do you find any association between the tempers of brothers and sisters from the data given below?

Good natured brothers and good natured sisters = 1230

Good natured brothers and sullen sisters = 850

Sullen brothers and good nature sisters = 530

Sullen brothers and sullen sisters = 980

8.12 Investigate the association between intelligence in fathers and sons from the following data

Attributes	intelligent sons	dull sons
intelligent fathers	240	80
dull fathers	90	570

8.13 Can vaccination be regarded as a preventive measure for small-pox from the data given below?

Of 1482 persons in a locality exposed to smell-pox, 368 in all were attacked. Of 1482 persons, 343 had been vaccinated and of these 35 were attacked.

8.14 Consider the data given in the following 2 × 2 contingency table and find out whether vaccination is effective in preventing the attack of B hepatitis disease?

Attributes	attacked	not attacked
vaccinated	11	538
not vaccinated	70	464

Test by applying χ^2-test at 1% level of significance.

8.15 A random sample of size 1024 was classified according to gender and seat belt usage as shown below.

Attributes	use seat belt	don't use seat belt
male	272	192
female	276	284

Do the data suggest an association between gender and seat belt used? Use $\alpha = 0.01$

8.16 Perform chi square test of independence to decide whether smoking is a cause of lung cancer by considering the date given in the following 2×2 category table.

Attributes	cancer	no cancer
smoking	75	34
not smoking	28	112

Test at 5% level of significance.

8.17 Explain the following:

i. Manifold classification

ii. $(r \times c)$ contingency table

iii. Chi-square distribution.

8.18 Find the chi-square to test the hypothesis that there is association between height of fathers and height of sons (at $\alpha = 0.01$):

Sons \ Fathers	very tall	tall	short
very tall	600	280	300
tall	400	700	400
short	250	400	800