15 ASSOCIATION

Many experiments, particularly in social sciences, result in observations that are only classified into categories so that the data can consist of frequency count for the categories. For example, the classification of people into income groups as very rich, moderate, or poor; manufactured items may be classified as being excellent, good, poor, or scrap condition; in a survey of job compatibility employed persons may be classified as being satisfied, neutral, or dissatisfied with their jobs; in plant breeding, the offsprings of a cross fertilization may be grouped into several genotypes; rainfall may be classified heavy, moderate, or light; each household may be classified as owning no cars, one car, or two or more cars. Our aim here is to present some inferential procedures that can be used to study data that are classified into multiple categories.

15.1 MULTINOMIAL POPULATIONS

When each element of a population is assigned to one and only one of more than two attribute categories, the population is called a *multinomial population*.

15.2 ATTRIBUTE (*QUALITATIVE VARIABLE*)

A characteristic which varies only in quality from one individual to another, is called an *attribute*. Examples of attributes are: marital status, education level, blindness, smoking, richness, beauty *etc*. It is not possible to measure an attribute quantitatively. The quantitative data relating to an attribute may be obtained simply by noting its presence or absence in the objects, and then counting that how many do or do not possess that attribute.

15.2.1 Class and Class Frequency. A *class* is a set of the objects which are sharing a given characteristic. A *class frequency* is the number of observations (or objects) which are distributed in a class.

15.2.2 Classification of Objects. The objects (or individuals) can be divided into two distinct, mutually exclusive and complementary classes according to whether the objects do or do not possess a particular attribute. This process of dividing the objects into two mutually exclusive classes is called *dichotomy*.

If several attributes are noted, the process of classification may be continued indefinitely. The objects that are classified according to as they do or do not possess the first attribute can further be subdivided according to as they do or do not possess the second attribute and the objects of each of these subclasses can still further be subdivided according to as they do or do not possess the third attribute, and so on, every class being divided into two subclasses at each step. For example, the members of the population of district Lahore may be classified according to sex as males or females; the members of each sex may be further subdivided according to marital status as married or unmarried; that results into the married males, unmarried males, married females or unmarried females; the members of these four classes may be still further subdivided according to educational status as literate or illiterate.

15.2.3 Notations and Terminology. For theoretical study it is necessary to have some notations to represent different classes and their class frequencies. The capital Latin letters A, B, $\cdots$ are used to denote the attributes and their presence. The Greek letters α, β, $\cdots$ are used to denote the absence of these attributes. Thus A will denote that the object possesses the attribute A and α will denote that the object does not possess the attribute A; B will denote that the object possesses the attribute B, and β will denote that the object does not possess the attribute B. Hence " α " means "not A", "β" means " not B ".

Class frequencies will be denoted by enclosing the class by symbols in brackets. Thus (A) denotes the number of objects possessing the attribute A; (αB) denotes the number of objects possessing the attribute B but not the attribute A.

The attributes denoted by A, B, $\cdots$ are called *positive attributes* and their contraries denoted by α, β, $\cdots$ are called *negative attributes*. Thus the classes A, B and AB represented by positive attributes are called *positive classes*; the classes α, β and $\alpha\beta$ represented by negative attributes are called *negative classes*; and the classes $A\beta$, αB, *etc.* represented by both positive as well as negative attributes are called *contrary classes*.

15.2.4 Order of Classes. *Order of class* is known by the number of attributes specifying the class, *e. g.*, a class specified by one attribute is known as the class of order 1, the classes specified by two attributes are called as the classes of order 2; and the classes specified by three attributes are known as the classes of order 3. The total number of observations denoted by n is called the frequency of the class of order zero since no attributes are specified.

In the study of only one attribute A, we have the following frequencies

Frequency of the class of order zero : n

Frequencies of the classes of order 1 : (A), (α)

In the study of two attributes A and B, we have the following frequencies

Frequency of the class of order zero : n

Frequencies of the classes of order 1 : (A), (α), (B), (β)

Frequencies of the classes of order 2 : (AB), $(A\beta)$, (αB), $(\alpha\beta)$

These observed frequencies can be expressed in the form of a 2 × 2 table as

	Attribute B		
Attribute A	B	β	Total
A	(AB)	$(A\beta)$	(A)
α	(αB)	$(\alpha\beta)$	(α)
Total	(B)	(β)	n

15.2.5 Number of Class Frequencies. If we include the total number of observations n as a frequency of the class of order zero, then in general, for k attributes the total number of class frequencies would be $(3)^k$. Thus in case of only one attribute the total number of class frequencies would be $(3)^1 = 3$; for two attributes it is $(3)^2 = 9$, and so on.

15.2.6 Ultimate Class Frequency. The frequencies of classes of the highest order are called *ultimate class frequencies.* The number of ultimate class frequencies for k attributes is given by $(2)^k$. Thus in case of two attributes the number of ultimate classes is $(2)^2 = 4$, and so on.

If n is included as a positive class, then for k attributes the number of positive classes is the same as the number of ultimate classes. For two attributes, the positive classes are n, (A), (B), (AB) and the ultimate classes are (AB), $(A\beta)$, (αB), $(\alpha\beta)$.

It is interesting to note a very simple result that any class frequency can always be expressed in terms of the class frequencies of higher order. Any class can always be expressed as a sum of its two subclasses produced by dichotomizing it for the study of a new characteristic. For example, in the study of two attributes, we may have:

$$\begin{array}{rcl|rcl}
n & = & (A) + (\alpha) & n & = & (B) + (\beta) \\
(A) & = & (AB) + (A\beta) & (B) & = & (AB) + (\alpha B) \\
(\alpha) & = & (\alpha B) + (\alpha\beta) & (\beta) & = & (A\beta) + (\alpha\beta)
\end{array}$$

15.2.7 Consistence of data. The class frequencies that have been observed in one and the same population are said to be *consistent*, if they conform with one another and do not conflict each other. In the study of attributes, no class frequency can ever be negative. If any class frequency is negative the data are said to be inconsistent. Inconsistency may be due to wrong counting, inaccurate additions or subtractions or due to misprints. The necessary and sufficient condition for the consistence of a set of class frequencies is that no ultimate class frequency should be negative. To test the consistence of data, we calculate the ultimate class frequencies from the given data and if any of the ultimate class frequencies turns out to be negative, data will be *inconsistent.* If no ultimate class frequency is negative, the data are consistent. It is however important to note that the consistence of data is no proof of accurate count, accurate additions or subtractions or the absence of misprints.

15.3 INDEPENDENCE OF ATTRIBUTES

If in a sample of size n, the class frequencies of attributes A, B and AB are represented by (A), (B) and (AB). Then we have

$$\text{Proportion of individuals possessing } A = \frac{(A)}{n}$$

$$\text{Proportion of individuals possessing } B = \frac{(B)}{n}$$

$$\text{Proportion of individuals possessing } AB = \frac{(AB)}{n}$$

The two attributes A and B are said to be independent if,

$$\text{Proportion of } AB = (\text{Proportion of } A)(\text{Proportion of } B)$$

$$\frac{(AB)}{n} = \frac{(A)}{n} \times \frac{(B)}{n}$$

$$(AB) = \frac{(A)(B)}{n}$$

In case of independence of attributes A and B, the 2×2 table must have the form

	Attribute B		
Attribute A	B	β	Total
A	$\frac{(A)(B)}{n}$	$\frac{(A)(\beta)}{n}$	(A)
α	$\frac{(\alpha)(B)}{n}$	$\frac{(\alpha)(\beta)}{n}$	(α)
Total	(B)	(β)	n

***Example* 15.1** *If there are* **144** *A's and* **384** *B's in* **1024** *observations. How many (i) AB's and (ii) $\alpha\beta$'s will there be for A and B* **being independent.**

Solution. We have $n = 1024, \quad (A) = 144, \quad (B) = 384$

For A and B being independent, we must have

(*i*) $$(AB) = \frac{(A)(B)}{n} = \frac{(144)(384)}{1024} = 54$$

(*ii*) $$(\alpha) = n - (A) = 1024 - 144 = 880$$

$$(\beta) = n - (B) = 1024 - 384 = 640$$

$$(\alpha\beta) = \frac{(\alpha)(\beta)}{n} = \frac{(880)(640)}{1024} = 550$$

***Example* 15.2** *If the A's are* **60%**, *the B's are* **40%**, *of the whole number of observations, what must be the percentage of AB's in order that we may conclude that A and B are independent?*

Solution. Let $n = 100$, then $(A) = 60, \quad (B) = 40$

For A and B being independent, we must have

$$(AB) = \frac{(A)(B)}{n} = \frac{60 \times 40}{100} = 24$$

There must be 24% AB's to justify the conclusion that A and B are independent.

***Example* 15.3** *Given the following data. Find whether A and B are independent or associated.*

(*i*) $n = 150, \quad (A) = 30, \quad (B) = 60, \quad (AB) = 12$

(*ii*) $(AB) = 256, \quad (\alpha\beta) = 144, \quad (A\beta) = 48, \quad (\alpha B) = 768$

Solution.

(*i*) Observed frequency of AB's $= (AB) = 12$

$$\text{Expected frequency of } AB\text{'s} = \frac{(A)(B)}{n} = \frac{(30)(60)}{150} = 12$$

Since $(AB) = \frac{(A)(B)}{n}$, the attributes A and B are independent.

(ii) We have the 2×2 table as

Attribute A	Attribute B B	β	Total
A	$(AB) = 256$	$(A\beta) = 48$	$(A) = 304$
α	$(\alpha B) = 768$	$(\alpha\beta) = 144$	$(\alpha) = 912$
Total	$(B) = 1024$	$(\beta) = 192$	$n = 1216$

Observed frequency of AB's $= (AB) = 256$

Expected frequency of AB's $= \dfrac{(A)(B)}{n} = \dfrac{(304)(1024)}{1216} = 256$

Since $(AB) = \dfrac{(A)(B)}{n}$, the attributes A and B are independent.

15.4 ASSOCIATION OF ATTRIBUTES (*CORRELATION OF QUALITATIVE VARIABLES*)

The two attributes A and B are said to be *associated* if they are not independent, *i. e.,*

$$(AB) \neq \frac{(A)(B)}{n}$$

Association of attributes may be classified as positive or negative.

15.4.1 Positive Association. The two attributes A and B are *positively associated* or *simply associated*, if

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

15.4.2 Negative Association. The two attributes A and B are *negatively associated* or *simply disassociated*, if

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

It should be noted that disassociation does not imply independence.

15.4.3 Complete Association and Disassociation. There will be complete (or perfect positive) association between two attributes A and B if one of them cannot occur without the other, though the other may occur without the one, that is, if

(i) $(A) = (B)$ ⇒ all A's are B's and all B's are A's

(ii) $(A) < (B)$ ⇒ all A's are B's

(iii) $(B) < (A)$ ⇒ all B's are A's

There will be complete disassociation (or perfect negative association) between two attributes A and B if none of A's is B's and none of α's is β's.

15.4.4 Coefficient of Association. The strength of association, between two attributes A and B, is known as *coefficient of association.*

The Yule's coefficient of association, denoted by Q, is defined as :

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

This coefficient lies between -1 and $+1$.

If $Q = 0$, the two attributes are independent.

If $Q = 1$, the two attributes are completely associated.

If $Q = -1$, the two attributes are completely disassociated.

***Example* 15.4** *Given the following* :

$(AB) = 110, \quad (\alpha B) = 90, \quad (A\beta) = 290, \quad (\alpha\beta) = 510$

Discuss association.

Solution. We have the 2×2 table as

Attribute A	Attribute B B	 β	Total
A	$(AB) = 110$	$(A\beta) = 290$	$(A) = 400$
α	$(\alpha B) = 90$	$(\alpha\beta) = 510$	$(\alpha) = 600$
Total	$(B) = 200$	$(\beta) = 800$	$n = 1000$

Observed frequency of AB's $= (AB) = 110$

$$\text{Expected frequency of } AB\text{'s} = \frac{(A)(B)}{n} = \frac{(400)(200)}{1000} = 80$$

Since $(AB) > \dfrac{(A)(B)}{n}$, the attributes A and B are positively associated.

***Example* 15.5** *1660 candidates appeared for a competitive examination and 422 were successful. 256 had attended a coaching class and of these 150 came out successful. Find the coefficient of association between success and coaching a class.*

Solution. Let A represent success and B represent attending coaching class, then we have

$n = 1660, \quad (A) = 422, \quad (B) = 256, \quad (AB) = 150$

Attribute A	Attribute B B	 β	Total
A	$(AB) = 150$	$(A\beta) = 422 - 150 = 272$	$(A) = 422$
α	$(\alpha B) = 256 - 150 = 106$	$(\alpha\beta) = 1404 - 272 = 1132$	$(\alpha) = 1660 - 422 = 1238$
Total	$(B) = 256$	$(\beta) = 1660 - 256 = 1404$	$n = 1660$

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

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

Exercise 15.1

1. **(a)** Distinguish between attribute and variable. Define positive classes; negative classes and ultimate classes.

(b) Given the following ultimate class frequencies, find the frequencies of the positive and negative classes and the whole number of observations n.

$(AB) = 95, \quad (A\beta) = 55, \quad (\alpha B) = 85, \quad (\alpha\beta) = 45$

$\{ n = 280, \ (A) = 150, \ (B) = 180, \ (\alpha) = 130, \ (\beta) = 100 \}$

2. **(a)** Given the following frequencies of the positive classes, find the frequencies of ultimate classes.

$n = 250, \quad (A) = 80, \quad (B) = 100, \quad (AB) = 70.$

$\{ (A\beta) = 10, \ (\alpha B) = 30, \ (\alpha\beta) = 140, \ (AB) = 70 \}$

(b) Measurements are made on a thousand husbands and a thousand wives. If the measurements of husbands exceed the measurements of the wives in 800 cases for one measurements, in 700 cases for another, and in 660 cases for both measurements, in how many cases will both the measurements on the wife exceed the measurement on the husband?

(160)

(c) Given that $(A) = (\alpha) = (B) = (\beta) = n/2$, show that

(*i*) $(AB) = (\alpha\beta)$ (*ii*) $(A\beta) = (\alpha B)$

3. **(a)** Define the consistence of the data.

(b) Find whether the data given below in each case are consistent?

(*i*) $n = 120, \ (A) = 82, \ (AB) = 90$

(*ii*) $n = 50, \ (A) = 40, \ (B) = 32, \ (AB) = 15$

(*iii*) $n = 1000, \ (AB) = 200, \ (A\beta) = 350, \ (\alpha B) = 500$

{ (*i*) Not consistent since $(A\beta) = -8$, (*ii*) Not consistent since $(\alpha\beta) = -7$, (*iii*) Not consistent since $(\alpha\beta) = -50$ }

(c) Comment on the following data contained in a report: 100 students appeared in a test of whom 80 passed in Statistics: 70 passed in Mathematics and 48 passed in both the subjects.

{ Not consistent, since $(\alpha\beta) = -2$ }

4. **(a)** What is meant by independence of attributes.

(b) There is 240 A's and 270 B's in 600 observations. What would be the number of AB if A and B are independent.

$\{ (AB) = 108 \}$

(c) If A's are 60% and B's are 40% of the whole number of observations, what must be the percentage of AB's in order that we conclude that A and B are independent.

{ AB's are 24% }

5. **(a)** When are two attributes independent, positively associated, negatively associated?

(b) Given the following data, determine the nature of association between the attributes A and B, *i. e.*, find whether A and B are independent, positively associated or negatively associated.

(*i*) $(A) = 30, \ (B) = 60, \ (AB) = 12, \ n = 150$

(*ii*) $(AB) = 110, \ (\alpha B) = 90, \ (A\beta) = 290, \ (\alpha\beta) = 510$

(*iii*) $(A) = 415, \ (AB) = 147, \ (\alpha) = 285, \ (\alpha\beta) = 170$

{ (*i*) Independent, (*ii*) Positively associated, (*iii*) Negative associated }

6. (*a*) What is meant by association of attributes?

(*b*) Explain the difference between the following with examples.

(*i*) Attribute and variable,

(*ii*) Correlation and association,

(*iii*) Positive association and negative association

7. (*a*) Find the association between injection against typhoid and exemption from attack from the following contingency table

Attribute	Attacked	Not attacked
Inoculated	528	25
Not inoculated	790	175

$(Q = 0.65)$

(*b*) Calculate the coefficient of the association between the intelligence of fathers and sons in the following data:

Intelligent fathers with intelligent sons = 265

Intelligent fathers with dull sons = 100

Dull fathers with intelligent sons = 95

Dull fathers with dull sons = 450

$(Q = 0.85)$

(*c*) Find if there is any association between the tempers of bothers and sisters from the following data :

Good natured bothers and good natured sisters = 1230

Good natured bothers and sullen sisters = 850

Sullen bothers and good natured sisters = 530

Sullen bothers and sullen sisters = 980

$(Q = 0.46)$

8. (*a*) 750 students appeared in an examination and 470 were successful. 465 had attended classes and 58 of them failed. Calculate the coefficient of association to discuss association between attending classes and success.

($Q = 0.92$, highly positive association)

(*b*) 100 students appeared in an examination, and 50 failed in Mathematics, 60 failed in Statistics and 40 failed in both. Find if there is any association between the failing in Mathematics and Statistics.

$(Q = 0.71)$

(*c*) Can vaccination be regarded as preventive measure for small pox from the following data: "Of 1482 persons in a locality exposed to small pox, 368 in all were attacked. Of 1482 persons, 343 persons, had been vaccinated and of these 35 were attacked".

$(Q = -0.57)$

15.5 TWO DIMENSIONAL COUNT DATA: CONTINGENCY TABLE

A simple random sample of n elements selected from a bivariate multinomial population that has been classified into r categories $A_1, A_2, \cdots, A_r$ of attribute A and c categories $B_1, B_2, \cdots, B_c$ of attribute B will produce a two-way frequency table which is called an $r \times c$ *contingency table* — a name due to Karl Pearson. A contingency table is made up of the observed frequencies relative to the two attributes and their categories which is generally presented in the following tabular form, with rows representing the r categories $A_1, A_2, \cdots, A_r$ of attribute A and columns representing c categories $B_1, B_2, \cdots, B_c$ of attribute B.

15.5.1 Cell Frequency. The number of observations falling in a particular cell is called the *cell frequency*.

An $r \times c$ Contingency Table

	Attribute B					
Attribute A	B_1	B_2	B_j	$\cdots$	B_c	Row total
A_1	o_{11}	o_{12}	o_{1j}	$\cdots$	o_{1c}	$o_{1\cdot}$
A_2	o_{21}	o_{22}	o_{2j}	$\cdots$	o_{2c}	$o_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
A_i	o_{i1}	o_{i2}	o_{ij}	$\cdots$	o_{ic}	$o_{i\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
A_r	o_{r1}	o_{r2}	o_{rj}	$\cdots$	o_{rc}	$o_{r\cdot}$
Column total	$o_{\cdot 1}$	$o_{\cdot 2}$	$o_{\cdot j}$	$\cdots$	$o_{\cdot c}$	n

The table shows, in all, $k = r\,c$ cells or categories. The symbol O_{ij} denotes the number of sample observations in the (i, j) category of attributes A and B, respectively, for $i = 1, 2, \cdots, r$ and $j = 1, 2, \cdots, c$. The entries in the table represent the realizations o_{ij} the observed frequencies of the random variables O_{ij}. Note that the i-th row total is the observed frequency of the i-th category of attribute A summed over all categories of attribute B. Similarly, the j-th column total is the observed frequency of the j-th category of attribute B summed over all categories of attribute A. Let

$$o_{i\cdot} = \sum_{j=1}^{c} o_{ij} \qquad \text{for } i = 1, 2, \cdots, r$$

$$o_{\cdot j} = \sum_{i=1}^{r} o_{ij} \qquad \text{for } j = 1, 2, \cdots, c$$

denote the row and column sums, respectively, where the "dot" notation indicates the subscript over which summation has taken place. That is

o_{ij} = Observed frequency of $A_i \cap B_j$

$o_{i\cdot}$ = Observed frequency of A_i, *i. e.*, *i*-th row total

$o_{\cdot j}$ = Observed frequency of B_j, *i. e.*, *j*-th column total

$$\sum_{i=1}^{r}\sum_{j=1}^{c} o_{ij} = \sum_{i=1}^{r} o_{i\cdot} = \sum_{j=1}^{c} o_{\cdot j} = n$$

15.6 TEST FOR STATISTICAL INDEPENDENCE

In analyzing bivariate multinomial populations, the first-step of a typical inferential aspect of interest usually is whether the two attributes are statistically independent or whether certain levels of one attribute tend to be associated or contingent with some levels of another attribute. If they are independent, we know that there is no relationship between them. If it turns out that they are not independent and a relationship does exit between the two attributes, the next step in the analysis then is to study the nature of the relationship. We begin with the first step of the analysis, testing whether or not the two attributes are independent.

We are concerned with testing the null hypothesis that the two criteria of classification are independent. Recall, if two classifications are independent of each other, a cell probability will equal the product of its respective row and column probabilities in accordance with multiplicative law of probability. Therefore, the null hypothesis stating that the events $A_1, A_2, \cdots, A_r$ are independent of events $B_1, B_2, \cdots, B_c$ can be rephrased

$$P(A_i \cap B_j) = P(A_i)P(B_j) \qquad \text{for all } i = 1, 2, \cdots, r \text{ and } j = 1, 2, \cdots, c.$$

Thus the null and alternative hypotheses for a test of statistical independence are

Null hypothesis H_0: A_i and B_j are independent for all cells (i, j)

Alternative hypothesis H_1: A_i and B_j are not independent for some (i, j)

Here, H_0 represents statistical independence and H_1 represent statistical dependence.

The problem now becomes testing the goodness of fit for the model of independence. We compare the observed frequencies o_{ij} with the expected frequencies $E(O_{ij})$ that are expected if the attributes are independent. Under the null hypothesis of independence of attributes the estimate of expected frequency $E(O_{ij})$ is

$$e_{ij} = \frac{o_{i\cdot}\, o_{\cdot j}}{n} = \frac{(i\text{-th row total})(j\text{-th column total})}{\text{number of observations}}$$

The test statistic then becomes

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

which has an approximate chi-square distribution with $\nu = (r - 1)(c - 1)$ degrees of freedom for large n.

In this case, we base the test statistic on the expected number of elements e_{ij} in the sample from each category if H_0 is true and the postulated proportions hold. The farther the observed frequency o_{ij} departs in either direction from the expected frequency e_{ij}, the large is $(o_{ij} - e_{ij})^2$ and hence the larger is χ^2. On the other hand, if there is perfect agreement between the observed and expected frequencies, *i. e.*, o_{ij} and e_{ij} are identical for all classes, $\chi^2 = 0$ because each $(o_{ij} - e_{ij})^2 = 0$. If all the observed frequencies o_{ij} are close to the expected frequencies e_{ij} supporting H_0, the value of χ^2 will be near to zero; if o_{ij} are far from e_{ij}, indicating rejection of H_0, the χ^2 will assume a large positive value. It follows, therefore, that for a given level of significance α the critical region is the upper tail of chi-square distribution with $\nu = (r - 1)(c - 1)$ degrees of freedom, *i. e.*,

Critical region: $\chi^2 > \chi^2_{\nu;1-\alpha}$

On the question how large a sample size should be, we know that this test is based on the normal approximation to the binomial, a fairly conservative rule of thumb is that the approximation is adequate if each $e_{ij} \geq 5$. If there are not at least 5 items, the value of chi-square is inflated because squared differences are divided by a very small size expected frequency in $\chi^2 = \Sigma\{(o_{ij} - e_{ij})^2 / e_{ij}\}$. However, if the cells have too small expected frequencies the condition of at least 5 items in each expected frequency class can also be accomplished by combining neighbouring row or column class, but for pair of rows or columns that is combined the number of rows or columns for degrees of freedom is reduced by one.

15.6.1 Assumptions. To conduct a valid test of hypothesis for independence using data from a contingency table, the following conditions must be met.

(*i*) A simple random sample of size n has been selected from a bivariate multinomial population.

(*ii*) The sample size n is reasonably large so that for each cell, the estimated expected frequency must be at least 5.

15.6.2 Yates' Correction. To improve the approximation to the χ^2 distribution and thus be able to obtain a more exact probability value from the χ^2 table, F. Yates has proposed a *correction for continuity*, applicable when the criterion has a single degree of freedom. The correction is intended to make the actual distribution of the criterion, as calculated from discrete data, more nearly like the χ^2 distribution based on normal deviations. The relation $Z^2 = \chi^2$ between Z and χ^2 holds only for a single degree of freedom. The approximation calls for the absolute value of each deviation to be decreased by $1/2$, because for two celled tables, the deviations are always equal in magnitude but opposite in sign. Therefore

$$\text{Adjusted } \chi^2 = \sum \frac{(|o - e| - 0.5)^2}{e}$$

Thus Yates' correction is analogous to the continuity correction which is applied in the normal approximation to the binomial distribution. There is a tendency to under estimate the probability, which means that the probability of rejecting the hypothesis will be increased. Adjustment results in a lower chi-square. Consequently, in testing hypothesis, it is worthwhile only when unadjusted χ^2 is greater than tabulated χ^2 at the desired probability level. When n (or e) is large continuity correction has little effect, but when e's are small, it should be applied. When $|o - e|$ is less than 0.5, the continuity correction should be omitted.

15.6.3 Coefficient of Contingency. The *coefficient of contingency* is a measure of the strength of association on a numerical scale as an index of association between two criteria of classification. When the test for statistical independence leads to the conclusion of dependence, we may wish to measure the strength of association between two criteria of classification. Insofar as the χ^2 statistic represents an over all deviation from the model of independence, it is intuitively reasonable to use this statistic to gauge the strength of this association. We may call χ^2 as the "square contingency". But in applying the χ^2 statistic as a measure of association the limitation is that the number of degrees of freedom attached to this statistic depends upon the dimensionality of the contingency table. A χ^2 value of 16.5 in a 2×2 contingency table would reflect a significant association, but this would not be so in a 6×8 contingency table. Several measures of association have been proposed to adjust the χ^2 statistic to a common scale that is irrespective of the dimensionality of the contingency table. We then write

$$\phi^2 = \frac{\chi^2}{n}$$

and call ϕ^2 as the "*mean square contingency*". In the following are two commonly used formulas, large values of a measure indicate a strong association and small values of a measure indicate a weak association between the two criteria of classification.

Pearson's coefficient of mean square contingency:

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

where q represents the number of rows or columns, whichever is smaller, and n indicates the sample size.

***Example* 15.6** *Four hundred and ninety two candidates for scientific posts gave particulars of their university degrees and their hobbies. The degrees were in either mathematics, chemistry or physics and the hobbies could be classified roughly as music, craftwork, reading or drama. The data are presented concisely in the following contingency table.*

	Degree		
Hobby	*Mathematics*	*Chemistry*	*Physics*
Music	24	83	17
Craftwork	11	62	28
Reading	32	121	34
Drama	10	26	44

Discuss the association between the two criteria of classification, i.e., the degrees and hobbies. If the null hypothesis of independence is rejected, calculate the Pearson's coefficient of mean square contingency. What could be its maximum value for this contingency table.

Solution. The elements of the one-sided right tail test of hypothesis are

Null hypothesis H_0: The degree and hobby are independent.

Alternative hypothesis H_1: The degree and hobby are not independent.

Level of significance: $\alpha = 0.05 \quad \Rightarrow \quad 1 - \alpha = 0.95$

Test statistic: $\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$ follows an approximate chi-square distribution under H_0 with

Degrees of freedom: $\nu = (r - 1)(c - 1) = (4 - 1)(3 - 1) = 6$,

Critical value: $\chi^2_{\nu;1-\alpha} = \chi^2_{6;0.95} = 12.59$ (From Table 11)

Critical region: $\chi^2 > 12.59$

Decision rule: Reject H_0 if $\chi^2 > 12.59$, otherwise do not reject H_0.

Observed value: The observed frequencies o_{ij} are :

Hobby	Degree Mathematics: B_1	Chemistry: B_2	Physics: B_3	Row total
Music: A_1	24	83	17	$o_{1\cdot} = 124$
Craftwork: A_2	11	62	28	$o_{2\cdot} = 101$
Reading: A_3	32	121	34	$o_{3\cdot} = 187$
Drama: A_4	10	26	44	$o_{4\cdot} = 80$
Column total	$o_{\cdot 1} = 77$	$o_{\cdot 2} = 292$	$o_{\cdot 3} = 123$	$n = 492$

The expected frequencies e_{ij} under the null hypothesis of independence are

$$e_{ij} = \frac{o_{i\cdot}\, o_{\cdot j}}{n} = \frac{(i\text{-th row total})(j\text{-th column total})}{\text{number of observations}}$$

which are given in the following table. Only $(r - 1)(c - 1) = (4 - 1)(3 - 1) = 6$ expected frequencies are obtained through this procedure. We could work through this procedure to give the other expected frequencies, but this is unnecessary, as the remaining frequencies can be found by using the fact that the sub-totals and totals must agree with those in observed data:

Hobby	Degree Mathematics: B_1	Chemistry: B_2	Physics: B_3	Row total
Music: A_1	$\frac{(124)(77)}{492} = 19.4$	$\frac{(124)(292)}{492} = 73.6$	31.0	124
Craftwork: A_2	$\frac{(101)(77)}{492} = 15.8$	$\frac{(101)(292)}{492} = 59.9$	25.3	101
Reading: A_3	$\frac{(187)(77)}{492} = 29.3$	$\frac{(187)(292)}{492} = 111.0$	46.7	187
Drama: A_4	12.5	47.5	20.0	80
Column total	77	292	123	492

The χ^2 -statistic is calculated as under

Cell (i, j)	Observed frequency o_{ij}	Expected frequency e_{ij}	$\frac{(o_{ij} - e_{ij})^2}{e_{ij}}$
A_1B_1	24	19.4	1.09
A_1B_2	83	73.6	1.20
A_1B_3	17	31.0	6.32
A_2B_1	11	15.8	1.46
A_2B_2	62	59.9	0.07
A_2B_3	28	25.3	0.29
A_3B_1	32	29.3	0.25
A_3B_2	121	111.0	0.90
A_3B_3	34	46.7	3.45
A_4B_1	10	12.5	0.50
A_4B_2	26	47.5	9.73
A_4B_3	44	20.0	28.80
Total	492	492	$\chi^2 = 54.06$

Conclusion: Since $\chi^2 = 54.06 > 12.59$, we reject H_0 and conclude that the two criteria of classification are association.

Pearson's coefficient of mean square contingency:

$$C = \sqrt{\frac{\chi^2}{n + \chi^2}} = \sqrt{\frac{54.06}{492 + 54.06}} = 0.315$$

Maximum value of C for this contingency table:

$$\sqrt{\frac{q-1}{q}} = \sqrt{\frac{3-1}{3}} = 0.8165$$

Example 15.7 *Discuss the resemblances of stature of parents and off-springs for the following data*

	Parents			
Off-springs	*Very tall*	*Tall*	*Medium*	*Short*
Very tall	20	30	20	2
Tall	14	125	85	12
Medium	3	140	165	125
Short	3	37	68	151

Solution. The elements of the one-sided right tail test of hypothesis are

Null hypothesis: H_0: The stature of off-springs is independent of the stature of parents.

Alternative hypothesis: H_1: The stature of off-springs is not independent of the stature of parents.

Level of significance: $\alpha = 0.05 \Rightarrow 1 - \alpha = 0.95$

Test statistic: $\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$ follows an approximate chi-square distribution under H_0 with

Degrees of freedom: $\nu = (r - 1)(c - 1) = (4 - 1)(3 - 1) = 6$, since first two column are pooled because $e_{11} = 2.88 < 5$.

Critical value: $\chi^2_{\nu;\,1-\alpha} = \chi^2_{6;\,0.95} = 12.59$ (From Table 11)

Critical region: $\chi^2 > 12.59$

Decision rule: Reject H_0 if $\chi^2 > 12.59$, otherwise do not reject H_0.

Observed value: The observed frequencies o_{ij} are given in the following contingency table:

	Parents				
Off-springs	Very tall: B_1	Tall: B_2	Medium: B_3	Short: B_4	Row total
Very tall: A_1	20	30	20	2	$o_{1\cdot} = 72$
Tall: A_2	14	125	85	12	$o_{2\cdot} = 236$
Medium: A_3	3	140	165	125	$o_{3\cdot} = 433$
Short: A_4	3	37	68	151	$o_{4\cdot} = 259$
Column total	$o_{\cdot 1} = 40$	$o_{\cdot 2} = 332$	$o_{\cdot 3} = 338$	$o_{\cdot 4} = 290$	$n = 1000$

The expected frequencies e_{ij} under the null hypothesis of independence are

$$e_{ij} = \frac{o_{i\cdot}\; o_{\cdot j}}{n} = \frac{(\text{ } i\text{-th row total) (} j\text{-th column total)}}{\text{number of observations}}$$

which are given in the following table. Only $(r - 1)(c - 1) = (4 - 1)(4 - 1) = 9$ expected frequencies are obtained through this procedure. We could work through this procedure to give the other expected frequencies, but this is unnecessary, as the remaining frequencies can be found by using the fact that the sub-totals and totals must agree with those in observed data.

	Parents				
Off-springs	Very tall: B_1	Tall: B_2	Medium: B_3	Short B_4	Row total
Very tall: A_1	$\frac{(72)(40)}{1000} = 2.88$	$\frac{(72)(332)}{1000} = 23.90$	$\frac{(72)(338)}{1000} = 24.34$	20.88	72
Tall: A_2	$\frac{(236)(40)}{1000} = 9.44$	$\frac{(236)(332)}{1000} = 78.35$	$\frac{(236)(338)}{1000} = 79.77$	68.44	236
Medium: A_3	$\frac{(433)(40)}{1000} = 17.32$	$\frac{(433)(332)}{1000} = 143.76$	$\frac{(433)(338)}{1000} = 146.35$	125.57	433
Short: A_4	10.36	85.99	87.54	75.11	259
Column total	40	332	338	290	1000

Combining the first and second columns of the expected frequencies, we get

	Parents		
Off spring	Tall: B_1	Medium: B_2	Short: B_3
Very tall: A_1	2.88 + 23.90 = 26.78	24.34	20.88
Tall: A_2	9.44 + 78.35 = 87.79	79.77	68.44
Medium: A_3	17.32 + 143.76 = 161.08	146.35	125.57
Short: A_4	10.36 + 85.99 = 96.35	87.54	75.11

Accordingly, we combined the observed frequencies as under

	Parents		
Off spring	Tall: B_1	Medium: B_2	Short: B_3
Very tall: A_1	20 + 30 = 50	20	2
Tall: A_2	14 + 125 = 139	85	12
Medium: A_3	3 + 140 = 143	165	125
Short: A_4	3 + 37 = 40	68	151

The χ^2 statistic is calculated as under

Cell (i,j)	Observed frequency o_{ij}	Expected frequency e_{ij}	$\frac{(o_{ij}-e_{ij})^2}{e_{ij}}$
$A_1 B_1$	50	26.78	20.13
$A_1 B_2$	20	24.34	0.77
$A_1 B_3$	2	20.88	17.07
$A_2 B_1$	139	87.79	29.87
$A_2 B_2$	85	79.77	0.34
$A_2 B_3$	12	68.44	46.54
$A_3 B_1$	143	161.08	2.03
$A_3 B_2$	165	146.35	2.38
$A_3 B_3$	125	125.57	0.00
$A_4 B_1$	40	96.35	32.96
$A_4 B_2$	68	87.54	4.36
$A_4 B_3$	151	75.11	76.68
Total	1000	1000	$\chi^2 = 233.13$

Conclusion: Since $\chi^2 = 233.13 > 12.59$, we reject H_0 and conclude that the Stature of off-springs is not independent of the stature of the parents.

***Pearson's coefficient of mean square contingency*:**

$$C = \sqrt{\frac{\chi^2}{n+\chi^2}} = \sqrt{\frac{233.13}{1000 + 233.13}} = 0.435$$

Example 15.8 *A random sample of 30 adults is classified according to the sex and the number of hours they watch television during a week :*

	Sex	
Time watching television	*Male*	*Female*
Over 25 hours	5	8
Under 25 hours	10	7

Using $\alpha = 0.01$ *test the hypothesis that a person's sex and time watching television are independent.*

Solution. The elements of the one-sided right tail test of hypothesis are

Null hypothesis H_0: The sex and time watching television are independent.

Alternative hypothesis H_1: The sex and time watching television are not independent.

Level of significance: $\alpha = 0.01 \quad \Rightarrow \quad 1 - \alpha = 0.99$

Test statistic: $$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(\left| o_{ij} - e_{ij} \right| - 0.5)^2}{e_{ij}}$$ follows an approximate chi-square distribution under H_0 with

Degrees of freedom: $\nu = (r-1)(c-1) = (2-1)(2-1) = 1$

Critical value: $\chi^2_{\nu;\,1-\alpha} = \chi^2_{1;\,0.99} = 6.63$

Critical region: $\chi^2 > 6.63$

Decision rule: Reject H_0 if $\chi^2 > 6.63$, otherwise do not reject H_0.

Observed value: The observed frequencies o_{ij} are given in the following table:

	Sex		
Time watching television	Male: B_1	Female: B_2	Row total
Over 25 hours: A_1	5	8	$o_{1\cdot} = 13$
Under 25 hours: A_2	10	7	$o_{2\cdot} = 17$
Column total	$o_{\cdot 1} = 15$	$o_{\cdot 2} = 15$	$n = 30$

The expected frequencies e_{ij} under the null hypothesis are

$$e_{ij} = \frac{o_{i\cdot}\; o_{\cdot j}}{n} = \frac{(i\text{-th row total})(j\text{-th column total})}{\text{number of observations}}$$

	Sex		
Time watching television	Male: B_1	Female: B_2	Row total
Over 25 hours: A_1	$\dfrac{(13)(15)}{30} = 6.5$	6.5	13
Under 25 hours: A_2	8.5	8.5	17
Column total	15	15	30

Now we calculate the χ^2 statistic as under

Category (i, j)	Observed frequency o_{ij}	Expected frequency e_{ij}	$\dfrac{(\lvert o_{ij} - e_{ij} \rvert - 0.5)^2}{e_{ij}}$
$A_1\ B_1$	5	6.5	0.154
$A_1\ B_2$	8	6.5	0.154
$A_2\ B_1$	10	8.5	0.118
$A_2\ B_2$	7	8.5	0.118
Total	30	30.0	$\chi^2 = 0.544$

Conclusion: Since $\chi^2 = 0.544 < 6.63$, we do not reject H_0: $\pi_{ij} = \pi_{i.}\,\pi_{.j}$ for all (i, j) against H_1: $\pi_{ij} \neq \pi_{i.}\,\pi_{.j}$ for at least one (i, j).

Exercise 15.2

1. (*a*) Define contingency table and cell frequency. What is a 2×2 contingency table.

(*b*) In an investigation into eye-colour and left or right handedness of a person, the following results were obtained:

	Handedness	
Eye colour	Left	Right
Blue	15	85
Brown	20	80

Do these results indicate, at the 5% level of significance, an association between eye colour and left or right handedness.

(Since *Adj* $\chi^2 = 0.56 < 3.84 = \chi^2_{1;\,0.95}$, we do not reject H_0: There is no association between eye colour and left or right handedness against H_1: There is association between eye colour and handedness.)

(*c*) An investigation into colour-blindness and sex of a person gave the following results:

	Colourblindness	
Sex	Colourblind	Not colourblind
Male	36	964
Female	19	981

Is there evidence, at the 5% level, of an association between the sex of a person and whether or not they are colourblind?

(Since *Adj* $\chi^2 = 4.79 > 3.84 = \chi^2_{1;\,0.95}$, we reject H_0: There is no association between sex of a person and colour-blindness in favour of H_1: There is association between sex of a person and colour-blindness.)

2. (*a*) A driving school examined the results of 100 candidates who were taking their driving test for the first time. They found that out of the 40 men, 28 passed and out of the 60 women, 34 passed. Do these results indicate, at the 5% level of significance, a

relationship between the sex of a candidate and the ability to pass first time ?

(Since *Adj* $\chi^2 = 1.290 < 3.84 = \chi^2_{1;\,0.95}$, we do not reject H_0: There is no relationship between the sex of a candidate and the ability to pass first time against H_1: There is relationship between the sex and ability to pass.)

(*b*) Out of 1350 persons, 450 were literate and 600 had traveled beyond the limits of their district, 300 of the literates were among those who had traveled. Find out by calculating (*i*) coefficient of association, (*ii*) the value of chi-square, if there is any association between traveling and literacy.

($Q = 0.6$, Since *Adj* $\chi^2 = 133.65 > 3.84 = \chi^2_{1;\,0.95}$, we reject H_0: There is no association between traveling and literacy in favour of H_1: There is association between traveling and literacy.)

3. (*a*) The following are the data on a random sample of 150 chickens, divided into two groups according to breed and into three group according to yield of eggs.

	Yield		
Breed	High	Medium	Low
Rhode Red	46	29	28
Leghorn White	27	14	6

Are these data consistent with the hypothesis that yield is not affected by the type of breed?

(Since $\chi^2 = 4.07 < 5.99 = \chi^2_{2;\,0.95}$, we do not reject H_0: There is no association between chicken breed and yield of eggs against H_1: There is association between chicken breed and yield of eggs.)

(*b*) The students of a college took three courses : arts, commerce and science. The students were classified according to the sex. The data on these students are given as follows :

	Course of study		
Sex	Arts	Commerce	Science
Male	200	300	100
Female	100	200	100

Use chi-square test whether there is any association between sex and choice of course of study.

(Since $\chi^2 = 13.888 > 5.99 = \chi^2_{2;\,0.95}$, we reject H_0: There is no association between sex and course of study in favour of H_1: There is association between sex and course of study.)

4. (*a*) The following table shows liking of three colours: pink, white and blue in samples of males and females:

	Sex	
Colour	Male	Female
Pink	20	40
White	40	20
Blue	60	20

Test whether there is any relation between sex and colour.

(Since $\chi^2 = 26.3889 > 5.99 = \chi^2_{2;\,0.95}$, we reject H_0: There is no relation between sex and liking of colours in favour of H_1: There is relation between sex and liking of colours.)

(b) The following table gives the condition at home and condition of the children.

	Condition at home	
Condition of children	Clean	Not clean
Clean	175	143
Fairly clean	136	116
Dirty	125	145

Test for the association between the conditions at home and condition of children.

(Since $\chi^2 = 5.027 < 5.99 = \chi^2_{2;\,0.95}$, we do not reject H_0: There is no association between conditions at home and condition of children against H_1: There is no association between conditions at home and condition of children.)

5. (a) The table given below shows the relation between the performance of students in economics and statistics. Test the hypothesis that the performance in economics is independent of the performance in statistics using 5% level of significance :

	Grade in statistics		
Grade in economics	High	Medium	Low
High	56	96	28
Medium	48	168	24
Low	16	86	78

(Since $\chi^2 = 89.2112 > 9.49 = \chi^2_{4;\,0.95}$, we reject H_0: There is no association between the performance of students in economics and statistics against H_1: There is association between the performance in economics and statistics.)

(b) A thousand households are taken at random and divided into three groups A, B and C, according to the total weekly income. The following table shows the numbers in each group having a colour television receive, a black and white receiver, or no television at all.

	Income group		
Television type	A	B	C
Colour television	56	51	93
Black and white	118	207	375
None	26	42	32

Calculate the expected frequencies if there is no association between total income and television ownership. Apply a test to find whether the observed frequencies suggest that there is such an association.

(Since $\chi^2 = 26.6 > 9.49 = \chi^2_{4;\,0.95}$, we reject H_0: There is no association between television type and income group in favour of H_1: There is association between television type and income group.)

6. (*a*) A random sample of 200 married men, all retired, were classified according to education and number of children as indicated below :

	Number of children		
Education	0 — 1	2 — 3	Over 3
Elementary	13	37	35
Secondary	19	42	14
College	12	17	11

Test the hypothesis, at 5% of significance, that the size of family is independent of the level of education attained by the father.

(Since $\chi^2 = 11.7194 > 9.49 = \chi^2_{4;\,0.95}$, we reject H_0: There is no association between education and number of children in favour, of H_1: There is association between education and number of children.)

(*b*) A survey of 200 families known to be regular television viewers was undertaken. They were asked which of the three television channel they watched most during an average week. A summary of their replies is given in the following table, together with the region in which they lived.

	Region			
Channel	North	East	South	West
PTV 1	29	16	42	23
PTV 2	6	11	26	7
STN	15	3	12	10

Test the hypothesis that there is no association between the channel watched most and the region.

(Since $\chi^2 = 13.446 > 12.59 = \chi^2_{6;\,0.95}$, we reject H_0: There is no association between the channel and region in favour of H_1: There is association between the channel and region.)

(*c*) From the following table showing the number of employees and condition of factory.

Condition of premises	Number of persons employed			
	Under 50	51 — 150	151 — 250	Over 250
A_1	84	133	49	62
A_2	87	82	20	25
A_3	26	9	9	5

Discuss the association between the condition of premises and the number of persons employed. Compute the coefficient of contingency.

(Since $\chi^2 = 30.06 > 12.59 = \chi^2_{6;\,0.95}$, we reject H_0: There is no association between the condition of premises and the number of persons employed in favour of H_1: There is association between the condition of premises and the number of persons employed. $C = 0.22$)

15.7 RANK CORRELATION.

The correlation between ranks of individuals for both the variables X and Y is called *rank correlation*. A special case of correlation is when both the variables X and Y consist of sets of ranks. Suppose, for example, that two judges have ranked the same set of n objects according to some characteristic of interest. We are interested in determining whether the ranks assigned to the objects by one judge are related to or show any agreement with ranks assigned to the same objects by another judge.

15.7.1 Derivation of Spearman's Coefficient of Rank Correlation. Suppose that we have a sample of n individuals from a continuous bivariate population and two measurements for variables X and Y are made on each individual. We have n pairs of observations (a_1, b_1), $(a_2, b_2), \cdots, (a_n, b_n)$. These values for two variables can be ranked in separate ordered series. Let $x_1, x_2, \cdots, x_n$ be the ranks of $a_1, a_2, \cdots, a_n$ and $y_1, y_2, \cdots, y_n$ be the ranks of $b_1, b_2, \cdots, b_n$. The coefficient of rank correlation r_r is the ordinary correlation coefficient between the two sets of ranks. Then the coefficient of rank correlation is

$$r_r = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

The r_r always lies between -1 and $+1$. This formula is called Spearman's coefficient of rank correlation, in the honour of Charles Edward Spearman. Spearman's rank correlation coefficient is equivalent to Pearson's product moment correlation coefficient computed for ranks rather than the original observations. This nonparametric procedure can be useful in correlation analysis even when the basic data are not available in the form of numerical magnitudes but when the ranks can be assigned. The ranks may be assigned in order from high to low, with 1 representing the highest, 2 the next highest, *etc.* (or in order from low to high, with 1 representing the lowest, 2 the next lowest, *etc.*).

***Example* 15.9** *Using Spearman's formula calculate coefficient of rank correlation for the following data giving ranks to the measured quantities.*

a_i	4.7	2.9	6.4	2.5	4.9	7.3
b_i	8.6	5.4	6.2	4.9	8.3	7.2

Solution. The coefficient of rank correlation is obtained as

Measurements		Ranks			
a_i	b_i	x_i	y_i	$d_i = x_i - y_i$	d_i^2
4.7	8.6	4	1	3	9
2.9	5.4	5	5	0	0
6.4	6.2	2	4	-2	4
2.5	4.9	6	6	0	0
4.9	8.3	3	2	1	1
7.3	7.2	1	3	-2	4
Sum					18

$$r_r = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6(18)}{6\{(6)^2-1\}} = 0.4857$$

Exercise 15.3

1. (*a*) What is rank correlation?

(*b*) The following table shows how 10 students, arranged in alphabetical order, were ranked according to their achievements in both laboratory and lecture portions of a statistics course. Find the coefficient of mark correlation.

Laboratory	8	3	9	2	7	10	4	6	1	5
Lecture	9	5	10	1	8	7	3	4	2	6

(r_r = 0.8545)

(*c*) The ranks of the same 10 students in Mathematics and Economics were as follows: (1, 6); (2, 5); (3, 1); (4, 4); (5, 2); (6, 7); (7, 8); (8, 10); (9, 3); (10, 9); the two numbers within brackets denoting the ranks of the same students in Mathematics, and Economics respectively. Calculate the rank correlation coefficient for proficiencies of this group in two subjects.

(r_r = 0.45)

3. (*a*) Five sacks of coal *A*, *B*, *C*, *D* and *E* have different weights, with *A* being heavier than *B*, *B* being heavier than *C*, and so on. A weight lifter ranks the sacks (heaviest first) in the order *A*, *D*, *B*, *E*, *C*. Calculate a coefficient of rank correlation.

(r_r = 0.5)

(*b*) Seven army recruits *A*, *B*, *C*, *D*, *E*, *F* and *G* were given two separate aptitude tests. Their orders of merit in each test were

Order of merit	1st	2nd	3rd	4th	5th	6th	7th
First test	*G*	*F*	*A*	*D*	*B*	*C*	*E*
Second test	*D*	*F*	*E*	*B*	*G*	*C*	*A*

Find Spearman's coefficient of rank correlation between the two orders and comment briefly on the correlation obtained.

(r_r = − 0.036, Very little negative correlation)

4. (*a*) Ten competitors in a beauty contest are ranked by three judges in the following order

Competitor	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*
Judge *X*	1	6	5	10	3	2	4	9	7	8
Judge *Y*	3	5	8	4	7	10	2	1	6	9
Judge *Z*	6	4	9	8	1	2	3	10	5	7

Use Spearman's rank correlation coefficient to discuss which pair of judges have the nearest approach to common tastes in beauty.

(r_{xy} = − 0.21, r_{yz} = − 0.30, r_{xz} = 0.64; This indicates that judges the *X* and *Z* have the nearest approach to common tastes in beauty.)

(*b*) The following table shows the grade point average awarded to six children in a competition by two different judges.

Child	*A*	*B*	*C*	*D*	*E*	*F*
Judge *X*	6.8	7.3	8.1	9.8	7.1	9.2
Judge *Y*	7.8	9.4	7.9	9.6	8.9	6.9

Calculate coefficient of rank correlation by Spearman's formula.
(r_r = 0.26)

(*c*) The following table shows the marks of six candidates in two subjects.

Candidate	*A*	*B*	*C*	*D*	*E*	*F*
Mathematics x_i	38	62	56	42	59	48
Statistics y_i	64	89	84	60	73	69

(*i*) Calculate the coefficient of rank correlation.

(*ii*) Comment on the value of your result.

{ (*i*) 0.886, (*ii*) High positive correlation }

Exercise 15.4
Objective Questions

1. Fill in the blanks.

(*i*) A characteristic which varies in quantity from one individual to another is called a ———. (*variable*)

(*ii*) A characteristic which varies in quality from one individual to another is called an ———. (*attribute*)

(*iii*) The observations made on objects regarding an attribute are called ——— data. (*qualitative*)

(*iv*) ——— is a process of dividing the objects into two mutually exclusive classes of an attribute. (*Dichotomy*)

(*v*) The degree of linear relationship between the two variables is called ———. (*correlation*)

(*vi*) The degree of relationship between the two attributes is called ———. (*association*)

(*vii*) The two attribures *A* and *B* are ———, if

$$(AB) = \frac{(A)(B)}{n}$$ (*independent*)

(*viii*) The two attribures *A* and *B* are ———, if

$$(AB) \neq \frac{(A)(B)}{n}$$ (*associated*)

(*ix*) The two attributes A and B are ——— associated, if
$$(AB) > \frac{(A)(B)}{n}$$ (*positively*)

(*x*) The two attributes A and B are ——— associated, if
$$(AB) < \frac{(A)(B)}{n}$$ (*negatively*)

(*xi*) The coefficient of association, denoted by Q, is a measure of association between the two ———. (*attributes*)

(*xii*) If the coefficient of association equals 0, the two attributes A and B are ———. (*independent*)

(*xiii*) If the coefficient of association is not equal to 0, the two attributes A and B are ———. (*associated*)

(*xiv*) If the coefficient of association equals -1, the two attributes A and B are completely ———. (*dissociated*)

(*xv*) If the coefficient of association equals 1, the two attributes A and B are completely ———. (*associated*)

(*xvi*) A ——— table consisting of r rows and c columns is made up of the observed frequencies relative to two attributes and their categories. (*contingency*)

(*xvii*) The two attributes are said to be ———, if for every cell of a contingency table the observed frequency o_{ij} is equal to expected frequency e_{ij}. (*independent*)

(*xx*) For an $r \times c$ contingency table, the χ^2-statistic has degrees of freedom $\nu =$ ———. $(r-1)(c-1)$

(*xxiv*) The larger are the difference between the observed and expected frequencies, the larger will be the value of χ^2 which leads to the ——— of H_0 of independence. (*rejection*)

(*xxv*) The rejection of H_0 of independence indicates that the two criteria of classification are ———. (*associated*)

(*xxvi*) In a chi-square test for independence, no expected frequency should be ——— than 5. (*less*)

2. Mark off the following statements as false or true.

(*i*) A characteristic which varies in quantity from one individual to another is called an attribute. (*false*)

(*ii*) The quantitative data relating to an attribute may be obtained simply by noting its presence or absence in the objects. (*true*)

(*iii*) The presence of attributes is denoted by capital Latin letters and their absence by Greek or small letters. (*true*)

(*iv*) The class frequencies of the highest order are called ultimate class frequencies. (*true*) (*v*) The two attributes A and B are associated, if. (*false*)

$$(AB) = \frac{(A)(B)}{n}$$

(*vi*) The two attributes A and B are positively associated, if (*false*)

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

(*vii*) The coefficient of correlation, denoted by r, is a measure of the strength of linear relationship between two variables. (*true*)

(*viii*) The coefficient of association, denoted by Q, is a measure of association between the two attributes. (*true*)

(*ix*) The coefficient of association always lies between -1 and 1. (*true*)

(*x*) A contingency table consisting of r rows and c columns is made up of the observed frequencies relative to two attributes and their categories. (*true*)

(*xi*) The disassociation of two attributes means their independence (*false*)

(*xii*) A measure of the discrepancy between the observed and expected frequencies is called a chi-square (χ^2) test of independence. (*true*)

(*xiii*) The value of χ^2-statistic is always non-negative. (*true*)

(*xiv*) The larger are the differences between the observed and expected frequencies, the larger will be the value of χ^2 which leads to rejection of H_0 of independence. (*true*)

(*xv*) The rejection of H_0 of independence indicates that two criteria of classification are associated. (*true*)