Chapter 15

ASSOCIATION

15.1 VARIABLE AND ATTRIBUTE

There are 4 persons and their heights in inches are 55, 56, 72 and 74. Here height is a characteristic and the figures 55, 56, 72 and 74 are the values of a variable. These figures are the result of measurements. You know that the measurements generate the continuous variable. Thus the variable on heights is a continuous variable. Suppose we select 4 bulbs from a certain lot and inspect them. The lot contains good as well as defective bulbs. The sample may contain 0, 1, 2, 3, 4 defective bulbs. The values 0, 1, 2, 3 and 4 are the values of a discrete variable.

Out of 4 persons whose heights are given above, 2 are tall with heights 72 and 74 inches and 2 are short with heights 55 and 56 inches. When we use the words, tall and short, any variable is not under consideration. We do not make any measurements. We only see who is tall and who is short. Here level of height *tall* or *short* is not a variable, it is called an *attribute*. Out of 4 bulbs 2 are good and 2 are defective. Here also any variable is not under consideration. We only count the defective bulbs and good bulbs. We examine whether the quality of being defective is present in a bulb or not. The status of the bulb is an attribute with two outcomes good and defective. Thus attribute is a quality and the data is collected to see how many objects possess the quality of being defective and how many elements do not possess this quality. Other famous examples of the attributes are level of education, level of smoking, level of social work, level of income, religion and colour etc. The data on the attribute is the result of recording the presence and absence of a certain quality (attribute) in the individuals. The data on the variables are called the quantitative data whereas the data on the attributes are called qualitative data or count data. As the data on the variables is collected for the purpose of analysis of data and for inference about the population parameters, similarly the data on the attribute or attributes is collected for the purpose of analysis of data and for testing of hypotheses about the attributes. We shall discuss the hypothesis testing about attributes in the subsequent topic in this Chapter.

15.1.1 NOTATION FOR ATTRIBUTES

For a single variable we use the symbol X and if there are two variables, we use the symbols X and Y for them. When there is a single attribute like height, the word, 'tall' may be denoted by A and 'short' may be denoted by α. If the tall and the short persons are divided into intelligent and 'non-intelligent' persons, then 'intelligent' may be denoted by B and β may be used for the opposite attribute 'non-intelligent'. It may be noted that the word *attribute* is used for the main group like intelligence and the sub-groups 'intelligent' and 'non-intelligent' are also called attributes.

15.1.2 ONE ATTRIBUTE

Suppose that there are 100 individuals in a certain sample, the sample size is denoted by n. These 100 individuals are divided into two mutually exclusive groups on the basis of the attribute of height. Out of 100, 60 are tall and 40 are short. If 'tall' are denoted by A and short are denoted by α, we can write:

A	α	
60	40	n = 100

There are two groups and we say that there are two classes A and α and the class frequency under A is 60. It is written as (A) = 60, similarly the number of individuals under α is written as (α) = 40. Thus the attributes written within the brackets show their class frequencies. In this example the sample is divided into two groups i.e; two classes 'tall' and 'short'. Dividing the data into two groups is called *dichotomy* which means *cutting into two*. In this example a single attribute '*height*' divides the data in two groups. As only one attribute is involved, the data is called *one-way classification*. We can make a small table as below:

One-Way Classification

A	α	
60 = (A)	40 = (α)	n = 100

Clearly (A) + (α) = n

The symbols (A) and (α) are used to denote the frequency of individuals who possess A and who do not possess A (α means not 'A'). It may be noted that the symbol 'A' is not necessarily fixed for 'tall'. In some other discussion 'short' may be denoted by A.

15.1.3 TWO ATTRIBUTES

The tall and short persons may further be divided into intelligent and non-intelligent persons. Intelligence may be denoted by B and β may be used for non-intelligence. The following table shows different attributes and their combinations. When two attributes are involved, the division of the sample as below is called two-way classification.

Table 15.1.

Two-Way Classification

	A	α	Total
B	(AB)	(αB)	(B)
β	(Aβ)	(αβ)	(β)
Total	(A)	(α)	n

The column totals are denoted by (A) and (α) and the row totals are denoted by (B) and (β). The above table contains 2 rows and 2 columns and is therefore called 2 × 2 *contingency table* or 2 × 2 cross-tabulation briefly written as 2 × 2 cross-table.

There may be more than two attributes. The symbols A,. B, C are used for the attributes and α, β, γ are used for the absence of the attributes A, B, C. Thus α means not A and β means not B and γ means not C.

Suppose that out of 60 tall persons, 30 are intelligent and out of 40 short persons, 20 are intelligent. We can write these frequencies in the following 2 × 2 contingency Table 15.2.

Table 15.2.

2 × 2 Contingency Table

	A	α	Total
B	(AB) = 30	(αB) = 20	(B) = 50
β	(Aβ) = 30	(αβ) = 20	(β) = 50
Total	(A) = 60	(α) = 40	n = 100

From table 15.2. we can write some relations immediately.

(i) $(A) + (\alpha) = n$ (ii) $(B) + (\beta) = n$

(iii) $(A) = (AB) + (A\beta)$ (iv) $(\alpha) = (\alpha B) + (\alpha\beta)$

(v) $(B) = (AB) + (\alpha B)$ (vi) $(\beta) = (A\beta) + (\alpha\beta)$

15.1.4 POSITIVE AND NEGATIVE CLASSES

The classes A, B, AB are called positive classes because they contain all positive attributes, the classes α, β, αβ are called negative classes because they have negative attributes. The classes αB and Aβ contain both positive and negative attributes, they are called mixed or contrary classes.

If we have three attributes A, B, C with their opponents or complements as α, β and γ, then we can write the different class frequencies as below in Table 15.3.

Table 15.3.

	A		α		
	C	γ	C	γ	Total
B	(ABC)	(ABγ)	(αBC)	(αBγ)	(B)
β	(AβC)	(Aβγ)	(αβC)	(αβγ)	(β)
	(AC)	(Aγ)	(αC)	(αγ)	
Total	(A)		(α)		n

In this table the positive classes are A, B, C, AB, AC, BC and ABC whereas the negative classes are α, β, γ, $\alpha\beta$, $\alpha\gamma$, $\beta\gamma$, $\alpha\beta\gamma$. All other classes are mixed.

15.1.5 ORDER OF CLASSES

The order of the class depends upon the number of attributes going into that class. If a certain class can give us information about only one attribute, it is called class of order one.

The classes A, α, B and β are classes of the order one and the frequencies (A), (α), (B) and (β) are the frequencies of order one. The classes AB, Aβ, αB and $\alpha\beta$ are the classes of order two and the frequencies (AB), (Aβ), (αB) and ($\alpha\beta$) are the frequencies of order two. The classes ABC, ABγ, αBC, αBγ, AβC A$\beta\gamma$, $\alpha\beta$C and $\alpha\beta\gamma$ are the classes of order three and (ABC), (ABγ) ··· ($\alpha\beta\gamma$) are the frequencies of order three. The sample size n does not contain any attribute and is therefore called frequency of order zero.

15.1.6 ULTIMATE CLASS FREQUENCIES

In a certain given situation, the *ultimate class frequencies* are the frequencies with the highest order. For two attributes, the ultimate class frequencies are (AB), (Aβ), (αB) and ($\alpha\beta$).

For three attributes, the ultimate class frequencies are of order 3 which are (ABC), (ABγ), (αBC), (αBγ), (AβC), (A$\beta\gamma$) ($\alpha\beta$C) and ($\alpha\beta\gamma$).

15.1.7 LOWER ORDER FREQUENCIES IN TERMS OF HIGHER ORDER FREQUENCIES

Let us discuss the relation of the lower order frequencies in terms of higher order frequencies. Let us consider the different cases.

(i) Single Attribute

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

(ii) Two Attributes

Let us consider 2 × 2 contingency Table 15.2. for the frequencies of the two attributes. Clearly

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

(iii) Three Attributes

Let us take help from Table 15.3. to write lower order frequencies into higher order frequencies. Clearly

$$n = (A) + (\alpha) \qquad n = (B) + (\beta)$$

$$(A) = (AC) + (A\gamma). \quad \text{But } (AC) = (ABC) + (A\beta C) \text{ and } (A\gamma) = (AB\gamma) + (A\beta\gamma)$$

$$\text{Thus} \quad (A) = (ABC) + (A\beta C) + (AB\gamma) + (A\beta\gamma)$$

$$\text{Similarly } (\alpha) = (\alpha C) + (\alpha\gamma) = (\alpha BC) + (\alpha\beta C) + (\alpha B\gamma) + (\alpha\beta\gamma)$$

$$n = (A)+(\alpha) = (ABC)+(A\beta C)+(AB\gamma)+(A\beta\gamma)+(\alpha BC)+(\alpha\beta C)+(\alpha B\gamma)+(\alpha\beta\gamma)$$

$$(B) = (AB) + (\alpha B) = (ABC) + (AB\gamma) + (\alpha BC) + (\alpha B\gamma)$$

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

$$\text{and} \quad n = (ABC) + (AB\gamma) + (\alpha BC) + (\alpha B\gamma) + (A\beta C) + (A\beta\gamma) + (\alpha\beta C) + (\alpha\beta\gamma)$$

With the help of the Tables 15.2. and 15.3. we can easily write any lower order frequency in terms of higher orders.

15.1.8 HIGHER ORDER FREQUENCIES INTO LOWER ORDER FREQUENCIES

Sometimes we have to express the frequency of a higher order into frequencies of lower order. For this purpose we use the following operators. The frequency (A) is written as n · A as if n · A means A's out of n. Similarly the frequency (α) is written as n . α and (AB) is written as n . AB and (ABC) is written as n . ABC.

We know $(A) + (\alpha) \;=\; n$

$$n \cdot A + n \cdot \alpha = n \qquad \cdots\cdots \quad \text{equation (1)}$$

using the operators, we shall assume that algebraic operations are applicable on these operators. Dividing equation (1) by n, we get

$$A + \alpha = 1 \quad \text{or} \quad A = 1 - \alpha \quad \text{and} \quad \alpha = 1 - A$$

Similarly we can establish with the help of operators that

$$B = 1 - \beta \text{ and } \beta = 1 - B \qquad C = 1 - \gamma \text{ and } \gamma = 1 - C$$

Example 15.1.

Express (AB) in terms of lower order frequencies with the help of operators.

Solution:

We write $(AB) = n \cdot AB$

Putting $A = 1 - \alpha$ and $B = 1 - \beta$

$$(AB) = n(1 - \alpha)(1 - \beta) = n\,[1 - \beta - \alpha + \alpha\beta] = n - n\beta - n\alpha + n\alpha\beta$$

Writing the original symbols for nβ, nα and n$\alpha\beta$, we have

$$(AB) = n - (\beta) - (\alpha) + (\alpha\beta) \quad \text{or} \quad (AB) = n - (\alpha) - (\beta) + (\alpha\beta)$$

It is to be noted that the left hand side contains positive attributes and all attributes on the right side are negative except one frequency which is n. Any attribute on the left side does not appear on the right side in this type of relation.

Example 15.2.

Express ($\alpha\beta\gamma$) in terms of lower order frequencies.

Solution:

($\alpha\beta\gamma$) can be written as n · $\alpha\beta\gamma$. Thus $(\alpha\beta\gamma) = n \cdot \alpha\beta\gamma$

Using the relations $\alpha = 1 - A,\ \beta = 1 - B,\ \gamma = 1 - C$ we get

$$\begin{aligned} (\alpha\beta\gamma) &= n(1 - A)(1 - B)(1 - C) \\ &= n - n \cdot A - n \cdot B - n \cdot C + n \cdot AB + n \cdot AC + n \cdot BC - n \cdot ABC \\ (\alpha\beta\gamma) &= n - (A) - (B) - (C) + (AB) + (AC) + (BC) - (ABC) \end{aligned}$$

The attributes on the left side are negative and all attributes on the right side are positive except one frequency of order zero that is n.

Example 15.3.

Given the following frequencies: n = 100, (AB) = 30, (A) = 40, (B) = 70. Calculate all the remaining frequencies.

Solution:

We know $(A) + (\alpha) = n$. Thus $40 + (\alpha) = 100$ or $(\alpha) = 60$

We know $(B) + (\beta) = n$, hence $70 + (\beta) = 100$ or $(\beta) = 30$

Also $(B) = (AB) + (\alpha B)$, hence $70 = 30 + (\alpha B)$ or $(\alpha B) = 40$

Also $(AB) + (A\beta) = (A)$, hence $30 + (A\beta) = 40$ or $(A\beta) = 10$

Also $(\beta) = (A\beta) + (\alpha\beta)$, hence $30 = 10 + (\alpha\beta)$ or $(\alpha\beta) = 20$

These frequencies can be calculated very easily if the given frequencies are substituted in the 2 × 2 contingency table. The unknown frequencies can be caleulated by simple addition or subtraction. Thus

	A	α	Total
B	(AB) = 30	$(\alpha B) = \boxed{40}$	(B) = 70
β	$(A\beta) = \boxed{10}$	$(\alpha\beta) = \boxed{20}$	$(\beta) = \boxed{30}$
Total	(A) = 40	$(\alpha) = \boxed{60}$	n = 100

The unknown frequencies within the rectangles have been calculated by simple subtraction to complete the table.

Example 15.4.

There are three attributes and their ultimate class frequencies are:

$(ABC) = 10$ $(AB\gamma) = 30$ $(\alpha BC) = 15$ $(\alpha B\gamma) = 60$

$(A\beta C) = 20$ $(A\beta\gamma) = 15$ $(\alpha\beta C) = 40$ $(\alpha\beta\gamma) = 70$

Calculate all the negative class frequencies of order one and order two.

Solution:

$(\alpha) = (\alpha BC) + (\alpha B\gamma) + (\alpha\beta C) + (\alpha\beta\gamma) = 15 + 60 + 40 + 70 = 185$

$(\beta) = (A\beta C) + (A\beta\gamma) + (\alpha\beta C) + (\alpha\beta\gamma) = 20 + 15 + 40 + 70 = 145$

$(\gamma) = (AB\gamma) + (A\beta\gamma) + (\alpha B\gamma) + (\alpha\beta\gamma) = 30 + 15 + 60 + 70 = 175$

$(\alpha\beta) = (\alpha\beta C) + (\alpha\beta\gamma) = 40 + 70 = 110$

$(\alpha\gamma) = (\alpha B\gamma) + (\alpha\beta\gamma) = 60 + 70 = 130$

$(\beta\gamma) = (A\beta\gamma) + (\alpha\beta\gamma) = 15 + 70 = 85$

These unknown frequencies can be calculated with the help of the following table.

	A		α		Total
	C	γ	C	γ	
B	(ABC) = 10	(ABγ) = 30	(αBC) = 15	(αBγ) = 60	(B) = 115
β	(AβC) = 20	(Aβγ) = 15	(αβC) = 40	(αβγ) = 70	(β) = 145
	(AC) = 30	(Aγ) = 45	(αC) = 55	(αγ) = 130	
Total	(A) = 75		(α) = 185		n = 260

Clearly $(\alpha) = 185$ $\qquad (\beta) = 145$

$(\gamma) = 30 + 15 + 60 + 70 = 175$ $\qquad (\alpha\beta) = 40 + 70 = 110$

$(\alpha\gamma) = 60 + 70 = 130$ $\qquad (\beta\gamma) = 15 + 70 = 85$

15.2 CONSISTENCY

If the class frequencies are observed in a certain sample data and all class frequencies are recorded correctly then there will be no error in them and they will be called consistent. But sometimes the class frequencies are not recorded correctly and their column total and row total do not agree with the grand total. If there is some error in any class frequency, then we say that the frequencies are inconsistent. If one class frequency is wrong, it will affect some other frequencies as well. A simple test of consistency is that all frequencies should be positive. If any frequency is negative, it means that there is inconsistency in the sample data. If the data is consistent, all the ultimate class frequencies will be positive.

Example 15.5.

Given the frequencies: n = 115, (B) = 45, (A) = 50 and (AB) = 50.

Check for consistency of the data.

Solution:

The data is called consistent if all the ultimate class frequencies are positive. Let us calculate some frequencies of order two.

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

Here $(A) = 50$ and $(AB) = 50$

Thus $50 = 50 + (A\beta)$ or $(A\beta) = 0$

It does not indicate inconsistency because some frequency can be zero.

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

$45 = 50 + (\alpha B)$ or $(\alpha B) = -5$

The data is inconsistent. It means the given frequencies are wrong. If we make a table of (2 × 2), we get

	A	α	Total
B	(AB) = 50	(αB) = – 5	(B) = 45
β	(Aβ) = 0	(αβ) = 70	(β) = 70
Total	(A) = 50	(α) = 65	n = 115

One frequency (αB) is negative in the table. Thus the sample data is inconsistent.

Example 15.6.

In a certain big college, 600 students of intermediate level were interviewed. They were asked to give their opinion about liking or disliking in the subjects of Mathematics, Statistics and Physics. The sample data sent by the enumerator was:

300 liked Mathematics.	350 liked Statistics.
340 liked Physics.	130 liked Mathematics and Statistics.
160 liked Mathematics and Physics.	180 liked Physics and Statistics.

100 liked all the three subjects. Examine the data for consistency.

Solution:

All the given frequencies can be written in the form of attributes. Let A, B, C denote liking Mathematics, Statistics and Physics respectively and α, β, γ are their opponents for disliking of the subjects. We are given

$$n = 600 \quad (A) = 300 \quad (B) = 350 \quad (C) = 340$$

$$(AB) = 130 \quad (AC) = 160 \quad (BC) = 180 \quad (ABC) = 100$$

All the given frequencies are positive, we can therefore calculate a negative class frequency of order three which is (αβγ).

$$\begin{aligned} \text{Now } (\alpha\beta\gamma) &= n \cdot \alpha\beta\gamma \\ &= n(1 - A)(1 - B)(1 - C) \\ &= n - (A) - (B) - (C) + (AB) + (AC) + (BC) - (ABC) \\ &= 600 - 300 - 350 - 340 + 130 + 160 + 180 - 100 = -20 \end{aligned}$$

A negative frequency indicates that the sample data sent by the enumerator is incorrect (inconsistent).

15.3 INDEPENDENCE OF ATTRIBUTES

Let us consider certain examples before we discuss the *independence* in a formal manner.

Example 15.7.

Consider the following sample data on the liking of males and females for fish.

Gender

	Males	Females	Total
Like Fish	80	80	160
Do not like Fish	20	20	40
Total	100	100	200

Discussion: There are 100 males out of which 80 like fish and out of 100 females 80 like fish. Males and females have the same liking for fish. We say that there is independence between gender and liking or disliking of fish. Another way of saying the same thing is that there is no relation between the gender and liking for fish.

Example 15.8.

Consider the following sample data on smoking by adult males and adult females:

Gender

	Males	Females	Total
Smokers	**20**	**1**	**21**
Non-smokers	80	99	179
Total	100	100	200

There are 100 males out of which 20 are smokers and out of 100 females there is only one smoker. It means that the smokers in males are 20 times more than the smokers among females. Males have a strong relation or association with smoking. Thus males and smoking are strongly associated. We say that there is positive association between males and smoking. There are 99 females who are non-smokers as compared to 80 male-non smokers. Thus females are inclined towards non-smoking. The association between females and non-smoking is also of positive type. There is only 1 female smoker as compared to 20 male smokers. Thus there is negative association between females and smoking and there is also negative association between males and non - smoking. Thus in a certain contingency table, when there is positive association between two attributes, then in the same table there exists the negative association between some other pairs of attributes. If there is positive association between A and B, then α and β are also positively associated. In this case there is negative association between A and β, and between α and B.

The data in the Example 15.8. may be written as

	Males A	Females α	Total
Non-smokers, B	80 (AB)	99 (αB)	179
Smokers, β	20 (Aβ)	1 ($\alpha\beta$)	21
Total	100	100	200

In this table 80 is less than 99 and 20 is greater than 1 (or 1 is less than 20). There is negative association between A and B and between α and β. There is positive association between A and β and between α and B. If the attributes in the one diagonal have positive association, then the attributes in the other diagonal have negative association.

15.3.1 DEFINITION OF INDEPENDENCE

We know that in probability, the two events A and B are called independent if the joint probability of A ∩ B is equal to the product of the marginal probabilities of A and B. Thus for independence of A and B

$$P(A \cap B) = P(A)\,P(B)$$

The same logic applies for defining independence of attributes. The two attributes are called independent if the probability of (AB) is equal to the product of the probability A and the probability of B. Consider a 2 × 2 contingency table as below:

	A	α	Total
B	(AB)	(αB)	(B)
β	(Aβ)	(αβ)	(β)
Total	(A)	(α)	n

If one individual is selected out of this table, then

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

For independence $P(AB) = P(A)\,.\,P(B)$

$$\frac{(AB)}{n} = \frac{(A)}{n}\,.\,\frac{(B)}{n} \quad \text{or} \quad (AB) = \frac{(A)\,(B)}{n}$$

This is called rule or criterion of independence of two attributes A and B. The class frequency (AB) is called observed frequency and $\frac{(A)\,(B)}{n}$ is called expected frequency when A and B are independent. For independence of A and B, the rule is $(AB) = \frac{(A)\,(B)}{n}$. But this rule is applicable only on the attributes A and B. Similarly for independence of other attributes, we have the rules:

$$(\alpha B) = \frac{(\alpha)\,(B)}{n} \qquad (A\beta) = \frac{(A)\,(\beta)}{n} \quad \text{and} \quad (\alpha\beta) = \frac{(\alpha)\,(\beta)}{n}$$

When $(AB) > \frac{(A)\,(B)}{n}$, then there is positive association between A and B.

Positive association between A and B means that proportion of A's in B's is greater than the proportion of A's in β's.

When $(AB) < \frac{(A)\,(B)}{n}$, there is negative association between A and B.

Negative association between A and B means that proportion of A's in B's is less than the proportion of A's in β's.

It is important to note that if A and B are associated in a positive manner, then α and β are also associated in the positive manner and other pairs Aβ and αB will have negative association.

15.3.2 ANOTHER DEFINITION OF INDEPENDENCE

The two attributes A and B are called independent if the proportion of A's in B's is the same as in non B's (β's).

$$\text{Proportion of A's in B's} = \frac{(AB)}{(B)} \quad \text{Proportion of A's in } \beta\text{'s} = \frac{(A\beta)}{(\beta)}$$

For independence these two proportions are equal.

$$\text{Thus} \quad \frac{(AB)}{(B)} = \frac{(A\beta)}{(\beta)} \quad \left[\text{If } \frac{a}{b} = \frac{c}{d}, \text{ then } \frac{a}{b} = \frac{c}{d} = \frac{a+c}{b+d}\right]$$

$$\text{Therefore} \quad \frac{(AB)}{(B)} = \frac{(A\beta)}{(\beta)} = \frac{(AB)+(A\beta)}{(B)+(\beta)} = \frac{(A)}{n}.$$

$$\text{Thus} \quad \frac{(AB)}{(B)} = \frac{(A)}{n} \text{ or } (AB) = \frac{(A)\,(B)}{n}.$$

This is called a simple rule of independence between A and B.

If A and B are independent then all the other pairs in the table are also independent. But if there is positive association between two pairs AB and αβ, then the other two pairs Aβ and αB will have negative association as explained earlier.

Let us consider the data of Example 15.7.

	A	α	Total
B	(AB) = 80	(αB) = 80	(B) = 160
β	(Aβ) = 20	(αβ) = 20	(β) = 40
Total	(A) = 100	(α) = 100	n = 200

$$\text{Here} \quad (AB) = 80 \quad \text{and} \quad \frac{(A)\,(B)}{n} = \frac{100 \times 160}{200} = 80$$

$$\text{Thus} \quad (AB) = \frac{(A)\,(B)}{n}.$$ Hence A and B are independent.

It also implies independence between A and β, α and B and α and β. Let us check another pair.

$$(\alpha\beta) = 20 \quad \text{and} \quad \frac{(\alpha)\,(\beta)}{n} = \frac{100 \times 40}{200} = 20$$

$$(\alpha\beta) = \frac{(\alpha)\,(\beta)}{n},$$ there is independence between α and β.

The students may check the other classes. The independence in this table means that men and women have the same liking for fish.

Example 15.9.

Men and women go to a certain store for buying the articles. They make the payment in cash or purchase on credit (loan). Investigate if there is any relation between mode of payment and the sex of the customer. Given the data below:

	Payment	
Sex	Cash	Credit
Males	80	40
Females	20	60

Solution:

Let us write the table along with the symbols

	A	α	Total
B	$(AB) = 80$	$(\alpha B) = 40$	$(B) = 120$
β	$(A\beta) = 20$	$(\alpha\beta) = 60$	$(\beta) = 80$
Total	$(A) = 100$	$(\alpha) = 100$	$n = 200$

$$(AB) = 80 \quad \text{and} \quad \frac{(A)\,(B)}{n} = \frac{100 \times 120}{200} = 60, \ (AB) > \frac{(A)\,(B)}{n}$$

There is positive association between A and B. It means that males make the cash payments with greater frequency than the females. If we check the pair (Aβ), we will find negative association.

$$(A\beta) = 20 \text{ and } \frac{(A)\,(\beta)}{n} = \frac{100 \times 80}{200} = 40, \ (A\beta) < \frac{(A)\,(\beta)}{n}$$

Thus there is negative association between A and β. Females are less inclined to make the cash payments. It is also clear from the given data. Out of 120 males, 80 make the payment on cash. 80 out of 120 means that $\frac{80}{120} \times 100 = 66.7\ \%$ males make cash payment. 20 out of 80 means that $\frac{20}{80} \times 100 = 25\ \%$ females make the cash payment. Thus males and cash payment go together with high frequency and are called positively related or associated.

Example 15.10.

We wish to determine if there is any difference in the popularity of football between college educated males and non college educated males. A sample of 100 college educated males showed that 55 were football fans. A sample of 200 non college educated males revealed that 125 were football fans. Is there any evidence of a difference in football popularity between college educated and non college educated males.

Solution:

We put the data in the following table.

	College educated males A	Non college educated males α	Total
Football fans, B	55 = (AB)	125	180 = (B)
Not football fans, β	45	75	120 = (β)
Total	100 (A)	200 (α)	300 = n

Here $(AB) = 55, \quad \frac{(A)\,(B)}{n} = \frac{100 \times 180}{300} = 60. \quad (AB) < \frac{(A)\,(B)}{n}$

Thus there is negative association between A and B. College-educated males show less of interest for football as compared to non college educated males. There is positive association between α and B. More of non college educated males are football fans as compared to college-educated males. Thus football is more popular among non college educated males. But here we are comparing only one observed frequency with the corresponding expected frequency. In Example 15.12, we shall compare all the observed frequencies with the corresponding expected frequencies. In Example 15.12, our inference will be different and we shall decide that whether there is independence between the attributes or not.

15.4 COEFFICIENT OF ASSOCIATION

When it is desired to calculate the level of association, we can calculate coefficient of association denoted by Q, where

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

This is called Yule's coefficient of association. It lies between − 1 and + 1. It is explained in the same manner as the coefficient of correlation r_{XY}. between the two random variable X and Y.

If $Q = -1$ it is perfect negative association between the attributes on the top left corner in the 2 × 2 cross table.

If $Q = 0$ it means independence

If $Q = 1$ it means perfect positive association between attributes.

Let us calculate Q from the data given in Example 15.10.

$$Q = \frac{(AB)\,(\alpha\beta) - (A\beta)\,(\alpha B)}{(AB)\,(\alpha\beta) + (A\beta)\,(\alpha B)} = \frac{55 \times 75 - 45 \times 125}{55 \times 75 + 45 \times 125} = \frac{4125 - 5625}{4125 + 5625} = \frac{-1500}{9750} = -0.15$$

This indicates negative association between A and B. It is the same result as obtained earlier in Example 15.10.

15.5 χ^2–DISTRIBUTION

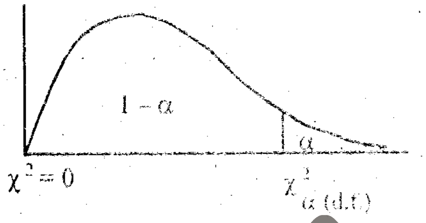


Figure 15.1

Chi-square written as χ^2 is a statistic which has a positively skewed distribution as shown below. The value of χ^2 varies from 0 to ∞. χ^2 cannot take any negative value. The shape of the distribution depends upon the degrees of freedom which is calculated from the given sample. χ^2-distribution can be used for various purposes. One of the applications of χ^2 is to test the independence between the attributes.

15.5.1 TEST OF INDEPENDENCE

With the help of χ^2-distribution, we can test whether the attributes are independent or there is association between them. The procedure runs as below:

1. The null hypothesis H_0 is framed

   We assume that there is independence between the attributes.

   The alternative hypothesis H_1 is that there is association between the attributes

2. Level of significance α is decided.

3. Test-statistic used is $\chi^2 = \Sigma\left(\frac{(f_0 - f_e)^2}{f_e}\right)$

   where f_0 stands for observed frequency and f_e stands for expected frequency calculated under the assumption that attributes are independent.

4. **Computations:**

   The χ^2-statistic can be used to check the independence in a table of attributes containing any number of columns and rows. Let us first explain the application of χ^2 on a 2 × 2 contingency table. The observed frequencies are given below in the form of a table. It is only for our convenience that we write the class frequencies in the form of a table having columns and rows.

2 × 2 Contingency Table

	A	α	Total
B	(AB)	(αB)	(B)
β	(Aβ)	(αβ)	(β)
Total	(A)	(α)	n

Our null hypothesis is that the attributes are independent. If A and B are independent then the observed frequency (AB) is equal to $\frac{(A)(B)}{n}$. By using this approach, we calculate the expected frequencies for all the four classes (AB), (Aβ), B, and (αβ). The expected frequencies are calculated under the assumption that null hypothesis is true.

Expected Frequencies Calculated

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

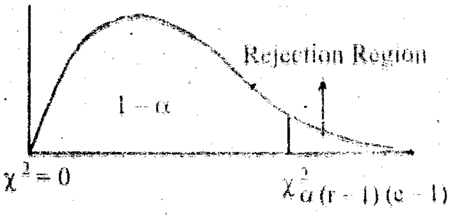
It may be noted that the column and row totals in the table of observed frequencies are the same as in the table of expected frequencies. The observed frequencies are denoted by f_0 and the expected frequencies are denoted by f_e. For the calculation of χ^2 we write the expected frequencies corresponding to their observed frequencies. The necessary calculations are done as shown in the following columns:

Observed frequencies f_0	Expected frequencies f_e	$f_0 - f_e$	$(f_0 - f_e)^2$	$\frac{(f_0 - f_e)^2}{f_e}$
(AB)	$\frac{(A)(B)}{n}$			
(Aβ)	$\frac{(A)(\beta)}{n}$			
(αB)	$\frac{(\alpha)(B)}{n}$			
(αβ)	$\frac{(\alpha)(\beta)}{n}$			
n	n			$\chi^2 = \Sigma\left(\frac{(f_0 - f_e)^2}{f_e}\right)$

5. Critical region:

The critical region in this test always lies in the right side of the distribution. It depends upon the level of significance α and the degrees of freedom. In tests of independence, the degree of freedom is calculated as below:

degrees of freedom (d.f.) = (r − 1) (c − 1)



where r is the number of rows and c is the number of columns in the contingency table. The critical value of χ^2 is seen from the table of χ^2. For level of significance α, and degrees of freedom (r − 1) (c − 1), the table value is denoted by $\chi^2_{\alpha(r-1)(c-1)}$. In a χ^2-table, under the column heading α and against d.f. = (r − 1) (c − 1) given in the left column, we read the value of $\chi^2_{\alpha\ (d.f.)}$. When $\alpha = 0.05$, d.f. = 1, then $\chi^2_{0.05(1)} = 3.841$.

6. Conclusion:

The hypothesis of independence is rejected if the calculated value of χ^2 lies in the rejection region. The rejection of hypothesis means that the attributes are associated.

Example 15.11.

Calculate χ^2 by using the data given in Example 15.7. to test the independence between the gender and liking for fish. Use $\alpha = 0.05$.

Solution:

The data of Example 15.7. is reproduced here

	Males	Females	Total
Like Fish	80	80	160
Do not like Fish	20	20	40
Total	100	100	200

We write the hypotheses as below:

1. H_0: There is independence between gender and liking for fish.
   H_1: There is association between gender and liking for fish.
2. Level of significance α is given, $\alpha = 0.05$.
3. Test − statistic used is χ^2 where $\chi^2 = \Sigma\left(\frac{(f_0 - f_e)^2}{f_e}\right)$
4. Computations:
   The given table of observed frequencies is written as

	A	α	Total
B	(AB) = 80	(αB) = 80	(B) = 160
β	(Aβ) = 20	($\alpha\beta$) = 20	(β) = 40
Total	(A) = 100	(α) = 100	n = 200

The corresponding expected frequencies are calculated as below:

	A	α	Total
B	$\frac{(A)\,(B)}{n} = \frac{100 \times 160}{200} = 80$	$\frac{(\alpha)\,(B)}{n} = \frac{100 \times 160}{200} = 80$	(B) = 160
β	$\frac{(A)\,(\beta)}{n} = \frac{100 \times 40}{200} = 20$	$\frac{(\alpha)\,(\beta)}{n} = \frac{100 \times 40}{200} = 20$	(β) = 40
Total	(A) = 100	(α) = 100	n = 200

The necessary columns are as below:

f_o	f_e	$f_o - f_e$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
80	80	0	0	0
20	20	0	0	0
80	80	0	0	0
20	20	0	0	0
200	200	0	0	0

Here d.f. = $(r-1)(c-1) = (2-1)(2-1) = 1$

5. **Critical region**: $\chi^2 > \chi^2_{0.05\,(1)} = 3.841$

6. **Conclusion:** The calculated value of $\chi^2 = 0$ which falls in the acceptance region. Thus hypothesis H_0 of independence is accepted. When $\chi^2 = 0$, it means perfect independence between the attributes. Males and females have exactly equal liking for eating fish.

Example 15.12.

Let us consider the data of Example 15.10. and calculate χ^2 to examine the independence between college education and liking for football.

Solution:

The data of Example 15.10. is written as below:

	College educated males A	Non college educated males α	Total
Football fans, B	$(AB) = 55$	$(\alpha B) = 125$	$(B) = 180$
Not football fans, β	$(A\beta) = 45$	$(\alpha\beta) = 75$	$(\beta) = 120$
Total	$(A) = 100$	$(\alpha) = 200$	$n = 300$

1. We frame the hypotheses as:

   H_0: There is independence between type of education and interest for football.

   H_1: There is association between type of education and their liking for football.

2. Level of significance, α is decided. Let $\alpha = 0.05$

3. Test – statistic used is χ^2 where $\chi^2 = \Sigma\left(\frac{(f_o - f_e)^2}{f_e}\right)$

4. Computations:

The expected frequencies under the assumption that H_o is true are calculated as below:

$$(AB) = \frac{(A)\,(B)}{n} = \frac{100 \times 180}{300} = 60 \qquad (A\beta) = \frac{(A)\,(\beta)}{n} = \frac{100 \times 120}{300} = 40$$

$$(\alpha B) = \frac{(\alpha)\,(B)}{n} = \frac{200 \times 180}{300} = 120 \qquad (\alpha\beta) = \frac{(\alpha)\,(\beta)}{n} = \frac{200 \times 120}{300} = 80$$

Calculation of χ^2

f_o	f_e	$(f_o - f_e)$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
55	60	– 5	25	0.4167
45	40	5	25	0.6250
125	120	5	25	0.2083
75	80	– 5	25	0.3125
300	300	0		$\chi^2 = 1.5625$

5. Critical region: $\chi^2 > \chi^2_{0.05\,(1)} = 3.841$

6. Conclusion: The calculated value of $\chi^2 = 1.5625$ is less than the critical value. Thus the hypothesis of independence is accepted. It means that college-educated males and non college educated males have the same liking for football. This result is different from the result given in Example 15.10. In Example 15.12. only one observed frequency of (AB) = 55 was compared with its corresponding expected frequency $\frac{(A) \times (B)}{n} = \frac{100 \times 180}{300} = 60$. The difference between 55 and 60 is not very large. They are very close. In χ^2 all the expected frequencies are compared with their observed frequencies. The χ^2-test is a very powerful test for test of independence. We shall admit the result or conclusion based on the χ^2-test. Thus H_o is accepted.

15.5.2 DIRECT FORMULA FOR CALCULATING χ^2 IN 2 × 2 CONTINGENCY TABLE

In a 2 × 2 contingency table the value of χ^2 can be calculated without calculating the expected frequencies. Suppose a 2 × 2 contingency table has four cell frequencies as distributed below:

	1st Attribute		Total
2nd Attribute	a	b	a + b
	c	d	c + d
Total	a + c	b + d	a + b + c + d

The value of χ^2 can be calculated directly by using the formula:

$$\chi^2 = \frac{(a + b + c + d)\,(ad - bc)^2}{(a + b)\,(b + d)\,(c + d)\,(a + c)}$$

The proof of this formula is beyond the level of this book.

Let us calculate χ^2 by using the above formula from the data given in Example 15.12. From the data given in example 15.12, we have

$a = 55, \quad b = 125, \quad c = 45 \quad$ and $\quad d = 75$

$$\text{Thus } \chi^2 = \frac{(55 + 125 + 45 + 75)\,(55 \times 75 - 125 \times 45)^2}{(55 + 125)\,(125 + 75)\,(45 + 75)\,(55 + 45)}$$

$$= \frac{(300)\,(2250000)}{(180)\,(200)\,(120)\,(100)} = \frac{675}{432} = 1.5625$$

This answer is the same as calculated in Example 15.12.

15.6 CONTINGENCY TABLE OF HIGHER ORDER

Sometimes a certain characteristic or attribute has more than two categories. For example when we are taking about heights of persons, the population or sample can be divided into four classes or categories like, very tall, tall, medium and short. In general if the attribute is A, then its different levels are denoted by $A_1, A_2, \ldots, A_r$ if it has r categories. The same population or sample may also be divided according to another characteristic say B with its levels $B_1, B_2, \ldots, B_c$ with c categories. The sample data on two attributes can be written in the form of two-way classification as below:

Table 15.4.
Two-way Classification

Attribute A	Attribute B						Row Totals
	B_1	B_2	...	B_j	...	B_c	
A_1	(A_1B_1)	(A_1B_2)	...	$(A_1 B_j)$	...	(A_1B_c)	(A_1)
A_2	(A_2B_1)	(A_2B_2)	...	$(A_2 B_j)$	...	(A_2B_c)	(A_2)
⋮	⋮	⋮		⋮		⋮	⋮
A_i	$(A_i B_1)$	$(A_i B_2)$	...	(A_iB_j)	...	$(A_i B_c)$	(A_i)
⋮	⋮	⋮		⋮		⋮	⋮
A_r	(A_rB_1)	(A_rB_2)	...	$(A_r B_j)$	...	(A_rB_c)	(A_r)
Column Totals	(B_1)	(B_2)	...	(B_j)	...	(B_c)	n

Table 15.4. contains r rows and c columns, it is therefore called r × c contingency table. Each frequency in the table is called cell frequency. It is the extension of 2 × 2 contingency table and χ^2-statistic is used to test the independence between the attributes given in the rows and columns.

The procedure is the same as explained earlier. For each observed frequency in the sample data, the corresponding expected frequency is calculated. It is calculated on the assumption that there is independence between the two characteristics. For each observed frequency (A_i B_j) the expected frequency is $\frac{(A_i)\,(B_j)}{n}$ where (A_i) is the total of the row A_i and (B_j) is the total of the column B_j. For expected frequency E, a more general formula may be written as

$$E = \frac{R \times C}{n}$$ where R is the row total and C is the column total.

χ^2 is calculated by the formula

$$\Sigma\left(\frac{(\text{Observed frequency} - \text{Expected frequency})^2}{\text{Expected frequency}}\right) = \Sigma\left(\frac{(f_o - f_e)^2}{f_e}\right)$$

15.7 LIMITATIONS OF χ^2

The χ^2-test of independence gives very good results or conclusions when all the cell frequencies are very large. For small cell frequencies the test is not very reliable. χ^2-test should not be used if any expected frequency is less than 5. If any expected frequency is less than 5, then something is to be done about it. One column containing the small frequency/frequencies is added to the adjacent column before calculating χ^2. Similarly if some row has expected frequencies less than 5, the entire row is added to the adjacent row by adding the corresponding cell frequencies. If we have the choice to reduce the number of rows or columns, we should choose that column or row which we think is least important in the given data and this column or row should be added to the adjacent column or row.

Example 15.13.

In a public opinion survey, 2000 persons were interviewed to give their opinion. The individuals interviewed are classified according to their attitude on a certain social scheme and according to sex. The data is given in the table below:

	Favour	Oppose	Undecided	Total
Men	600	320	280	1200
Women	450	280	70	800
Total	1050	600	350	2000

Calculate χ^2 to examine whether men and women differ in their opinion about the social scheme.

Solution:

1. The null hypothesis H_0 is that there is independence between the sex and their attitude towards the social scheme.

   The alternative hypothesis H_1 is that there is association between the two characteristics.

2. Level of significance: Let $\alpha = 0.05$

3. Test-statistic: $\chi^2 = \Sigma\left(\frac{(f_o - f_e)^2}{f_e}\right)$

4. Computations: Let A_1 and A_2 denote the rows and B_1, B_2 and B_3 denote the columns. The given table can be written as:

	B_1	B_2	B_3	Total
A_1	600	320	280	$(A_1) = 1200$
A_2	450	280	70	$(A_2) = 800$
Total	$(B_1) = 1050$	$(B_2) = 600$	$(B_3) = 350$	$n = 2000$

Expected frequencies f_e are calculated as below:

	B_1	B_2	B_3	Total
A_1	$\frac{1050 \times 1200}{2000}$ $= 630$	$\frac{600 \times 1200}{2000}$ $= 360$	$\frac{350 \times 1200}{2000}$ $= 210$	$(A_1) = 1200$
A_2	$\frac{1050 \times 800}{2000}$ $= 420$	$\frac{600 \times 800}{2000}$ $= 240$	$\frac{350 \times 800}{2000}$ $= 140$	$(A_2) = 800$
Total	$(B_1) = 1050$	$(B_2) = 600$	$(B_3) = 350$	$2000 = n$

It is important to note that the column and row totals are equal in the original table and table of expected frequencies.

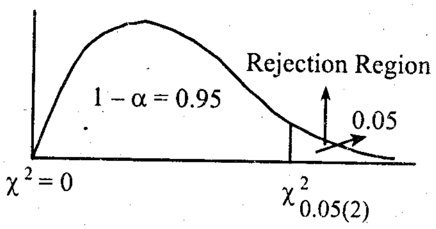
χ^2-Calculated

f_o	f_e	$(f_o - f_e)$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
600	630	– 30	900	1.43
450	420	30	900	2.14
320	360	– 40	1600	4.44
280	240	40	1600	6.67
280	210	70	4900	23.33
70	140	– 70	4900	35.00
2000	2000	0		$\chi^2 = 73.01$

5. Region of rejection:

d.f. $= (r - 1)(c - 1) = (2 - 1)(3 - 1) = 2$

$\chi^2 > \chi^2_{0.05(2)} = 5.991$



6. Conclusion: The calculated value of χ^2 is 73.01 and the critical value of χ^2 is 5.991. The χ^2 calculated from the sample data falls in the rejection region. Thus hypothesis of independence is rejected. It means that men and women have different opinions about the social scheme. Sex is associated with the attitude towards the social scheme.

Example 15.14.

Given the following table. Calculate χ^2 to examine whether there is evidence of relationship between the intelligence level of fathers and sons. Use $\alpha = 0.05$

	Fathers			
Sons	Very Intelligent	Average	Non-Intelligent	Total
Very Intelligent	10	35	5	50
Average	150	140	15	305
Non–Intelligent	40	95	20	155
Total	200	270	40	510

Solution:

1. The null hypothesis to be tested is that there is no relationship between the intelligence of fathers and sons.
   The alternative hypothesis is that there is relationship (association) between the intelligence level of fathers and sons.
2. Level of significance: $\alpha = 0.05$
3. Test-statistic: $\chi^2 = \Sigma\left(\frac{(f_o - f_e)^2}{f_e}\right)$
4. Computations:
   Let A_1, A_2, A_3 be used for the rows and B_1, B_2, B_3 be used for columns headings.
   Table of expected frequencies calculated

	B_1	B_2	B_3	Total
A_1	$\frac{200 \times 50}{510} = 19.6$	$\frac{270 \times 50}{510} = 26.5$	$\frac{40 \times 50}{510} = 3.9$	$(A_1) = 50$
A_2	$\frac{200 \times 305}{510} = 119.6$	$\frac{270 \times 305}{510} = 161.5$	$\frac{40 \times 305}{510} = 23.9$	$(A_2) = 305$
A_3	$\frac{200 \times 155}{510} = 60.8$	$\frac{270 \times 155}{510} = 82.0$	$\frac{40 \times 155}{510} = 12.2$	$(A_3) = 155$
Total	$(B_1) = 200$	$(B_2) = 270$	$(B_3) = 40$	$n = 510$

One expected frequency under the column B_3 and against row A_1 is 3.9 which is less than 5. This frequency cannot be used in the calculation of χ^2. Now we have two options (i) column B_3 is added to column B_2 (ii) Row A_1 is added to row A_2. But the total of column B_3 is 40 which is minimum of all the column and row totals. It means column B_3 is less important as compared to row A_1. Thus column B_3 is added to column B_2. This is equivalent tc combining a small sample data with a large sample data. Thus the tables of observed frequencies and the expected frequencies would become:

Observed Frequencies

	B_1	$B_2 + B_3$	Total
A_1	10	35 + 5 = 40	50
A_2	150	140 + 15 = 155	305
A_3	40	95 + 20 = 115	155
Total	200	310	510

Expected Frequencies

	B_1	$B_2 + B_3$	Total
A_1	19.6	26.5 + 3.9 = 30.4	50
A_2	119.6	161.5 + 23.9 = 185.4	305
A_3	60.8	82.0 + 12.2 = 94.2	155
Total	200	310.0	510

χ^2–Calculated

f_o	f_e	$(f_o - f_e)$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
10	19.6	− 9.6	92.16	4.70
150	119.6	30.4	924.16	7.73
40	60.8	− 20.8	432.64	7.12
40	30.4	9.6	92.16	3.03
155	185.4	− 30.4	924.16	4.98
115	94.2	20.8	432.64	4.59
510	510	0		$\chi^2 = 32.15$

5. Region of rejection:

$$\text{d.f.} = (r - 1)(c - 1) = (3 - 1)(2 - 1) = 2$$

Critical region is $\chi^2 > \chi^2_{0.05(2)} = 5.991$

6. Conclusion: Since the calculated value of $\chi^2 = 32.15$ which is greater than the critical value of 5.991, the null hypothesis H_0 is rejected and H_1 is accepted. It means that there is relationship (association) between the intelligence levels of fathers and sons. Intelligent fathers have usually intelligent sons. This is what the sample data indicates through the χ^2 as test of independence.

15.8 RANK CORRELATION:

We are often confronted with situations where the basic data are not available in numerical magnitudes but where the rankings can be developed and used to examine the relationship between data sets. To calculate the Spearman's rank correlation coefficient, we first rank the X's among themselves, giving rank 1 to the largest or smallest value, rank 2 to the second largest or second smallest, and so on; then we rank the Y's similarly among themselves. The Spearman's rank correlation coefficient, r_s, is given by the following formula

$$r_s = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)}$$

where d = difference between the ranks for the paired observations,

n = number of paired observations

When there are tied observations, the mean rank is given to each observation in the set of ties. For example, if the fourth and fifth largest values of a variable are the same, we assign each the rank (4+5)/2 = 4.5, and if the sixth, seventh and eighth largest values of a variable are the same, we assign each the rank = (6+7+8) / 3 = 7. The possible range of values for Spearman's rank correlation coefficient $\mathbf{r_s}$ is −1 to +1. If $\mathbf{r_s} = +1$, there is perfect positive rank correlation and if $\mathbf{r_s} = -1$, there is perfect negative rank correlation. If X and Y are independent of each other, there is no relationship and thus the rank correlation coefficient $\mathbf{r_s} = 0$.

Example 15.15.

The following were the "performance under stress" rankings of 10 honor students before and after mid-semester:

Student	A	B	C	D	E	F	G	H	I	J
Rank before	1	2	3	4	5	6	7	8	9	10
Rank after	6	5	8	9	3	4	10	1	7	2

Compute the Spearman's rank correlation coefficient for this data set.

Solution:

Rank before (X)	Rank after (Y)	d = X – Y	d^2
1	6	–5	25
2	5	–3	9
3	8	–5	25
4	9	–5	25
5	3	+2	4
6	4	+2	4
7	10	–3	9
8	1	+7	49
9	7	+2	4
10	2	+8	64
			$\Sigma d^2 = 218$

Spearman's rank correlation coefficient, $r_s = 1 - \dfrac{6\Sigma d^2}{n(n^2 - 1)}$

$$= 1 - \frac{6(218)}{10(100 - 1)} = 1 - 1.32 = -0.32$$

Example 15.16.

A Statistics instructor wants to know whether there is a correlation between students midterm averages and their final examination scores. The instructor takes a random sample of nine students from previous Statistics courses and obtains the following data:

Midterm average X	72	96	86	77	67	92	90	74	60
Final examination score Y	49	97	80	73	71	86	95	48	52

(i) Determine the rank correlation coefficient, r_s, of the data.

(ii) Interpret the value of r_s obtained in part (i).

Solution:

Midterm average (X)	Final examination score (Y)	Rank of X	Rank of Y	d = X – Y	d^2
72	49	3	2	1	1
96	97	9	9	0	0
86	80	6	6	0	0
77	73	5	5	0	0
67	71	2	4	–2	4
92	86	8	7	1	1
90	95	7	8	–1	1
74	48	4	1	3	9
60	52	1	3	–2	4
					$\Sigma d^2 = 20$

(i) $r_s = 1 - \frac{6\Sigma d^2}{n(n^2-1)} = 1 - \frac{6(20)}{9(81-1)} = 1 - 0.17 = 0.83$

(ii) The rank correlation coefficient, $r_s = 0.83$ suggests that there is a strong positive correlation between midterm average and final-examination score in Statistics courses.

Example 15.17.

The number of hours of study for an examination and the grades received by a random sample of 10 students are:

Number of hours studied, X	8	5	11	13	10	5	18	15	2	8
Grade in examination, Y	56	44	79	72	70	54	94	85	33	65

Compute and interpret the Spearman's rank correlation coefficient.

Solution:

Number of hours studied (X)	Grade in examination (Y)	Rank of X	Rank of Y	d = X − Y	d^2
8	56	4.5	4	0.5	0.25
5	44	2.5	2	0.5	0.25
11	79	7	8	−1.0	1.00
13	72	8	7	1.0	1.00
10	70	6	6	0.0	0.00
5	54	2.5	3	−0.5	0.25
18	94	10	10	0.0	0.00
15	85	9	9	0.0	0.00
2	33	1	1	0.0	0.00
8	65	4.5	5	−0.5	0.25
					$\Sigma d^2 = 3$

Spearman's rank correlation coefficient, $r_s = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$

$$= 1 - \frac{6(3)}{10(100-1)} = 1 - \frac{18}{990} = 1 - 0.02 = 0.98$$

$r_s = 0.98$, indicating strong positive correlation between the number of hours of study and the grade in examination.