Indira Gandhi
National Open University
School of Sciences

MST-002
DESCRIPTIVE STATISTICS

Block

1

ANALYSIS OF QUANTITATIVE DATA

UNIT 1
Measures of Central Tendency 7

UNIT 2
Measures of Dispersion 37

UNIT 3
Moments 53

UNIT 4
Skewness and Kurtosis 67

Curriculum and Course Design Committee

Prof. K. R. Srivathasan
Pro-Vice Chancellor
IGNOU, New Delhi

Prof. Parvin Sinclair
Pro-Vice Chancellor
IGNOU, New Delhi

Prof. Geeta Kaicker
Director, School of Sciences
IGNOU, New Delhi

Prof. Jagdish Prasad
Department of Statistics
University of Rajasthan, Jaipur

Prof. R. M. Pandey
Department of Bio-Statistics
All India Institute of Medical Sciences
New Delhi

Prof. Rahul Roy
Math. and Stat. Unit
Indian Statistical Institute, New Delhi

Dr. Diwakar Shukla
Department of Mathematics and Statistics
Dr. Hari Singh Gaur University, Sagar

Prof. Rakesh Srivastava
Department of Statistics
M. S. University of Baroda, Vadodara

Prof. G. N. Singh
Department of Applied Mathematics
I. S. M. Dhanbad

Dr. Gulshan Lal Taneja
Department of Mathematics
M. D. University, Rohtak

Faculty members of School of Sciences, IGNOU

Statistics
Dr. Neha Garg
Dr. Nitin Gupta
Mr. Rajesh Kaliraman
Dr. Manish Trivedi

Mathematics
Dr. Deepika Garg
Prof. Poornima Mital
Prof. Sujatha Varma
Dr. S. Venkataraman

Block Preparation Team

Content Editor
Dr. Rajesh Tailor
School of Studies in Statistics
Vikram University, Ujjain (MP)

Language Editor
Dr. Nandini Sahu
School of Humanities, IGNOU

Secretarial Support
Mr. Deepak Singh

Course Writers
Dr. Soubhik Chakraborti (Unit 1& 2)
Department of Applied Mathematics
B. I. T. Mesra, Ranchi (JH)

Dr. Manish Trivedi (Unit 3 & 4)
School of Sciences, IGNOU

Formatted By
Dr. Manish Trivedi
Mr. Prabhat Kumar Sangal
School of Sciences, IGNOU

Programme and Course Coordinator: Dr. Manish Trivedi

Block Production

Mr. Y. N. Sharma, SO (P.)
School of Sciences, IGNOU

Acknowledgement: We gratefully acknowledge to Prof. Geeta Kaicker, Director, School of Sciences for her great support and guidance.

December, 2011

ISBN-978-81-266-

Further information on the Indira Gandhi National Open University may be obtained from University's Office at Maidan Garhi, New Delhi-110068 or visit University's website http://www.ignou.ac.in

Printed and published on behalf of the Indira Gandhi National Open University, New Delhi by the Director, School of Sciences.

Laser Typeset by: Tessa Media & Computers, C-206, A.F.E.-II, Okhla, New Delhi

Printed at:

DESCRIPTIVE STATISTICS

In MST-001, we have discussed some mathematical techniques to make the learners able to cope up with the derivational and numerical part of some statistical techniques which are discussed in this course and other core and elective courses of this programme. We also discussed some basic methods of collection, organization and representation of data in that course.

After collection and organization, the next step is to proceed for data analysis to explore the properties of the data. The statistical tools which describe the properties of the data are known as descriptive statistics. The techniques of the descriptive statistics are frequently used for analysing the data in various fields. The purpose of discussing those techniques is to make you aware of the three major properties that describe the data. These properties are:

1. The numerical value of an observation (commonly referred as central value) around which most of the other numerical values show a tendency to concentrate or group, called central tendency.
2. The degree to which numerical values are scattered or dispersed around the central value, called dispersion**.**
3. The extent of departure of numerical values from symmetrical distribution (taken as normal distribution) around the central value, called skewness.

In Block 1, we have discussed the measures based on the above three properties. These measures can also be used to compare two distributions. The development of such types of measures was based on univariate distributions.

The next step in this direction is to study simultaneously two or more variables on the same unit of the population. This kind of analysis is important for drawing inferences from the co-variation between variables in a given data. In this regard the concept of statistical relationship between two variables is introduced in Block 2 and quantitative measures of relationship between two variables for analysing the strength of relationship are developed.

As a sequel of Block 2, the average relationship between two variables in terms of regression analysis is elaborated in Block 3. In Block 3, (i) the quantitative measure of the degree of association between a dependent variable and two or more independent variables taken together in a group, known as multiple correlation coefficient; (ii) the quantitative measure of the degree of association between a dependent variable and any one of the independent variables included in the analysis, while the effect of other independent variables included in the analysis is held constant, known as partial correlation coefficient**,** have also been discussed.

The Block 4 is mainly concerned with the qualitative characteristics and analysis of qualitative data. Such type of data arises when a sample from some population is classified with respect to two or more qualitative variables.

Notations and Symbols

x_i	:	Value of i^{th} observation of variable X
f_i	:	Frequency of i^{th} class
$\overline{X}$	:	Arithmetic mean
A	:	Assumed mean
$\sum$	:	Sum of observations
W_i	:	Weight of i^{th} observation
$\overline{X}_w$	:	Weighted mean
$N=\sum_{i=1}^{k} f_i$	:	Total number of observations in data
GM	:	Geometric Mean
HM	:	Harmonic Mean
Q_i	:	i^{th} Quartile
D_i	:	i^{th} Decile
P_i	:	i^{th} Percentile
QD	:	Quartile Deviation
MD	:	Mean Deviation
$\mathrm{Var}(X)=\sigma_x^2$	:	Variance of X
SD = σ	:	Standard Deviation
RMSD	:	Root Mean Square Deviation
$\mu_r^{'}$	:	r^{th} Moment about arbitrary point
μ_r	:	r^{th} Central moment
S_k	:	Coefficient of skewness
β_1	:	Measures of skewness
β_2	:	Measures of kurtosis
γ_1	:	Derivative of β_1
γ_2	:	Derivative of β_2

ANALYSIS OF QUANTITATIVE DATA

A raw data after collection are not suitable to draw conclusions about the mass or population from which it has been collected. Some inferences about the mass can be drawn from the frequency distribution which condenses and reduces the bulk of data. In general, a distribution can be categorized by two parameters, i.e., (i) Measures of location and (ii) Measures of dispersion.

Generally, the data are condensed into a single value around which the most of values tend to cluster in finding central value. Such a value lies in the center of the data and is known as central tendency. A central value explores an idea of whole mass. But the information so obtained is neither exhaustive nor compressive as the measure of central tendency does not provide the information about the scatterness of the observations. This leads us to conclude that a measure of central tendency alone is not enough to have a clear picture of data, one need to have a measure of dispersion or variation.

Moments are statistical measures that describe certain characteristics of the distribution. Measures of Skewness and Kurtosis give us the direction and the magnitude of the lack of symmetry and peakedness or flatness of data.

In this block, we shall discuss some statistical tools which are used to analyse the quantitative data. In Unit 1, a detailed idea has been explored about the measures of Central Tendency. In Unit 2, we have discussed about the variation of data and some useful measures of dispersion. Then, in Unit 3 we have described various types of moments and their uses. In Unit 4, we have discussed the Skewness and Kurtosis and their coefficients.

References:

- Agrawal, B. L.; Basic Statistics, New Age International (P) Ltd. Publishers, New Delhi, 3rd edn., 1996
- Goon, A. M., Gupta, M. K. and Das Gupta, B.; Fundamentals of Statistics Vol-I; World Press Culcutta.
- Gupta, M. P. and Gupta, S. P.; Business Statistics; Sultan Chand & Sons Publications.
- Gupta S. C. and Kapoor, V. K.; Fundamentals of Mathematical Statistics, Sultan Chand & Sons Publications.
- Mukhopadhayaya, P.; Mathematical Statistics, Books & Allied (p) Ltd., Kolkata
- Hoel, P. G.; Introduction to Mathematical Statistics, Wiley Series in Probability and Statistics.

UNIT 1 MEASURES OF CENTRAL TENDENCY

Structure

1.1 Introduction

Objectives

1.2 Measures of Central Tendency

1.3 Arithmetic Mean

1.4 Weighted Mean

1.5 Median

1.6 Mode

1.7 Geometric Mean

1.8 Harmonic Mean

1.9 Partition Values

Quartiles

Deciles

Percentiles

1.10 Summary

1.11 Solutions /Answers

1.1 INTRODUCTION

As we know that after the classification and tabulation of data one often finds too much detail for many uses that may be made of information available. We, therefore, need further analysis of the tabulated data to draw inference. In this unit we are going to discuss about measures of central tendencies. For the purpose of analysis, very important and powerful tool is a single average value that represents the entire mass of data.

The term average in Statistics refers to a one figure summary of a distribution. It gives a value around which the distribution is concentrated. For this reason that average is also called the measure of central tendency. For example, suppose Mr. X drives his car at an average speed of 60 km/hr. We get an idea that he drives fast (on Indian roads of course!). To compare the performance of two classes, we can compare the average scores in the same test given to these two classes. Thus, calculation of average condenses a distribution into a single value that is supposed to represent the distribution. This helps both individual assessments of a distribution as well as in comparison with another distribution.

This unit comprises some sections as the Section 1.2 gives the definition of measures of central tendency. The significance and properties of a good measure of central tendency are also described in Sub-sections 1.2.1 and 1.2.2. In sub sequent Sections 1.3, 1.4, 1.5 and 1.6, direct and indirect methods for calculating Arithmetic mean, Weighted mean, Median and mode, respectively are explained with their merits and demerits, whereas in Sections 1.7 and 1.8 methods for calculating Geometric mean and Harmonic mean for ungrouped and grouped data, respectively are explained with their merits and demerits. The concepts and methods of calculating the partition values are described in Section 1.9.

Objectives

After studying this unit, you would be able to

- define an average;
- explain the significance of a measure of central tendency;
- explain the properties of a good average;
- calculate the different types of measures of central tendency;
- describe the merits and demerits different types of measures of central tendency; and
- describe the methods of calculation of partition values.

1.2 MEASURES OF CENTRAL TENDENCY

According to **Professor Bowley**, averages are “statistical constants which enable us to comprehend in a single effort the significance of the whole”. They throw light as to how the values are concentrated in the central part of the distribution.

For this reason as on last page that they are also called the measures of central tendency, an average is a single value which is considered as the most representative for a given set of data. Measures of central tendency show the tendency of some central value around which data tend to cluster.

1.2.1 Significance of the Measure of Central Tendency

The following are two main reasons for studying an average:

1. **To get a single representative**

   Measure of central tendency enables us to get a single value from the mass of data and also provide an idea about the entire data. For example it is impossible to remember the heights measurement of all students in a class. But if the average height is obtained, we get a single value that represents the entire class.

2. **To facilitate comparison**

   Measures of central tendency enable us to compare two or more than two populations by reducing the mass of data in one single figure. The comparison can be made either at the same time or over a period of time. For example, if a subject has been taught in more than two classes so by obtaining the average marks of those classes, comparison can be made.

1.2.2 Properties of a Good Average

The following are the properties of a good measure of average:

1. **It should be simple to understand**

   Since we use the measures of central tendency to simplify the complexity of a data, so an average should be understandable easily otherwise its use is bound to be very limited.

2. **It should be easy to calculate**

   An average not only should be easy to understand but also should be simple to compute, so that it can be used as widely as possible.

3. **It should be rigidly defined**

   A measure of central tendency should be defined properly so that it has an appropriate interpretation. It should also have an algebraic formula so that if different people compute the average from same figures, they get the same answer.

4. **It should be liable for algebraic manipulations**

   A measure of central tendency should be liable for the algebraic manipulations. If there are two sets of data and the individual information is available for both set, then one can be able to find the information regarding the combined set also then something is missing.

5. **It should be least affected by sampling fluctuations**

   We should prefer a tool which has a sampling stability. In other words, if we select 10 different groups of observations from same population and compute the average of each group, then we should expect to get approximately the same values. There may be little difference because of the sampling fluctuation only.

6. **It should be based on all the observations**

   If any measure of central tendency is used to analyse the data, it is desirable that each and every observation is used for its calculation.

7. **It should be possible to calculate even for open-end class intervals**

   A measure of central tendency should able to be calculated for the data with open end classes.

8. **It should not be affected by extremely small or extremely large observations**

   It is assumed that each and every observation influences the value of the average. If one or two very small or very large observations affect the average i.e. either increase or decrease its value largely, then the average cannot be consider as a good average.

1.2.3 Different Measures of Central Tendency

The following are the various measures of central tendency:

1. Arithmetic Mean
2. Weighted Mean
3. Median
4. Mode
5. Geometric Mean
6. Harmonic Mean

1.2.4 Partition Values

1. Quartiles
2. Deciles
3. Percentiles

1.3 ARITHMETIC MEAN

Arithmetic mean (also called mean) is defined as the sum of all the observations divided by the number of observations. Arithmetic mean (AM) may be calculated for the following two types of data:

1. For Ungrouped Data

For ungrouped data, arithmetic mean may be computed by applying any of the following methods:

(1) Direct Method

Mathematically, if x_1, x_2,…, x_n are the n observations then their mean is

$$\overline{X}=\frac{(x_1+x_2+x_3+...+x_n)}{n}$$

$$\overline{X}=\frac{\sum_{i=1}^{n}x_i}{n}$$

If f_i is the frequency of x_i (i=1, 2,…, k), the formula for arithmetic mean would be

$$\overline{X}=\frac{(f_1x_1+f_2x_2+...+f_kx_k)}{(f_1+f_2+...+f_k)}$$

$$\overline{X}=\frac{\sum_{i=1}^{k}f_ix_i}{\sum_{i=1}^{k}f_i}$$

(2) Short-cut Method

The arithmetic mean can also be calculated by taking deviations from any arbitrary point "A", in which the formula shall be

$$\overline{X}=A+\frac{\sum_{i=1}^{n}d_i}{n} \qquad \text{where, } d_i = x_i - A$$

If f_i is the frequency of x_i (i=1, 2,…, k), the formula for arithmetic mean would be

$$\overline{X}=A+\frac{\sum_{i=1}^{k}f_id_i}{\sum_{i=1}^{k}f_i}, \qquad \text{where, } d_i = x_i - A$$

Here, k is the number of distinct observations in the distribution.

Note: Usually the short-cut method is used when data are large.

Example 1: Calculate mean of the weights of five students

54, 56, 70, 45, 50 (in kg)

Solution: If we denote the weight of students by x then mean is obtained by

$$\overline{X} = \frac{x_1 + x_2 + ... + x_n}{n}$$

Thus, $$\overline{X} = \frac{54 + 56 + 70 + 45 + 50}{5} = \frac{275}{5} = 55$$

Therefore, average weight of students is 55 kg.

Example 2: Compute arithmetic mean of the weight of students for the given data in Example 1 by using shortcut method.

Solution: For shortcut method, we use following formula

$$\overline{X} = A + \frac{\sum d_i}{n}, \qquad \text{where } d_i = x_i - A$$

If 50 is taken as the assumed value A in the given data in Example 1 then, for the calculation of d_i we prepare following table:

x	**d = x - A**
54	54–50 = 4
56	56–50 = 6
70	70–50 =20
45	45–50 = – 5
50	50–50=0
	$\sum_{i=1}^{n} d_i = 25$

We have A = 50 then,

$$\overline{X} = A + \frac{\sum_{i=1}^{n} d_i}{n} = 50 + \frac{25}{5} = 50 + 5 = 55$$

Example 3: Calculate arithmetic mean for the following data:

x	20	30	40
f	5	6	4

Solution: We have the following frequency distribution:

x	**f**	**fx**
20	5	100
30	6	180
40	4	160
	$\sum_{i=1}^{k} f_i = 15$	$\sum_{i=1}^{k} f_i x_i = 440$

Arithmetic Mean, $\overline{X} = \dfrac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i}$

$$\overline{X} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i} = \frac{440}{15} = 29.3$$

E1) Find the arithmetic mean of the following observations:

5, 8, 12, 15, 20, 30.

E2) For the following discrete frequency distribution find arithmetic mean:

Wages (in Rs)	20	25	30	35	40	50
No. of workers	5	8	20	10	5	2

2 For Grouped Data

Direct Method

If f_i is the frequency of x_i (i =1, 2,…, k) where x_i is the mid value of the i^{th} class interval, the formula for arithmetic mean would be

$$\overline{X} = \frac{(f_1x_1 + f_2x_2 + \ldots + f_kx_k)}{(f_1 + f_2 + \ldots + f_k)}$$

$$\overline{X} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i} = \frac{\sum fx}{\sum f} = \frac{\sum fx}{N},$$

where, $N = f_1 + f_2 + ... + f_k$

Short-cut Method

$$\overline{X} = A + \frac{\sum_{i=1}^{k} f_i d_i}{\sum_{i=1}^{k} f_i}, \qquad \text{where, } d_i = x_i - A$$

Here, f_i would be representing the frequency of the i^{th} class, x_i is the mid-value of the i^{th} class and k is the number of classes.

Example 4: For the following data, calculate arithmetic mean using direct method:

Class Interval	0-10	10-20	20-30	30-40	40-50
Frequency	3	5	7	9	4

Solution: We have the following distribution:

Class Interval	**Mid Value x**	**Frequency f**	**fx**
0-10	05	03	15
10-20	15	05	75
20-30	25	07	175
30-40	35	09	315
40-50	45	04	180
		$\sum_{i=1}^{k} f_i$ =N=28	$\sum_{i=1}^{k} f_i x_i = 760$

$$\text{Mean} = \frac{\sum_{i=1}^{k} f_i x_i}{N} = 760/28 = 27.143$$

Now let us solve one exercise.

E3) Find arithmetic mean of the distribution of marks given below:

Marks	0-10	10-20	20-30	30-40	40-50
No. of students	6	9	17	10	8

1.3.1 Properties of Arithmetic Mean

Arithmetic mean fulfills most of the properties of a good average except the last two. It is particularly useful when we are dealing with a sample as it is least affected by sampling fluctuations. It is the most popular average and should always be our first choice unless there is a strong reason for not using it.

Three algebraic properties of mean are given below:

Property 1: Sum of deviations of observations from their mean is zero. Deviation is also called dispersion that will be discussed in detail in Unit 2 of this block.

Proof: We have to prove ∑(x – mean) = 0

The sum of deviations of observations x_1, x_2, …, x_n from their mean is

$$\sum_{i=1}^{n}\left(x_i - \overline{x}\right) = \sum_{i=1}^{n} x_i - n\,\overline{x}$$

$$\Rightarrow \sum_{i=1}^{n} x_i - n\frac{1}{n}\sum_{i=1}^{n} x_i = 0$$

Property 2: Sum of squares of deviations taken from mean is least in comparison to the same taken from any other average.

Proof: We have

$$\sum_{i=1}^{n}\left(x_i - A\right)^2 = \sum_{i=1}^{n}\left(x_i - \overline{x} + \overline{x} - A\right)^2$$

where, A is an assumed mean / Median / Mode

$$\Rightarrow \sum_{i=1}^{n}\left(x_i - A\right)^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2 + n\,(\overline{x} - A)^2 + 2(\overline{x} - A)\sum_{i=1}^{n}(x_i - \overline{x})$$

$$\Rightarrow \sum_{i=1}^{n}\left(x_i - A\right)^2 - \sum_{i=1}^{n}(x_i - \overline{x})^2 = n\,(\overline{x} - A)^2 + 0 \qquad \text{(By Property 1)}$$

$$\Rightarrow \sum_{i=1}^{n}(x_i - A)^2 - \sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 \geq 0$$

$$\Rightarrow \sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 \leq \sum_{i=1}^{n}(x_i - A)^2$$

That means the sum of squares of deviations taken from mean is least in comparison to the same taken from any other average.

Property 3: Arithmetic mean is affected by both the change of origin and scale.

Proof: If $u_i = \frac{x_i - a}{h}$,

where, a and h are constant. Then

$$x_i = a + h\,u_i$$

$$\sum_{i=1}^{n} x_i = na + h\sum_{i=1}^{n} u_i$$

$$\frac{1}{n}\sum_{i=1}^{n} x_i = a + h\frac{1}{n}\sum_{i=1}^{n} u_i$$

$$\overline{X} = a + h\ \overline{U}$$

1.3.2 Merits and Demerits of Arithmetic Mean

Merits of Arithmetic Mean

1. It utilizes all the observations;
2. It is rigidly defined;
3. It is easy to understand and compute; and
4. It can be used for further mathematical treatments.

Demerits of Arithmetic Mean

1. It is badly affected by extremely small or extremely large values;
2. It cannot be calculated for open end class intervals; and
3. It is generally not preferred for highly skewed distributions.

1.4 WEIGHTED MEAN

Weight here refers to the importance of a value in a distribution. A simple logic is that a number is as important in the distribution as the number of times it appears. So, the frequency of a number can also be its weight. But there may be other situations where we have to determine the weight based on some other reasons. For example, the number of innings in which runs were made may be considered as weight because runs (50 or 100 or 200) show their importance. Calculating the weighted mean of scores of several innings of a player, we may take the strength of the opponent (as judged by the proportion of matches lost by a team against the opponent) as the corresponding weight. Higher the proportion stronger would be the opponent and hence more would be the weight. If x_i has a weight w_i, then weighted mean is defined as:

$$\overline{X}_W = \frac{\sum_{i=1}^{k} x_i w_i}{\sum_{i=1}^{k} w_i} \qquad \text{for all } i = 1, 2, 3, \ldots, k.$$

1.5 MEDIAN

Median is that value of the variable which divides the whole distribution into two equal parts. Here, it may be noted that the data should be arranged in ascending or descending order of magnitude. When the number of observations is odd then the median is the middle value of the data. For even number of observations, there will be two middle values. So we take the arithmetic mean of these two middle values. Number of the observations below and above the median, are same. Median is not affected by extremely large or extremely small values (as it corresponds to the middle value) and it is also not affected by open end class intervals. In such situations, it is preferable in comparison to mean. It is also useful when the distribution is skewed (asymmetric). Skewness will be discussed in Unit 4 of this block.

1. Median for Ungrouped Data

Mathematically, if $x_1, x_2, \ldots, x_n$ are the n observations then for obtaining the median first of all we have to arrange these n values either in ascending order or in descending order. When the observations are arranged in ascending or descending order, the middle value gives the median if n is odd. For even number of observations there will be two middle values. So we take the arithmetic mean of these two values.

$$M_d = \left(\frac{n+1}{2}\right)^{th} \text{observation} \text{ ; (when n is odd)}$$

$$M_d = \frac{\left(\frac{n}{2}\right)^{th} \text{observation} + \left(\frac{n}{2}+1\right)^{th} \text{observation}}{2} \text{; (when n is even)}$$

Example 5: Find median of following observations:

6, 4, 3, 7, 8

Solution: First we arrange the given data in ascending order as

3, 4, 6, 7, 8

Since, the number of observations i.e. 5, is odd, so median would be the middle value that is 6.

Example 6: Calculate median for the following data:

7, 8, 9, 3, 4, 10

Solution: First we arrange given data in ascending order as

3, 4, 7, 8, 9, 10

Here, Number of observations (n) = 6 (even). So we get the median by

$$M_d = \frac{\left(\frac{n}{2}\right)^{th} \text{observation} + \left(\frac{n}{2}+1\right)^{th} \text{observation}}{2}$$

$$= \frac{\left(\frac{6}{2}\right)^{th} \text{observation} + \left(\frac{6}{2}+1\right)^{th} \text{observation}}{2}$$

$$M_d = \frac{3^{rd} \text{observation} + 4^{th} \text{observation}}{2}$$

$$= \frac{7+8}{2} = \frac{15}{2} = 7.5$$

E4) Find the median of the following values:

(i) 10, 6, 15, 2, 3, 12, 8

(ii) 10, 6, 2, 3, 8, 15, 12, 5

For Ungrouped Data (when frequencies are given)

If x_i are the different value of variable with frequencies f_i then we calculate cumulative frequencies from f_i then median is defined by

$M_d =$ Value of variable corresponding to $\left(\frac{\sum f}{2}\right)^{th} = \left(\frac{N}{2}\right)^{th}$ cumulative frequency.

Note: If N/2 is not the exact cumulative frequency then value of the variable corresponding to next cumulative frequencies is the median.

Example 7: Find Median from the given frequency distribution

x	20	40	60	80
f	7	5	4	3

Solution: First we find cumulative frequency

x	**f**	**c.f.**
20	7	7
40	5	12
60	4	16
80	3	19
	$\sum_{i=1}^{k} f_i = 19$	

M_d = Value of the variable corresponding to the $\left(\frac{19}{2}\right)^{th}$ cumulative frequency

= Value of the variable corresponding to 9.5 since 9.5 is not among c.f.

So, the next cumulative frequency is 12 and the value of variable against 12 cumulative frequency is 40. So median is 40.

E5) Find the median of the following frequency distribution:

Mid Values	5	10	15	20	25	30	35	40	45	50
No.of students	2	5	10	14	16	20	13	9	7	4

2. Median for Grouped Data

For class interval, first we find cumulative frequencies from the given frequencies and use the following formula for calculating the median:

$$\text{Median} = L + \frac{\left(\frac{N}{2} - C\right)}{f} \times h$$

where, L = lower class limit of the median class,

N = total frequency,

C = cumulative frequency of the pre-median class,

f = frequency of the median class, and

h = width of the median class.

Median class is the class in which the $(N/2)^{th}$ observation falls. If N/2 is not among any cumulative frequency then next class to the N/2 will be considered as median class.

Example 8: Calculate median for the data given in Example 4.

Solution: We first form a cumulative frequency table (cumulative frequency of a class gives the number of observations less than the upper limit of the class; strictly speaking, this is called cumulative frequency of less than type; we also have cumulative frequency of more than type which gives the number of observations greater than or equal to the lower limit of a class):

Class Interval	Frequency f	Cumulative Frequency (< type)
0-10	3	3
10-20	5	8
20-30	7	15
30-40	9	24
40-50	4	28

$$\sum_{i=1}^{k} f_i = N = 28 \qquad \Rightarrow \qquad \frac{N}{2} = \frac{28}{2} = 14$$

Since 14 is not among the cumulative frequency so the class with next cumulative frequency i.e. 15, which is 20-30, is the median class.

We have L = lower class limit of the median class = 20

N = total frequency = 28

C = cumulative frequency of the pre median class =8

f = frequency of the median class = 7

h = width of median class = 10

Now substituting all these values in the formula of Median

$$\text{Median} = L + \frac{\left(\frac{N}{2} - C\right)}{f} \times h$$

$$M_d = 20 + \frac{14-8}{7} \times 10 = 28.57$$

Therefore, median is 28.57.

E6) Find Median for the following frequency distribution:

Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70
No.of students	5	10	15	20	12	10	8

E7) Find the missing frequency when median is given as Rs 50.

Expenditure (Rs)	0-20	20-40	40-60	60-80	80-100
No. of families	5	15	30	--	12

1.5.1 Merits and Demerits of Median

Merits of Median

1. It is rigidly defined;
2. It is easy to understand and compute;
3. It is not affected by extremely small or extremely large values; and
4. It can be calculated even for open end classes (like “less than 10” or “50 and above”).

Demerits of Median

1. In case of even number of observations we get only an estimate of the median by taking the mean of the two middle values. We don’t get its exact value;
2. It does not utilize all the observations. The median of 1, 2, 3 is 2. If the observation 3 is replaced by any number higher than or equal to 2 and if the number 1 is replaced by any number lower than or equal to 2, the median value will be unaffected. This means 1 and 3 are not being utilized;

3. It is not amenable to algebraic treatment; and
4. It is affected by sampling fluctuations.

1.6 MODE

Highest frequent observation in the distribution is known as mode. In other words, mode is that observation in a distribution which has the maximum frequency. For example, when we say that the average size of shoes sold in a shop is 7 it is the modal size which is sold most frequently.

For Ungrouped Data

Mathematically, if $x_1, x_2, \ldots, x_n$ are the n observations and if some of the observation are repeated in the data, say x_i is repeated highest times then we can say the x_i would be the mode value.

Example 9: Find mode value for the given data

2, 2, 3, 4, 7, 7, 7, 7, 9, 10, 12, 12

Solution: First we prepare frequency table as

x	2	3	4	7	9	10	12
f	2	1	1	4	1	1	2

This table shows that 7 have the maximum frequency. Thus, mode is 7.

E8) Find the model size for the following items:

4, 7, 6, 5, 4, 7, 8, 3, 7, 2, 7, 6, 1, 2, 5

For Grouped Data:

Data where several classes are given, following formula of the mode is used

$$M_0 = L + \frac{|f_1 - f_0|}{|f_1 - f_0| + |f_1 - f_2|} \times h$$

where, L = lower class limit of the modal class,

f_1 = frequency of the modal class,

f_0 = frequency of the pre-modal class,

f_2 = frequency of the post-modal class, and

h = width of the modal class.

Modal class is that class which has the maximum frequency.

Example 10: For the data given in Example 4, calculate mode.

Solution: Here the frequency distribution is

Class Interval	Frequency
0-10	3
10-20	5
30-40	7
40-50	9
50-60	4

Corresponding to highest frequency 9 model class is 40-50 and we have

$L = 40,\ f_1 = 9,\quad f_o = 7,\ f_2 = 4,\ h = 10$

Applying the formula,

$$\text{Mode} \;=\; 40 + \frac{(9-7)}{(2\times 9-7-4)}\times 10$$

$$= 42.86$$

E9) Calculate mode from the data given in E6)

1.6.1 Relationship between Mean, Median and Mode

For a symmetrical distribution the mean, median and mode coincide. But if the distribution is moderately asymmetrical, there is an empirical relationship between them. The relationship is

Mean – Mode = 3 (Mean – Median)

Mode = 3 Median – 2 Mean

Note: Using this formula, we can calculate mean/median/mode if other two of them are known.

E10) In an asymmetrical distribution the mode and mean are 35.4 and 38.6 respectively. Calculate the median.

1.6.2 Merits and Demerits of Mode

Merits of Mode

1. Mode is the easiest average to understand and also easy to calculate;
2. It is not affected by extreme values;
3. It can be calculated for open end classes;
4. As far as the modal class is confirmed the pre-modal class and the post modal class are of equal width; and
5. Mode can be calculated even if the other classes are of unequal width.

Demerits of Mode

1. It is not rigidly defined. A distribution can have more than one mode;
2. It is not utilizing all the observations;
3. It is not amenable to algebraic treatment; and
4. It is greatly affected by sampling fluctuations.

1.7 GEOMETRIC MEAN

The geometric mean (GM) of n observations is defined as the n-th root of the product of the n observations. It is useful for averaging ratios or proportions. It is the ideal average for calculating index numbers (index numbers are economic barometers which reflect the change in prices or commodity consumption in the current period with respect to some base period taken as standard). It fails to give the correct average if an observation is zero or negative.

1. **For Ungrouped Data**

If $x_1, x_2, ..., x_n$ are the n observations of a variable X then their geometric mean is

$$GM = \sqrt[n]{x_1 x_2 ... x_n}$$

$$GM = (x_1 x_2 ... x_n)^{\frac{1}{n}}$$

Taking log of both sides

$$\log GM = \frac{1}{n} \log (x_1 x_2 ... x_n)$$

$$\log GM = \frac{1}{n} (\log x_1 + \log x_2 + ... + \log x_n)$$

$$\Rightarrow GM = \text{Antilog} \left(\frac{1}{n} \sum_{i=1}^{n} \log x_i \right)$$

Example 11: Find geometric mean of 2, 4, 8.

Solution: $GM = (x_1\, x_2 \ldots x_n)^{\frac{1}{n}}$

$$GM = (2 \times 4 \times 8)^{\frac{1}{3}} = (64)^{\frac{1}{3}} = 4$$

Thus, geometric mean is 4.

E11) Find the GM of 4, 8, 16.

E12) Calculate GM of 5, 15, 25, 35.

---

2. For Grouped data

If $x_1 x_2, ..., x_k$ are k values (or mid values in case of class intervals) of a variable X with frequencies $f_1, f_2, ..., f_k$ then

$$GM = \left(x_1^{f_1} x_2^{f_2} ... x_k^{f_k} \right)^{\frac{1}{\sum f_i}}$$

$$GM = \left(x_1^{f_1} x_2^{f_2} ... x_k^{f_k} \right)^{\frac{1}{N}}$$

where $N = f_1 + f_2 + ... + f_k$

Taking log of both sides

$$\log GM = \frac{1}{N}\log\left(x_1^{f_1}x_2^{f_2}\ldots x_k^{f_k}\right)$$

$$\log GM = \frac{1}{N}\left(f_1\log x_1 + f_2\log x_2 + \ldots + f_k\log x_k\right)$$

$$\Rightarrow GM = \text{Antilog}\left(\frac{1}{N}\sum_{i=1}^{k} f_i \log x_i\right)$$

Example 12: For the data in Example 4, calculate geometric mean.

Solution:

Class	Mid Value (x)	Frequency (f)	log x	f× log x
0-10	5	3	0.6990	2.0970
10-20	15	5	1.1761	5.8805
20-30	25	7	1.3979	9.7853
30-40	35	9	1.5441	13.8969
40-50	45	4	1.6532	6.6128
		N = 28		$\sum f_i \log x_i = 38.2725$

Using the formula

$$GM = \text{Antilog}\left(\frac{1}{N}\sum_{i=1}^{k} f_i \log x_i\right)$$

$$= \text{Antilog}\left(\frac{38.2725}{28}\right)$$

$$= \text{Antilog}\ (1.3669)$$

$$= 23.28$$

E13) Calculate GM of the following distribution

Class	0-10	10-20	20-30	30-40	40-50
Frequency	12	15	25	18	10

1.7.1 Merits and Demerits of Geometric Mean

Merits of Geometric Mean

1. It is rigidly defined;
2. It utilizes all the observations;
3. It is amenable to algebraic treatment (the reader should verify that if GM_1 and GM_2 are Geometric Means of two series-Series 1 of size n and Series 2 of size m respectively, then Geometric Mean of the combined series is given by

   Log GM = (n GM_1 + m GM_2) / (n + m);
4. It gives more weight to small items; and
5. It is not affected greatly by sampling fluctuations.

Demerits of Geometric Mean

1. Difficult to understand and calculate; and
2. It becomes imaginary for an odd number of negative observations and becomes zero or undefined if a single observation is zero.

1.8 HARMONIC MEAN

The harmonic mean (HM) is defined as the reciprocal (inverse) of the arithmetic mean of the reciprocals of the observations of a set.

1. For Ungrouped Data

If $x_1, x_2, ..., x_n$ are the n observations of a variable X, then their harmonic mean is

$$\text{HM} = \frac{1}{\frac{1}{n}\left[\frac{1}{x_1}+\frac{1}{x_2}+...+\frac{1}{x_n}\right]}$$

$$\text{HM} = \frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}}$$

Example 13: Calculate the Harmonic mean of 1, 3, 5, 7

Solution: Formula for harmonic mean is

$$\text{HM} = \frac{1}{\frac{1}{n}\left[\frac{1}{x_1}+\frac{1}{x_2}+...+\frac{1}{x_n}\right]}$$

$$\text{HM} = \frac{1}{\frac{1}{4}\left[\frac{1}{1}+\frac{1}{3}+\frac{1}{5}+\frac{1}{7}\right]}$$

= 4 / (1.0000 + 0.3333 + 0.2000 + 0.1428)

= 4/1.6761 = 2.39

E14) Calculate the Harmonic Mean of 2, 4, 6.

2. For Grouped Data

If $x_1 x_2, ..., x_k$ are k values (or mid values in case of class intervals) of a variable X with their corresponding frequencies $f_1, f_2, ..., f_k$, then

$$\text{HM} = \frac{1}{\frac{1}{N}\left[\frac{f_1}{x_1}+\frac{f_2}{x_2}+...+\frac{f_k}{x_k}\right]}$$

$$HM = \frac{N}{\sum_{i=1}^{k} \frac{f_i}{x_i}} \qquad \text{where, } N = \sum_{i=1}^{k} f_i.$$

When equal distances are travelled at different speeds, the average speed is calculated by the harmonic mean. It cannot be calculated if an observation is zero.

Example 14: For the data given in Example 4, calculate harmonic mean.

Solution:

Class	Mid Value (x)	Frequency (f)	f/x
0-10	5	3	0.600
10-20	15	5	0.330
20-30	25	7	0.280
30-40	35	9	0.257
40-50	45	4	0.088
		N = ∑f = 28	∑f/x = 1.555

Using the formula,

$$HM = \frac{N}{\sum_{i=1}^{k} \frac{f_i}{x_i}}$$

$$= 28/1.555 = 17.956$$

E15) Calculate harmonic mean for the given data:

Class	0-10	10-20	20-30	30-40	40-50
Frequency	3	5	10	8	4

1.8.1 Merits and Demerits of Harmonic Mean

Merits of Harmonic mean

1. It is rigidly defined;
2. It utilizes all the observations;
3. It is amenable to algebraic treatment; and
4. It gives greater importance to small items.

Demerits of Harmonic Mean

1. Difficult to understand and compute.

1.8.2 Relations between AM, GM and HM

Relation 1: AM ≥ GM ≥ HM

Proof: Let x_1 and x_2 be two real numbers which are non-zero and non negative. Then

$$AM = \frac{x_1 + x_2}{2}$$

$$GM = \sqrt{x_1.x_2}$$

$$HM = \frac{2}{\frac{1}{x_1} + \frac{1}{x_2}}$$

Consider $\left(\sqrt{x_1} - \sqrt{x_2}\right)^2 \geq 0$

$$x_1 + x_2 - 2\sqrt{x_1}x_2 \geq 0$$

$$\frac{x_1 + x_2}{2} \geq \sqrt{x_1 x_2} \quad \text{so} \quad AM \geq GM \qquad \text{... (1)}$$

Again $\left(\frac{1}{\sqrt{x_1}} - \frac{1}{\sqrt{x_2}}\right)^2 \geq 0$

$$\frac{1}{x_1} + \frac{1}{x_2} - \frac{2}{\sqrt{x_1}.\sqrt{x_2}} \geq 0$$

$$\sqrt{x_1 x_2} \geq \frac{2}{\frac{1}{x_1} + \frac{1}{x_2}} \quad \text{so} \quad GM \geq HM \qquad \text{… (2)}$$

So by equations (1) & (2)

$$AM \geq GM \geq HM$$

Relation 2: GM = $\sqrt{AM.HM}$

Proof: Let x_1 and x_2 be two real numbers which are non-zero and non negative. Then

$$AM \times HM = \frac{x_1 + x_2}{2} \cdot \frac{2}{\frac{1}{x_1} + \frac{1}{x_2}}$$

$$= \left(x_1 + x_2\right) \frac{x_1 x_2}{\left(x_1 + x_2\right)}$$

$$= x_1 x_2$$

$$\left(GM\right)^2 = AM \times HM$$

So GM = $\sqrt{AM \times HM}$

1.9 PARTITION VALUES

Partition values are those values of variable which divide the distribution into a certain number of equal parts. Here it may be noted that the data should be arranged in ascending or descending order of magnitude. Commonly used partition values are quartiles, deciles and percentiles. For example, quartiles divide the data into four equal parts. Similarly, deciles and percentiles divide the distribution into ten and hundred equal parts, respectively.

1.9.1 Quartiles

Quartiles divide whole distribution in to four equal parts. There are three quartiles- 1st Quartile denoted as Q_1, 2nd Quartile denoted as Q_2 and 3rd Quartile as Q_3, which divide the whole data in four parts. 1st Quartile contains the ¼ part of data, 2nd Quartile contains ½ of the data and 3rd Quartile contains the ¾ part of data. Here, it may be noted that the data should be arranged in ascending or descending order of magnitude.

For Ungrouped Data

For obtaining the quartiles first of all we have to arrange the data either in ascending order or in descending order of their magnitude. Then, we find the $(N/4)^{th}$, $(N/2)^{th}$ and $(3N/4)^{th}$ placed item in the arranged data for finding out the 1st Quartile (Q_1) , 2nd Quartile (Q_2) and 3rd Quartile (Q_3) respectively. The value of the $(N/4)^{th}$ placed item would be the 1st Quartile (Q_1) , value of $(N/2)^{th}$ placed item would be the 2nd Quartile (Q_2) and value of $(3N/4)^{th}$ placed item would be the 3rd Quartile (Q_3) of that data.

Mathematically, if $x_1 x_2,...,x_N$ are N values of a variable X then 1st Quartile (Q_1) , 2nd Quartile (Q_2) and 3rd Quartile (Q_3) are defined as:

First Quartile (Q_1) = $\frac{N}{4}$th placed item in the arranged data

Second Quartile (Q_2) = $\frac{N}{2}$th placed item in the arranged data

Third Quartile (Q_3) = $3\left(\frac{N}{4}\right)$th placed item in the arranged data

For Grouped Data

If $x_1, x_2,..., x_k$ are k values (or mid values in case of class intervals) of a variable X with their corresponding frequencies $f_1, f_2,..., f_k$, then first of all we form a cumulative frequency distribution. After that we determine the i^{th} quartile class as similar as we do in case of median.

The i^{th} quartile is denoted by Q_i and defined as

$$Q_i = L + \frac{\left(\frac{iN}{4} - C\right)}{f} \times h \qquad \text{for} \quad i = 1, 2, 3$$

where, L = lower class limit of i^{th} quartile class,

h = width of the i^{th} quartile class,

N = total frequency,

C = cumulative frequency of pre i^{th} quartile class, and

f = frequencies of i^{th} quartile class.

“i” denotes i^{th} quartile class. It is the class in which $\left(\frac{i \times N}{4}\right)^{th}$ observation falls in cumulative frequency. It is easy to see that the second quartile (i = 2) is the median.

1.9.2 Deciles

Deciles divide whole distribution in to ten equal parts. There are nine deciles. D_1, D_2,...,D_9 are known as 1st Decile, 2nd Decile,...,9th Decile respectively and i^{th} Decile contains the $(iN/10)^{th}$ part of data. Here, it may be noted that the data should be arranged in ascending or descending order of magnitude.

For Ungrouped Data

For obtaining the deciles first of all we have to arrange the data either in ascending order or in descending order of their magnitude. Then, we find the $(1N/10)^{th}$, $(2N/10)^{th}$,,.., $(9N/10)^{th}$ placed item in the arranged data for finding out the 1st decile (D_1), 2nd decile (D_2), ..., 9th decile (D_9) respectively. The value of the $(N/10)^{th}$ placed item would be the 1st decile, value of $(2N/10)^{th}$ placed item would be the 2nd decile. Similarly, the value of $(9N/10)^{th}$ placed item would be the 9th decile of that data.

Mathematically, if $x_1 x_2,...,x_N$ are N values of a variable X then the i^{th} decile is defined as:

i^{th} Decile (D_i) = $\frac{i\,N}{10}$ th placed item in the arranged data (i = 1, 2, 3, … ,9)

For Grouped Data

If $x_1,\ x_2,\ ...,x_k$ are k values (or mid values in case of class intervals) of a variable X with their corresponding frequencies $f_1, f_2,..., f_k$, then first of all we form a cumulative frequency distribution. After that we determine the i^{th} deciles class as similar as we do in case of quartiles.

The i^{th} decile is denoted by D_i and given by

$$D_i = L + \frac{\left(\frac{iN}{10} - C\right)}{f} \times h \qquad \text{for} \quad i = 1, 2, ..., 9$$

where, L = lower class limit of i^{th} decile class,

h = width of the i^{th} decile class,

N = total frequency,

C = cumulative frequency of pre i^{th} decile class; and

f = frequency of i^{th} decile class.

“i” denotes i^{th} decile class. It is the class in which $\left(\frac{i \times N}{10}\right)^{th}$ observation falls in cumulative frequency. It is easy to see that the fifth quartile (i =5) is the median.

1.9.3 Percentiles

Percentiles divide whole distribution in to 100 equal parts. There are ninety nine percentiles. P_1, P_2, …,P_{99} are known as 1st percentile, 2nd percentile,…,99th percentile and ith percentile contains the $(iN/100)^{th}$ part of data. Here, it may be noted that the data should be arranged in ascending or descending order of magnitude.

For Ungrouped Data

For obtaining the percentiles first of all we have to arrange the data either in ascending order or in descending order of their magnitude. Then, we find the $(1N/100)^{th}$, $(2N/100)^{th}$,..., $(99N/100)^{th}$ placed item in the arranged data for finding out the 1st percentile (P_1) , 2nd percentile (P_2),...,99th percentile (P_{99}) respectively. The value of the $(N/100)^{th}$ placed item would be the 1st percentile, value of $(2N/100)^{th}$ placed item would be the 2nd percentile. Similarly, the value of $(99N/100)^{th}$ placed item would be the 99th percentile of that data.

Mathematically, if $x_1 x_2, ..., x_N$ are N values of a variable X then the ith percentile is defined as:

$$i^{th} \text{ Percentile } (P_i) = \frac{i\,N}{100} \text{th placed item in the arranged data} (i = 1, 2, \dots ,99)$$

For Grouped Data

If $x_1, x_2, ..., x_k$ are k values (or mid values in case of class intervals) of a variable X with their corresponding frequencies $f_1, f_2, ..., f_k$, then first of all we form a cumulative frequency distribution. After that we determine the ith percentile class as similar as we do in case of median.

The ith percentile is denoted by P_i and given by

$$P_i = L + \frac{\left(\frac{iN}{100} - C\right)}{f} \times h \qquad \text{for} \quad i = 1, 2, ..., 99$$

where, L = lower limit of ith percentile class,

h = width of the ith percentile class,

N = total frequency,

C = cumulative frequency of pre ith prcentile class; and

f = frequency of ith percentile class.

"i" denotes ith percentile class. It is the class in which $\left(\frac{i \times N}{100}\right)^{th}$ observation falls in cumulative frequency.

It is easy to see that the fiftieth percentile (i =50) is the median.

Example 15: For the data given in Example 4, calculate the first and third quartiles.

Solution: First we find cumulative frequency given in the following cumulative frequency table:

Class Interval	**Frequency**	**Cumulative Frequency (< type)**
0-10	3	3
10-20	5	8
20-30	7	15
30-40	9	24
40-50	4	28
	$N=\sum f = 28$	

Here, N/4 = 28/4 = 7. The 7th observation falls in the class 10-20. So, this is the first quartile class. 3N/4 = 21th observation falls in class 30-40, so it is the third quartile class.

For first quartile L = 10, f = 5, C = 3, N = 28

$$Q_1 = 10+\frac{(7-3)}{5}\times 10 = 18$$

For third quartile L = 30, f = 9, C = 15

$$Q_3 = 30+\frac{(21-15)}{9}\times 10 = 36.67$$

E16) Calculate the first and third quartiles for the data given in E6)

1.10 SUMMARY

In this unit, we have discussed:

1. How to describe an average;
2. The utility of an average;
3. The properties of a good average;
4. The different types of averages along with their merits and demerits; and
5. The different kinds of partition values.

1.11 SOLUTIONS / ANSWERS

E1) For calculating the arithmetic mean, we add all the observations and divide by 6 as follows:

$$\overline{X} = \frac{\sum x}{n} = \frac{5+8+12+15+20+30}{6} = 15$$

Using short-cut method suppose the assumed mean A = 15.

x	**d = x–A**
5	–10
8	–7
12	–3
15	0
20	+5
30	+15
	$\sum d_i = 0$

$$\overline{X} = A + \frac{\sum d}{n} \quad = 15 + \frac{0}{6} \quad = 15$$

E2) We have the following frequency distribution:

Wages	**No. of Workers**	**fx**
20	5	100
25	8	200
30	20	600
35	10	350
40	5	200
50	2	100
	50	$\sum xf = 1550$

$$\overline{X} = \frac{\sum_{i=1}^{6} f_i x_i}{\sum_{i=1}^{6} f_i} \quad = \frac{1550}{50} = 31$$

Using short cut method with assumed mean A = 30

x	**f**	**d = x - 30**	**fd**
20	5	–10	–50
25	8	–5	–40
30	20	0	0
35	10	5	50
40	5	10	50
50	2	20	40
	N = 50		$\sum f\,d = 50$

$$\overline{X} = A + \frac{\sum_{i=1}^{6} f_i d_i}{N}$$

$$= 30 + \frac{50}{50} = 30 + 1 = 31$$

E3) We have the following frequency distribution:

Marks	**No. of Students (f)**	**Mid Points x**	**fx**
0-10	6	5	30
10-20	9	15	135
20-30	17	25	425
30-40	10	35	350
40-50	8	45	360
	$\sum f_i = 50$		$\sum f_i x_i = 1300$

$$\overline{X} = \frac{\sum_{i=1}^{5} f_i x_i}{\sum_{i=1}^{5} f_i} = \frac{1300}{50} = 26$$

Using short-cut method

Marks	**f**	**x**	$\mathbf{d = \frac{x-25}{10}}$	**f d**
0-10	6	5	–2	–12
10-20	9	15	–1	–9
20-30	17	25=A	0	0
30-40	10	35	+1	+10
40-50	8	45	+2	+16
	$\sum f = 50$			$\sum fd = 5$

Now $$\overline{X} = A + \frac{\sum fd}{N} \times h$$

$$\overline{X} = 25 + \frac{5}{50} \times 10$$

$$= 26$$

E4) (1) Here n = 7, after arranging in ascending order, we get

2, 3, 6, 8, 10, 12, 15 and the Median value will be

$$\text{Median} = \text{value of } \left(\frac{n+1}{2}\right)^{th} \text{ item}$$

$$= \text{value of } 4^{th} \text{ item} = 8$$

(2) Here n = 8, so arranging in ascending order we get the values as 2, 3, 5, 6, 8, 10, 12, 15 and therefore

$$M_d = \frac{\left(\frac{n}{2}\right)^{th} \text{observation} + \left(\frac{n}{2}+1\right)^{th} \text{observation}}{2}$$

$$= \frac{4^{th}\text{ Observation} + 5^{th}\text{ Observation}}{2}$$

$$M_d = \frac{6+8}{2} = 7$$

E5)

Marks	5	10	15	20	25	30	35	40	45	50
No. of students	2	5	10	14	16	20	13	9	7	4
Cumulative Frequency	2	7	17	31	47	67	80	89	96	100

Now, Median = Value of the variable corresponding to the $\left(\frac{N}{2}\right)^{th}$ cumulative frequency

= Value of the variable corresponding to the $\left(\frac{100}{2}\right)^{th}$ cumulative frequency

= Value of the variable corresponding to the 50th cumulative frequency,

Since 50 is not among cumulative frequency so the next cumulative frequency is 67 and the value of variable against 67 is 30. Therefore 30 is the median.

E6) First we shall calculate the cumulative frequency distribution

Marks	**f**	**Cumulative Frequency**
0-10	5	5
10-20	10	15
20-30	15	30 = C
30-40	**20 = f**	**50**
40-50	12	62
50-60	10	72
60-70	8	80
	N= 80	

Here $\frac{N}{2} = \frac{80}{2} = 40$,

Since, 40 is not in the cumulative frequency so, the class corresponding to the next cumulative frequency 50 is median class. Thus 30-40 is median class.

$$\text{Median} = L + \frac{\left(\frac{N}{2} - C\right)}{f} \times h$$

$$= 30 + \frac{40-30}{20} \times 10$$

$$= 35$$

E7) First we shall calculate the cumulative frequency distribution

Class	**f**	**Cumulative Frequency**
0-20	5	5
20-40	15	20
40-60	30	50
60-80	f_4	$50+f_4$
80-100	8	$58+f_4$

Median class is 40-60 since median value is given 50,

$$\text{Median} = L + \frac{\left(\frac{N}{2} - C\right)}{f} \times h$$

$$50 = 40 + \frac{\left(\frac{58+f_4}{2} - 20\right)}{30} \times 20$$

$$50 - 40 = \frac{58+f_4-40}{3}$$

$$18 + f_4 = 30$$

$$f_4 = 30 - 18 = 12$$

E8) First we shall form the frequency distribution

Size	1	2	3	4	5	6	7	8
Frequency	1	2	1	2	2	2	4	1

From the frequency distribution we see that size 7 occurs with maximum frequency of 4 hence mode is 7.

E9) First we shall form the frequency distribution

Marks	f
0-10	5
10-20	10
2-30	15
30-40	20
40-50	12
50-60	10
60-70	8

Here mode class is 30-40 corresponding to the highest frequency 20.

Mode $$M_0 = L + \frac{(f_1 - f_0)}{(2f_1 - f_0 - f_2)} \times h$$

$$= 30 + \frac{20-15}{40-15-12} \times 10$$

$$= 30 + \frac{5}{13} \times 10$$

$$= 30 + 3.84 = 33.84$$

E10) We have

Mode = 3 Median – 2 Mean

35.4 = 3 Median – 2 (38.6)

3 Median = 35.4 + 77.2

Median = 112.6/3 = 37.53

E11) We have

$$GM = \sqrt[3]{4.8.16}$$

$$\Rightarrow \sqrt[3]{2^2.2.4.4^2}$$

$$\Rightarrow \sqrt[3]{2^3.4^3}$$

$$\Rightarrow 2 \times 4 = 8$$

E12) First we shall form the frequency distribution

x	log x
5	0.6990
15	1.1761
25	1.3979
35	1.5441
	$\sum \log x = 4.7971$

$$\text{Now GM} = \text{antilog}\left[\frac{\sum \log x}{n}\right]$$

$$= \text{antilog}\left[\frac{4.7971}{4}\right]$$

$$= \text{antilog}\ (1.1993) = 15.82$$

E13) We have the following frequency distribution:

Class	**f**	**x**	**log x**	**flog x**
0-10	12	5	0.6690	8.0280
10-20	15	15	1.1761	17.6415
20-30	25	25	1.3979	34.9475
30-40	18	35	1.5441	27.7938
40-50	10	45	1.6532	16.5320
	80			$\sum f_i \log x_i = 104.9428$

$$\text{GM} = \text{antilog}\left[\frac{\sum f \log x}{N}\right]$$

$$= \text{antilog}\left[\frac{104.9428}{80}\right]$$

$$= \text{antilog}\ (1.3118) = 20.52$$

E14) We have the formulae of the Harmonic mean

$$\text{HM} = \frac{1}{\frac{1}{n}\left[\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_n}\right]}$$

By putting the given values

$$\text{HM} = \frac{1}{\frac{1}{3}\left[\frac{1}{2} + \frac{1}{4} + \frac{1}{6}\right]}$$

$$= \frac{3 \times 12}{[6 + 3 + 2]}$$

$$= \frac{3 \times 12}{11} = 3.27$$

E15) We have given the following frequency distribution:

Class	**f**	**x**	**f/x**
0-10	3	5	0.600
10-20	5	15	0.333
20-30	10	25	0.400
30-40	8	35	0.225
40-50	4	45	0.088
	N = 30		$\sum \frac{f}{x} = 1.646$

Therefore by putting the values in formulae, we get

$$\text{HM} = \frac{N}{\sum_{i=1}^{5} \frac{f_i}{x_i}}$$

$$= 3\ 0 / 1.646 = 19.22$$

E16) First we shall calculate the cumulative frequency distribution

Marks	**f**	**Cumulative Frequency**
0-10	5	5
10-20	10	15
20-30	15	30
30-40	20	50
40-50	12	62
50-60	10	72
60-70	8	80
	N= 80	

$$\text{Here } \frac{N}{2} = \frac{80}{2} = 40,$$

Here, N/4 = 80/4 = 20. The 20th observation falls in the class 20-30. So, this is the first quartile class. 3N/4 = 3×80/4 = 60th observation falls in class 40-50, so it is the third quartile class.

For first quartile L = 20, f = 15, C = 15, N = 80

$$Q_1 = 20 + \frac{(20-15)}{15} \times 10 = 23.33$$

For third quartile L = 40, f = 12, C = 50

$$Q_3 = 40 + \frac{(60-50)}{12} \times 10 = 48.33$$

UNIT 2 MEASURES OF DISPERSION

Structure

2.1 Introduction

Objectives

2.2 Measures of Dispersion

2.3 Range

2.4 Quartile Deviation

2.5 Mean Deviation

2.6 Variance and Standard Deviation

2.7 Root Mean Square Deviation

2.8 Coefficient of Variation

2.9 Summary

2.10 Solutions / Answers

2.1 INTRODUCTION

Different measures of central tendency, discussed in Unit 1 of this block, give a value around which the data is concentrated. But it gives no idea about the nature of scatter or spread. For example, the observations 10, 30 and 50 have mean 30 while the observations 28, 30, 32 also have mean 30. Both the distributions are spread around 30. But it is observed that the variability among units is more in the first than in the second. In other words, there is greater variability or dispersion in the first set of observations in comparison to other. Measure of dispersion is calculated to get an idea about the variability in the data.

In Section 2.2, the concepts of measures of dispersion are described. Significance and properties of a good measure of dispersion are also explored in Sub-sections 2.2.1 and 2.2.2. The basic idea about the range is explained in Section 2.3 whereas the measures of quartile deviation are introduced in Section 2.4. Mean deviation is described in Section 2.5. Methods of calculation of variance and standard deviation for grouped and ungrouped data are explained in Section 2.6 whereas in Section 2.7 the basis idea about root mean square deviation is provided. Coefficient of variation, which is a relative measure of variation is described is Section 2.8.

Objectives

After reading this unit, you would be able to

- conceptualize dispersion;
- explain the utility of a measure of dispersion;
- explain the properties of a good measure of dispersion;
- explain methods of calculation of different types of measures of dispersion along with their merits and demerits; and
- solve the numerical problems related to the measures of dispersion.

2.2 MEASURES OF DISPERSION

According to Spiegel, the degree to which numerical data tend to spread about an average value is called the variation or dispersion of data. Actually, there are two basic kinds of a measure of dispersion (i) Absolute measures and (ii) Relative measures. The absolute measures of dispersion are used to measure the variability of a given data expressed in the same unit, while the relative measures are used to compare the variability of two or more sets of observations. Following are the different measures of dispersion:

1. Range
2. Quartile Deviation
3. Mean Deviation
4. Standard Deviation and Variance

2.2.1 Significance of Measures of Dispersion

Measures of dispersion are needed for the following four basic purposes:

1. Measures of dispersion determine the reliability of an average value. In other words, measures of variation are pointed out as to how far an average is representative of the entire data. When variation is less, the average closely represents the individual values of the data and when variation is large; the average may not closely represent all the units and be quite unreliable.
2. Another purpose of measuring variation is to determine the nature and causes of variations in order to control the variation itself. For example the variation in the quality of product in the process form of production can be checked by quality control department by identifying the reason for the variations in the quality of product. Thus, measurements of dispersion are helpful to control the causes of variation.
3. The measures of dispersion enable us to compare two or more series with regard to their variability. The relative measures of dispersion may also determine the uniformity or consistency. Smaller value of relative measure of dispersion implies greater uniformity or consistency in the data.
4. Measures of dispersion facilitate the use of other statistical methods. In other words, many powerful statistical tools in statistics such as correlation analysis, the testing of hypothesis, the analysis of variance, techniques of quality control, etc. are based on different measures of dispersion.

2.2.2 Properties of Good Measure of Dispersion

The properties of a good measure of dispersion are similar to the properties of a good measure of average. So, a good measure of dispersion should possess the following properties:

1. It should be simple to understand;
2. It should be easy to compute;
3. It should be rigidly defined;
4. It should be based on each and every observations of data;
5. It should be amenable to further algebraic treatment;

6. It should have sampling stability; and
7. It should not be unduly affected by extreme observations.

The detailed description of the properties has been discussed in Unit 1 of this block. Different measures of dispersions are discussed in following sections.

2.3 RANGE

Range is the simplest measure of dispersion. It is defined as the difference between the maximum value of the variable and the minimum value of the variable in the distribution. Its merit lies in its simplicity. The demerit is that it is a crude measure because it is using only the maximum and the minimum observations of variable. However, it still finds applications in Order Statistics and Statistical Quality Control. It can be defined as

$$R = X_{Max} - X_{Min}$$

where, X_{Max}: Maximum value of variable and

X_{Min}: Minimum value of variable.

Example 1: Find the range of the distribution 6, 8, 2, 10, 15, 5, 1, 13.

Solution: For the given distribution, the maximum value of variable is 15 and the minimum value of variable is 1. Hence range = 15 -1 = 14.

E1) Calculate the range for the following data:

60, 65, 70, 12, 52, 40, 48

E2) Calculate range for the following frequency distribution:

Class	0-10	10-20	20-30	30-40	40-50
Frequency	2	6	12	7	3

2.3.1 Merits and Demerits of Range

Merits of Range

1. It is the simplest to understand;
2. It can be visually obtained since one can detect the largest and the smallest observations easily and can take the difference without involving much calculations; and
3. Though it is crude, it has useful applications in areas like order statistics and statistical quality control.

Demerits of Range

1. It utilizes only the maximum and the minimum values of variable in the series and gives no importance to other observations;
2. It is affected by fluctuations of sampling;

3. It is not very suitable for algebraic treatment;
4. If a single value lower than the minimum or higher than the maximum is added or if the maximum or minimum value is deleted range is seriously affected; and
5. Range is the measure having unit of the variable and is not a pure number. That's why sometimes coefficient of range is calculated by

$$\text{Cofficient of Range} = \frac{X_{Max} - X_{Min}}{X_{Max} + X_{Min}}$$

2.4 QUARTILE DEVIATION

As you have already studied in Unit 1 of this block that Q_1 and Q_3 are the first quartile and the third quartile respectively. $(Q_3 - Q_1)$ gives the inter quartile range. The semi inter quartile range which is also known as Quartile Deviation (QD) is given by

$$\text{Quartile Deviation (QD)} = (Q_3 - Q_1) / 2$$

Relative measure of Q.D. known as Coefficient of Q.D. and is defined as

$$\text{Cofficient of QD} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Example 2: For the following data, find the quartile deviation:

Class Interval	0-10	10-20	20-30	30-40	40-50
Frequency	03	05	07	09	04

Solution: We have N/4 = 28/4 = 7 and 7th observation falls in the class 10-20. This is the first quartile class.

Similarly, 3N/4 = 21 and 21st observation falls in the interval 30-40. This is the third quartile class.

Class Interval	**Frequency**	**Cumulative Frequency**
0-10	3	3
10-20	5	8
20-30	7	15
30-40	9	24
40-50	4	28

Using the formulae of first quartile and third quartile we found

$$Q_1 = 10 + \frac{(7-3)}{5} \times 10 = 18$$

$$Q_3 = 30 + \frac{(21-15)}{9} \times 10 = 36.67$$

Hence

$$\text{Quartile Deviation} = (36.67 - 18)/2 = 9.335$$

E3) Calculate the Quartile Deviation for the following data:

Class	0-5	5-10	10-15	15-20	20-25	25-30	30-35	35-40
Frequency	6	8	12	24	36	32	24	8

2.5 MEAN DEVIATION

Mean deviation is defined as average of the sum of the absolute values of deviation from any arbitrary value viz. mean, median, mode, etc. It is often suggested to calculate it from the median because it gives least value when measured from the median.

The deviation of an observation x_i from the assumed mean A is defined as (x_i – A).

Therefore, the mean deviation can be defined as

$$MD = \frac{1}{n}\sum_{i=1}^{n}|x_i - A|$$

The quantity $\sum | x_i - A |$ is minimum when A is median.

We accordingly define mean deviation from mean as

$$MD = \frac{\sum_{i=1}^{n}|x_i - \overline{x}|}{n}$$

and from the median as

$$MD = \frac{\sum_{i=1}^{n}|x_i - \text{median}|}{n}$$

For frequency distribution, the formula will be

$$MD = \frac{\sum_{i=1}^{k} f_i|x_i - \overline{x}|}{\sum_{i=1}^{k} f_i}$$

$$MD = \frac{\sum_{i=1}^{k} f_i|x_i - \text{median}|}{\sum_{i=1}^{k} f_i}$$

where, all symbols have usual meanings.

Example 3: Find mean deviation for the given data

1, 2, 3, 4, 5, 6, 7

Solution: First of all we find Mean

$$\overline{x} = \frac{1+2+3+4+5+6+7}{7} = \frac{28}{7} = 4$$

Then, we will find $|x_i - \overline{x}|$: 3, 2, 1, 0 1, 2, 3

So, $\sum |x_i - \overline{x}| = 12$

Therefore,

$$MD = \frac{12}{7} = 1.71$$

Example 4: Find mean deviation from mean for the following data:

x	1	2	3	4	5	6	7
f	3	5	8	12	10	7	5

Solution: First of all we have to calculate the mean from the given data

x	**f**	**fx**	$\lvert \mathbf{x} - \overline{\mathbf{x}} \rvert$	$\mathbf{f}\lvert \mathbf{x} - \overline{\mathbf{x}} \rvert$
1	3	3	3.24	9.72
2	5	10	2.24	11.20
3	8	24	1.24	9.92
4	12	48	0.24	2.88
5	10	50	0.76	7.60
6	7	42	1.76	12.32
7	5	35	2.76	13.80
	50	212	12.24	67.44

$$\overline{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{212}{50} = 4.24$$

$$MD = \frac{\sum_{i=1}^{k} f_i |x_i - \overline{x}|}{\sum_{i=1}^{k} f_i} = \frac{67.44}{50} = 1.348$$

Example 5: Consider the data given in Example 2 and find the Mean deviation from median.

Solution:

Class	**Mid Value (x)**	**Frequency (f)**	**C.F.**	$\mathbf{f_i\,|x_i - median|}$
0-10	5	3	3	70.71
10-20	15	5	8	67.85
20-30	25	7	15	24.99
30-40	35	9	24	57.87
40--50	45	4	28	65.72
		$\sum f = 28$		$\sum f_i\,|x_i - median| = 287.14$

We have N/2 = 28/2 = 14

The 14th observation falls in the class 20-30. This is therefore the median class.

Using the formula of median, Median = $20+\frac{14-8}{7}\times 10$ = 28.57

Mean Deviation from Median $= \frac{\sum_{i=1}^{k} f_i \left|x_i - \text{median}\right|}{\sum_{i=1}^{k} f_i} =$ 10.255

E4) Following are the marks of 7 students in Statistics:

16, 24, 13, 18, 15, 10, 23

Find the mean deviation from mean.

E5) Find mean deviation for the following distribution:

Class	0-10	10-20	20-30	30-40	40-50
Frequency	5	8	15	16	6

2.5.1 Merits and Demerits of Mean Deviation

Merits of Mean Deviation

1. It utilizes all the observations;
2. It is easy to understand and calculate; and
3. It is not much affected by extreme values.

Demerits of Mean Deviation

1. Negative deviations are straightaway made positive;
2. It is not amenable to algebraic treatment; and
3. It can not be calculated for open end classes.

2.6 VARIANCE AND STANDARD DEVIATION

In the previous section, we have seen that while calculating the mean deviation, negative deviations are straightaway made positive. To overcome this drawback we move towards the next measure of dispersion called variance. Variance is the average of the square of deviations of the values taken from mean. Taking a square of the deviation is a better technique to get rid of negative deviations.

Variance is defined as

$$\text{Var}(x) = \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2$$

and for a frequency distribution, the formula is

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{k} f_i\left(x_i - \bar{x}\right)^2$$

where, all symbols have their usual meanings.

It should be noted that sum of squares of deviations is least when deviations are measured from the mean. This means $\sum(x_i - A)^2$ is least when A = Mean.

Example 6: Calculate the variance for the data given in Example 3.

Solution:

x	$(\mathbf{x}-\bar{\mathbf{x}})$	$(\mathbf{x}-\bar{\mathbf{x}})^2$
1	-3	9
2	-2	4
3	-1	1
4	0	0
5	1	1
6	2	4
7	3	9

We have $\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = 28$

Therefore, $\mathrm{Var}(x) = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = \frac{1}{7}\times 28 = 4$

Example 7: For the data given in Example 4 of Unit 1, compute variance.

Solution: We have the following data:

Class	**Mid Value**	**Frequency (f)**	$\mathbf{f}(\mathbf{x}-\bar{\mathbf{x}})^2$
0-10	5	3	1470.924
10-20	15	5	737.250
20-30	25	7	32.1441
30-40	35	9	555.606
40-50	45	4	1275.504
		$\sum f = 28$	$\sum f\left(x_i - \bar{x}\right)^2$ =4071.428

$$\text{Variance} = \frac{\sum_{i=1}^{k} f_i \left(x_i - \bar{x}\right)^2}{\sum_{i=1}^{k} f_i}$$

$$= 4071.429 / 28 = 145.408$$

2.6.1 Merits and Demerits of Variance

Merits of Variance

1. It is rigidly defined;
2. It utilizes all the observations;
3. Amenable to algebraic treatment;
4. Squaring is a better technique to get rid of negative deviations; and
5. It is the most popular measure of dispersion.

Demerits of Variance

1. In cases where mean is not a suitable average, standard deviation may not be the coveted measure of dispersion like when open end classes are present. In such cases quartile deviation may be used;
2. It is not unit free;
3. Although easy to understand, calculation may require a calculator or a computer; and
4. Its unit is square of the unit of the variable due to which it is difficult to judge the magnitude of dispersion compared to standard deviation.

2.6.2 Variance of Combined Series

In Section 2.6, we have discussed how to calculate the variance for a single variable. If there are two or more populations and the information about the means and variances of those populations are available then we can obtain the combined variance of several populations. If n_1, n_2,..., n_k are the sizes, $\bar{x}_1$, $\bar{x}_2$,…, $\bar{x}_k$ are the means and σ_1^2, σ_2^2,…,σ_k^2 are the variances of k populations, then the combined variance is given by

$$\sigma^2 = \frac{1}{\left(n_1 + n_2 + \ldots + n_k\right)}\left[n_1\left(\sigma_1^2 + d_1^2\right) + n_2\left(\sigma_2^2 + d_2^2\right) + \ldots + n_k\left(\sigma_k^2 + d_k^2\right)\right]$$

where, $d_i = \bar{x}_i - \bar{x}$ and

$$\bar{x} = \frac{\sum_{i=1}^{k} n_i \bar{x}_i}{\sum_{i=1}^{k} n_i}$$ is the mean of the combined series.

Example 8: Suppose a series of 100 observations has mean 50 and variance 20. Another series of 200 observations has mean 80 and variance 40. What is the combined variance of the given series?

Solution: First we find the mean of the combined series

$$\overline{x} = \frac{(100\times 50+200\times 80)}{(100+200)}$$

$$= \frac{21000}{300} = 70$$

Therefore, $d_1 = 50 - 70 = -20$ and $d_2 = 80 - 70 = 10$

Variance of the combined series

$$= \frac{(100\times(20+400)+200\times(40+100))}{(100+200)}$$

$$\Rightarrow \frac{(42000+28000)}{(300)} = \frac{70000}{300} = 233.33$$

E6) For a group containing 100 observations the arithmetic mean and standard deviation are 16 and $\sqrt{21}$ respectively. For 50 observations, the mean and standard deviation are 20 and 2 respectively. Calculate mean and standard deviation of other half.

2.6.3 Standard Deviation

Standard deviation (SD) is defined as the positive square root of variance. The formula is

$$\text{SD} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}}$$

and for a frequency distribution the formula is

$$\text{SD} = \sqrt{\frac{\sum_{i=1}^{k} f_i (x_i - \overline{x})^2}{\sum_{i=}^{k} f_i}}$$

where, all symbols have usual meanings. SD, MD and variance cannot be negative.

Example 9: Find the SD for the data in Example 4 of Unit 1.

Solution: In the Example 7, we have already found the

Variance = 145.408

So

$$\text{SD} = +\sqrt{(145.408)} = 12.04$$

2.6.4 Merits and Demerits of Standard Deviation

Merits of Standard Deviation

1. It is rigidly defined;
2. It utilizes all the observations;
3. It is amenable to algebraic treatment;

4. Squaring is a better technique to get rid of negative deviations; and
5. It is the most popular measure of dispersion.

Demerits of Standard Deviation

1. In cases where mean is not a suitable average, standard deviation may not be the appropriate measure of dispersion like when open end classes are present. In such cases quartile deviation may be used;
2. It is not unit free; and
3. Although it is easy to understand but calculation may require a calculator or a computer.

E7) Find the standard deviation for the following numbers:

10 27 40 60 33 30 10

E8) Calculate standard deviation for the following data:

Class	0-10	10-20	20-30	30-40	40-50
Frequency	5	8	15	16	6

Remark:

SD is algebraically more amenable than MD. MD straightaway neglects the sign of observations which is also a useful information. In SD, the signs are considered but they get automatically eliminated due to squaring. However MD will be useful in situations where Median is more suitable than mean as an average.

2.7 ROOT MEAN SQUARE DEVIATION

As we have discussed in Section 2.6 that standard deviation is the positive square root of the average of the squares of deviations taken from mean. If we take the deviations from assumed mean then it is called Root Mean Square Deviation and it is defined as

$$\text{RMSD} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - A)^2}{n}}$$

where, A is the assumed mean.

For a frequency distribution, the formula is

$$\text{RMSD} = \sqrt{\frac{\sum_{i=1}^{k} f_i (x_i - A)^2}{\sum_{i=1}^{k} f_i}}$$

When assumed mean is equal to the actual mean $\overline{x}$ i.e. $A = \overline{x}$ root mean square deviation will be equal to the standard deviation.

2.8 COEFFICIENT OF VARIATION

Coefficient of Variation (CV) is defined as

$$CV = \frac{\sigma}{\overline{x}} \times 100$$

It is a relative measure of variability. If we are comparing the two data series, the data series having smaller CV will be more consistent. One should be careful in making interpretation with CV. For example, the series 10, 10, 10 has SD zero and hence CV is also zero. The series 50, 50 and 50 also has SD zero and hence CV is zero. But the second series has higher mean. So we shall regard the second series as more consistent than the first.

Example 10: Suppose batsman A has mean 50 with SD 10. Batsman B has mean 30 with SD 3. What do you infer about their performance?

Solution: A has higher mean than B. This means A is a better run maker.

However, B has lower CV (3/30 = 0.1) than A (10/50 = 0.2) and is consequently more consistent.

E9) If $n = 10, \overline{x} = 4, \quad \sum x^2 = 200$, find the coefficient of variation.

E10) For a distribution, the coefficient of variation is 70.6% and the value of arithmetic mean is 16. Find the value of standard deviation.

2.9 SUMMARY

In this unit, we have discussed:

1. The measures of dispersion;
2. The utility of a measure of dispersion;
3. The properties of a good measure of dispersion;
4. The different types of measures of dispersion along with their merits and demerits; and
5. Coefficient of Variation.

While average gives the value around which a distribution is scattered, measure of dispersion tells how it is scattered. So if one suitable measure of average and one suitable measure of dispersion is calculated (say mean and SD), we get a good idea of the distribution even if the distribution is large.

2.10 SOLUTIONS /ANSWERS

E1) Maximum value (X_{Max}) = 70
Minimum value (X_{Min}) = 12

Therefore, $R = X_{Max} - X_{Min}$

$$= 70 - 12 = 58$$

E2) In the given frequency distribution, we have

Lower limit of first class = 0

Upper limit of last class = 50

Therefore,

Range = 50 – 0 =50

E3) First we construct the following cumulative frequency distribution:

Class	f	Cumulative frequencies
0-5	6	6
5-10	8	14
10-15	12	26
15-20	24	50
20-25	36	86
25-30	32	118
30-35	24	142
35-40	8	150
	N =150	

$$Q_1 = L + \frac{\left(\frac{N}{4} - C\right)}{f} \times h \qquad \text{15-20 is the } Q_1 \text{ class}$$

$$= 15 + \frac{\left(\frac{150}{4} - 26\right)}{24} \times 5$$

$$= 15 + \frac{37.5 - 26}{24} \times 5 = 15 + \frac{57.5}{24}$$

$$= 17.40$$

$$Q_3 = L + \frac{\left(\frac{3N}{4} - C\right)}{f} \times h \qquad \text{25-30 is the } Q_3 \text{ class}$$

$$= 25 + \frac{\left(\frac{3 \times 150}{4} - 86\right)}{32} \times 5$$

$$= 25 + \frac{112.5 - 86}{24} \times 5 = 25 + \frac{132.5}{24}$$

$$= 30.52$$

Therefore, $QD = \frac{Q_3 - Q_1}{2} = \frac{30.52 - 17.40}{2} = \frac{13.12}{2} = 6.56$

E4) We have the following distribution:

x	**(x-17)**	$\|\mathbf{x}-\bar{\mathbf{x}}\|$
16	-1	1
24	+7	7
13	-4	4
18	+1	1
15	-2	2
10	-7	7
23	+6	6
$\sum x = 119$		$\sum \|x-\bar{x}\| = 28$

Then $$\bar{x} = \frac{\sum x}{n} = \frac{119}{7} = 17$$

Therefore, $$\text{MD} = \frac{\sum |x-\bar{x}|}{n} = \frac{28}{7} = 4$$

E5) First we construct the following frequency distribution:

Class	**x**	**f**	**xf**	$(\mathbf{x}-\bar{\mathbf{x}})$	$\|\mathbf{x}-\bar{\mathbf{x}}\|$	$\mathbf{f}\|\mathbf{x}-\bar{\mathbf{x}}\|$
0-10	5	5	25	-22	22	110
10-20	15	8	120	-12	12	96
20-30	25	15	375	-2	2	30f
30-40	35	16	560	8	8	128
40-50	45	6	270	18	18	108
						$\sum f\|(x-\bar{x})\| = 472$

Then $$\bar{x} = \frac{\sum xf}{\sum f} = \frac{1350}{50} = 27$$

and

$$\text{MD} = \frac{\sum f|x-\bar{x}|}{\sum f} = \frac{472}{50} = 9.44$$

E6) We have given

$$n = 100, \quad \bar{x} = 16, \quad \sigma = \sqrt{21}$$
$$n_1 = 50, \quad \bar{x}_1 = 20, \quad \sigma_1 = 2, \text{ and } \quad n_2 = 50$$

we have to find

$$\bar{x}_2 = ?, \quad \sigma_2 = ?$$

Now, $\overline{x} = \frac{n_i\,\overline{x}_1 + n_2\,\overline{x}_2}{n_1+n_2} = \frac{50\times 20 + 50\overline{x}_2}{100}$

$$16 = \frac{50\times 20 + 50\,\overline{x}_2}{100}$$

$$50\,\overline{x}_2 = 16\times 100 - 50\times 20$$

$$50\,\overline{x}_2 = 1600 - 1000 \;= 600$$

$$\Rightarrow \overline{x}_2 = 12$$

$d_1 = \overline{x}_1 - \overline{x} \;=\; 20-16 = 4 \Rightarrow d_1^2 = 16$

$d_2 = \overline{x}_2 - \overline{x} \;=\; 12-16 = -4 \Rightarrow d_2^2 = 16$

$$\sigma^2 = \frac{1}{n_1+n_2}\left[n_1\left(\sigma_1^2 + d_1^2\right) + n_2\left(\sigma_2^2 + d_2^2\right)\right]$$

$$\left(\sqrt{21}\right)^2 = \frac{\left[50\,(4+16) + 50\left(\sigma_2^2 + 16\right)\right]}{100}$$

$$21\times 100 - 20\times 50 \;=\; 50\sigma_2^2 + 50\times 16$$

$$50\,\sigma_2^2 \;=\; 2100 - 1000 - 800$$

$$50\sigma_2^2 \;=\; 2100 - 1000 - 800$$

$$50\sigma_2^2 \;=\; 300$$

$$\Rightarrow \sigma_2^2 \;=\; 6 \;\Rightarrow\; \sigma_2 = \sqrt{6}$$

E7) First we calculate the following distribution:

x	x²
10	100
27	729
40	1600
60	3600
33	1089
30	900
10	100
$\sum x = 210$	$\sum x^2 = 8118$

Then

$$\overline{x} = \frac{\sum x}{n} = \frac{210}{7} = 30$$

and therefore,

$$\sigma = \sqrt{\frac{1}{n}\sum x^2 - \left(\overline{x}\right)^2} = \sqrt{\frac{8118}{7} - 900}$$

$$=\sqrt{1159.7-900}=\sqrt{259.7}=16.115$$

E8) Let us take A = 25 and calculate the following frequency distribution:

Class	**x**	**f**	**d = x - A**	**f d**	**f d²**
0-10	5	5	-20	-100	2000
10-20	15	8	-10	-80	800
20-30	25	15	0	0	0
30-40	35	16	10	160	1600
40-50	45	6	20	120	2400
		N = 50		$\sum fd=100$	$\sum fd^2=6800$

Therefore,

$$\sigma=\sqrt{\frac{1}{N}\sum_{i=1}^{k}f_i d_i^2-\left(\frac{\sum_{i=1}^{k}f_i d_i}{N}\right)^2}=\sqrt{\frac{1}{50}(6800)-\left(\frac{100}{50}\right)^2}$$

$$=\sqrt{136-4}=\sqrt{132}=11.49$$

E9) We have

$$\sigma^2=\frac{1}{n}\sum_{i=1}^{n}x_i^2-\overline{x}^2$$

$$=\frac{1}{n}\times 200-(4)^2=20-16=4$$

$$\Rightarrow \sigma^2=4 \Rightarrow \sigma=2$$

Therefore, coefficient of variation

$$C.V(X)=\frac{\sigma}{\overline{x}}\times 100$$

$$=\frac{2}{4}\times 100 = 50\%$$

E10) We have the formulae for coefficient of variation

$$CV=\frac{\sigma}{\overline{x}}\times 100$$

By putting the given values, we have

$$\sigma = \frac{CV\times\overline{x}}{100} = \frac{70.6\times 16}{100}$$

$$= 11.296$$

UNIT 3 MOMENTS

Structure

3.1 Introduction

Objectives

3.2 Introduction to Moments

3.3 Methods for Calculation of Moments

Moments about Arbitrary Points

Moments about Origin

Moments about Mean

Relation between Moments about Mean and Moments about Arbitrary Points

Effect of Change of Origin and Scale on Moments

Sheppard's Corrections for Moments

Charlier's Check for Moments

3.4 Pearson's Beta and Gamma Coefficients

3.5 Summary

3.6 Solutions / Answers

3.1 INTRODUCTION

Moments are popularly used to describe the characteristic of a distribution. They represent a convenient and unifying method for summarizing many of the most commonly used statistical measures such as measures of tendency, variation, skewness and kurtosis. Moments are statistical measures that give certain characteristics of the distribution. Moments can be raw moments, central moments and moments about any arbitrary point. For example, the first raw moment gives mean and the second central moment gives variance. Although direct formulae exist for central moments even then they can be easily calculated with the help of raw moments. The r^{th} central moment of the variable x is h^r times the r^{th} central moment of u where $u = (x – A)/h$ is a new variable obtained by subjecting x to a change of origin and scale. Since A does not come into the scene so there is no effect of change of origin on moments.

The basic concepts of moments are described in Section 3.2. In Section 3.3, the methods of calculation of different kinds of moments are explored. In this section the methods of calculation of raw moments, moments about zero and central moments are explained. In Section 3.4, the Pearson's Beta and Gamma coefficients of skewness and kurtosis are described. We shall discuss the skewness and kurtosis in details in Unit 4 of this block.

Objectives

After studying this unit, you would be able to

- define moments;
- explain different types of moments and their uses;
- derive the relation between raw and central moments;

- describe the effect of change of origin and scale on moments;
- calculate the raw and central moments for grouped and ungrouped frequency distributions;
- define the Shephard's correction for moments; and
- calculate the Beta and Gamma coefficients of skewness and kurtosis.

3.2 INTRODUCTION TO MOMENTS

Moment word is very popular in mechanical sciences. In science moment is a measure of energy which generates the frequency. In Statistics, moments are the arithmetic means of first, second, third and so on, i.e. rth power of the deviation taken from either mean or an arbitrary point of a distribution. In other words, moments are statistical measures that give certain characteristics of the distribution. In statistics, some moments are very important. Generally, in any frequency distribution, four moments are obtained which are known as first, second, third and fourth moments. These four moments describe the information about mean, variance, skewness and kurtosis of a frequency distribution. Calculation of moments gives some features of a distribution which are of statistical importance. Moments can be classified in raw and central moment. Raw moments are measured about any arbitrary point A (say). If A is taken to be zero then raw moments are called moments about origin. When A is taken to be Arithmetic mean we get central moments. The first raw moment about origin is mean whereas the first central moment is zero. The second raw and central moments are mean square deviation and variance, respectively. The third and fourth moments are useful in measuring skewness and kurtosis.

3.3 METHODS OF CALCULATION OF MOMENTS

Three types of moments are:

1. Moments about arbitrary point,
2. Moments about mean, and
3. Moments about origin

3.3.1 Moments about Arbitrary Point

When actual mean is in fraction, moments are first calculated about an arbitrary point and then converted to moments about the actual mean. When deviations are taken from arbitrary point, the formula's are:

For Ungrouped Data

If $x_1, x_2, ..., x_n$ are the n observations of a variable X, then their moments about an arbitrary point A are

Zero order moment A $$\mu'_0 = \frac{\sum_{i=1}^{n}(x_i - A)^0}{n} = 1$$

First order moment $\mu_1' = \dfrac{\sum_{i=1}^{n}(x_i - A)^1}{n}$

Second order moment $\mu_2' = \dfrac{\sum_{i=1}^{n}(x_i - A)^2}{n}$

Third order moment $\mu_3' = \dfrac{\sum_{i=1}^{n}(x_i - A)^3}{n}$

and Fourth order moment $\mu_4' = \dfrac{\sum_{i=1}^{n}(x_i - A)^4}{n}$

In general the rth order moment about arbitrary point A is given by

$$\mu_r' = \frac{\sum_{i=1}^{n}(x_i - A)^r}{n}; \quad \text{for } r = 1, 2, \text{- - -}$$

For Grouped Data

If $x_1, x_2, ..., x_k$ are k values (or mid values in case of class intervals) of a variable X with their corresponding frequencies $f_1, f_2, ..., f_k$, then moments about an arbitrary point A are

Zero order moment $\mu_0' = \dfrac{\sum_{i=1}^{k} f_i(x_i - A)^0}{N}$ = 1; where N = $\sum_{i=1}^{k}$fi

First order moment $\mu_1' = \dfrac{\sum_{i=1}^{k} f_i(x_i - A)^1}{N}$

Second order moment $\mu_2' = \dfrac{\sum_{i=1}^{k} f_i(x_i - A)^2}{N}$

Third order moment $\mu_3' = \dfrac{\sum_{i=1}^{k} f_i(x_i - A)^3}{N}$

Fourth order moment $\mu_4' = \dfrac{\sum_{i=1}^{k} f_i(x_i - A)^4}{N}$

In general, the rth order moment about an arbitrary point A can be obtained as

$$\mu_r' = \frac{\sum_{i=1}^{k} f_i(x_i - A)^r}{N} \quad ; r = 1, 2, \ldots$$

In a frequency distribution, to simplify calculation we can use short-cut method.

If $d_i = \frac{(x_i - A)}{h}$ or $(x_i - A) = hd_i$ then, we get the moments about an arbitrary point A are

First order moment $$\mu_1' = \frac{\sum_{i=1}^{k} f_i d_i^1}{N} \times h$$

Second order moment $$\mu_2' = \frac{\sum_{i=1}^{k} f_i d_i^2}{N} \times h^2$$

Third order moment $$\mu_3' = \frac{\sum_{i=1}^{k} f_i d_i^3}{N} \times h^3$$

Fourth order moments $$\mu_4' = \frac{\sum_{i=1}^{k} f_i d_i^4}{N} \times h^4$$

Similarly, r^{th} order moment about A is given by

$$\mu_r' = \frac{\sum_{i=1}^{k} f_i d_i^r}{N} \times h^r; \quad \text{for } r = 1, 2, \ldots$$

3.3.2 Moments about Origin

In case, when we take an arbitrary point A = 0 then, we get the moments about origin.

For Ungrouped Data

The r^{th} order moment about origin is defined as:

r^{th} order moment $$\mu_r' = \frac{\sum_{i=1}^{n} (x_i - 0)^r}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i^r$$

First order moment $$\mu_1' = \frac{\sum_{i=1}^{n} (x_i - 0)}{n} = \overline{x}$$

Second order moment $$\mu_2' = \frac{\sum_{i=1}^{n} (x_i - 0)^2}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i^2$$

Third order moment $$\mu_3' = \frac{\sum_{i=1}^{n} (x_i - 0)^3}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i^3$$

Forth order moment $$\mu_4' = \frac{\sum_{i=1}^{n} (x_i - 0)^4}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i^4$$

For Grouped Data

r th order moment $\mu_r' = \dfrac{\sum_{i=1}^{k} f_i\left(x_i - 0\right)^r}{N} = \dfrac{1}{N}\sum_{i=1}^{k} f_i x_i^r$

First order moment $\mu_1' = \dfrac{\sum_{i=1}^{k} f_i\,(x_i - 0)}{N} = \dfrac{1}{N}\sum_{i=1}^{k} f_i x_i$

Second order moment $\mu_2' = \dfrac{\sum_{i=1}^{k} f_i\,(x_i - 0)^2}{N} = \dfrac{1}{N}\sum_{i=1}^{k} f_i x_i^2$

Third order moment $\mu_3' = \dfrac{\sum_{i=1}^{k} f_i(x_i - 0)^3}{N} = \dfrac{1}{N}\sum_{i=1}^{k} f_i x_i^3$

Fourth order moment $\mu_4' = \dfrac{\sum_{i=1}^{k} f_i\,(x_i - 0)^4}{N} = \dfrac{1}{N}\sum_{i=1}^{k} f_i x_i^4$

3.3.3 Moments about Mean

When we take the deviation from the actual mean and calculate the moments, these are known as moments about mean or central moments. The formulae are:

For Ungrouped Data

Zero order moment $\mu_0 = \dfrac{\sum_{i=1}^{n} (x_i - \overline{x})^0}{n} = 1$

First order moment $\mu_1 = \dfrac{\sum_{i=1}^{n} (x_i - \overline{x})^1}{n} = 0$

Thus first order moment about mean is zero, because the algebraic sum of the deviation from the mean is zero $\sum_{i=1}^{n} \left(x_i - \overline{x}\right) = 0$.

Second order moment $\mu_2 = \dfrac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n} = \sigma^2$ (variance)

Therefore, second order moment about mean is variance. These results, viz, $\mu_0 = 1$, $\mu_1 = 0$ and $\mu_2 = \sigma^2$ are found very important in statistical theory and practical.

Third order moment $\mu_3 = \dfrac{\sum_{i=1}^{n} (x_i - \overline{x})^3}{n}$

and Fourth order moment $\mu_4 = \dfrac{\sum_{i=1}^{n} (x_i - \overline{x})^4}{n}$

In general, the r^th order moment of a variable about the mean is given by

$$\mu_r = \frac{\sum_{i=1}^{k} (x_i - \overline{x})^r}{N}; \quad \text{for } r = 0, 1, 2, \ldots$$

For Grouped Data

In case of frequency distribution the r^th order moment about mean is given by:

$$\mu_r = \frac{\sum_{i=1}^{k} f_i (x_i - \overline{x})^r}{N}; \quad \text{for } r = 0, 1, 2 \ldots$$

By substituting the different value of are we can gate different orders moment about mean as follows:

Zero order moment $\mu_0 = \dfrac{\sum_{i=1}^{k} f_i \ (x_i - \overline{x})^0}{N} = 1 \quad \text{as } N = \sum_{i=1}^{k} f_i$

First order moment $\mu_1 = \dfrac{\sum_{i=1}^{k} f_i \ (x_i - \overline{x})^1}{N} = 0$

Because $\sum_{i=1}^{k} f_i \left(x_i - x\right) = 0$

Second order moment $\mu_2 = \dfrac{\sum_{i=1}^{k} f_i \ (x_i - \overline{x})^2}{N} = \text{Variance } (\sigma^2)$

Third order moment $\mu_3 = \dfrac{\sum_{i=1}^{k} f_i \ (x_i - \overline{x})^3}{N}$

and Fourth order moment $\mu_4 = \dfrac{\sum_{i=1}^{k} f_i \ (x_i - \overline{x})^4}{N}$

3.3.4 Relation between Moments about Mean and Moments about Arbitrary Point

The r^th moment about mean is given by

$$u_r = \frac{1}{N} \sum_{i=1}^{k} f_i \ (x_i - \overline{x})^r \ ;\text{where } r = 0, 1, \ldots$$

$$u_r = \frac{1}{N} \sum_{i=1}^{k} f_i \ (x_i - A + A - \overline{x})^r$$

$$u_r = \frac{1}{N}\sum_{i=1}^{k} f_i \left\{(x_i - A) - (\bar{x} - A)\right\}^r \qquad \ldots (1)$$

If $d_i = x_i - A$ then,

$$x_i = A + d_i$$

$$\frac{1}{n}\sum x_i = A + \frac{1}{n}\sum d_i$$

$$\bar{x} = \left(A + \mu_1^{'}\right)$$

$$\because \mu_1^{'} = \frac{1}{n}\sum d_i$$

$$\mu_1^{'} = \bar{x} - A$$

Therefore, we get from equation (1)

$$u_r = \frac{1}{N}\sum_{i=1}^{k} f_i \,(d_i - \mu_1^{'})^r$$

$$\Rightarrow \frac{1}{N}\sum_{i=1}^{k} f_i \left\{d_i^r - {}^rC_1 d_1^{r-1}\mu_1^{'} + {}^rC_2 d_i^{r-2}(\mu_1^{'})^2 - {}^rC_3 d_i^{r-3}(\mu_1^{'})^3 + \ldots + (-1)^r(\mu_1^{'})^r\right\}$$

$$\Rightarrow \frac{1}{N}\sum_{i=1}^{k} f_i d_i^r - {}^rC_1\mu_1^{'}\frac{1}{N}\sum_{i=1}^{k} f_i d_i^{r-1} + {}^rC_2\mu_1^{'2}\frac{1}{N}\sum_{i=1}^{k} f_i d_i^{r-2} - {}^rC_3\mu_1^{'3}\frac{1}{N}\sum_{i=1}^{k} f_i d_i^{r-3} \ldots + (-1)^r\mu_1^{'r}$$

Then,

$$\mu_r = \mu_r^{'} - {}^r\,C_1\mu_{r-1}^{'}\mu_1^{'} + {}^r\,C_2\mu_{r-2}^{'}\mu_1^{'2} - {}^r\,C_3\mu_{r-3}^{'}\mu_1^{'3} + \ldots + (-1)^r\,\mu_1^{'r} \qquad \ldots (2)$$

In particular on putting r = 2, 3 and 4 in equation (2), we get

$$\mu_2 = \mu_2^{'} - \mu_1^{'2}$$

$$\mu_3 = \mu_3^{'} - 3\mu_2^{'}\mu_1^{'} + 2\mu_1^{'3}$$

$$\mu_4 = \mu_4^{'} - 4\mu_3^{'}\mu_1^{'} + 6\mu_2^{'}\mu_1^{'2} - 3\mu_1^{'4}$$

3.3.5 Effect of Change of Origin and Scale on Moments

Let $u_i = \dfrac{x_i - A}{h}$ so that $x_i = A + hu_i$ and $\left(x_i - \bar{x}\right) = h\left(u_i - \bar{u}\right)$

Thus, r^th moment of x about arbitrary point x = A is given by

$$\mu_r^{'}(x) = \frac{1}{N}\sum_{i=1}^{k} f_i\,(x_i - A)^r$$

$$= \frac{1}{N}\sum_{i=1}^{k} f_i\,(hu_i)^r$$

$$\mu_r^{'}(x) = h^r\frac{1}{N}\sum_{i=1}^{k} f_i u_i^r = h^r\mu_r^{'}(u)$$

and, r^{th} moment of x about mean is given by

$$\mu_r(x) = \frac{1}{N}\sum_{i=1}^{k} f_i\,(x_i - \overline{x})^r$$

$$= \frac{1}{N}\sum_{i=1}^{k} f_i\,\{h(u_i - \overline{u})\}^r$$

$$\mu_r(x) = h^r\,\frac{1}{N}\sum_{i=1}^{k} f_i\,(u_i - \overline{u})^r = h^r\mu_r(u)$$

Thus, the r^{th} moment of the variable x about mean is h^r times the r^{th} moment of the new variable u about mean after changing the origin and scale.

3.3.6 Sheppard's Corrections for Moments

The fundamental assumption that we make in farming class intervals is that the frequencies are uniformly distributed about the mid points of the class intervals. All the moment calculations in case of grouped frequency distributions rely on this assumption. The aggregate of the observations or their powers in a class is approximated by multiplying the class mid point or its power by the corresponding class frequency. For distributions that are either symmetrical or close to being symmetrical, this assumption is acceptable. But it is not acceptable for highly skewed distributions or when the class intervals exceed about $1/20^{th}$ of the range. In such situations, W. F. Sheppard suggested some corrections to be made to get rid of the so called "grouping errors" that enter into the calculation of moments.

Sheppard suggested the following corrections known as Sheppard's corrections in the calculation of central moments assuming continuous frequency distributions if the frequency tapers off to zero in both directions

$$\mu_2\,(\text{corrected}) = \mu_2 - \frac{h^2}{12}$$

$$\mu_3\,(\text{corrected}) = \mu_3$$

$$\mu_4\,(\text{corrected}) = \mu_4 - \frac{1}{2}h^2\,\mu_2 + \frac{7}{240}h^4$$

where, h is the width of class interval.

3.3.7 Charlier's Checks for Moments

The following identities

$$\sum f\,(x+1) = \sum fx + N$$

$$\sum f\,(x+1)^2 = \sum fx^2 + 2\sum fx + N$$

$$\sum f\left(x+1\right)^3 = \sum fx^3 + 3\sum fx^2 + 3\sum fx + N$$

$$\sum f\left(x+1\right)^4 = \sum fx^4 + 4\sum fx^3 + 6\sum fx^2 + 4\sum fx + N$$

are used to check the calculations done for finding moments.

3.4 PEARSON'S BETA AND GAMMA COEFFICIENTS

Karl Pearson defined the following four coefficients, based upon the first four central moments:

1. β_1 is defined as

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$$

It is used as measure of skewness. For a symmetrical distribution, β_1 shall be zero.

β_1 as a measure of skewness does not tell about the direction of skewness, i.e. positive or negative. Because μ_3 being the sum of cubes of the deviations from mean may be positive or negative but $\mu^2{}_3$ is always positive. Also μ_2 being the variance always positive. Hence β_1 would be always positive. This drawback is removed if we calculate Karl Pearson's coefficient of skewness γ_1 which is the square root of β_1 ,i. e.

$$\gamma_1 = \pm\sqrt{\beta_1} = \frac{\mu_3}{\left(\mu_2\right)^{3/2}} = \frac{\mu 3}{\sigma^2}$$

Then the sign of skewness would depend upon the value of μ_3 whether it is positive or negative. It is advisable to use γ_1 as measure of skewness.

2. β_2 measures kurtosis and it is defined by

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

and similarly, coefficient of kurtosis γ_2 is defined as

$$\gamma_2 = \beta_2 - 3$$

Example 1: For the following distribution calculate first four moments about mean and also find β_1, β_2, γ_1 and γ_2:

Marks	5	10	15	20	25	30	35
Frequency	4	10	20	36	16	12	2

Solution: First we construct the following frequency distribution for calculation of moments:

Marks (x)	**f**	$\mathbf{d = \frac{(x-20)}{5}}$	**fd**	$\mathbf{fd^2}$	$\mathbf{fd^3}$	$\mathbf{fd4}$
5	4	-3	-12	36	-108	324
10	10	-2	-20	40	-80	160
15	20	-1	-20	20	-20	20
20	36	0	0	0	0	0
25	16	1	16	16	16	16
30	12	2	24	48	96	192
35	2	3	6	18	54	162
			$\sum fd = -6$	$\sum fd^2 = 178$	$\sum fd^3 = -42$	$\sum fd^4 = 874$

Then

$$\mu_1' = \frac{\sum fd}{N} \times h = \frac{-6}{100} \times 5 = -0.3$$

$$\mu_2' = \frac{\sum fd^2}{N} \times h^2 = \frac{178}{100} \times 25 = 44.5$$

$$\mu_3' = \frac{\sum fd^3}{N} \times h^3 = \frac{-42}{100} \times 125 = -52.5$$

$$\mu_4' = \frac{\sum fd^4}{N} \times h^4 = \frac{874}{100} \times 625 = 5462.5$$

Moments about mean

$$\mu_2 = \mu_2' - \mu_1'^2 = 44.5 - 0.09 = 44.41 = \sigma^2$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + \mu_1'^3$$

$$= -52.5 - 3 \times 44.5 \times -0.3 + 2(-0.3)^3$$

$$= -52.5 + 40.05 - 0.054 = -12.504$$

$$\mu_4 = \mu_4' - 4\mu_1'\mu_3' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4$$

$$= 5462.5 - 4\,(-0.3 \times -52.5) + 6\,(44.5)\,(-0.3)^2 - 3\,(-0.3)^4$$

$$= 5462.5 - 63 + 24.03 - 0.0243$$

$$= 5423.5057$$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(-12.504\;)^2}{(44.41)^3} = 0.001785$$

$$\gamma_1 = \frac{\mu_3}{\sigma^3} = \frac{-12.504}{(6.6641)^3} = -0.0422$$

$$\beta_2 = \frac{\mu_4}{\sigma_2^2} = \frac{5423.5057}{(44.41)^2} = 02.7499$$

$$\gamma_2 = \beta_2 - 3 = 2.7499 - 3 = -\ 0.2501$$

E1) The first four moments of a distribution about the value 5 of a variable are 1, 10, 20 and 25. Find the central moments, β_1 and $\beta_{2.}$

E2) For the following distribution, find central moments, β_1 and β_2:

Class	1.5-2.5	2.5-3.5	3.5-4.5	4.5-5.5	5.5-6.5
Frequency	1	3	7	3	1

E3) Wages of workers are given in the following table:

Weekly Wages	10-12	12-14	14-16	16-18	18-20	20-22	22-24
Frequency	1	3	7	12	12	4	3

Find the first four central moment and β_1 and $\beta_{2.}$

3.5 SUMMARY

In this unit, we have discussed:

1. What is moments;
2. The different types of moments and their uses;
3. The relation between raw and central moments;
4. The effect of change of origin and scale on moments;
5. How to calculate the raw and central moments for the given frequency distribution;
6. Shephard's Correction for Moments; and
7. The Beta and Gamma coefficients.

3.6 SOLUTIONS / ANSWERS

E1) We have given

$$\mu_1^{'} = 1,\ \mu_2^{'} = 10,\ \mu_3^{'} = 20 \text{ and } \mu_4^{'} = 25$$

Now we have to find out moments about mean

$$\mu_2 = \mu_2' - \mu_1'^2 = 10 - (1)^2 = 9$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1^1 + 2\mu_1'^3 = 20 - 3\times 10 \times 1 + 2(1)^3$$

$$= 20 - 30 + 2 = -8$$

$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4$$

$$= 25 - 4\times 20\times 1 + 6\times 10\times 1^2 - 3\times 1^4 = 2$$

So therefore, $\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(-8)^2}{(9)^3} = \frac{64}{729} = 0.0877$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{2}{81} = 0.0247$$

E2) For calculation of moments

x	**f**	$\mathbf{d} = (\mathbf{x} - \bar{\mathbf{x}})$	**fd**	**fd^2**	**fd^3**	**fd^4**
2	1	–2	–2	4	–8	16
3	3	–1	–3	3	–3	3
4	7	0	0	0	0	0
5	3	1	3	3	3	3
6	1	2	2	4	8	16
	N =15		$\sum fd = 0$	$\sum fd^2 = 14$	$\sum fd^3 = 0$	$\sum fd^4 = 38$

We therefore have,

$$\mu_1 = \frac{\sum fd}{N} = \frac{0}{15} = 0$$

$$\mu_2 = \frac{\sum fd^2}{N} = \frac{14}{150} = 0.933$$

$$\mu_3 = \frac{\sum fd^3}{N} = \frac{0}{15} = 0$$

$$\mu_4 = \frac{\sum fd^4}{N} = \frac{38}{15} = 2.533$$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(0)}{(0.933)^2}$$

Since β_1=0 that means the distributions is symmetrical.

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{(2.53)}{(0.933)^2} = 2.91$$

As, $\beta_2 < 3$ that means curve is platykurtic.

E3) Calculation of first four moments

Wages	**f**	**x**	$\mathbf{d' = \frac{x-17}{2}}$	**fd'**	**fd'^{2}**	**fd'^{3}**	**fd'^{4}**
10-12	1	11	–3	–3	9	–27	81
12-14	3	13	–2	–6	12	–24	48
14-16	7	15	–1	–7	7	–7	7
16-18	20	17	0	0	0	0	0
18-20	12	19	1	12	12	12	12
20-22	4	21	2	8	16	32	64
22-24	3	23	3	9	27	81	243
				$\sum fd'$ $=13$	$\sum fd'^2$ $=27$	$\sum fd'^3$ $=67$	$\sum fd'^4$ $=455$

Therefore, using formula

$$\mu_1' = \frac{\sum fd'}{N} \times h = \frac{13}{50} \times 2 = 0.52$$

$$\mu_2' = \frac{\sum fd'^2}{N} \times h^2 = \frac{27}{50} \times 4 = 2.16$$

$$\mu_3' = \frac{\sum fd'^3}{N} \times h^3 = \frac{67}{50} \times 8 = +10.72$$

$$\mu_4' = \frac{\sum fd'^4}{N} \times h^4 = \frac{455}{50} \times 16 = 145.6$$

So, we have

$$\mu_1 = 0,\ \mu_2 = \mu_2' - \mu_1'^2 = 2.16 - 0.2704 = 1.8896$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3$$

$$\mu_3 = 10.72 - 3 \times 2.16 \times 0.52 + (0.52)^3$$
$$= 10.72 - 3.3696 + 0.1406$$

$$= 7.491$$

$$\mu_4 = \mu_4^{'} - 4\mu_1^{'}\mu_3^{'} + 6\mu_2^{'}\mu_1^{'2} - 3\mu_1^{'4}$$

$$\mu_4 = 145.6 - 4 \times 0.52 \times 10.72 + 6 \times 2.56 \times 0.2704$$
$$- 3 \times 0.07312$$

$$= 145.6 - 22.2976 + 3.5043 - 0.2193$$

$$= 126.5874$$

Thus, the β coefficients

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(7.491)^2}{(1.8896)^3} = \frac{56.11508}{6.7469} = 8.317$$

and $\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{(126.5874)}{(1.8896)^2} = \frac{126.5874}{3.5706} = 35.4527$

UNIT 4 SKEWNESS AND KURTOSIS

Structure

4.1 Introduction

Objectives

4.2 Concept of Skewness

4.3 Various Measures of Skewness

4.4 Concept of Kurtosis

Measures of Kurtosis

4.5 Summary

4.6 Solutions/Answers

4.1 INTRODUCTION

In Units 1 and 2, we have talked about average and dispersion. They give the location and scale of the distribution. In addition to measures of central tendency and dispersion, we also need to have an idea about the shape of the distribution. Measure of skewness gives the direction and the magnitude of the lack of symmetry whereas the kurtosis gives the idea of flatness.

Lack of symmetry is called skewness for a frequency distribution. If the distribution is not symmetric, the frequencies will not be uniformly distributed about the centre of the distribution. Here, we shall study various measures of skewness and kurtosis.

In this unit, the concepts of skewness are described in Section 4.2 whereas the various measures of skewness are given with examples in Section 4.3. In Section 4.4, the concepts and the measures of kurtosis are described.

Objectives

On studying this unit, you would be able to

- describe the concepts of skewness;
- explain the different measures of skewness;
- describe the concepts of kurtosis;
- explain the different measures of kurtosis; and
- explain how skewness and kurtosis describe the shape of a distribution.

4.2 CONCEPT OF SKEWNESS

Skewness means lack of symmetry. In mathematics, a figure is called symmetric if there exists a point in it through which if a perpendicular is drawn on the X-axis, it divides the figure into two congruent parts i.e. identical in all respect or one part can be superimposed on the other i.e mirror images of each other. In Statistics, a distribution is called symmetric if mean, median and mode coincide. Otherwise, the distribution becomes asymmetric. If the right tail is longer, we get a positively skewed distribution for which mean > median > mode while if the left tail is longer, we get a negatively skewed distribution for which mean < median < mode.

The example of the Symmetrical curve, Positive skewed curve and Negative skewed curve are given as follows:

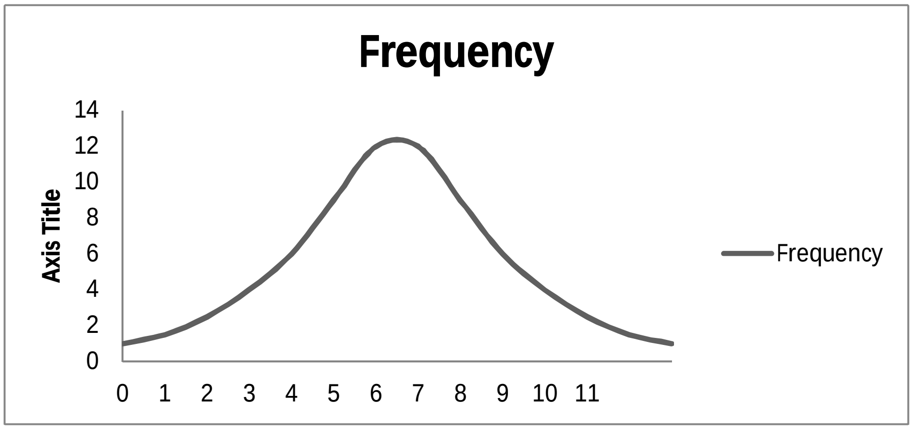


Fig. 4.1: Symmetrical Curve

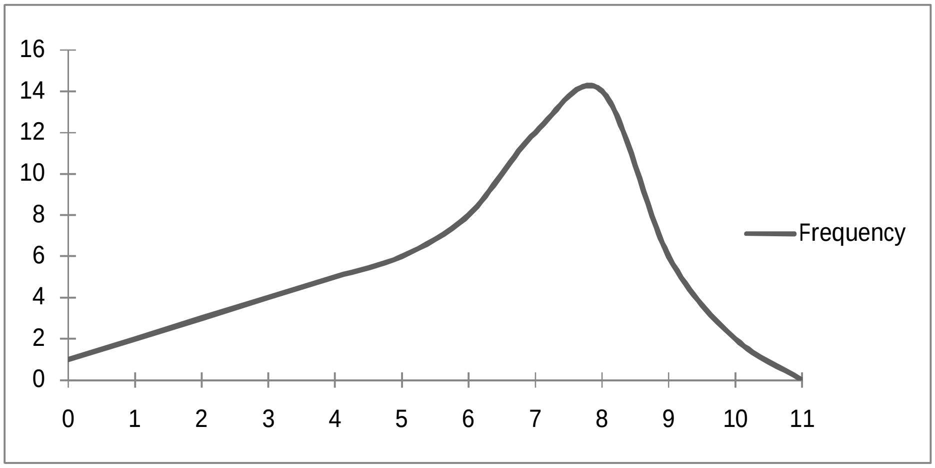


Fig. 4.2: Negative Skewed Curve

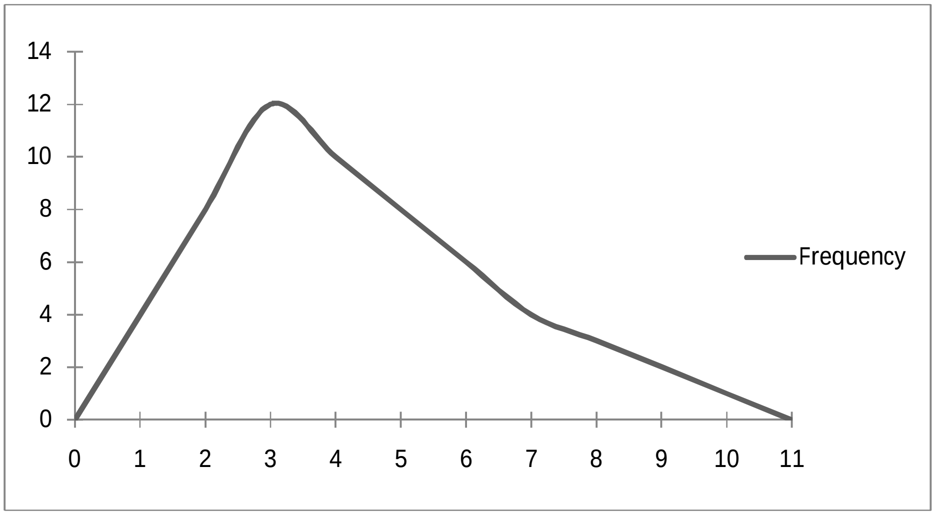


Fig. 4.3: Positive Skewed Curve

4.2.1 Difference between Variance and Skewness

The following two points of difference between variance and skewness should be carefully noted.

1. Variance tells us about the amount of variability while skewness gives the direction of variability.
2. In business and economic series, measures of variation have greater practical application than measures of skewness. However, in medical and life science field measures of skewness have greater practical applications than the variance.

4.3 VARIOUS MEASURES OF SKEWNESS

Measures of skewness help us to know to what degree and in which direction (positive or negative) the frequency distribution has a departure from symmetry. Although positive or negative skewness can be detected graphically depending on whether the right tail or the left tail is longer but, we don't get idea of the magnitude. Besides, borderline cases between symmetry and asymmetry may be difficult to detect graphically. Hence some statistical measures are required to find the magnitude of lack of symmetry. A good measure of skewness should possess three criteria:

1. It should be a unit free number so that the shapes of different distributions, so far as symmetry is concerned, can be compared even if the unit of the underlying variables are different;
2. If the distribution is symmetric, the value of the measure should be zero. Similarly, the measure should give positive or negative values according as the distribution has positive or negative skewness respectively; and
3. As we move from extreme negative skewness to extreme positive skewness, the value of the measure should vary accordingly.

Measures of skewness can be both absolute as well as relative. Since in a symmetrical distribution mean, median and mode are identical more the mean moves away from the mode, the larger the asymmetry or skewness. An absolute measure of skewness can not be used for purposes of comparison because of the same amount of skewness has different meanings in distribution with small variation and in distribution with large variation.

4.3.1 Absolute Measures of Skewness

Following are the absolute measures of skewness:

1. Skewness (S_k) = Mean – Median
2. Skewness (S_k) = Mean – Mode
3. Skewness (S_k) = $(Q_3 - Q_2) - (Q_2 - Q_1)$

For comparing to series, we do not calculate these absolute mearues we calculate the relative measures which are called coefficient of skewness. Coefficient of skewness are pure numbers independent of units of measurements.

4.3.2 Relative Measures of Skewness

In order to make valid comparison between the skewness of two or more distributions we have to eliminate the distributing influence of variation. Such elimination can be done by dividing the absolute skewness by standard deviation. The following are the important methods of measuring relative skewness:

1. β and γ Coefficient of Skewness

Karl Pearson defined the following β and γ coefficients of skewness, based upon the second and third central moments:

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$$

It is used as measure of skewness. For a symmetrical distribution, β_1 shall be zero. β_1 as a measure of skewness does not tell about the direction of skewness, i.e. positive or negative. Because μ_3 being the sum of cubes of the deviations from mean may be positive or negative but μ_3^2 is always positive. Also, μ_2 being the variance always positive. Hence, β_1 would be always positive. This drawback is removed if we calculate Karl Pearson's Gamma coefficient $\gamma_1$which is the square root of β_1 i. e.

$$\gamma_1 = \pm\sqrt{\beta_1} = \frac{\mu_3}{(\mu_2)^{3/2}} = \frac{\mu_3}{\sigma^3}$$

Then the sign of skewness would depend upon the value of μ_3 whether it is positive or negative. It is advisable to use γ_1 as measure of skewness.

2. Karl Pearson's Coefficient of Skewness

This method is most frequently used for measuring skewness. The formula for measuring coefficient of skewness is given by

$$S_k = \frac{\text{Mean} - \text{Mode}}{\sigma}$$

The value of this coefficient would be zero in a symmetrical distribution. If mean is greater than mode, coefficient of skewness would be positive otherwise negative. The value of the Karl Pearson's coefficient of skewness usually lies between ± 1 for moderately skewed distubution. If mode is not well defined, we use the formula

$$S_k = \frac{3(\text{Mean} - \text{Median})}{\sigma}$$

By using the relationship

Mode = (3 Median – 2 Mean)

Here, $-3 \le S_k \le 3$. In practice it is rarely obtained.

3. Bowleys's Coefficient of Skewness

This method is based on quartiles. The formula for calculating coefficient of skewness is given by

$$S_k = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_1)}$$

$$= \frac{(Q_3 - 2Q_2 + Q_1)}{(Q_3 - Q_1)}$$

The value of S_k would be zero if it is a symmetrical distribution. If the value is greater than zero, it is positively skewed and if the value is less than zero it is negatively skewed distribution. It will take value between +1 and -1.

4. Kelly's Coefficient of Skewness

The coefficient of skewness proposed by Kelly is based on percentiles and deciles. The formula for calculating the coefficient of skewness is given by

Based on Percentiles

$$S_k = \frac{(P_{90} - P_{50}) - (P_{50} - P_{10})}{(P_{90} - P_{10})}$$

$$= \frac{(P_{90} - 2P_{50} + P_{10})}{(P_{90} - P_{10})}$$

where, P_{90}, P_{50} and P_{10} are 90th, 50th and 10th Percentiles.

Based on Deciles

$$S_k = \frac{(D_9 - 2D_5 + D_1)}{D_9 - D_1}$$

where, D_9, D_5 and D_1 are 9th, 5th and 1st Decile.

Example1: For a distribution Karl Pearson's coefficient of skewness is 0.64, standard deviation is 13 and mean is 59.2 Find mode and median.

Solution: We have given

$$S_k = 0.64, \sigma = 13 \text{ and Mean} = 59.2$$

Therefore by using formulae

$$S_k = \frac{\text{Mean} - \text{Mode}}{\sigma}$$

$$0.64 = \frac{59.2 - \text{Mode}}{13}$$

$$\text{Mode} = 59.20 - 8.32 = 50.88$$

$$\text{Mode} = 3 \text{ Median} - 2 \text{ Mean}$$

$$50.88 = 3 \text{ Median} - 2\,(59.2)$$

$$\text{Median} = \frac{50.88 + 118.4}{3} = \frac{169.28}{3} = 56.42$$

E1) Karl Pearson's coefficient of skewness is 1.28, its mean is 164 and mode 100, find the standard deviation.

E2) For a frequency distribution the Bowley's coefficient of skewness is 1.2. If the sum of the 1st and 3rd quarterlies is 200 and median is 76, find the value of third quartile.

E3) The following are the marks of 150 students in an examination. Calculate Karl Pearson's coefficient of skewness.

Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
No. of Students	10	40	20	0	10	40	16	14

Remarks about Skewness

1. If the value of mean, median and mode are same in any distribution, then the skewness does not exist in that distribution. Larger the difference in these values, larger the skewness;

2. If sum of the frequencies are equal on the both sides of mode then skewness does not exist;

3. If the distance of first quartile and third quartile are same from the median then a skewness does not exist. Similarly if deciles (first and ninth) and percentiles (first and ninety nine) are at equal distance from the median. Then there is no asymmetry;

4. If the sums of positive and negative deviations obtained from mean, median or mode are equal then there is no asymmetry; and

5. If a graph of a data become a normal curve and when it is folded at middle and one part overlap fully on the other one then there is no asymmetry.

4.4 CONCEPT OF KURTOSIS

If we have the knowledge of the measures of central tendency, dispersion and skewness, even then we cannot get a complete idea of a distribution. In addition to these measures, we need to know another measure to get the complete idea about the shape of the distribution which can be studied with the help of Kurtosis. Prof. Karl Pearson has called it the "Convexity of a Curve". Kurtosis gives a measure of flatness of distribution.

The degree of kurtosis of a distribution is measured relative to that of a normal curve. The curves with greater peakedness than the normal curve are called **"Leptokurtic".** The curves which are more flat than the normal curve are called **"Platykurtic"**. The normal curve is called "**Mesokurtic**." The Fig.4 describes the three different curves mentioned above:

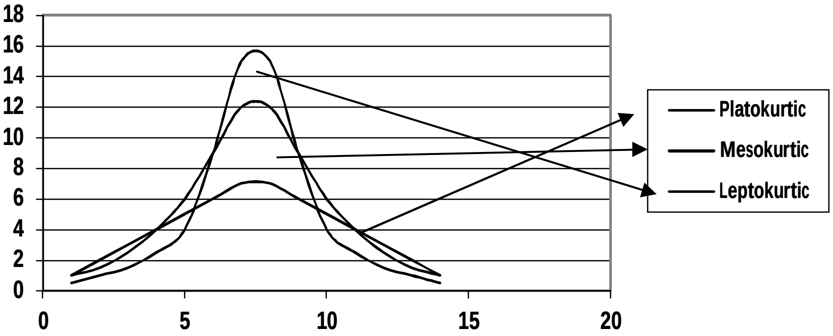


Fig.4.4: Platykurtic Curve, Mesokurtic Curve and Leptokurtic Curve

4.4.1 Measures of Kurtosis

1. Karl Pearson's Measures of Kurtosis

For calculating the kurtosis, the second and fourth central moments of variable are used. For this, following formula given by Karl Pearson is used:

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

or $$\gamma_2 = \beta_2 - 3$$

where, μ_2 = Second order central moment of distribution

μ_4 = Fourth order central moment of distribution

Description:

1. If $\beta_2 = 3$ or $\gamma_2 = 0$, then curve is said to be mesokurtic;

2. If $\beta_2 < 3$ or $\gamma_2 < 0$, then curve is said to be platykurtic;

3. If $\beta_2 > 3$ or $\gamma_2 > 0$, then curve is said to be leptokurtic;

2. Kelly's Measure of Kurtosis

Kelly has given a measure of kurtosis based on percentiles. The formula is given by

$$\beta_2 = \frac{P_{75} - P_{25}}{P_{90} - P_{10}}$$

where, P_{75}, P_{25}, P_{90}, and P_{10} are 75th, 25th , 90th and 10th percentiles of dispersion respectively.

If $\beta_2 > 0.26315$, then the distribution is platykurtic.

If $\beta_2 < 0.26315$, then the distribution is leptokurtic.

Example 2: First four moments about mean of a distribution are 0, 2.5, 0.7 and 18.75. Find coefficient of skewness and kurtosis.

Solution: We have $\mu_1 = 0$, $\mu_2 = 2.5$, $\mu_3 = 0.7$ and $\mu_4 = 18.75$

Therefore, Skewness, $\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(0.7)^2}{(2.5)^3} = 0.031$

$$\text{Kurtosis, } \beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{18.75}{(2.5)^2} = \frac{18.75}{6.25} = 3.$$

As β_2 is equal to 3, so the curve is mesokurtic.

E4) The first four raw moments of a distribution are 2, 136, 320, and 40,000. Find out coefficients of skewness and kurtosis.

4.5 SUMMARY

In this unit, we have discussed:

1. What is skewness;
2. The significance of skewness;
3. The types of skewness exists in different kind of distrbutions;
4. Different kinds of measures of skewness;
5. How to calculate the coefficients of skewness;
6. What is kurtosis;
7. The significance of kurtosis;
8. Different measures of kurtosis; and
9. How to calculate the coefficient of kurtosis.

4.6 SOLUTIONS/ANSWERS

E1) Using the formulae, we have

$$S_k = \frac{\text{Mean} - \text{Mode}}{\sigma}$$

$$1.28 = \frac{164 - 100}{\sigma}$$

$$\sigma = \frac{64}{1.28} = 50$$

E2) We have given $S_k = 1.2$

$$Q_1 + Q_3 = 200$$

$$Q_2 = 76$$

then $$S_k = \frac{(Q_3 + Q_1 - 2Q_2)}{(Q_3 - Q_1)}$$

$$1.2 = \frac{(200 - 2 \times 76)}{(Q_3 - Q_1)}$$

$$Q_3 - Q_1 = \frac{48}{1.2} = 40$$

$$Q_3 - Q_1 = 40 \qquad \ldots (1)$$

and it is given $Q_1 + Q_3 = 200$

$$Q_1 = 200 - Q_3$$

Therefore, from equation (1)

$$Q_3 - (200 - Q_3) = 40$$

$$2Q_3 = 240$$

$$Q_3 = 120$$

E3) Let us calculate the mean and median from the given distribution because mode is not well defined.

Class	**f**	**x**	**CF**	$\mathbf{d' = \frac{x-35}{10}}$	**fd'**	**fd'²**
0-10	10	5	10	-3	-30	90
10-20	40	15	50	-2	-80	160
20-30	20	25	70	-1	-20	20
30-40	0	35	70	0	0	0
40-50	10	45	80	+1	10	10
50-60	40	55	120	+2	80	160
60-70	16	65	136	+3	48	144
70-80	14	75	150	+4	56	244
					$\sum fd'$ $= 64$	$\sum fd'^2$ $= 828$

$$\text{Median} \quad = L + \frac{\left(\frac{N}{2} - C\right)}{f} \times h$$

$$= 40 + \frac{75-70}{10} \times 10 = 45$$

$$\text{Mean} \quad (\overline{x}) = A + \frac{\sum_{i=1}^{k} fd'}{N} \times h$$

$$= 35 + \frac{64}{150} \times 10 = 39.27$$

$$\text{Standard Deviation } (\sigma) \quad = \quad h \times \sqrt{\frac{\sum fd'^2}{N} - \left(\frac{\sum fd'}{N}\right)^2}$$

$$= 10 \times \sqrt{\frac{828}{150} - \left(\frac{64}{150}\right)^2}$$

$$= 10 \times \sqrt{5.33} = 23.1$$

Therefore, coefficient of skewness:

$$S_k = \frac{3(\text{Mean} - \text{Median})}{\sigma}$$

$$= \frac{3(39.27 - 45)}{23.1} = -0.744$$

E4) Given that $\mu'_1 = 2$, $\mu'_2 = 136$, $\mu'_3 = 320$ and $\mu'_4 = 40{,}000$

First of all we have to calculate the first four central moments

$$\mu_1 = \mu'_1 - \mu'_1 = 0$$

$$\mu_2 = \mu'_2 - (\mu'_1)^2 = 136 - 2^2 = 132$$

$$\mu_3 = \mu'_3 - 3\mu'_2\mu'_1 + 2\mu_1'^3$$

$$= 320 - 3 \times 132 \times 2 + 2\,(2)^3$$

$$= 320 - 792 + 16 = -456$$

$$\mu_4 = \mu'_4 - 4\mu'_1\mu'_3 + 6\mu'_2\mu_1'^2 - 3\mu_1'^4$$

$$= 40{,}000 - 4 \times 2 \times 320 + 6 \times 2^2 \times 136 - 3 \times 2^4$$

$$= 40{,}000 - 2560 + 3{,}264 - 48 = 40656$$

Skewness, $\beta_1 = \dfrac{\mu_3^2}{\mu_2^3} = \dfrac{(-456)^2}{(132)^3} = 0.0904$

Kurtosis, $\beta_2 = \dfrac{\mu_4}{\mu_2^2} = \dfrac{40656}{(132)^2} = 2.333$

Indira Gandhi
National Open University
School of Sciences

MST-002
DESCRIPTIVE STATISTICS

Block

2

CORRELATION FOR BIVARIATE DATA

UNIT 5
Fitting of Curves **5**

UNIT 6
Correlation Coefficient **25**

UNIT 7
Rank Correlation **45**

UNIT 8
Intra-Class Correlation **61**

Curriculum and Course Design Committee

Prof. K. R. Srivathasan
Pro-Vice Chancellor
IGNOU, New Delhi

Prof. Parvin Sinclair
Pro-Vice Chancellor
IGNOU, New Delhi

Prof. Geeta Kaicker
Director, School of Sciences
IGNOU, New Delhi

Prof. Jagdish Prasad
Department of Statistics
University of Rajasthan, Jaipur

Prof. R. M. Pandey
Department of Bio-Statistics
All India Institute of Medical Sciences
New Delhi

Prof. Rahul Roy
Math. and Stat. Unit
Indian Statistical Institute, New Delhi

Dr. Diwakar Shukla
Department of Mathematics and Statistics
Dr. Hari Singh Gaur University, Sagar

Prof. Rakesh Srivastava
Department of Statistics
M. S. University of Baroda, Vadodara

Prof. G. N. Singh
Department of Applied Mathematics
I. S. M. Dhanbad

Dr. Gulshan Lal Taneja
Department of Mathematics
M. D. University, Rohtak

Faculty members of School of Sciences, IGNOU

Statistics
Dr. Neha Garg
Dr. Nitin Gupta
Mr. Rajesh Kaliraman
Dr. Manish Trivedi

Mathematics
Dr. Deepika Garg
Prof. Poornima Mital
Prof. Sujatha Varma
Dr. S. Venkataraman

Block Preparation Team

Content Editor
Dr. Meenakshi Srivastava
Department of Statistics
Institute of Social Sciences
Dr. B. R. Ambedkar University, Agra

Language Editor
Dr. Nandini Sahu
School of Humanities, IGNOU

Secretarial Support
Mr. Deepak Singh

Course Writer
Dr. Rajesh Tailor
School of Studies in Statistics
Vikram University, Ujjain

Formatted By
Dr. Manish Trivedi
Mr. Prabhat Kumar Sangal
School of Sciences, IGNOU

Programme and Course Coordinator: Dr. Manish Trivedi

Block Production

Mr. Y. N. Sharma, SO (P.)
School of Sciences, IGNOU

Acknowledgement: We gratefully acknowledge to Prof. Geeta Kaicker, Director, School of Sciences for her great support and guidance.

December, 2011

ISBN-978-81-266-

Further information on the Indira Gandhi National Open University may be obtained from University's Office at Maidan Garhi, New Delhi-110068 or visit University's website http://www.ignou.ac.in

Printed and published on behalf of the Indira Gandhi National Open University, New Delhi by the Director, School of Sciences.

Laser Typeset by: Tessa Media & Computers, C-206, A.F.E.-II, Okhla, New Delhi

Printed at:

CORRELATION FOR BIVARIATE DATA

In Block 1 of this course, you have studied the analysis of quantitative data mainly dealt with the quantitative techniques which describes the one or more variables e.g. height, weight, sales, income, etc. independently. Those units were broadly classified as measures of central tendency, measures of dispersion, moments, skewness and kurtosis. Often we come across the situation where information on two or more variables, together like height and weight, income and expenditure, literacy and poverty, etc. are available and our interest is to study the relationship between these two variables. The present block deals with the situations having information on two variables.

Unit 1 describes the fitting of various curves including straight line, second degree of parabola, power curves and exponential curves for the given set of data using principle of least squares. With the help of fitting of the curves one can estimate the dependent variable for given value of independent variable.

Unit 2 gives the concept of correlation which studies the linear association between two variables. The concept of correlation and correlation coefficient would be very helpful in regression analysis.

Unit 3 describes the rank correlation which handles the situation where study characteristics are not measureable but can be presented in the form of ranks according to merit of individuals. In this unit, you will study the rank correlation coefficient with its properties.

Unit 4 deals with two different types of situations. First in which no linear association exists between two variables but they may have some other type of curvilinear relationship. In this situation correlation coefficient fails to determine the intensity of relationship and we use correlation ratio. Another situation, when we are interested in studying the relationship among the members of a group or family, leads us to intraclass correlation coefficient. This unit describes the coefficient of determination, correlation ratio and intra-class correlation coefficient.

Suggested Readings:

- Ansari, M. A., Gupta, O. P. and Chaudhari S. S.; Applied Statistics, Kedar Nath Ram Nath & Co., Meerut 1979.
- Arora, S. and Bansi Lal; New Mathematical Statistics, Satya Prakashan, New Delhi, 1989.
- Chaturvedi, J. C.; Elementary Statistics, Prakash Brothers, Agra, 1963
- Elhance, D. N.; Fundamentals of Statistics, Kitab Mahal, Allahabad, 1987
- Goon, A. M., Gupta, M. K. and Das Gupta, B.; Fundamentals of Statistics-Vol-I; World Press Culcutta.
- Gupta, M. P. and Gupta, S. P.; Business Statistics; Sultan Chand & Sons Publications.
- Gupta S. C. and Kapoor, V. K.; Fundamentals of Mathematical Statistics, Sultan Chand & Sons Publications.

Notations and Symbols

$\frac{\partial}{\partial a}$	: Partial derivative with respect to a
U	: Sum of squares of errors
$\sum_{i=1}^{n}$	: Sum over i from 1 to n
log x	: Logarithm of x at the base 10
r = Corr (x, y)	: Correlation coefficient between X and Y
Cov (x, y)	: Covariance between X and Y
$V(x) = \sigma_x^2$	: Variance of X
σ_x	: Standard deviation of X
$\overline{x}$	: Mean of X
A	: Assumed mean
r_s	: Rank correlation coefficient
R_x	: Rank of X
d_i	: Difference between R_x and R_y
r_c	: Concurrent deviation
C	: Number of concurrent deviations
r^2	: Coefficient of determination
η	: Correlation ratio
r_{ic}	: Intra-class correlation coefficient
σ_m^2	: Variance of means

UNIT 5 FITTING OF CURVES

Structure

5.1 Introduction

Objectives

5.2 Principle of Least Squares

5.3 Fitting of Straight Line

5.4 Fitting of Second Degree Parabola

5.5 Fitting of a Power Curve $Y = aX^b$

5.6 Fitting of Exponential Curve $Y = ab^X$

5.7 Fitting of Exponential Curve $Y = ae^{bX}$

5.8 Summary

5.9 Solutions / Answers

5.1 INTRODUCTION

All the methods that you have learnt in Block 1 of this course were based on the uni-variate distributions, i.e. all measures analysed only single variable. But many times we have data for two or more variables and our interest is to know the best functional relationship between variables that given data describes. In this unit, you will learn how to fit the various functions such as straight line, parabola of the second degree, power curve and exponential curves. Curve fitting has theoretical importance in regression and correlation analysis while practically it is used to present the relationship by simple algebraic expression. All these methods can be used to estimate the values of the dependent variable for the specific value of the independent variable.

Section 5.2 gives the basic idea of the principle of least squares and procedure of fitting any curve for any given set of data. Section 5.3 explains the fitting of straight line while Sections 5.4 and 5.5 give the fitting of second degree parabola and power curve respectively. Fitting of exponential curves is described in Sections 5.6 and 5.7. Fitting of all functions considered in this unit are explained with examples also.

Objectives

After reading this unit, you would be able to

- describe the principle of least squares;
- describe the procedure of fitting any curve or functional relationship;
- define and calculate the residuals;
- fit a straight line for the given data;
- fit the second degree parabola for given data;
- fit a power curve for the given data; and
- fit a exponential curves $Y = ab^x$ and $Y = ae^{bx}$.

5.2 PRINCIPLE OF LEAST SQUARES

Let Y and X be the dependent and independent variables respectively and we have a set of values $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$, i.e. observations are taken from n individuals on X and Y. We are interested in studying the function $Y = f(X)$. If Y_i is the estimated value of Y obtained by the function and y_i is the observed value of Y at x_i then we can define residual.

The difference between y_i and Y_i i.e. the difference between observed value and estimated value is called error of the estimate or residual for y_i.

Principle of least squares consists in minimizing the sum of squares of the residuals, i.e. according to principle of least squares

$$U = \sum_{i=1}^{n} (y_i - Y_i)^2 \text{ should be minimum.}$$

Let us consider a curve (function) of the type

$$Y = a + bX + cX^2 + ... + tX^k \qquad ... (1)$$

where, Y is dependent variable, X is independent variable and a, b, c,…,t are unknown constants. Suppose we have $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ values of two variables (X, Y) i.e. data for variables X and Y. These variables may be height and weight, sales and profit, rainfall and production of any crop, etc. In all these examples, first variables, i.e. height, sales and rainfall seem to be independent variables, while second variables, i.e. weight, profit and production of crop seem to be dependent variables.

With the given values $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$, curve (function) given in equation (1) produces set of n equations

$$\left.\begin{aligned} y_1 &= a + bx_1 + cx_1{}^2 + ... + tx_1{}^k \\ y_2 &= a + bx_2 + cx_2{}^2 + ... + tx_2{}^k \\ &. \\ &. \\ &. \\ y_n &= a + bx_n + cx_n{}^2 + ... + tx_n{}^k \end{aligned}\right\} \qquad … (2)$$

Our problem is to determine the constants a, b, c,…, t such that it represents the curve of best fit given by equation (1) of degree k.

If n = k+1, i.e. number of equations and number of unknown constants are equal, there is no problem in determining the unknown constants and error can be made absolutely zero. But more often n > k+1 i.e. number of equations is greater than the number of unknown constants and it is impossible to do away with all errors i.e. these equations cannot be solved exactly which satisfy set of equations (2).

Therefore, we try to determine the values of a, b, c,…, t which satisfy set of equations (2) as nearly as possible.

Substituting $x_1, x_2, ..., x_n$ for X in equation (1) we have

$$\left.\begin{aligned} Y_1 &= a + bx_1 + cx_1{}^2 + \ldots + tx_1{}^k \\ Y_2 &= a + bx_2 + cx_2{}^2 + \ldots + tx_2{}^k \\ &. \\ &. \\ &. \\ Y_n &= a + bx_n + cx_n{}^2 + \ldots + tx_n{}^k \end{aligned}\right\} \quad \ldots (3)$$

The quantities $Y_1, Y_2, \ldots, Y_n$ are called expected or estimated values of $y_1, y_2, \ldots, y_n$ (given values of Y) for the given values of $x_1, x_2, \ldots, x_n$. Here $y_1, y_2, \ldots, y_n$ are the observed values of Y.

Let us define a quantity U, the sum of squares of errors i.e.

$$U = \sum_{i=1}^{n} (y_i - Y_i)^2$$

$$U = \sum_{i=1}^{n} (y_i - a - bx_i - cx_i^2 - \cdots - tx_i^k)^2 \quad \ldots(4)$$

According to the principle of least squares the constant a, b,…, t are chosen in such a way that the sum of squares of residuals is minimum.

According to principle of maxima and minima (theorem of differential calculus), the extreme value (maximum or minimum) of the function U are obtained by

$$\frac{\partial U}{\partial a} = 0 = \frac{\partial U}{\partial b} = \frac{\partial U}{\partial c} = \cdots = \frac{\partial U}{\partial t}$$

(provided that the partial derivatives exist).

Let us take

$$\frac{\partial U}{\partial a} = 0 \Rightarrow \frac{\partial}{\partial a}\sum_{i=1}^{n}(y_i - Y_i)^2 = 0$$

$$\Rightarrow \frac{\partial}{\partial a}\sum_{i=1}^{n}(y_i - a - bx_i - cx_i^2 - \cdots - tx_i^k)^2 = 0$$

$$\Rightarrow 2\sum (y_i - a - bx_i - cx_i^2 - \cdots - tx_i^k)(-1) = 0$$

$$\Rightarrow \sum y_i - na - b\sum x_i - c\sum x_i^2 - \cdots - t\sum x_i^k = 0$$

$$\Rightarrow \sum y_i = na + b\sum x_i + c\sum x_i^2 + \cdots + t\sum x_i^k \quad \ldots(5)$$

$$\frac{\partial U}{\partial b} = 0 = \frac{\partial}{\partial b}\sum_{i=1}^{n}(y_i - Y_i)^2$$

$$\Rightarrow \frac{\partial}{\partial b}\sum_{i=1}^{n}(y_i - a - bx_i - cx_i^2 - \cdots - tx_i^k)^2 = 0$$

$$\Rightarrow 2\sum (y_i - a - bx_i - cx_i^2 - \cdots - tx_i^k)(-x_i) = 0$$

$$\Rightarrow \sum x_i(y_i - a - bx_i - cx_i^2 - \cdots - tx_i^k) = 0$$

$$\Rightarrow \sum x_i y_i - a\sum x_i - b\sum x_i^2 - c\sum x_i^3 - \cdots - t\sum x_i^{k+1} = 0$$

$$\Rightarrow \sum x_i y_i = a\sum x_i + b\sum x_i^2 + c\sum x_i^3 + \cdots + t\sum x_i^{k+1}$$

Therefore, by the conditions $\frac{\partial U}{\partial a} = 0 = \frac{\partial U}{\partial b} = \frac{\partial U}{\partial c} = \cdots = \frac{\partial U}{\partial t}$, ultimately we get the following (k+1) equations

$$\left.\begin{aligned}
&\sum y_i = na + b\sum x_i + c\sum x_i^2 + \cdots + t\sum x_i^k = 0\\
&\sum y_i x_i = a\sum x_i + b\sum x_i^2 + c\sum x_i^3 + \cdots + t\sum x_i^{k+1} = 0\\
&.\\
&.\\
&.\\
&\sum y_i x_i^k = a\sum x_i^k + b\sum x_i^{k+1} + c\sum x_i^{k+2} + \cdots + t\sum x_i^{2k} = 0
\end{aligned}\right\} \qquad \ldots (6)$$

where, summation extended to i from 1 to n.

In simple way equation (6) can be expressed as

$$\left.\begin{aligned}
&\sum y = na + b\sum x + \cdots + t\sum x^k\\
&\sum xy = a\sum x + b\sum x^2 + \cdots + t\sum x^{k+1}\\
&\sum x^2 y = a\sum x^2 + b\sum x^3 + \cdots + t\sum x^{k+2}\\
&.\\
&.\\
&.\\
&\sum x^k y = a\sum x^k + b\sum x^{k+1} + \cdots + t\sum x^{2k}
\end{aligned}\right\} \qquad \ldots (7)$$

These equations are known as normal equations for the curve in equation (1). These equations are solved as simultaneous equations and give the value of (k+1) constants a, b, c, …, t. Substitution of these values in second order partial derivatives gives positive value of the function. Positive value of the function indicates that the values of a, b, c,…, t obtained by solving the set of equations (6), minimize U which is sum of squares of residuals . With these values of a, b, c,…, t , curve in equation (1) is the curve of best fit.

5.3 FITTING OF STRAIGHT LINE

Section 5.2 described the procedure of fitting of any curve using principle of least squares. In this Section, we are fitting straight line for the given set of points (x_i, y_i) i = 1, 2, …, n, using principle of least squares and adopting the procedure given in Section 5.2.

Let $Y = a + bX$ … (8)

be an equation of straight line and we have a set of n points (x_i, y_i); i = 1, 2, …, n. Here, the problem is to determine a and b so that the straight line $Y = a + bX$ is the line of the best fit. With given n points (x_i, y_i), let the straight line be $y_i = a + bx_i$ where, y_i is the observed value of variable Y and $a + bx_i = Y_i$ is the estimated value of Y obtained by the straight line in equation (8). According to the principle of least squares, a and b are to be determined so that the sum of squares of residuals is minimum, i.e.

$$U = \sum_{i=1}^{n} (y_i - Y_i)^2$$

$$U = \sum_{i=1}^{n} (y_i - a - bx_i)^2 \qquad \text{… (9)}$$

is minimum. From the principle of maxima and minima we take partial derivatives of U with respect to a and b and equating to zero, i.e.

$$\frac{\partial U}{\partial a} = 0$$

$$\Rightarrow \frac{\partial}{\partial a} \sum_{i=1}^{n} (y_i - a - bx_i)^2 = 0$$

$$\Rightarrow 2\sum_{i=1}^{n} (y_i - a - bx_i)(-1) = 0$$

$$\Rightarrow \sum_{i=1}^{n} (y_i - a - bx_i) = 0$$

$$\Rightarrow \sum_{i=1}^{n} y_i - na - b\sum_{i=1}^{n} x_i = 0$$

$$\Rightarrow \sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i \qquad \text{… (10)}$$

and

$$\Rightarrow \frac{\partial U}{\partial b} = 0$$

$$\Rightarrow \frac{\partial}{\partial b} \sum_{i=1}^{n} (y_i - a - bx_i)^2 = 0$$

$$\Rightarrow 2\sum_{i=1}^{n} (y_i - a - bx_i)(-x_i) = 0$$

$$\Rightarrow \sum_{i=1}^{n} (y_i x_i - ax_i - bx_i^2) = 0$$

$$\Rightarrow \sum_{i=1}^{n} y_i x_i - a\sum_{i=1}^{n} x_i - b\sum_{i=1}^{n} x_i^2 = 0$$

$$\Rightarrow \sum_{i=1}^{n} y_i x_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 \qquad \text{… (11)}$$

Equations (10) and (11) are known as normal equations for straight line in equation (8) to determine a and b. In simple form, equations (10) and (11) can be written as

$$\sum y = na + b\sum x$$

$$\sum y\,x = a\sum x + b\sum x^2$$

Values of a and b are obtained by solving equations (10) and (11). With these value of a and b, straight line Y = a + bX is the line of best fit to the given set of points (x_i, y_i) where i =1, 2,…, n.

Let us discuss a problem of fitting of straight line numerically for the given set of data.

Example 1: Fit a straight line to the following data.

x	1	6	11	16	20	26
y	13	16	17	23	24	31

Solution: Let the straight line be $Y = a + bX$ and to obtain a and b for this straight line the normal equations are

$$\sum y = na + b\sum x \quad \text{and}$$

$$\sum xy = a\sum x + b\sum x^2$$

Here, there is need of $\sum y$, $\sum x$, $\sum xy$ and $\sum x^2$ which are obtained by the following table

x	**y**	**x^2**	**xy**
1	13	1	13
6	16	36	96
11	17	121	187
16	23	256	368
20	24	400	480
26	31	676	806
$\sum x = 80$	$\sum y = 124$	$\sum x^2 = 1490$	$\sum xy = 1950$

Substituting the values of $\sum y$, $\sum x$, $\sum xy$ and $\sum x^2$ in the normal equations, we get

$$124 = 6a + 80b \quad \ldots (12)$$

$$1950 = 80a + 1490b \quad \ldots (13)$$

Now we solve equations (12) and (12).

Multiplying equation (12) by 80 and equation (13) by 6, i.e.

$$124 = 6a + 80b \quad]\times 80$$

and

$$1950 = 80a + 1490b \quad]\times 6$$

we get,

$$9920 = 480\ a + 6400\ b \qquad \dots (14)$$

$$11700 = 480\ a + 8940\ b \qquad \dots (15)$$

Subtracting (14) from (15), we obtain

$$1780 = 2540\ b$$

$$\Rightarrow b = 1780/2540 = 0.7008$$

Substituting the value of b in equation (12), we get

$$124 = 6\ a + 80 \times 0.7008$$

$$124 = 6\ a + 56.064$$

$$67.936 = 6a$$

$$\Rightarrow a = 11.3227$$

with these values of a and b the line of best fit is Y = 11.3227 + 0.7008 X.

Now let us do one problem for fitting of straight line.

E 1) Fit a straight line to the following data:

x	6	7	8	9	11
y	5	4	3	2	1

5.4 FITTING OF SECOND DEGREE PARABOLA

Let $Y = a + bX + cX^2$ … (16)

be the second degree parabola and we have a set of n points (x_i, y_i); i =1, 2,…, n. Here the problem is to determine a, b and c so that the equation of second degree parabola given in equation (16) is the best fit equation of parabola. Let with given n points (x_i, y_i) the second degree parabola be

$$y_i = a + bx_i + cx_i^2 \qquad \dots (17)$$

Let $Y_i = a + bx_i + cx_i^2$ be the estimated value of Y. Then according to the principle of least squares, we have to determine a, b and c so that the sum of squares of residuals is minimum, i.e.

$$U = \sum_{i=1}^{n}(y_i - Y_i)^2 = \sum_{i=1}^{n}(y_i - a - bx_i - cx_i^2)^2 \qquad \dots (18)$$

is minimum. Using principle of maxima and minima, we take partial derivatives of U with respect to a, b and c and equating to zero, i.e.

$$\frac{\partial U}{\partial a} = 0 = \frac{\partial U}{\partial b} = \frac{\partial U}{\partial c} \qquad \dots (19)$$

Now

$$\frac{\partial U}{\partial a} = 0 \Rightarrow 2\sum_{i=1}^{n}(y_i - a - bx_i - cx_i^2)(-1) = 0$$

$$\Rightarrow -2\sum_{i=1}^{n}(y_i - a - bx_i - cx_i^2) = 0$$

$$\Rightarrow \sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} x^2 \qquad \dots (20)$$

$$\frac{\partial U}{\partial b} = 0 \Rightarrow -2\sum_{i=1}^{n}(y_i - a - bx_i - cx_i^2)x_i = 0$$

$$\Rightarrow \sum_{i=1}^{n} y_i x_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_i^3 \qquad \dots (21)$$

Similarly, $\frac{\partial U}{\partial c} = 0$ provides

$$-2\sum_{i=1}^{n}\left(y_i - a - bx_i - cx_i^2\right) x_i^2 = 0$$

$$\Rightarrow \sum_{i=1}^{n} y_i x_i^2 = a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i^3 + c\sum_{i=1}^{n} x_i^4 \qquad \dots (22)$$

Equations (20), (21) and (22) are known as normal equations for estimating a, b and c which can be written as

$$\sum y = na + b\sum x + c\sum x^2$$

$$\sum xy = a\sum x + b\sum x^2 + c\sum x^3$$

$$\sum x^2y = a\sum x^2 + b\sum x^3 + c\sum x^4$$

Values of a, b and c are obtained by solving equations (20), (21) and (22). With these values of a, b and c, the second degree parabola $Y = a + bX + cX^2$ is the best fit.

Now we solve a problem of fitting a second degree parabola.

Example 2: Fit a second degree parabola for the following data:

x	0	1	2	3	4
y	1	3	4	5	6

Solution: Let $Y = a + bX + cX^2$ be the second degree parabola and we have to determine a, b and c. Normal equations for second degree parabola are

$$\sum y = na + b\sum x + c\sum x^2,$$

$$\sum xy = a\sum x + b\sum x^2 + c\sum x^3, \text{ and}$$

$$\sum x^2y = a\sum x^2 + b\sum x^3 + c\sum x^4$$

To solve above normal equations, we need $\sum y$, $\sum x$, $\sum xy$, $\sum x^2y$, $\sum x^2$, $\sum x^3$ and $\sum x^4$ which are obtained from following table:

x	**y**	**xy**	**x^2**	**x^2y**	**x^3**	**x^4**
0	1	0	0	0	0	0
1	3	3	1	3	1	1
2	4	8	4	16	8	16
3	5	15	9	45	27	81
4	6	24	16	96	64	256
$\sum x = 10$	$\sum y = 19$	$\sum xy = 50$	$\sum x^2 = 30$	$\sum x^2 y = 160$	$\sum x^3 = 100$	$\sum x^4 = 354$

Substituting the values of $\sum y, \sum x, \sum xy, \sum x^2 y$, $\sum x^2, \sum x^3$ and $\sum x^4$ in above normal equations, we have

$$19 = 5a + 10b + 30c \quad \dots (23)$$

$$50 = 10a + 30b + 100c \quad \dots (24)$$

$$160 = 30a + 100b + 354c \quad \dots (25)$$

Now, we solve equations (23), (24) and (25).

Multiplying equation (23) by 2, we get

$$38 =10a + 20b + 60c \quad \dots (26)$$

Subtracting equation (26) from equation (24)

$$50 = 10a + 30b + 100c$$

$$38 =10a + 20b + 60c$$

$$12 = 10b + 40c \quad \dots (27)$$

Multiplying equation (24) by 3, we get

$$150 = 30a + 90b + 300c \quad \dots (28)$$

Subtracting equation (28) from equation (25), we get

$$160 = 30a + 100b + 354c$$

$$150 = 30a + 90b + 300c$$

$$10 = 10b + 54c \quad \dots (29)$$

Now we solve equation (27) and (29)

Subtracting equation (27) from equation (29), we get

$$10 = 10b + 54c$$

$$12 =10b + 40c$$

$$-2 = 14\,c$$

$$c = -\ 2/14$$

$$c = -\ 0.1429$$

Substituting the value of c in equation (29), we get

$$10 = 10\text{ b} + 54 \times (-0.1429)$$

$$10 = 10\text{ b} - 7.7166$$

$$17.7166 = 10\text{ b}$$

$$\text{b} = 1.7717$$

Substituting the value of b and c in equation (23), we get

$$19 = 5\text{a} + 10 \times (1.7717) + (-0.1429 \times 30)$$

$$19 = 5\text{a} + 17.717 - 4.287$$

$$\text{a} = 1.114$$

Thus, the second degree of parabola of best fit is

$$\text{Y} = 1.114 + 1.7717\text{ X} - 0.1429\text{ X}^2 \qquad \ldots (30)$$

Now let us solve a problem.

E2) Fit a second degree parabola to the following data:

x	2	4	6	8	10
y	1	2	3	4	5

5.5 FITTING OF A POWER CURVE $\mathbf{Y = aX^b}$

Let $\qquad Y = aX^b \qquad \ldots (31)$

be a power curve where a and b are constants. We have a set of n points (x_i, y_i) $i = 1, 2, \ldots, n$. Here, the problem is to determine a and b such that the power curve $Y = aX^b$ is the curve of best fit.

Taking log both sides of equation (31), we get

$$\log Y = \log(aX^b)$$

$$\log Y = \log a + \log X^b$$

$$\Rightarrow \log Y = \log a + b \log X \qquad \ldots (32)$$

Let $\log Y = U$, $\log a = A$ and $\log X = V$

Then equation (32) becomes

$$U = A + bV \qquad \ldots (33)$$

Now equation (33) is the linear form of the power curve (31).

Adopting the procedure of fitting of straight line, the normal equations for straight line equation (33) are

$$\sum u = nA + b\sum v \qquad \ldots (34)$$

$$\sum uv = A\sum v + b\sum v^2 \qquad \ldots (35)$$

equations (34) and (35) can be solved for A and b.

After getting A, we get a = antilog (A)

With these a and b, power curve equation (31) is the best fit equation of the curve to the given set of points.

Note: Here we are using log at the base 10.

Example 3: Fit power curve $Y = aX^b$ for the following data:

x	6	2	10	5	8
y	9	11	12	8	7

Solution: Let the power curve be $Y = aX^b$ and normal equations for estimating a and b are

$$\sum u = nA + b\sum v$$

$$\sum uv = A\sum v + b\sum v^2$$

where,

u = log y, v = log x and A = log a

Note: Here we are using log at the base 10.

To find the values of a and b from the above normal equations, we require $\sum u$, $\sum v$, $\sum uv$ and $\sum v^2$ which are being calculated in the following table:

x	**y**	**u=log y**	**v= log x**	**uv**	$\mathbf{v^2}$
6	9	0.9542	0.7782	0.7425	0.6055
2	11	1.0414	0.3010	0.3135	0.0906
10	12	1.0792	1.0000	1.0792	1.0000
5	8	0.9031	0.6990	0.6312	0.4886
8	7	0.8451	0.9031	0.7632	0.8156
31	47	4.8230	3.6813	3.5296	3.0003
		$\sum u = 4.8230$	$\sum v = 3.6813$	$\sum uv = 3.5296$	$\sum v^2 = 3.0003$

Substituting the values of $\sum u = 4.8230$, $\sum v = 3.6813$, $\sum uv = 3.5296$ and $\sum v^2 = 3.0003$ in above normal equations, we obtain

$$4.8230 = 5A + 3.6813\,b \qquad \dots (36)$$

$$3.5296 = 3.6813A + 3.0003b \qquad \dots (37)$$

Now, we solve the equation (36) and equation (37). Multiplying equation (36) by 3.6813 and equation (37) by 5, we have

$$17.7549 = 18.4065A + 13.5519\,b \qquad \dots (38)$$

$$17.6480 = 18.4065A + 15.0015\,b \qquad \dots (39)$$

Subtracting equation (39) from equation (38), we have

$$0.1069 = -1.4496\,b$$

$$\Rightarrow b = -0.0737$$

Substituting the value of b in equation (36), we get

$$A = 1.0216$$

Now a = antilog A = antilog (1.0216)

$$a = 10.5099$$

Thus, the power curve of the best fit is $Y = 10.5099\, X^{-0.0737}$.

Now let us solve a problem.

E 3) Fit a power curve $Y = aX^b$ to the following data:

x	5	6	9	8	11
y	2	5	8	11	15

5.6 FITTING OF THE EXPONENTIAL CURVE $Y = ab^X$

Let $\quad Y = ab^X$ … (40)

be an exponential curve and we have a set of n points $(x_i, y_i)\ i = 1, 2, ..., n$. We have to determine a and b such that equation (40) is the curve of best fit.

Taking log both sides of equation (40)

$$\log Y = \log a + \log b^X$$

$$\log Y = \log a + X \log b$$

Let, logY = U, log a = A and log b = B

Now, equation (40) comes in the linear form as

$$U = A + BX \qquad \dots (41)$$

which is the equation of straight line. Normal equations for equation (41) can be obtained as

$$\sum u = nA + B\sum x \qquad \cdots (42)$$

$$\sum ux = A\sum x + B\sum x^2 \qquad \cdots (43)$$

By solving equation (42) and equation (43), we obtain A and B and finally

a = antilog A and b = antilog B.

With these a and b, the exponential curve $Y = ab^X$ is the curve of best fit for the given set of data.

Note: Here we are using log base 10.

Now let us solve a problem of fitting of exponential curve $Y = ab^X$.

Example 4: Fit the exponential curve $Y = ab^X$ from the following data.

x	2	4	6	8	10
y	1	3	6	12	24

Solution: Let the exponential curve be $Y = ab^X$ and normal equations for estimating a and b are

$$\sum u = nA + B\sum x$$

$$\sum ux = A\sum x + B\sum x^2$$

where, a = antilog(A) and b = antilog (B)

x	**y**	**u=log y**	**ux**	$\mathbf{x^2}$
2	1	0.0000	0.0000	4
4	3	0.4771	1.9085	16
6	6	0.7782	4.6690	36
8	12	1.0792	8.6334	64
10	24	1.3802	13.8021	100
$\sum x = 30$	$\sum y = 46$	$\sum u = 3.7147$	$\sum ux = 29.0130$	$\sum x^2 = 220$

Substituting the values of $\sum x$, $\sum y$, $\sum u$, $\sum ux$ and $\sum x^2$ in the above normal equations, we get

3.7147 = 5A+30B … (44)

29.0130 = 30 A+220 B … (45)

Multiplying equation (44) by 6, we have

22.2882 = 30A+180B … (46)

Subtracting equation (46) from equation (45), we have

29.0130 = 30A + 220B

22.2882 = 30A + 180B

6.7248 = 40 B

B = 0.1681

Substituting the value of B in equation (44), we get

A= − 0.26566

Thus, a = antilog A = antilog (− 0.26566) =1.8436 and

b = antilog B = antilog (0.1681) =1.4727

Thus, exponential curve of best fit is $Y = 1.8436(1.4727)^X$

Now let us solve a problem.

E 4) Fit an exponential curve of the type $Y = ab^X$ for the following data

x	1	2	3	4	5
y	8	15	33	65	130

5.7 FITTING OF EXPONENTIAL CURVE $Y = ae^{bX}$

Let $Y = ae^{bX}$ … (47)

be an exponential curve and we have a set of n points (x_i, y_i) $i = 1, 2, ..., n$. Here problem is to determine a and b such that equation (47) is the curve of best fit.

Taking log of both side of equation (47)

$$\log Y = \log a + X\, b \log e$$

Let log Y =U, log a = A and b log e = B

Now, equation (47) can be written as

$$U = A + BX \quad \dots (48)$$

(which is the equation of straight line)

Normal equations for equation (48) can be obtained as

$$\sum u = nA + B\sum x \quad \dots (49)$$

$$\sum ux = A\sum x + B\sum x^2 \quad \dots (50)$$

We can get A and B from these normal equations. Then

$$a = \text{antilog } A \text{ and } b = \frac{B}{\log e}$$

With these a and b, the exponential curve $Y = ae^{bX}$ is the best fit equation of the curve for the given set of data.

Note: Here also we are using log base 10.

Now let us solve a problem of fitting of exponential curve of type $Y = ae^{bX}$.

Example 5: Fit an exponential curve of the type $Y = ae^{bX}$ from the following data.

x	1	2	4
y	5	10	30

Solution: To fit the exponential curve $Y = ae^{bX}$ normal equations are

$$\sum u = nA + B\sum x$$
$$\sum ux = A\sum x + B\sum x^2$$

x	**y**	**u =log y**	**u x**	**x^2**
1	5	0.6990	0.6990	1
2	10	1.0000	2.0000	4
4	30	1.4771	5.9085	16
$\sum x = 7$	$\sum y = 45$	$\sum u = 3.1761$	$\sum x^2 = 8.6075$	$\sum ux = 21$

Now the normal equations are

$$3A + 7\,B = 3.1761$$

$$7A + 21B = 8.6075$$

By solving these equations as simultaneous equations, we get

$$A = 0.4604 \text{ and } B = 0.2564$$

Then,

$$a = \text{antilog}\,(A) = \text{antilog}\,(0.4604) = 2.8867$$

$$b = \frac{B}{\log e} = \frac{0.2564}{\log(2.71828)} = \frac{0.2564}{0.43429} = 0.5904$$

Thus, the curve of best fit is $Y = 2.8867\ e^{0.5904\ x}$

5.8 SUMMARY

In this unit, we have discussed:

1. The purpose of fitting the curve;
2. Residual is the difference of observed value and estimated value;
3. The principle of least squares;
4. How to fit a straight line;
5. How to fit a second degree parabola;
6. How to fit a power curve; and
7. How to fit exponential curves.

5.9 SOLUTIONS / ANSWERS

E 1) Let the straight line be $Y = a + bX$ and to obtain a and b for this straight line the normal equations are

$$\sum y = na + b\sum x$$

and

$$\sum xy = a\sum x + b\sum x^2$$

S. No.	**x**	**y**	**x^2**	**xy**

1	6	5	36	30
2	7	4	49	28
3	8	3	64	24
4	9	2	81	18
5	11	1	121	11
Total	41	15	351	111

Here,

$$\sum y = 15,\ \sum x = 41,\ \sum xy = 111 \text{ and } \sum x^2 = 351$$

Then normal equations are

$$15 = 5a + 41b \qquad \dots (51)$$

$$111 = 41a + 351b \qquad \dots (52)$$

Now, we solve above two normal equations.

Multiplying equation (51) by 41 and equation (52) by 5, we have

$$15 \;=\; 5a + 41b \quad \big] \times 41$$

and

$$111 = 41 + 351b \quad \big] \times 5$$

we obtain

$$615 = 205\ a + 1681b \qquad \dots (53)$$

$$555 = 205\ a + 1755\ b \qquad \dots (54)$$

Subtracting equation (53) from equation (54), we have

$$-60 = 74\ b$$

$$\Rightarrow b = -\ 60/74 = -0.8108$$

Substituting the value of b in equation (51), we get

$$15 = 5\ a + 41 \times (-0.8108)$$

$$15 = 5\ a - 33.2428$$

$$\Rightarrow a = 9.6486$$

with these of a and b, the line of best fit is $Y = 9.6486 - 0.8108\ X$.

E2) Let $Y = a + bX + cX^2$ be the second degree parabola and we have to determine a , b and c. Normal equations for second degree parabola are

$$\sum y = na + b\sum x + c\sum x^2$$

$$\sum xy = a\sum x + b\sum x^2 + c\sum x^3$$

$$\sum x^2 y = a\sum x^2 + b\sum x^3 + c\sum x^4$$

To solve above normal equations, we need

$$\sum y, \sum x, \sum xy, \sum x^2 y, \sum x^3 \text{ and } \sum x^4$$

which are obtained from following table

x	**y**	**xy**	**x^2**	**x^2y**	**x^3**	**x^4**
2	1	2	4	4	8	16
4	2	8	16	32	64	256
6	3	18	36	108	216	1296
8	4	32	64	256	512	4096
10	5	50	100	500	1000	10000
$\sum x$ $= 30$	$\sum y$ $= 15$	$\sum xy$ $= 450$	$\sum x^2$ $= 220$	$\sum x^2 y$ $= 900$	$\sum x^3$ $= 1800$	$\sum x^4$ $= 15664$

Here, $\sum x = 30, \sum y = 15, \sum xy = 450, \sum x^2 y = 900,$
$\sum x^2 = 220 , \sum x^3 = 1800$ and $\sum x^4 = 15664.$

Substituting these values in above normal equations, we have

15 = 5a + 30b + 220c … (55)

110 = 30a + 220b + 1800c … (56)

900 = 220a + 1800b b + 15664c … (57)

Now we solve equations (55), (56) and (57).By multiply equation (55) by 6

90 = 30a + 180b + 1320c … (58)

Subtracting equation (58) from equation (56), we get

110 = 30a + 220b +1800c

90 = 30a +180b +1320c

----

20 = 40b + 480c … (59)

Multiplying equation (55) by 44, we have

660 = 220a +1320b + 9680c … (60)

Subtracting equation (59) from equation (57), we get

900 = 220a +1800b +15664c

660 = 220a +1320b + 9680c

----

240 = 480b + 5984c … (61)

Now we solve equation (59) and equation (61)

Multiplying equation (59) by 12, we get

240 = 480b + 5760c … (62)

Subtracting equation (61) from equation (62)

$$240 = 480b + 5760c$$

$$240 = 480b + 5984c$$

$$0 = -\,224\ c$$

$$\Rightarrow c = 0$$

Substituting the value of c in equation (62), we get

$$240 = 480b + 5760 \times 0 \Rightarrow b = \frac{240}{480} = 0.5$$

Putting the values of b and c in equation (55), we get

$$15 = 5a + 30 \times (0.5) + 220 \times (0)$$

$$15 = 5a + 15$$

$$\Rightarrow a = 0$$

Thus, the second degree parabola of best fit is

$$Y = 0 + 0.5\ X + 0\,X^2 \Rightarrow Y = 0.5X$$

E 3) Let the power curve be $Y = aX^b$ and normal equations for estimating a and b are

$$\sum u = nA + b\sum v$$

$$\sum uv = A\sum v + b\sum v^2$$

where,

$$U = \log Y,\ V = \log X \text{ and } A = \log a$$

To find the values of a and b from the above normal equations we require $\sum u$ $\sum v$, $\sum uv$ and $\sum v^2$ which are being calculated in the following table:

x	y	u=log y	v=log x	uv	v
5	2	0.3010	0.6990	0.2104	0.4886
6	5	0.6990	0.7782	0.5439	0.6055
9	8	0.9031	0.9542	0.8618	0.9106
8	11	1.0414	0.9031	0.9405	0.8156
11	15	1.1761	1.0414	1.2248	1.0845
		$\sum u$ $= 4.1206$	$\sum v$ $= 4.3759$	$\sum uv$ $= 3.7814$	$\sum v^2$ $= 3.9048$

Substituting the values $\sum u = 4.1206$, $\sum v = 4.3759$, $\sum uv$ =3.7814 and $\sum v^2 = 3.9048$ in above normal equations, we have

$4.1206 = 5A + 4.3759\ b$ … (63)

$3.7184 = 4.3759A + 3.9048\ b$ … (64)

Now we solve equation (63) and equation (64).

Multiplying equation (63) by 4.3759 and equation (64) by 5, we have

$18.0303 = 21.8795A + 19.1485\ b$ … (65)

$18.5920 = 21.8795A + 19.5240\ b$ … (66)

Subtracting equation (65) from equation (66), we have

$0.5619 = 0.3755b$

$\Rightarrow b = 1.4964$

Substituting the value of b in equation (63), we obtain

$A = -0.4571$

Now a = antilog (A) = antilog (−0.4571)

$a = 0.3491$

Thus, the power curve of the best fit is $Y = 0.3491\ X^{1.4964}$

E 4) Let the exponential curve be $Y = ab^X$ and normal equations for estimating a and be are

$$\sum u = nA + B\sum x$$

$$\sum ux = A\sum x + B\sum x^2$$

where, a = antilog(A) and b = antilog (B)

To find the values of a and b from the above normal equations we require $\sum u$ $\sum x$, $\sum ux$ and $\sum x^2$ which are being calculated in the following table:

x	**y**	**u = log y**	**ux**	**x^2**
1	8	0.9031	0.9031	1
2	15	1.1762	2.3522	4
3	33	1.5185	4.5555	9
4	65	1.8129	7.2517	16
5	130	2.1139	10.5697	25
$\sum x = 15$		$\sum u = 7.5246$	$\sum ux = 25.6322$	$\sum x^2 = 55$

Substituting the values $\sum x$, $\sum u$, $\sum ux$ and $\sum x^2$ in above normal equations, we have

$7.5246 = 5A + 15\ B$

$$25.6322 = 15\ A + 55\ B$$

After solving these equations we get B = − 0.3058

and A = 2.4224

Consequently, a = antilog (2.4224) = 264.5087 and

b = antilog (B) = antilog (− 0.3058) = 0.4494

Thus, the exponential curve of best fit is $Y = 264.5087(0.4494)^X$

UNIT 6 CORRELATION COEFFICIENT

Structure

6.1 Introduction

Objectives

6.2 Concept and Definition of Correlation

6.3 Types of Correlation

6.4 Scatter Diagram

6.5 Coefficient of Correlation

Assumptions for Correlation Coefficient

6.6 Properties of Correlation Coefficient

6.7 Short-cut Method for the Calculation of Correlation Coefficient

6.8 Correlation Coefficient in Case of Bivariate Frequency Distribution

6.9 Summary

6.10 Solutions / Answers

6.1 INTRODUCTION

In Block 1, you have studied the various measures such as measures of central tendency, measures of dispersion, moments, skewness and kurtosis which analyse variables separately. But in many situations we are interested in analysing two variables together to study the relationship between them. In this unit, you will learn about the correlation, which studies the linear relationship between the two or more variables. You would be able to calculate correlation coefficient in different situations with its properties. Thus before starting this unit you are advised to go through the arithmetic mean and variance that would be helpful in understanding the concept of correlation.

In Section 6.2, the concept of correlation is discussed with examples, that describes the situations, where there would be need of correlation study. Section 6.3 describes the types of correlation. Scatter diagrams which give an idea about the existence of correlation between two variables is explained in Section 6.4. Definition of correlation coefficient and its calculation procedure are discussed in Section 6.5. In this unit, some problems are given which illustrate the computation of the correlation coefficient in different situations as well as by different methods. Some properties of correlation coefficient with their proof are also given. In Section 6.6 the properties of the correlation coefficient are described whereas the shortcut method for the calculation of the correlation coefficient is explained in Section 6.7. In Section 6.8 the method of calculation of correlation coefficient in case of bivariate frequency distribution is explored.

Objectives

After reading this unit, you would be able to

- describe the concept of correlation;
- explore the types of correlation;
- describe the scatter diagram;

- interpret the correlation from scatter diagram;
- define correlation coefficient;
- describe the properties of correlation coefficient; and
- calculate the correlation coefficient.

6.2 CONCEPT AND DEFINITION OF CORRELATION

In many practical applications, we might come across the situation where observations are available on two or more variables. The following examples will illustrate the situations clearly:

1. Heights and weights of persons of a certain group;
2. Sales revenue and advertising expenditure in business; and
3. Time spent on study and marks obtained by students in exam.

If data are available for two variables, say X and Y, it is called bivariate distribution.

Let us consider the example of sales revenue and expenditure on advertising in business. A natural question arises in mind that is there any connection between sales revenue and expenditure on advertising? Does sales revenue increase or decrease as expenditure on advertising increases or decreases?

If we see the example of time spent on study and marks obtained by students, a natural question appears whether marks increase or decrease as time spent on study increase or decrease.

In all these situations, we try to find out relation between two variables and correlation answers the question, if there is any relationship between one variable and another.

When two variables are related in such a way that change in the value of one variable affects the value of another variable, then variables are said to be correlated or there is correlation between these two variables.

Now, let us solve a little exercise.

E1) What do you mean by Correlation?

6.3 TYPES OF CORRELATION

According to the direction of change in variables there are two types of correlation

1. Positive Correlation
2. Negative Correlation

1. Positive Correlation

Correlation between two variables is said to be positive if the values of the variables deviate in the same direction i.e. if the values of one variable increase (or decrease) then the values of other variable also increase (or decrease). Some examples of positive correlation are correlation between

1. Heights and weights of group of persons;
2. House hold income and expenditure;
3. Amount of rainfall and yield of crops; and
4. Expenditure on advertising and sales revenue.

In the last example, it is observed that as the expenditure on advertising increases, sales revenue also increases. Thus, the change is in the same direction. Hence the correlation is positive.

In remaining three examples, usually value of the second variable increases (or decreases) as the value of the first variable increases (or decreases).

2. Negative Correlation

Correlation between two variables is said to be negative if the values of variables deviate in opposite direction i.e. if the values of one variable increase (or decrease) then the values of other variable decrease (or increase). Some examples of negative correlations are correlation between

1. Volume and pressure of perfect gas;
2. Price and demand of goods;
3. Literacy and poverty in a country; and
4. Time spent on watching TV and marks obtained by students in examination.

In the first example pressure decreases as the volume increases or pressure increases as the volume decreases. Thus the change is in opposite direction.

Therefore, the correlation between volume and pressure is negative.

In remaining three examples also, values of the second variable change in the opposite direction of the change in the values of first variable.

Now, let us solve a little exercise.

E2) Explore some examples of positive and negative correlations.

6.4 SCATTER DIAGRAM

Scatter diagram is a statistical tool for determining the potentiality of correlation between dependent variable and independent variable. Scatter diagram does not tell about exact relationship between two variables but it indicates whether they are correlated or not.

Let (x_i, y_i); $(i = 1,2,...,n)$ be the bivariate distribution. If the values of the dependent variable Y are plotted against corresponding values of the independent variable X in the XY plane, such diagram of dots is called scatter diagram or dot diagram. It is to be noted that scatter diagram is not suitable for large number of observations.

6.4.1 Interpretation from Scatter Diagram

In the scatter diagram

1. If dots are in the shape of a line and line rises from left bottom to the right top (Fig.1), then correlation is said to be perfect positive.

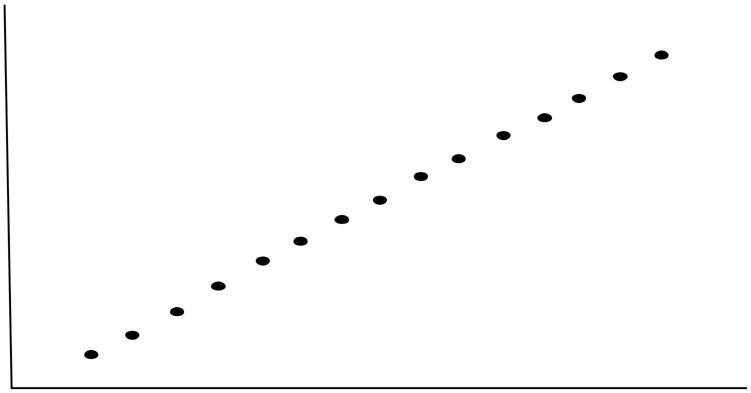


Fig. 1: Scatter diagram for perfect positive correlation

2. If dots in the scatter diagram are in the shape of a line and line moves from left top to right bottom (Fig. 2), then correlation is perfect negative.

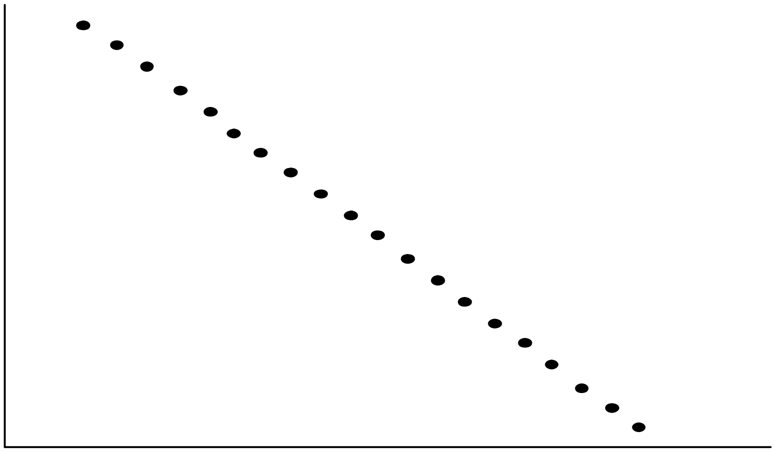


Fig. 2: Scatter diagram for perfect negative correlation

3. If dots show some trend and trend is upward rising from left bottom to right top (Fig.3) correlation is positive.

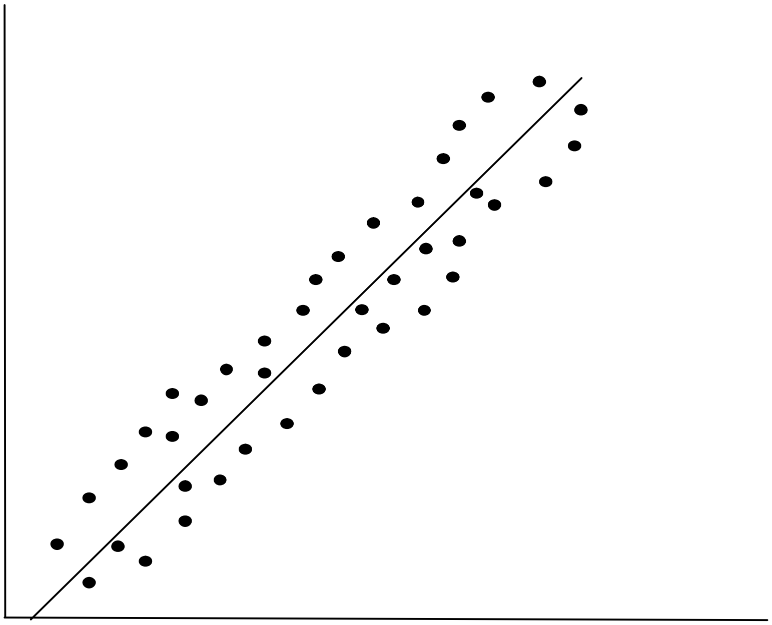


Fig. 3: Scatter diagram for positive correlation

4. If dots show some trend and trend is downward from left top to the right bottom (Fig.4) correlation is said to be negative.

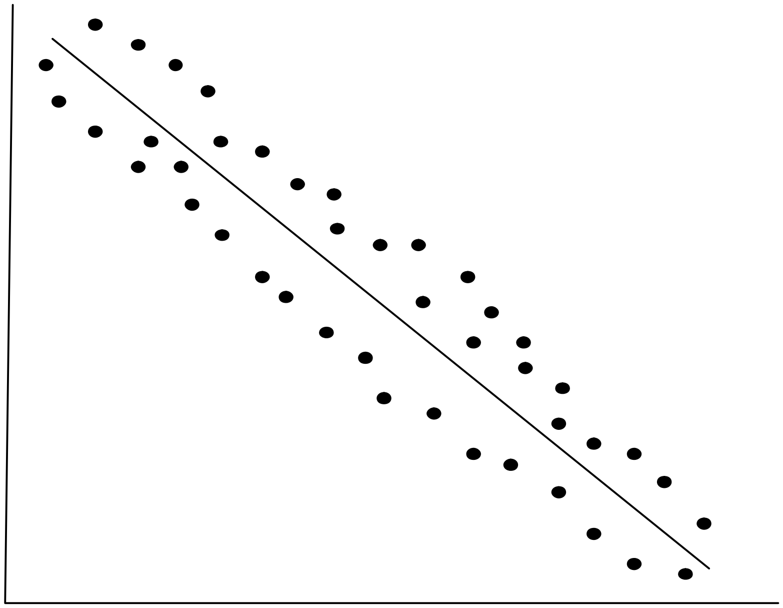


Fig. 4: Scatter diagram for negative correlation

5. If dots of scatter diagram do not show any trend (Fig. 5) there is no correlation between the variables.

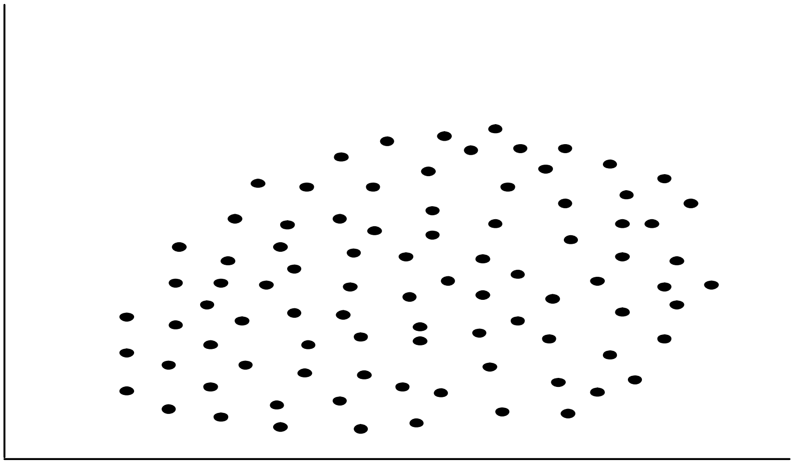


Fig. 5: Scatter diagram for uncorrelated data

6.5 COEFFICIENT OF CORRELATION

Scatter diagram tells us whether variables are correlated or not. But it does not indicate the extent of which they are correlated. Coefficient of correlation gives the exact idea of the extent of which they are correlated.

Coefficient of correlation measures the intensity or degree of linear relationship between two variables. It was given by British Biometrician Karl Pearson (1867-1936).

Note: Linear relationships can be expressed in such a way that the independent variable is multiplied by the slope coefficient, added by a constant, which determines the dependent variable. If Y is a dependent variable, X is an independent variable, b is a slope coefficient and a is constant then linear relationship is expressed as $Y = a + bX$.
In fact linear relationship is the relationship between dependent and independent variables of direct proportionality. When these variables plotted on a graph give a straight line.

If X and Y are two random variables then correlation coefficient between X and Y is denoted by r and defined as

$$r = \text{Corr}(x, y) = \frac{\text{Cov}(x, y)}{\sqrt{V(x)\,V(y)}} \qquad \ldots(1)$$

$\text{Corr}(x, y)$ is indication of correlation coefficient between two variables X and Y.

Where, $\text{Cov}(x, y)$ the covariance between X and Y which is defined as:

$$\text{Cov}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

and $V(x)$ the variance of X, is defined as:

$$V(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2,$$

Note: In the above expression $\sum_{i=1}^{n}$ denotes the sum of the values for i = 1 to n; For example $\sum_{i=1}^{n} x_i$ means sum of values of X for i = 1 to n. If n = 2 i.e. $\sum_{i=1}^{2} x_i$ which is equivalent to $x_1 + x_2$. If limits are not written the summation expression i.e. $\sum x$, which indicates the sum of all values of X. You may find the discussed formulae without limits in many books. We can also write $\frac{1}{N}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$ as $\frac{1}{N}\sum(x - \overline{x})(y - \overline{y})$ and both have same meaning.

Similarly,

$V(y)$ the variance of Y is defined by

$$V(y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2$$

where, n is number of paired observations.

Then, the correlation coefficient "r" may be defined as:

$$r = \text{Corr}(x, y) = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right)\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2\right)}} \qquad \ldots (2)$$

Karl Pearson's correlation coefficient r is also called product moment correlation coefficient. Expression in equation (2) can be simplified in various forms. Some of them are

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\left(\sum_{i=1}^{n}(x_i - \overline{x})^2\right)\left(\sum_{i=1}^{n}(y_i - \overline{y})^2\right)}} \qquad \ldots (3)$$

or

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}x_i y_i - \overline{x}\,\overline{y}}{\sqrt{\left\{\frac{\sum_{i=1}^{n}x_i^2}{n} - \overline{x}^2\right\}\left\{\frac{\sum_{i=1}^{n}y_i^2}{n} - \overline{y}^2\right\}}} \qquad \ldots (4)$$

or

$$r = \frac{\sum_{i=1}^{n}x_i y_i - n\overline{x}\,\overline{y}}{\sqrt{\left\{\sum_{i=1}^{n}x_i^2 - n\overline{x}^2\right\}\left\{\sum_{i=1}^{n}y_i^2 - n\overline{y}^2\right\}}} \qquad \ldots (5)$$

or

$$r = \frac{n\sum_{i=1}^{n}x_i y_i - \sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{\sqrt{\left\{n\sum_{i=1}^{n}x_i^2 - \left(\sum_{i=1}^{n}x_i\right)^2\right\}\left\{n\sum_{i=1}^{n}y_i^2 - \left(\sum_{i=1}^{n}y_i\right)^2\right\}}} \qquad \ldots (6)$$

6.5.1 Assumptions for Correlation Coefficient

1. **Assumption of Linearity**
   Variables being used to know correlation coefficient must be linearly related. You can see the linearity of the variables through scatter diagram.

2. **Assumption of Normality**
   Both variables under study should follow Normal distribution. They should not be skewed in either the positive or the negative direction.

3. **Assumption of Cause and Effect Relationship**
   There should be cause and effect relationship between both variables, for example, Heights and Weights of children, Demand and Supply of goods, etc. When there is no cause and effect relationship between variables then correlation coefficient should be zero. If it is non zero then correlation is termed as chance correlation or spurious correlation. For example, correlation coefficient between:

1. Weight and income of a person over periods of time; and
2. Rainfall and literacy in a state over periods of time.

Now, let us solve a little exercise.

E3) Define correlation coefficient.

6.6 PROPERTIES OF CORRELATION COEFFICIENT

Property 1: Correlation coefficient lies between -1 and +1.

Description: Whenever we calculate the correlation coefficient by any one of the formulae given in the Section 6.5 its value always lies between -1 and +1.

Proof: Consider

$$\frac{1}{n}\sum\left[\left(\frac{x-\bar{x}}{\sigma_x}\right)\pm\left(\frac{y-\bar{y}}{\sigma_y}\right)\right]^2 \geq 0$$

(Since square quantity is always greater than or equal to zero)

$$\Rightarrow \frac{1}{n}\sum\left(\frac{x-\bar{x}}{\sigma_x}\right)^2+\frac{1}{n}\sum\left(\frac{y-\bar{y}}{\sigma_y}\right)^2 \pm 2\frac{1}{n}\sum\left[\left(\frac{x-\bar{x}}{\sigma_x}\right)\left(\frac{y-\bar{y}}{\sigma_y}\right)\right] \geq 0$$

$$\Rightarrow \frac{\sigma_x^2}{\sigma_x^2}+\frac{\sigma_y^2}{\sigma_y^2}\pm\frac{2\text{Cov}(x,y)}{\sigma_x\sigma_y} \geq 0$$

Since $\frac{1}{n}\sum(x-\bar{x})^2 = \text{Variance of } X = \sigma_x^2,$

Similarly, $\frac{1}{n}\sum(y-\bar{y})^2 = \sigma_y^2$ and $\frac{1}{n}\sum(x-\bar{x})(y-\bar{y}) = \text{Cov}(x,y)$

Therefore, after putting the values

$$\Rightarrow 1+1\pm 2r \geq 0$$

$$\Rightarrow 2\pm 2r \geq 0$$

$$\Rightarrow 1\pm r \geq 0 \qquad \dots (7)$$

If we take positive sign in equation (7) then

$$r \geq -1$$

It shows that the correlation coefficient will always be greater than or equal to -1.

If we take negative sign in equation (7) then

$$1-r \geq 0$$

$$r \leq 1$$

It shows that correlation coefficient will always be less than or equal to +1. Thus

$$\Rightarrow -1 \leq r \leq 1$$

If $r = +1$, the correlation is perfect positive and if $r = -1$ correlation is perfect negative.

Property 2: Correlation coefficient is independent of change of origin and scale.

Description: Correlation coefficient is independent of change of origin and scale, which means that if a quantity is subtracted and divided by another quantity (greater than zero) from original variables , i.e. $U = \frac{X - a}{h}$ and $V = \frac{Y - b}{k}$ then correlation coefficient between new variables U and V is same as correlation coefficient between X and Y, i.e. $\text{Corr}(x, y) = \text{Corr}(u, v)$.

Proof: Suppose $u = \frac{x - a}{h}$ and $v = \frac{y - b}{k}$ then

$$x = a + h\,u \text{ and } \bar{x} = a + h\,\bar{u} \qquad \ldots (8)$$

$$\text{and} \quad y = a + k\,v \text{ and } \bar{y} = a + h\,\bar{v} \qquad \ldots (9)$$

where, a, b, h and k are constants such that a > 0, b > 0, h > 0 and k > 0.

We have to prove $\text{Corr}(x, y) = \text{Corr}(u, v)$ i.e. there is no change in correlation when origin and scale are changed.

$$\text{Cov}(x, y) = \frac{1}{n}\sum\left(x - \bar{x}\right)\left(y - \bar{y}\right)$$

$$= \frac{1}{n}\sum\left(a + h\,u - a - h\,\bar{u}\right)\left(b + k\,v - b - b\,\bar{v}\right)$$

$$= \frac{1}{n}hk\sum\left(u - \bar{u}\right)\left(v - \bar{v}\right),$$

$$\text{Cov}(x, y) = hk\text{Cov}(u, v)$$

and

$$V(x) = \frac{1}{n}\sum\left(x - \bar{x}\right)^2$$

$$= \frac{1}{n}\sum\left(a + hu - a - h\,\bar{u}\right)^2$$

$$= h^2\frac{1}{n}\sum\left(u - \bar{u}\right)^2$$

$$V(x) = h^2V(u)$$

Similarly,

$$V(y) = k^2V(v)$$

$$\text{Corr}(x, y) = \frac{\text{Cov}(x, y)}{\sqrt{V(x)V(y)}}$$

$$\text{Corr}(x, y) = \frac{hk\text{Cov}(u, v)}{\sqrt{h^2V(u)k^2V(v)}}$$

$$\text{Corr}(x, y) = \frac{\text{Cov}(u, v)}{\sqrt{V(u)V(v)}}$$

$$\text{Corr}(x, y) = \text{Corr}(u, v)$$

i.e. correlation coefficient between X and Y is same as correlation coefficient between U and V Thus, **c**orrelation coefficient is independent of change of origin and scale.

Property 3: If X and Y are two independent variables then correlation coefficient between X and Y is zero, i.e. $\text{Corr}(x, y) = 0$.

Proof: Covariance between X and Y is defined by

$$\text{Cov}(x, y) = \frac{1}{n}\sum(x - \bar{x})(y - \bar{y})$$

$$= \frac{1}{n}\sum(xy - y\bar{x} - x\bar{y} + \bar{x}\bar{y})$$

$$= \frac{1}{n}\sum xy - \bar{x}\frac{1}{n}\sum y - \bar{y}\frac{1}{n}\sum x + \bar{x}\bar{y}\frac{1}{n}\sum 1$$

$$= \frac{1}{n}\sum xy - \bar{x}\bar{y} - \bar{y}\bar{x} + \bar{x}\bar{y}$$

$$= \frac{1}{n}\sum xy - \bar{x}\bar{y}$$

$$= \bar{x}\bar{y} - \bar{x}\bar{y}$$

(if variables are independent then, $\frac{1}{n}\sum xy = \bar{x}\bar{y}$). Therefore,

$$\text{Cov}(x, y) = 0$$

Thus, correlation is

$$\text{Corr}(x, y) = \frac{\text{Cov}(x, y)}{\sqrt{V(x)V(y)}} = \frac{0}{\sqrt{V(x)V(y)}} = 0$$

As correlation measures the degree of linear relationship, different values of coefficient of correlation can be interpreted as below:

Value of correlation coefficient	**Correlation is**
+1	Perfect Positive Correlation
- 1	Perfect Negative Correlation
0	There is no Correlation
0 - 0.25	Weak Positive Correlation
0.75 - (+1)	Strong Positive Correlation
−0.25 - 0	Weak Negative Correlation
−0.75 - (−1)	Strong Negative Correlation

Let us discuss some problems of calculation of correlation coefficient.

Example 1: Find the correlation coefficient between advertisement expenditure and profit for the following data:

Advertisement expenditure	30	44	45	43	34	44
Profit	56	55	60	64	62	63

Solution: To find out the correlation coefficient between advertisement expenditure and profit, we have Karl Pearson's formula in many forms [(2), (3), (4), (5) and (6)] and any of them can be used. All these forms provide the same result. Let us take the form of equation (3) to solve our problem which is

$$r = \mathrm{Corr}(x, y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\left(\sum_{i=1}^{n}(x_i - \overline{x})^2\right)\left(\sum_{i=1}^{n}(y_i - \overline{y})^2\right)}}$$

Steps for calculation are as follow:

1. In columns 1 and 2, we take the values of variables X and Y respectively.
2. Find sum of the variables X and Y i.e.

$$\sum_{i=1}^{6} x_i = 240 \text{ and } \sum_{i=1}^{6} y_i = 360$$

3. Calculate arithmetic means of X and Y as

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum_{i=1}^{6} x_i}{6} = \frac{240}{6} = 40$$

and $$\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{\sum_{i=1}^{6} y_i}{6} = \frac{360}{6} = 60$$

4. In column 3, we take deviations of each observations of X from mean of X, i.e. 30 − 40 = −10, 44 − 40 = 4 and so on other values of the column can be obtained.
5. Similarly column 5 is prepared for variable Y i.e.

   56 – 60 = −4, 55 − 60 = −5

   and so on.
6. Column 4 is the square of column 3 and column 6 is the square of column 5.
7. Column 7 is the product of column 3 and column 5.
8. Sum of each column is obtained and written at the end of column.

To find out the correlation coefficient by above formula, we require the values of $\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$, $\sum_{i=1}^{n}(x_i - \overline{x})^2$ and $\sum_{i=1}^{n}(y_i - \overline{y})^2$ which are obtained by the following table:

x	y	$(\mathbf{x}-\overline{\mathbf{x}})$	$(\mathbf{x}-\overline{\mathbf{x}})^2$	$(\mathbf{y}-\overline{\mathbf{y}})$	$(\mathbf{y}-\overline{\mathbf{y}})^2$	$(\mathbf{x}-\overline{\mathbf{x}})(\mathbf{y}-\overline{\mathbf{y}})$
30	56	−10	100	−4	16	40
44	55	4	16	−5	25	−20
45	60	5	25	0	0	0
43	64	3	9	4	16	12
34	62	−6	36	2	4	−12
44	63	4	16	3	9	12
$\sum x_i$ $= 240$	$\sum_{i=1}^{6} y_i$ $= 360$	$\sum_{i=1}^{6}(x_i - \overline{x})$ $= 0$	$\sum_{i=1}^{6}(x_i - \overline{x})^2$ $= 202$	$\sum_{i=1}^{6}(y_i - \overline{y})$ $= 0$	$\sum_{i=1}^{7}(y_i - \overline{y})^2$ $= 70$	$\sum_{i=1}^{6}(x_i - \overline{x})(y_i - \overline{y})$ $= 32$

Taking the values of $\sum_{i=1}^{6}(x_i - \overline{x})(y_i - \overline{y})$, $\sum_{i=1}^{6}(x_i - \overline{x})^2$ and $\sum_{i=1}^{6}(y_i - \overline{y})^2$ from the table and substituting in the above formula we have the correlation coefficient

$$r = \text{Corr}(x, y) = \frac{\sum_{i=1}^{6}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\left\{\sum_{i=1}^{6}(x_i - \overline{x})^2\right\}\left\{\sum_{i=1}^{6}(y_i - \overline{y})^2\right\}}}$$

$$r = \text{Corr}(x, y) = \frac{32}{\sqrt{202 \times 70}} = \frac{32}{\sqrt{14140}} = \frac{32}{118.91} = 0.27$$

Hence, the correlation coefficient between expenditure on advertisement and profit is 0.27. This indicates that the correlation between expenditure on advertisement and profit is positive and we can say that as expenditure on advertisement increases (or decreases) profit increases (or decreases). Since it lies between 0.25 and 0.5 it can be considered as week positive correlation coefficient.

Example 2: Calculate Karl Pearson's coefficient of correlation between price and demand for the following data.

Price	17	18	19	20	22	24	26	28	30
Demand	40	38	35	30	28	25	22	21	20

Solution: In Example 1, we used formula given in equation (3) in which deviations were taken from mean. When means of x and y are whole number, deviations from mean makes calculation easy. Since, in Example 1, means x and y were whole number we preferred formula given in equation (3). When means are not whole numbers calculation by formula given in equation (3) becomes cumbersome and we prefer any formula given in equation (4) or (5) or (6). Since here means of x and y are not whole number, so we are preferring formula (6)

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left\{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right\}\left\{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right\}}}$$

Let us denote price by the variable X and demand by variable Y.

To find the correlation coefficient between price i.e.X and demand Y using formula given in equation (6), we need to calculate, $\sum_{i=1}^{n} x_i$, $\sum_{i=1}^{n} y_i$, $\sum_{i=1}^{n} x_i y_i$, $\sum_{i=1}^{n} x_i^2$ and $\sum_{i=1}^{n} y_i^2$ which are being obtained in the following table:

x	**y**	**x^2**	**y^2**	**xy**
17	40	289	1600	680
18	38	324	1444	684
19	35	361	1225	665
20	30	400	900	600
22	28	484	784	616
24	25	576	625	600
26	22	676	484	572
28	21	784	441	588
30	20	900	400	600
$\sum x = 204$	$\sum y = 259$	$\sum x^2 = 4794$	$\sum y^2 = 7903$	$\sum xy = 5605$

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left\{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right\}\left\{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right\}}}$$

$$r = \text{Corr}(x, y)$$

$$= \frac{(9 \times 5605) - (204)(259)}{\sqrt{\{(9 \times 4794) - (204 \times 204)\}\{(9 \times 7903) - (259 \times 259)\}}}$$

$$r = \text{Corr}(x, y) = \frac{50445 - 52836}{\sqrt{(43146 - 41616) \times (71127 - 67081)}}$$

$$r = \text{Corr}(x, y) = \frac{-2391}{\sqrt{1530 \times 4046}}$$

$$r = \text{Corr}(x, y) = \frac{-2391}{2488.0474}$$

$$r = \text{Corr}(x, y) = -\ 0.96$$

Note: We can use $\sum x$ instead of $\sum_{i=1}^{n} x_i$. Second expression indicates sum over x_i for i =1 to n. On the other hand first expression ($\sum x$) indicates sum over all values of X. In the current example we are using summation sign ($\sum$) without limit.

Now, let us solve the following exercises.

E4) Calculate coefficient of correlation between x and y for the following data:

x	1	2	3	4	5
y	2	4	6	8	10

E5) Find the coefficient of correlation for the following ages of husband and wife:

Husband’s age	23	27	28	29	30	31
Wife’s age	18	22	23	24	25	26

6.7 SHORT-CUT METHOD FOR THE CALCULATION OF CORRELATION COEFFICIENT

When values of variables are big and actual means of variables X and Y i.e. $\overline{x}$ and $\overline{y}$ are not whole number (in Example 1 mean of X and Y i.e. $\overline{x} = 40$ and $\overline{y} = 60$ were whole number) then calculation of correlation coefficient by the formula (2), (3), (4), (5) and (6) is somewhat cumbersome and we have shortcut method in which deviations are taken from assumed mean i.e. instead of actual means $\overline{x}$ and $\overline{y}$, we use assumed mean, hence $(x_i - \overline{x})$ and $(y_i - \overline{y})$ are replaced by $x_i - A_x = d_x$ and $y_i - A_y = d_y$ where A_x and A_y are assumed means of (Assumed mean may be any value of given variable of our choice) variables X and Y respectively. Formula for correlation coefficient by shortcut method is

$$r = \text{Corr}(x, y) = \frac{n\sum d_x d_y - \sum d_x \sum d_y}{\sqrt{\left\{n\sum d_x^2 - (\sum d_x)^2\right\}\left\{n\sum d_y^2 - (\sum d_y)^2\right\}}}$$

Here,

n = No. of pairs of observations,

A_x = Assumed mean of X,

A_y = Assumed mean of Y,

$\sum d_x = \sum (x - A_x)$: Sum of deviation from assumed mean A_x in X-series,

$\sum d_y = \sum (x - A_y)$: Sum of deviation from assumed mean A_y in Y-series,

$\sum d_x d_y = \sum (x - A_x)(y - A_y)$: Sum of product of deviations from assumed means A_X and A_Y in x and y series respectively,

$\sum d_x^2 = \sum (x - A_x)^2$: Sum of squares of the deviations from assumed mean A_x, in x series and

$\sum d_y^2 = \sum (y - A_y)^2$: Sum of squares of the deviations from assumed mean A_y in y series.

Note: Results from usual method and short-cut method are same.

Example 3: Calculate correlation coefficient from the following data by short-cut method:

x	10	12	14	18	20
y	5	6	7	10	12

Solution: By short-cut method correlation coefficient is obtained by

$$r = \text{Corr}(x, y) = \frac{n \sum d_x d_y - \sum d_x \sum d_y}{\sqrt{\left\{ n \sum d_x^2 - (\sum d_x)^2 \right\}\left\{ n \sum d_y^2 - (\sum d_y)^2 \right\}}}$$

$\sum d_x$, $\sum d_y$, $\sum d_x d_y$, $\sum d_x^2$ and $\sum d_y^2$ are being obtained from the following table.

Let A_x = Assumed mean of X =14 and A_y = Assumed mean of Y = 7

x	**y**	**d_x = x - 14**	**d_x^2**	**d_y= y - 7**	**d_y^2**	**$d_x d_y$**
1	5	10 - 14 = -4	16	5–7 = –2	4	8
12	6	12–14 = –2	4	6–7 = –1	1	2
14	7	14–14 = 0	0	7–7 = 0	0	0
18	10	18–14 = 4	16	10–7 = 3	9	12
20	12	20–14 = 6	36	12–7 = 5	25	30
$\sum x$ $= 74$	$\sum y$ $= 40$	$\sum d_x = 4$	$\sum d_x^2$ $= 72$	$\sum d_y$ $= 5$	$\sum d_y^2$ $= 39$	$\sum d_x d_y$ $= 52$

Putting the required values in above formula

$$r = \frac{(5 \times 52) - (4 \times 5)}{\sqrt{\{(5 \times 72) - (4 \times 4)\}\{(5 \times 39) - (5 \times 5)\}}}$$

$$r = \frac{260 - 20}{\sqrt{\{360 - 16\}\{195 - 25\}}}$$

$$r = \frac{240}{\sqrt{\{344\}\{170\}}} = \frac{240}{241.8264} = 0.99$$

Thus, there is a very high correlation between x and y.

Now, let us solve an exercise.

E6) Find correlation coefficient between the values of X and Y from the following data by short -cut method:

x	10	20	30	40	50
y	90	85	80	60	45

6.8 Correlation Coefficient in Case of Bivariate Frequency Distribution

In bivariate frequency distribution one variable is presented in row and another in column and corresponding frequencies are given in cells (See Example 4).

If we consider two variables X and Y where, the distribution of X is given in columns and the distribution of Y is given in row. In this case we adopt the following procedure to calculate correlation coefficient.

1. Other than the given bivariate frequency distribution make three columns in the right of the table ($f_x d_x, f_x d_x^2$ and $f_x d_x d_y$) two columns in left of the table (mid value for x and class interval of variable (x) d_x), three rows in the bottom of the table $\left(f_y d_y\ , f_y d_y^2\right.$ and $\left. f_y d_x d_y\right)$and two rows in the top of the table (mid value for y and d_y). Where f_x is the sum of all frequencies for the given x value i.e. f_x= $\sum_y f_{xy}$ and f_y is the sum of all frequencies for the given y values i.e. f_y= $\sum_x f_{xy}$
2. Find the mid value x and $d_x = x_i - A_x$ i.e. deviation from assumed mean A_x or step deviation i.e. $d_x = (x_i - A_x)/h$ where h is such that $(x_i - A_x)/h$ is a whole number.
3. Apply step (2) for variable Y also.
4. Find $f_x d_x$ by multiplying d_x by respective frequency f_x and get $\sum_{i=1}^{N} f_x d_x$.
5. Find $f_y d_y$ by multiplying d_y by respective frequency f_y and get $\sum_{i=1}^{N} f_y d_y$.
6. Find $f_y d_y^2$ by multiplying d_y^2 by respective frequency f_y and get $\sum_{i=1}^{N} f_y d_y^2$.
7. Find $f_x d_x^2$ by multiplying d_x^2 by respective frequency f_x and get $\sum_{i=1}^{N} f_x d_x^2$.
8. Multiply respective d_x and d_y for each cell frequency and put the figures in left hand upper corner of each cell.

9. Find $f_{xy}d_xd_y$ by multiplying f_{xy} with d_xd_y and put the figures in right hand lower corner of each cell and we apply the following formula:

$$r = \frac{\sum f_{xy}d_xd_y - \frac{\sum f_xd_x \sum f_yd_y}{N}}{\sqrt{\left\{\sum f_xd_x^2 - \frac{(\sum f_xd_x)^2}{N}\right\}\left\{\sum f_yd_y^2 - \frac{(\sum f_yd_y)^2}{N}\right\}}}$$

where, $N = \sum f_x = \sum y_y$.

Example 4: Calculate the correlation coefficient between ages of husbands and ages of wives for the following bivariate frequency distribution:

Ages of Husbands	Ages of Wives					Total
	10-20	20-30	30-40	40-50	50-60	
15-25	6	3	-	-	-	9
25-35	3	16	10	-	-	29
35-45	-	10	15	7	-	32
45-55	-	-	7	10	4	21
55-65	-	-	-	4	5	9
Total	9	29	32	21	9	100

Solution: Let, $d_y = (y - 35)/10$, where assumed mean $A_y = 35$ and $y = 10$.
$d_x = (x - 40)/10$, where assumed mean $A_x = 40$ and $h = 10$.

		CI	10-20	20-30	30-40	40-50	50-60				
		MV (y)	15	25	35	45	55				
CI	MV (x)	d_x / d_y	-2	-1	0	+1	+2	f_x	$f_x d_x$	$f_x d_x^2$	$f_{xy}d_xd_y$
15-25	20	-2	4 6 **24**	2 3 **6**	-	-	-	9	-18	36	30
25-35	30	-1	2 3 **6**	1 16 **16**	0 10 **0**	-	-	29	-29	29	22
35-45	40	0	-	0 10 **0**	0 15 0	0 7 **0**	-	32	0	0	0
45-55	50	+1	-	-	0 7 **0**	1 10 **10**	2 4 **8**	21	21	21	18
55-65	60	+2	-	-	-	2 4 **8**	4 5 **20**	9	18	36	28
		f_y	9	29	32	21	9	N = 100	$\sum f_x d_x$ = - 8	$\sum f_x d_x^2$ = 122	$\sum f_{xy}d_xd_y$ = 98
		f_yd_y	-18	-29	0	21	18	$\sum f_y d_y$ = -8			
		$f_yd_y^2$	36	29	0	21	36	$\sum f_y d_y^2$ = 122			
		$f_{xy}d_xd_y$	30	22	0	18	28	$\sum f_{xy} d_x d_v$ = 98			

$$r = \frac{98 - \frac{(-8\times -8)}{100}}{\sqrt{\left\{122 - \frac{(-8)^2}{100}\right\}\left\{122 - \frac{(-8)^2}{100}\right\}}}$$

$$r = \frac{98 - 0.64}{\sqrt{\{122 - 0.64\}\{122 - 0.64\}}} = 0.802$$

6.8 SUMMARY

In this unit, we have discussed:

1. Concept of correlation;
2. Types of correlation;
3. The scatter diagrams of different correlations;
4. Calculation of Karl Pearson's coefficient of correlation;
5. Short-cut method of calculation of correlation coefficient;
6. Properties of correlation coefficient; and
7. Calculation of correlation coefficient for bi-variate data.

6.9 SOLUTIONS / ANSWERS

E1) When two variables are related in such a way that change in the value of one variable affects the value of another variable, then variables are said to be correlated or there is correlation between these two variables.

E2) Positive correlation: Correlation between

(i) Sales and profit

(ii) Family Income and year of education

Negative correlation: Correlation between

(i) No. of days students absent in class and score in exam

(ii) Time spent in office and time spent with family

E3) **C**oefficient of correlation measures the intensity or degree of linear relationship between two variables. It is denoted by r. Formula for the calculation of correlation coefficient is

$$r = \mathrm{Corr}(x, y) = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2\right)}}$$

E4) We have some calculation in the following table:

x	**y**	$(\mathbf{x}-\bar{\mathbf{x}})$	$(\mathbf{x}-\bar{\mathbf{x}})^2$	$(\mathbf{y}-\bar{\mathbf{y}})$	$(\mathbf{y}-\bar{\mathbf{y}})^2$	$(\mathbf{x}-\bar{\mathbf{x}})(\mathbf{y}-\bar{\mathbf{y}}$
1	2	-2	4	-4	16	8
2	4	-1	1	-2	4	2
3	6	0	0	0	0	0
4	8	1	1	2	4	2
5	10	2	4	4	16	8
15	30	0	10	0	40	20
15	30	0	10	0	40	20

Here $\sum x = 15$

$$\Rightarrow \bar{x} = \sum x / n = 15/5 = 3$$

and

$$\sum y = 30$$

$$\Rightarrow \bar{y} = \sum y / n = 30/5 = 6$$

From the calculation table, we observe that

$$\sum (x-\bar{x})^2 = 10, \ \sum (y-\bar{y})^2 = 40 \text{ and } \sum (x-\bar{x})(y-\bar{y}) = 20$$

Substituting these values in the formula

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)\left(\sum_{i=1}^{n}(y_i-\bar{y})^2\right)}}$$

$$r = \text{Corr}(x, y) = \frac{20}{\sqrt{10\times 40}} = \frac{20}{20} = 1$$

Hence, there is perfect positive correlation between X and Y.

E5) Let us denote the husband's age as X and wife's age by Y

x	**y**	$\mathbf{x}^2$	$\mathbf{y}^2$	**xy**
23	18	529	324	414
27	22	729	484	594
28	23	784	529	644
29	24	841	576	696
30	25	900	625	750
31	26	961	676	806
$\sum x = 168$	$\sum y = 138$	$\sum x^2 = 4744$	$\sum y^2 = 3214$	$\sum xy = 3904$

Here,

$$\sum x = 168, \ \sum y = 138, \sum x^2 = 4744,$$

$$\sum y^2 = 3214 \sum xy = 3904$$

We use the formula

$$r = \text{Corr}(x, y) = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{\left\{n\sum x^2 - (\sum x)^2\right\}\left\{n\sum y^2 - (\sum y)^2\right\}}}$$

$$r = \frac{(6 \times 3904) - (168)(138)}{\sqrt{\{(6 \times 4744) - (168 \times 168)\}\{(6 \times 3214) - (138 \times 138)\}}}$$

$$r = \frac{23424 - 23184}{\sqrt{\{28464 - 28224\}\{19284 - 19044\}}}$$

$$r = \frac{23424 - 23184}{\sqrt{240 \times 240}} = 1$$

Hence there is perfect positive correlation between X and Y.

E6) By short-cut method correlation coefficient is obtained by

$$r = \text{Corr}(x, y) = \frac{n\sum d_x d_y - \sum d_x \sum d_y}{\sqrt{\{n\sum d_x^2 - (\sum d_x)^2\}\{n\sum d_y^2 - (\sum d_y)^2\}}}$$

$\sum d_x$, $\sum d_y$, $\sum d_x d_y$, $\sum d_x^2$ and $\sum d_y^2$ are being obtained through the following table.

Let A_x = assumed mean of X = 30 and A_y = Assumed mean of Y = 70

x	**y**	**d_x=x-30**	**d_x^2**	**d_y=y-70**	**d_y^2**	**$d_x d_y$**
10	90	-20	400	20	400	-400
20	85	-10	100	15	225	-150
30	80	0	0	10	100	0
40	60	10	100	-10	100	-100
50	45	20	400	-25	625	-500
$\sum x = 150$	$\sum y = 360$	$\sum d_x = 0$	$\sum d_x^2 = 1000$	$\sum d_y = 10$	$\sum d_y^2 = 1450$	$\sum d_x d_y = -1150$

Putting the required values in above formula

$$r = \frac{(5 \times -1150) - (0 \times 10)}{\sqrt{\{(5 \times 1000) - (0)\}\{(5 \times 1450) - (10 \times 10)\}}}$$

$$r = \frac{-5750}{\sqrt{\{5000\}\{7250 - 100\}}}$$

$$r = \frac{-5750}{5979.1304} = -0.96.$$

UNIT 7 RANK CORRELATION

Structure

7.1 Introduction

Objectives

7.2 Concept of Rank Correlation

7.3 Derivation of Rank Correlation Coefficient Formula

7.4 Tied or Repeated Ranks

7.5 Concurrent Deviation

7.6 Summary

7.7 Solutions / Answers

7.1 INTRODUCTION

In second unit of this block, we have discussed the correlation with its properties and also the calculation of correlation coefficient. In correlation coefficient or product moment correlation coefficient, it is assumed that both characteristics are measurable. Sometimes characteristics are not measurable but ranks may be given to individuals according to their qualities. In such situations rank correlation is used to know the association between two characteristics. In this unit, we will discuss the rank correlation and calculation of rank correlation coefficient with its merits and demerits. We will also study the method of concurrent deviation.

In Section 7.2, you will know the concept of rank correlation while Section 7.3 gives the derivation of Spearman's rank correlation coefficient formula. Merits and demerits of the rank correlation coefficient are discussed in Sub-section 7.3.1. There might be a situation when two items get same rank. This situation is called tied or repeated rank which is described in Section 7.4. You will learn the method of concurrent deviation in Section 7.5.

Objectives

After reading this unit, you would be able to

- explain the concept of rank correlation;
- derive the Spearman's rank correlation coefficient formula;
- describe the merits and demerits of rank correlation coefficient;
- calculate the rank correlation coefficient in case of tied or repeated ranks; and
- describe the method of concurrent deviation.

7.2 CONCEPT OF RANK CORRELATION

For the calculation of product moment correlation coefficient characters must be measurable. In many practical situations, characters are not measurable. They are qualitative characteristics and individuals or items can be ranked in order of their merits. This type of situation occurs when we deal with the qualitative study such as honesty, beauty, voice, etc. For example, contestants of a singing competition may be ranked by judge according to their performance. In another example, students may be ranked in different subjects according to their performance in tests.

Arrangement of individuals or items in order of merit or proficiency in the possession of a certain characteristic is called ranking and the number indicating the position of individuals or items is known as rank.

If ranks of individuals or items are available for two characteristics then correlation between ranks of these two characteristics is known as rank correlation.

With the help of rank correlation, we find the association between two qualitative characteristics. As we know that the Karl Pearson's correlation coefficient gives the intensity of linear relationship between two variables and Spearman's rank correlation coefficient gives the concentration of association between two qualitative characteristics. In fact Spearman's rank correlation coefficient measures the strength of association between two ranked variables. Derivation of the Spearman's rank correlation coefficient formula is discussed in the following section.

7.3 DERIVATION OF RANK CORRELATION COEFFICIENT FORMULA

Suppose we have a group of n individuals and let $x_1, x_2, ..., x_n$ and $y_1, y_2, ..., y_n$ be the ranks of n individuals in characteristics A and B respectively.

It is assumed that no two or more individuals have the same rank in either characteristics A or B. Suppose both characteristics X and Y are taking rank values 1, 2, 3,…, n. Then sum of ranks of characteristics A is

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + ... + x_n$$

$$= 1 + 2 + ... + n \qquad \text{(since X is taking values 1, 2,…,n)}$$

$$\sum_{i=1}^{n} x_i = \frac{n(n+1)}{2} \qquad \text{…(1)}$$

From, the formula of sum of n natural numbers.

Mean of variable X is

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\frac{n(n+1)}{2}$$

$$\bar{x} = \frac{(n+1)}{2}$$

Since both variables are taking same values 1, 2, …, n then

$$\bar{x} = \bar{y} = \frac{(n+1)}{2}$$

If variance of X is denoted by $\sigma_x{}^2$ then

$$\sigma_x{}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

$$\sigma_x{}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i^2 + \overline{x}^2 - 2x_i\overline{x})$$

$$\sigma_x{}^2 = \frac{1}{n}(\sum_{i=1}^{n}x_i^2 + \sum_{i=1}^{n}\overline{x}^2 - 2\overline{x}\sum_{i=1}^{n}x_i)$$

$$\sigma_x{}^2 = \frac{1}{n}(\sum_{i=1}^{n}x_i^2 + n\overline{x}^2 - 2n\overline{x}^2)$$

$$\sigma_x{}^2 = \frac{1}{n}(\sum_{i=1}^{n}x_i^2 - n\overline{x}^2)$$

$$\sigma_x{}^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \overline{x}^2 = \frac{1}{n}(x_1^2 + x_2^2 + \ldots + x_n^2) - \overline{x}^2 \qquad \ldots (2)$$

Substituting the value of $\overline{x}$ in equation (2), we have

$$\sigma_x{}^2 = \frac{1}{n}(1^2 + 2^2 + \ldots + n^2) - \left\{\frac{n+1}{2}\right\}^2 \text{ (Since X is taking values 1,2,…,n)}$$

$$\sigma_x{}^2 = \frac{1}{n}\left\{\frac{n(n+1)(2n+1)}{6}\right\} - \left\{\frac{n+1}{2}\right\}^2$$

(From the formula of sum of squares of n natural numbers)

$$\sigma_x{}^2 = \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4}$$

$$\sigma_x{}^2 = (n+1)\left\{\frac{2n+1}{6} - \frac{(n+1)}{4}\right\}$$

$$\sigma_x{}^2 = (n+1)\left\{\frac{2(2n+1) - 3(n+1)}{12}\right\}$$

$$\sigma_x{}^2 = (n+1)\left\{\frac{4n+2-3n-3}{12}\right\}$$

$$\sigma_x{}^2 = (n+1)\left\{\frac{n-1}{12}\right\}$$

$$\sigma_x{}^2 = \frac{n^2-1}{12} \qquad \text{(from the formula } (a - b)(a + b) = a^2 - b^2\text{)}$$

Since both variables X and Y are taking same values, they will have same variance, thus

$$\sigma_y^2 = \sigma_x^2 = \frac{n^2-1}{12}$$

Let d_i be the difference of the ranks of i^{th} individual in two characteristics, then

$$d_i = x_i - y_i$$

$$d_i = x_i - y_i - \overline{x} + \overline{y} \qquad\qquad \text{Since } \overline{x} = \overline{y}$$

$$d_i = (x_i - \overline{x}) - (y_i - \overline{y})$$

Squaring and summing d_i^2 over i = 1 to n, we have

$$\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} \{(x_i - \overline{x}) - (y_i - \overline{y})\}^2$$

$$\Rightarrow \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} \{(x_i - \overline{x})^2 + (y_i - \overline{y})^2 - 2(x_i - \overline{x})(y_i - \overline{y})\}$$

$$\Rightarrow \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (x_i - \overline{x})^2 + \sum_{i=1}^{n} (y_i - \overline{y})^2 - 2\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) \qquad \ldots (3)$$

Dividing equation (3) by n, we have

$$\frac{1}{n}\sum_{i=1}^{n} d_i^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^2 + \frac{1}{n}\sum_{i=1}^{n} (y_i - \overline{y})^2 - 2\frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

$$\Rightarrow \frac{1}{n}\sum_{i=1}^{n} d_i^2 = \sigma_x^2 + \sigma_y^2 - 2\text{Cov}(x, y) \qquad \ldots (4)$$

We know that, $r = \dfrac{\text{Cov}(x, y)}{\sigma_x \sigma_y}$, which implies that $\text{Cov}(x, y) = r\sigma_x\sigma_y$.

Substituting $\text{Cov}(x, y) = r\sigma_x\sigma_y$ in equation (4), we have

$$\frac{1}{n}\sum_{i=1}^{n} d_i^2 = \sigma_x^2 + \sigma_y^2 - 2r\sigma_x\sigma_y$$

Since, $\sigma_x^2 = \sigma_y^2$, then

$$\frac{1}{n}\sum_{i=1}^{n} d_i^2 = \sigma_x^2 + \sigma_x^2 - 2r\sigma_x\sigma_x$$

$$\frac{1}{n}\sum_{i=1}^{n} d_i^2 = 2\sigma_x^2 - 2r\sigma_x^2$$

$$\frac{1}{n}\sum_{i=1}^{n} d_i^2 = 2\sigma_x^2(1 - r)$$

$$\frac{1}{2n\sigma_x^2}\sum_{i=1}^{n} d_i^2 = (1 - r)$$

$$r = 1 - \frac{\sum_{i=1}^{n} d_i^2}{2n\sigma_x^2}$$

$$r = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \qquad (\text{Since } \sigma_x{}^2 = \frac{n^2 - 1}{12})$$

We denote rank correlation coefficient by r_s, and hence

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \qquad \ldots (5)$$

This formula was given by Spearman and hence it is known as Spearman's rank correlation coefficient formula.

Let us discuss some problems on rank correlation coefficient.

Example 1: Suppose we have ranks of 8 students of B.Sc. in Statistics and Mathematics. On the basis of rank we would like to know that to what extent the knowledge of the student in Statistics and Mathematics is related.

Rank in Statistics	1	2	3	4	5	6	7	8
Rank in Mathematics	2	4	1	5	3	8	7	6

Solution: Spearman's rank correlation coefficient formula is

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

Let us denote the rank of students in Statistics by R_x and rank in Mathematics by R_y. For the calculation of rank correlation coefficient we have to find $\sum_{i=1}^{n} d_i^2$ which is obtained through the following table:

Rank in Statistics (R_x)	**Rank in Mathematics** (R_y)	**Difference of Ranks** $(d_i = R_x - R_y)$	d_i^2
1	2	−1	1
2	4	−2	4
3	1	2	4
4	5	−1	1
5	3	2	4
6	8	−2	4
7	7	0	0
8	6	2	4
			$\sum d_i^2 = 22$

Here, n = number of paired observations = 8

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)} = 1 - \frac{6 \times 22}{8 \times 63} = 1 - \frac{132}{504} = \frac{372}{504} = 0.74$$

Thus there is a positive association between ranks of Statistics and Mathematics.

Example 2: Suppose we have ranks of 5 students in three subjects Computer, Physics and Statistics and we want to test which two subjects have the same trend.

Rank in Computer	2	4	5	1	3
Rank in Physics	5	1	2	3	4
Rank in Statistics	2	3	5	4	1

Solution: In this problem, we want to see which two subjects have same trend i.e. which two subjects have the positive rank correlation coefficient.

Here we have to calculate three rank correlation coefficients

$r_{12s} =$ Rank correlation coefficient between the ranks of Computer and Physics

$r_{23s} =$ Rank correlation coefficient between the ranks of Physics and Statistics

$r_{13s} =$ Rank correlation coefficient between the ranks of Computer and Statistics

Let R_1, R_2 and R_3 be the ranks of students in Computer, Physics and Statistics respectively.

Rank in Computer (R_1)	**Rank in Physics (R_2)**	**Rank in Statistics (R_3)**	**$d_{12} =$ $R_1 - R_2$**	**d_{12}^2**	**$d_{23} =$ $R_2 - R_3$**	**d_{23}^2**	**$d_{13} =$ $R_1 - R_3$**	**d_{13}^2**
2	5	2	−3	9	3	9	0	0
4	1	3	3	9	−2	4	1	1
5	2	5	3	9	−3	9	0	0
1	3	4	−2	4	−1	1	−3	9
3	4	1	−1	1	−3	9	2	4
Total				32		32		14

Thus,

$$\sum d_{12}^2 = 32, \ \sum d_{23}^2 = 32 \ \text{ and } \ \sum d_{13}^2 = 14.$$

Now

$$r_{12s} = 1 - \frac{6\sum d_{12}^2}{n(n^2-1)} = 1 - \frac{6 \times 32}{5 \times 24} = 1 - \frac{8}{5} = -\frac{3}{5} = -0.6$$

$$r_{23s} = 1 - \frac{6\sum d_{23}^2}{n(n^2-1)} = 1 - \frac{6 \times 32}{5 \times 24} = 1 - \frac{8}{5} = -\frac{3}{5} = -0.6$$

$$r_{13s} = 1 - \frac{6\sum d_{13}^2}{n(n^2-1)} = 1 - \frac{6 \times 14}{5 \times 24} = 1 - \frac{7}{10} = \frac{3}{10} = 0.3$$

r_{12s} is negative which indicates that Computer and Physics have opposite trend. Similarly, negative rank correlation coefficient r_{23s} shows the opposite trend in Physics and Statistics. $r_{13s} = 0.3$ indicates that Computer and Statistics have same trend.

Sometimes we do not have rank but actual values of variables are available. If we are interested in rank correlation coefficient, we find ranks from the given values. Considering this case we are taking a problem and try to solve it.

Example 3: Calculate rank correlation coefficient from the following data:

x	78	89	97	69	59	79	68
y	125	137	156	112	107	136	124

Solution: We have some calculation in the following table:

x	**y**	**Rank of x (R_x)**	**Rank of y (R_y)**	**d = R_x - R_y**	**d^2**
78	125	4	4	0	0
89	137	2	2	0	0
97	156	1	1	0	0
69	112	5	6	-1	1
59	107	7	7	0	0
79	136	3	3	0	0
68	124	6	5	1	1
					$\sum_{i=1}^{n} d_i^2 = 2$

Spearman's Rank correlation formula is

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

$$r_s = 1 - \frac{6 \times 2}{7(49 - 1)} = 1 - \frac{12}{7 \times 48}$$

$$= 1 - \frac{1}{28} = \frac{27}{28} = 0.96$$

Now, let us solve a little exercise.

E1) Calculate Spearman's rank correlation coefficient from the following data:

x	20	38	30	40	50	55
y	17	45	30	35	40	25

7.3.1 Merits and Demerits of Rank Correlation Coefficient

Merits of Rank Correlation Coefficient

1. Spearman's rank correlation coefficient can be interpreted in the same way as the Karl Pearson's correlation coefficient;
2. It is easy to understand and easy to calculate;

3. If we want to see the association between qualitative characteristics, rank correlation coefficient is the only formula;
4. Rank correlation coefficient is the non-parametric version of the Karl Pearson's product moment correlation coefficient; and
5. It does not require the assumption of the normality of the population from which the sample observations are taken.

Demerits of Rank Correlation Coefficient

1. Product moment correlation coefficient can be calculated for bivariate frequency distribution but rank correlation coefficient cannot be calculated; and
2. If n >30, this formula is time consuming.

7.4 TIED OR REPEATED RANKS

In Section 7.3, it was assumed that two or more individuals or units do not have same rank. But there might be a situation when two or more individuals have same rank in one or both characteristics, then this situation is said to be tied.

If two or more individuals have same value, in this case common ranks are assigned to the repeated items. This common rank is the average of ranks they would have received if there were no repetition. For example we have a series 50, 70, 80, 80, 85, 90 then 1st rank is assigned to 90 because it is the biggest value then 2nd to 85, now there is a repetition of 80 twice. Since both values are same so the same rank will be assigned which would be average of the ranks that we would have assigned if there were no repetition. Thus, both 80 will receive the average of 3 and 4 i.e. (Average of 3 & 4 i.e. (3 + 4) / 2= 3.5) 3.5 then 5th rank is given to 70 and 6th rank to 50. Thus, the series and ranks of items are

Series	50	70	80	80	85	90
Ranks	6	5	3.5	3.5	2	1

In the above example 80 was repeated twice. It may also happen that two or more values are repeated twice or more than that.

For example, in the following series there is a repetition of 80 and 110. You observe the values, assign ranks and check with following.

Series	50	70	80	90	80	120	110	110	110	100
Ranks	10	9	7.5	6	7.5	1	3	3	3	5

When there is a repetition of ranks, a correction factor $\frac{m(m^2-1)}{12}$ is added to $\sum d^2$ in the Spearman's rank correlation coefficient formula, where m is the number of times a rank is repeated. It is very important to know that this correction factor is added for every repetition of rank in both characters.

In the first example correction factor is added once which is 2(4 - 1)/12 = 0.5, while in the second example correction factors are 2(4 - 1)/12 = 0.5 and

3 (9 - 1)/12 = 2 which are aided to $\sum d^2$.

Thus, in case of tied or repeated rank Spearman's rank correlation coefficient formula is

$$r_s = 1 - \frac{6\left\{\sum d^2 + \frac{m(m^2-1)}{12} + ...\right\}}{n(n^2-1)}$$

Example 4: Calculate rank correlation coefficient from the following data:

Expenditure on advertisement	10	15	14	25	14	14	20	22
Profit	6	25	12	18	25	40	10	7

Solution: Let us denote the expenditure on advertisement by x and profit by y

x	**Rank of x (R_x)**	**y**	**Rank of y (R_y)**	**d = R_x - R_y**	**d^2**
10	8	6	8	0	0
15	4	25	2.5	1.5	2.25
14	6	12	5	1	1
25	1	18	4	−3	9
14	6	25	2.5	3.5	12.25
14	6	40	1	5	25
20	3	10	6	−3	9
22	2	7	7	−5	25
					$\sum d^2 = 83.50$

$$r_s = 1 - \frac{6\left\{\sum d^2 + \frac{m(m^2-1)}{12} + ...\right\}}{n(n^2-1)}$$

Here rank 6 is repeated three times in rank of x and rank 2.5 is repeated twice in rank of y, so the correction factor is

$$\frac{3(3^2-1)}{12} + \frac{2(2^2-1)}{12}$$

Hence rank correlation coefficient is

$$r_s = 1 - \frac{6\left\{83.50 + \frac{3(3^2-1)}{12} + \frac{2(2^2-1)}{12}\right\}}{8(64-1)}$$

$$r_s = 1 - \frac{6\left\{83.50 + \frac{3\times 8}{12} + \frac{2\times\ 3}{12}\right\}}{8\,X\,63}$$

$$r_s = 1 - \frac{6(83.50 + 2.50)}{504}$$

$$r_s = 1 - \frac{516}{504}$$

$$r_s = 1 - 1.024 = -\ 0.024$$

There is a negative association between expenditure on advertisement and profit.

Now, let us solve the following exercises.

E2) Calculate rank correlation coefficient from the following data:

x	10	20	30	30	40	45	50
y	15	20	25	30	40	40	40

E3) Calculate rank correlation coefficient from the following data:

x	70	70	80	80	80	90	100
y	90	90	90	80	70	60	50

7.5 CONCURRENT DEVIATION

Sometimes we are not interested in the actual amount of correlation coefficient but we want to know the direction of change i.e. whether correlation is positive or negative, coefficient of concurrent deviation serves our purpose. In this method correlation is calculated between the direction of deviations and not their magnitudes. Coefficient of concurrent deviation is denoted by r_c and given by

$$r_c = \pm\sqrt{\pm\frac{(2c-k)}{k}} \qquad \ldots (6)$$

where, c is the number of concurrent deviation or the number of + sign in the product of two deviations, $k = n - 1$ i.e. total number of paired observation minus one. This is also called coefficient of correlation by concurrent deviation method. Steps for the calculation of concurrent deviation (see the Example 5 simultaneously) are:

1. The first value of series x is taken as a base and it is compared with next value i.e. second value of series x. If second value is greater than first value, ‘+’ sign is assigned in the Column titled D_x. If second value is less than the first value then ‘-’ sign is assigned in the column D_x.
2. If first and second values are equal then ‘=’ sign is assigned.
3. Now second value is taken as base and it is compared with the third value of the series. If third value is less than second ‘-’ is assigned against the

third value. If the third value is greater than the second value '+' is assigned. If second and third values or equal than '=' sign is assigned.

4. This procedure is repeated upto the last value of the series.
5. Similarly, we obtain column D_y for series y.
6. We multiply the column D_x and D_y and obtain column $D_x D_y$. Multiplication of same sign results '+'sign and that of different sign is '-' sign.
7. Finally number of '+' sign are counted in the column $D_x D_y$, it is called c and we get coefficient concurrent deviation by the formula (6).
8. In the formula, inside and outside the square root, sign '+' and '-' depends on the value of $(2c-k)$. If this value is positive than '+' sign is taken at both places if $(2c-k)$ is negative '-'sign is considered at both the places.

Let us discuss the following problem.

Example 5: We have data of income and expenditure of 11 workers of an organization in the following table:

Income	65	40	35	75	63	79	35	20	80	60	50
Expenditure	60	55	50	66	30	71	40	35	80	75	80

Find whether correlation is positive or negative by coefficient of concurrent deviation.

Solution: Coefficient of concurrent deviation is given by

$$r_c = \pm\sqrt{\pm\frac{(2c-k)}{k}}$$

Let us denote the income by x and expenditure by y and we calculate c by the following table:

x	Change of direction sign for x (D_x)	y	Change of direction sign for y (D_y)	$D_x D_y$
65		60		
40	−	55	−	+
35	−	50	−	+
75	+	66	+	+
63	−	30	−	+
79	+	71	+	+
35	−	40	−	+
20	−	35	−	+
80	+	81	+	+
60	−	75	−	+
50	−	80	+	−
				c = 9

Here, c = 9 and k = n −1 = 10 then we have

$$r = \pm\sqrt{\pm\frac{2\times9-10}{10}} \qquad = +\sqrt{+\frac{8}{10}}$$

(Both signs are + because 2c − k is positive)

$$= \sqrt{0.8} = 0.89$$

Thus correlation is positive.

Now, let us solve the following exercises.

E 4) Find the coefficient of correlation between supply and price by concurrent deviation method for the following data:

Year	1970	1971	1972	1973	1974	1975
Supply	114	127	128	121	120	124
Price	108	104	105	106	100	99

E5) Calculate coefficient of concurrent deviation for the following data:

x	368	384	385	360	347	384
y	122	121	124	125	122	126

7.6 SUMMARY

In this unit, we have discussed:

1. The rank correlation which is used to see the association between two qualitative characteristics;
2. Derivation of the Spearman's rank correlation coefficient formula;
3. Calculation of rank correlation coefficient in different situations- (i) when values of variables are given, (ii) when ranks of individuals in different characteristics are given and (iii) when repeated ranks are given;
4. Properties of rank correlation coefficient; and
5. Concurrent deviation which provides the direction of correlation.

7.7 SOLUTIONS /ANSWERS

E1) We have some calculations in the following table:

x	**y**	**Rank of x (R_x)**	**Rank of y (R_y)**	**d = R_x - R_y**	**d^2**
20	17	6	6	0	0
38	45	4	1	3	9
30	30	5	4	1	1
40	35	3	3	0	0
50	40	2	2	0	0
55	25	1	5	−4	16
					$\sum_{i=1}^{n} d_i^2 = 26$

$$r_s = 1 - \frac{6 \times 26}{6(36-1)}$$

$$= 1 - \frac{26}{35} = \frac{9}{35} = 0.26$$

E2) We have some calculations in the following table:

x	Rank of x (R_x)	y	Rank of y (R_y)	d = R_x - R_y	d^2
10	7	15	7	0	0
20	6	20	6	0	0
30	4.5	25	5	−0.5	0.25
30	4.5	30	4	0.5	0.25
40	3	40	2	1	1
45	2	40	2	0	0
50	1	40	2	−1	1
					$\sum d^2 = 2.5$

$$r_s = 1 - \frac{6\left\{\sum d^2 + \frac{m(m^2-1)}{12} + ...\right\}}{n(n^2-1)}$$

Here, rank 4.5 is repeated twice in rank of x and rank 2 is repeated thrice in rank of y so the correction factor is

$$\frac{2(2^2-1)}{12} + \frac{3(3^2-1)}{12}$$

and therefore, rank correlation coefficient is

$$r_s = 1 - \frac{6\left\{2.5 + \frac{2(2^2-1)}{12} + \frac{3(3^2-1)}{12}\right\}}{7(49-1)}$$

$$r_s = 1 - \frac{6\left\{2.5 + \frac{2 \times 3}{12} + \frac{3 \times 8}{12}\right\}}{7 \times 48}$$

$$r_s = 1 - \frac{6(2.5 + 2.5)}{336}$$

$$r_s = 1 - \frac{30}{336} = \frac{306}{336}$$

$$r_s = 0.91$$

E3) We have some calculations in the following table:

x	Rank of x (R_x)	y	Rank of y (R_y)	d = R_x - R_y	d^2
70	6.5	90	2	4.5	20.25
70	6.5	90	2	4.5	20.25
80	4	90	2	2	4
80	4	80	4	0	0
80	4	70	5	−1	1
90	2	60	6	−4	16
100	1	50	7	−6	36
					$\sum d^2 = 97.5$

Rank correlation coefficient is

$$r_s = 1 - \frac{6\left\{\sum d^2 + \frac{m(m^2-1)}{12} + \ldots\right\}}{n(n^2-1)}$$

Here, rank 4 and 6.5 is repeated thrice and twice respectively in rank of x and rank 2 is repeated thrice in rank of y, so the correction factor is

$$\frac{3(3^2-1)}{12} + \frac{2(2^2-1)}{12} + \frac{3(3^2-1)}{12}$$

and therefore, rank correlation coefficient is

$$r_s = 1 - \frac{6\left\{97.5 + \frac{3(3^2-1)}{12} + \frac{2(2^2-1)}{12} + \frac{3(3^2-1)}{12}\right\}}{7(49-1)}$$

$$r_s = 1 - \frac{6\{97.5 + 4.5\}}{7 \times 48}$$

$$r_s = 1 - \frac{6(102)}{336}$$

$$r_s = 1 - \frac{102}{56} = -0.82$$

E4) Coefficient of concurrent deviation is given

$$r_c = \pm\sqrt{\pm\frac{(2c-k)}{k}}$$

Let us denote the supply by x and price by y and we calculate c by the following table:

x	Change of Direction sign for x (D_x)	y	Change of Direction sign for y (D_y)	D_xD_y
114		108		
127	+	104	−	−
128	+	105	+	+
121	−	106	+	−
120	−	100	−	+
124	+	99	−	−
				c = 2

Now c = 2 and k = n - 1 = 5

$$r = \pm\sqrt{\pm\frac{2\times2-5}{5}} \quad = -\sqrt{\frac{1}{5}}$$

(Both signs are '-' because 2c − k is negative)

$= -\,0.45$

Thus, correlation is negative.

E5) Coefficient of concurrent deviation is given

$$r_c = \pm\sqrt{\pm\frac{(2c-k)}{k}}$$

Let us denote the supply by x and price by y and we calculate c by the following table:

x	**Change of Direction sign for x (D_x)**	**y**	**Change of Direction sign for y (D_y)**	**D_xD_y**
368		122		
384	+	121	−	−
385	+	124	+	+
360	−	125	+	−
347	−	122	−	+
384	+	126	+	+
				c = 3

Now c = 3 and k = n −1= 5

$$r = \pm\sqrt{\pm\frac{2\times3-5}{5}} \quad = \sqrt{\frac{1}{5}}$$

(Both signs are '+' because 2c−k is positive)

r = 0.45

Thus, correlation is positive.

UNIT 8 INTRA-CLASSES CORRELATION

Structure

8.1 Introduction

Objectives

8.2 Coefficient of Determination

8.3 Correlation Ratio

Characteristics of Correlation Ratio

8.4 Intra-class Correlation Coefficient

Limits of Intra-class Correlation Coefficient

8.5 Summary

8.6 Solutions / Answers

8.1 INTRODUCTION

In Unit 5 of this block, you acquired the knowledge of fitting of various curves including straight line that shows the linear relationship between two variables. This linear relationship is measured by correlation coefficient that gives the strength of linear relationship. When characteristics are not measurable but ranks may be assigned to individuals according to their qualities, rank correlation is used to see the association between characteristics. Sometimes researchers are interested in studying the correlation among the members of the same family. In such cases intra-class correlation is used.

During the investigation more often it is observed that the two variables are not linearly related but they have some other type of curvilinear relationship. In this case correlation coefficient fails to give the strength of relationship and we use correlation ratio.

In this unit, you will study the coefficient of determination, correlation ratio and intra-class correlation. To understand the concept of correlation ratio and intra-class correlation, you are advised go through the concept and properties of correlation coefficient discussed in Unit 6 of this block.

Section 8.2 of this unit describes the coefficient of determination whereas correlation ratio is discussed with its properties in Section 8.3. Intra-class correlation coefficient is explained in Section 8.4. A derivation of limits of intra-class correlation coefficient is given in Sub-section 8.4.1.

Objectives

After reading this unit, you would be able to

- define the coefficient of determination;
- describe the correlation ratio;
- describe properties of correlation ratio;
- define the intra-class correlation coefficient; and
- describe the limits of intra-class correlation coefficient.

8.2 COEFFICIENT OF DETERMINATION

A measure which is used in statistical model analysis to assess how well a model explains and predicts future outcomes is known as coefficient of determination. The coefficient of determination is the measure of variation in the dependent variable that is explained by the regression function. In other words, the coefficient of determination indicates how much of the total variation in the dependent variable can be accounted by the regression function. This is the square of the correlation coefficient and denoted by r^2. It is expressed as a percentage. If coefficient of determination is 0.64, it implies that 64% of the variation in dependent variable Y is by the regression function or explained by the independent variable. Since it is the square of the correlation coefficient it ranges from 0 to 1.

The value of $r^2 = 0$ normal indicate that the dependent variable cannot be predicted from the independent variable, whereas the value of $r^2 =1$ is the indication that the dependent variable can be predicted from the independent variable without error.

In analysis of variance, r^2 is the proportion of the total sum of squares which has been explained by linear regression. $(1- r^2)$ is called the coefficient of non-determination.

Now solve the following exercises.

E1) If $r^2 = 0.75$, interpret the result.

E2) What would you conclude if $r^2 = 1$?

E3) What is the interpretation of $r^2 = 0$?

8.3 CORRELATION RATIO

Correlation coefficient measures the intensity or degree of linear relationship between two variables i.e. the extent of linear relationship can be explained by correlation coefficient if two variables are linearly related. If variables are not linearly related and show some curvilinear relationship then correlation coefficient is not a suitable measure to show the extent of relationship. In this type of cases, we study correlation ratio which is appropriate tool to know the degree of relationship between two variables i.e. concentration of points around the curve of best fit.

Correlation ratio is also used to measure the degree of association between a quantitative variable and another variable which may be qualitative or quantitative. Correlation ratio is determined only by the observations of dependent variable. Since quantitative scale for independent variable is not necessary so it is used only to classify dependent variable.

When regression is linear, the correlation coefficient and correlation ratio both produce the same results i.e. $r = \eta$, where η is correlation ratio.

So far we were dealing the situations where there was single value of Y corresponding to any value of X for example, data in the form

x	2	3	4	5
y	4	7	9	7

But in practice, there might be a situation where we have more than one values of y for each value of x. For example heights of 20 sons according to height of their fathers are given below:

Height of Fathers (in inches)	Height of sons (in inches)				
65	66	66	67	68	65
68	68	69	69	72	70
70	70	72	73	74	73
72	74	75	73	74	75

If we consider father's height by X and son's height by Y, in the above example more than one values of Y are available for each value of X.

In general X and Y may be in the following form

x_1: y_{11}, $y_{12}, \ldots, y_{1j}, \ldots, y_{1n}$

x_2 : y_{21}, $y_{22}, \ldots, y_{2j}, \ldots, y_{2n}$

.

x_i : y_{i1}, $y_{i2}, \ldots, y_{ij}, \ldots, y_{in}$

.

x_m : y_{m1}, $y_{m2}, \ldots, y_{mj}, \ldots, y_{mn}$

Let us suppose that for each value of $x_i \, (i = 1, 2, \ldots, m)$, variable Y has n values $y_{ij}(i = 1, 2, \ldots, m; j = 1, 2, \ldots, n)$ then mean of variable Y for i^th array is defined as

$$\overline{y}_i = \frac{1}{n}\sum_{j=1}^{n} y_{ij}$$

Then correlation ratio η is obtained by

$$\eta_{yx}^2 = \frac{\sum_{i=1}^{m} n_i(\overline{y}_i - \overline{y})^2}{\sum_{i=1}^{m}\sum_{j=1}^{n}(y_{ij} - \overline{y})^2}$$

Now we present the above distribution with frequencies i.e.

x_1 $y_{11}\,f_{11}$, $y_{12}\,f_{12}, \ldots, y_{1j}f_{1j}, \ldots, y_{1n}f_{1n}$

x_2 $y_{21}f_{21}$, $y_{22}f_{22}, \ldots, y_{2j}f_{2j}, \ldots, y_{2n}f_{2n}$

.

x_i $y_{i1}f_{i1}$, $y_{i2}f_{i2}, \ldots, y_{ij}f_{ij}, \ldots, y_{in}f_{in}$

.

x_m $y_{m1}f_{m1}$, $y_{m2}f_{m2}, \ldots, y_{mj}f_{mj}, \ldots, y_{mn}f_{mn}$

It means for $x_i (i = 1, 2, \ldots, m)$, Y takes values $y_{ij}(j = 1, 2, \ldots, n)$ and frequency f_{ij} is attached with y_{ij}.

Note: You might have studied the frequency distribution earlier. Frequency is the number of repetitions of a value. If in a series of data, 2 is repeated 5 times then we say frequency of 2 is 5. And frequency distribution is the arrangement of values of variable with its frequencies.

In above frequency distribution $\sum_{j=1}^{n} f_{ij} = n_i$ and $\sum_{i=1}^{m} f_{ij} = n_j$

Total frequency $N = \sum_{i}^{m} n_i = \sum_{j}^{n} n_j$ i^{th} row can be considered as i^{th}array.

Then, mean of the i^{th} array can be defined as

$$\bar{y}_i = \frac{\sum_{j=1}^{n} f_{ij} y_{ij}}{\sum_{j=1}^{n} f_{ij}} = \frac{\sum_{j=1}^{n} f_{ij} y_{ij}}{n_i} = \frac{T_i}{n_i}$$

and over all mean

$$\bar{y} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} y_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}} = \frac{\sum_{i=1}^{m} \bar{y}_i n_i}{\sum_{i=1}^{m} n_i} = \frac{T}{N}$$

Then, correlation ratio of Y on X is denoted by η_{yx} and defined as:

$$\eta_{yx}^2 = 1 - \frac{\sigma_e^2}{\sigma_y^2}$$

where,

$$\sigma_e^2 = \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}(y_{ij} - \bar{y}_i)^2$$

$$\sigma_y^2 = \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}(y_{ij} - \bar{y})^2$$

On simplifying, we get

$$\eta_{yx}^2 = \frac{\sum_{i=1}^{m} n_i(\bar{y}_i - \bar{y})^2}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}(y_i - \bar{y})^2}$$

Some more simplification gives

$$\eta_{yx}^2 = \left[\sum_{i=1}^{m}\left(\frac{T_i^2}{n_i}\right) - \frac{T^2}{N}\right] \Big/ N\sigma_y^2$$

Example 1: Compute η_{yx}^2 for the following table:

x y	47	52	57	62	67
57	4	4	2	0	0
62	4	8	8	1	0
67	0	7	12	1	4
72	0	3	1	8	5
77	0	0	3	5	6

Solution: It is known that

$$\eta_{yx}^2 = \left[\sum_{i=1}^{m}\left(\frac{T_i^2}{n_i}\right) - \frac{T^2}{N}\right] / N\sigma_y^2$$

X / Y	47	52	57	62	67	$f_{i.}$	$f_{ij}\,y_{ij}^2$
57	4	4	2	0	0	10	32490
62	4	8	8	1	0	21	80724
67	0	7	12	1	4	24	107736
72	0	3	1	8	5	17	88128
77	0	0	3	5	6	14	83006
n_i	8	22	26	15	15	N = 86	$\sum_{i=1}^{m} f_{ij} y_{ij}^2 =$ 392084
T_i	476	1406	1717	1090	1090	$T = \sum_{i=1}^{m} T_i$ $= 5782$	$\bar{y} = \frac{T}{N} =$ $\frac{5782}{86} = 67.23$
$\frac{T_i^2}{n_i}$	15705.07	48215.51	60165.08	44003.70	44003.70	$\sum_{i=1}^{m} \frac{T_i^2}{n_i} =$ 211493.07	
$\bar{y}_i$	59.50	64.05	66.04	72.67	72.67		

$$\sigma_y^2 = \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}(y_{ij} - \bar{y})^2$$

$$\sigma_y^2 = \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} y_{ij}^2 - \bar{y}^2$$

$$= \frac{392084}{86} - (67.23)^2 = 38.90$$

So,

$$\eta_{yx}^2 = \left[\sum_{i=1}^{m}\left(\frac{T_i^2}{n_i}\right) - \frac{T^2}{N}\right] / N\sigma_y^2$$

$$= \frac{211493.07 - (5782)^2/86}{86 \times 38.90}$$

$$= \frac{1624.76}{86 \times 38.90} = 0.485$$

E1) Compute η_{yx}^2 for the following table:

x / y	5	10	15	20	25
5	8	8	4	0	0
10	7	15	15	1	0
20	0	6	1	15	11
25	0	0	5	10	8

8.3.1 Characteristics of Correlation Ratio

1. Absolute value of correlation ratio is always less than or equal to one.
2. Absolute value of correlation ratio can not be less than the value of correlation coefficient.
3. η_{YX} is independent of change of origin and scale.
4. Correlation coefficient of X on Y i.e. r_{XY} and correlation coefficient of Y on X, r_{YX} are equal but correlation ratio of Y on X, η_{yx} and correlation ratio of X on Y, η_{xy} are different.
5. Difference ($\eta_{yx}^2 - r^2$) measures the extent to which the true regression of Y on X departs from linearity.
6. It is based on only the values of dependent variable. Auxiliary variable is used just for classify the observations of independent variable.
7. When correlation is linear and forming a straight line then $\eta_{yx}^2 = r^2$ are same.
8. When scatter diagram does not show any trend then both η_{yx} and r are zero.
9. If scatter diagram shows straight line and dots lie precisely on a line, then the correlation coefficient and the correlation ratio, both are 1, i.e. $\eta_{yx} = r = 1$.
10. If $\eta_{yx} > r$ then, scatter diagram shows curved trend line.

8.4 INTRA-CLASS CORRELATION COEFFICIENT

In product moment correlation, both variables measure different characteristics e.g. one variable is price and another is demand. Similarly one variable is expenditure on advertisement and another may be sales revenue. But in many practical situations specially in medical, agriculture and biological field one may be interested to know the correlation among the units of a group or a family. For example, in agricultural experiment scientist may be interested to know the correlation among the yields of the plots of the block given same fertilizer.

In the study of heights of brothers, one may be interested in correlation between the heights of brothers of a family. In such correlation both variables measure the same characteristics, i.e. yield and height. By this correlation, we mean the extent to which the members of the same family resemble each other with respect to the considered characteristic. This correlation is called intra-class correlation.

Let us consider n families $F_1, F_2, ..., F_n$ with number of members $k_1, k_2, ..., k_n$ respectively. Let $x_{ij}\ (i = 1, 2, ..., n;\ j = 1, 2, ..., k_i)$ be the value of j^{th} member of the i^{th} family.

In i^{th} family there would be $k_i(k_i-1)$ pairs of observation. So the total number of pairs are $\sum_{i=1}^{n} k_i(k_i-1) = N$.

Note: If in a family there are three members 1, 2 and 3 then there would be (1, 2) (2, 1) (1, 3) (3, 1) (2, 3) and (3, 2) i.e. 6 pairs of observation, here number k = 3 so k (k - 1) = 3(3 - 1) =3×2 = 6 pairs.

Table giving the values of N pairs of observations is called intra-class correlation table and the product moment correlation coefficient calculated from $\sum_{i=1}^{n} k_i(k_i-1) = N$ pairs of observations is called intra-class correlation coefficient. Since the value of each member of the i^{th} family occurs (k_i-1) times as a value of the X variable as well as a value of the Y variable. Then mean of variable X and Y are same and

$$\bar{x} = \bar{y} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{k_i}(k_i-1)x_{ij}}{\sum_{i=1}^{n} k_i(k_i-1)} = \frac{1}{N}\left[\sum_{i=1}^{n}(k_i-1)\sum_{j=1}^{k_i} x_{ij}\right]$$

Similarly, both variable X and Y have same variance which is

$$\sigma_x^2 = \sigma_y^2 = \frac{1}{N}\sum_{i=1}^{n}(k_i-1)\sum_{j=1}^{k_i}(x_{ij}-\bar{x})^2 ,$$

and covariance between X and Y is

$$\text{Cov}(x,y) = \frac{1}{N}\sum_{i}\sum_{j\neq 1} (x_{ij}-\bar{x})(x_{il}-\bar{x})$$

Finally, formula for intra-class correlation coefficient is

$$r_{ic} = \frac{\sum_{i=1}^{n} k_i^2(\bar{x}_i-\bar{x})^2 - \sum_{i=1}^{n}\sum_{j=1}^{n}(x_{ij}-\bar{x})^2}{\sum_{i=1}^{n}\sum_{j=1}^{k_i}(k_i-1)(x_{ij}-\bar{x})^2}$$

If $k_i = k$ i.e. each family have equal number of members then

$$r_{ic} = \frac{nk^2\sigma_m^2 - nk\sigma_x{}^2}{(k-1)nk\sigma_x{}^2}$$

$$r_{ic} = \frac{1}{(k-1)}\left\{\frac{k\sigma_m^2}{\sigma_x{}^2}-1\right\}$$

where, $\sigma_x{}^2$ is the variance of X and σ_m^2 is the variance of means of families.

8.4.1 Limits of Intra-class Correlation Coefficient

Intra-class correlation coefficient is

$$r_c = \frac{1}{(k-1)}\left\{\frac{k\sigma_m^2}{\sigma_x{}^2}-1\right\}$$

$$\Rightarrow r_c(k-1) = \left\{\frac{k\sigma_m^2}{\sigma_x{}^2}-1\right\}$$

$$\Rightarrow r_c(k-1)+1=\frac{k\sigma_m^2}{\sigma_x^2}\geq 0$$

$$\Rightarrow r_c \geq -\frac{1}{k-1}$$

Thus, lower limit of intra-class correlation coefficient is $-\dfrac{1}{k-1}$

also

$$1+(k-1)\,r_c \leq k \quad \text{as} \quad \frac{\sigma_m^2}{\sigma_x^2}\leq 1$$

$$r_c \leq 1$$

Thus, $\quad -\dfrac{1}{k\ -1} \leq\ r_c\ \leq 1$

8.5 SUMMARY

In this unit, we have discussed:

1. The coefficient of determination;
2. Properties of the coefficient of determination;
3. Concept of correlation ratio;
4. Properties of correlation ratio; and
5. The intra-class correlation coefficient.

8.6 SOLUTIONS / ANSWERS

E1) It is known that

$$\eta_{yx}^2=\left[\sum_{i=1}^{m}\left(\frac{T_i^2}{n_i}\right)-\frac{T^2}{N}\right]/\,N\sigma_y^2$$

Y \ X	5	10	15	20	25	$f_{i.}$	$f_{ij}\,y_{ij}^2$
5	8	8	4	0	0	20	500
10	7	15	15	1	0	38	3800
15	0	6	1	15	11	45	10125
20	0	0	5	10	8	33	13200
25	5	10	15	20	25	23	14375
n_i	15	41	49	27	27	$\sum_{i=1}^{m} n_i = N$ $=159$	$\sum_{i=1}^{m} f_{ij}y_{ij}^2 = 42000$
T_i	110	490	675	575	540	$T=\sum_{i=1}^{m} T_i = 2390$	$\bar{y}=\frac{T}{N}=\frac{2390}{159}$ $= 264.1509$
$\frac{T_i^2}{n_i}$	806.66	5856.1	9298.47	12245.37	10800	$\sum_{i=1}^{m}\frac{T_i^2}{n_i}=3900660$	
$\bar{y}_i$	7.33	11.9	13.77	21.296	20.00		

$$\sigma_y^2 = \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}(y_{ij} - \overline{y})^2$$

$$\sigma_y^2 = \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} y_{ij}^2 - \overline{y}^{\ 2}$$

$$= \frac{42000}{159} - (246.1509)^2$$

$$= 38.2066$$

So,

$$\eta_{yx}^2 = \left[\sum_{i=1}^{m}\left(\frac{T_i^{\ 2}}{n_i}\right) - \frac{T^2}{N}\right] / N\sigma_y^2$$

$$= \frac{39006.6040 - (2390)^2/159}{159 \times 38.2066}$$

$$= 0.9211$$

Indira Gandhi
National Open University
School of Sciences

MST-002
DESCRIPTIVE STATISTICS

Block

3

REGRESSION AND MULTIPLE CORRELATION

UNIT 9
Linear Regression 5

UNIT 10
Plane of Regression 23

UNIT 11
Multiple Correlation 37

UNIT 12
Partial Correlation 55

Curriculum and Course Design Committee

Prof. K. R. Srivathasan
Pro-Vice Chancellor
IGNOU, New Delhi

Prof. Parvin Sinclair
Pro-Vice Chancellor
IGNOU, New Delhi

Prof. Geeta Kaicker
Director, School of Sciences
IGNOU, New Delhi

Prof. Jagdish Prasad
Department of Statistics
University of Rajasthan, Jaipur

Prof. R. M. Pandey
Department of Bio-Statistics
All India Institute of Medical Sciences
New Delhi

Prof. Rahul Roy
Math. and Stat. Unit
Indian Statistical Institute, New Delhi

Dr. Diwakar Shukla
Department of Mathematics and Statistics
Dr Hari Singh Gaur University, Sagar

Prof. Rakesh Srivastava
Department of Statistics
M. S. University of Baroda, Vadodara

Prof. G. N. Singh
Department of Applied Mathematics
I. S. M. Dhanbad

Dr. Gulshan Lal Taneja
Department of Mathematics
M. D. University, Rohtak

Faculty members of School of Sciences, IGNOU

Statistics
Dr. Neha Garg
Dr. Nitin Gupta
Mr. Rajesh Kaliraman
Dr. Manish Trivedi

Mathematics
Dr. Deepika Garg
Prof. Poornima Mital
Prof. Sujatha Varma
Dr. S. Venkataraman

Block Preparation Team

Content Editor
Dr. Meenakshi Srivastava
Department of Statistics
Institute of Social Sciences
Dr. B. R. Ambedkar University, Agra

Language Editor
Dr. Nandini Sahu
School of Humanities, IGNOU

Secretarial Support
Mr. Deepak Singh

Course Writer
Dr. Rajesh Tailor
School of Studies in Statistics
Vikram University, Ujjain

Formatted By
Dr. Manish Trivedi
Mr. Prabhat Kumar Sangal
Dr. Neha Garg
School of Sciences, IGNOU

Programme and Course Coordinator: Dr. Manish Trivedi

Block Production

Mr. Y. N. Sharma, SO (P.)
School of Sciences, IGNOU

Acknowledgement: We gratefully acknowledge to Prof. Geeta Kaicker, Director, School of Sciences for her great support and guidance.

December, 2011

ISBN-978-81-266-

Further information on the Indira Gandhi National Open University may be obtained from University's Office at Maidan Garhi, New Delhi-110068 or visit University's website http://www.ignou.ac.in

Printed and published on behalf of the Indira Gandhi National Open University, New Delhi by the Director, School of Sciences.

Laser Typeset by : Tessa Media & Computers, C-206, A.F.E.-II, Okhla, New Delhi

Printed at:

REGRESSION AND MULTIPLE CORRELATION

In Block 2 of this course, you have studied different methods including curve fitting using principle of least-squares, correlation, correlation coefficient, rank correlation coefficient and intra-class correlation coefficient.

Unit 9 of this block discusses the linear regression which considers two variables, one as a dependent variable and another one as an independent variable. It describes the two regression lines, regression coefficient and draws distinction between correlation and regression. Angle between two regression lines is also discussed.

If the informations on three or more variables are available and someone is interested to study the relation between them, then one has to use the concepts of plane of regression. Unit 10 provides the basic concepts about the plane of regression considering only three variables. In this unit, you will also learn Yule's notation and concept as well as the properties and variance of residuals.

Unit 11 describes the multiple correlation coefficients with its properties. Correlation coefficient measures the linear strength of association between two variables whereas multiple correlation coefficients measure the strength of association between a dependent variable and joint effect of two or more independent variables.

Sometimes one considers the correlation between dependent variable and one independent variable while ignoring the effect of others when we are considering two or more variables. This type of situation is handled by partial correlation. Unit 12 is mainly focused on partial correlation coefficient. Multiple correlation coefficients are also explained in terms of partial and total correlation coefficient in this unit.

Suggested Readings:

- Goon, A. M., Gupta, M. K. and Das Gupta, B.; Fundamentals of Statistics Vol-I; World Press, Culcutta.
- Gupta, M. P. and Gupta, S. P.; Business Statistics; Sultan Chand & Sons Publications.
- Gupta S. C. and Kapoor, V. K.; Fundamentals of Mathematical Statistics, Sultan Chand & Sons Publications.
- Ray, N. and Sharma, H. S.; Mathematical Statistics, Ram Prasad & Sons, Agra, 7th edn., 1983
- Sancheti, D. C. and Kapoor, V. K.; Statistics, Sultanchand & Sons, New Delhi, 7th edn., 1991
- Varshney, R. P.; Advanced Statistics, Jawahar Publication, Agra, 28th edn., 2003-2004

Notations and Symbols

:

y_i	: Observed values of y $(i=1,2,3,...,n)$
Y_i	: Expected values of y $(i=1,2,3,...,n)$
U	: Sum of squares of errors
$\frac{\partial U}{\partial a}$	: Partial derivatives of U with respect to 'a
$\mathrm{Cov}(x, y)$	: Covariance between x and y
$V(x)=\sigma_x^2$	: Variance of x
r = Corr (x, y)	: Correlation coefficient between x and y
b_{yx}	: Regression coefficient of y on x
b_{xy}	: Regression coefficient of x on y θ
θ_1	: Acute angle
θ_2	: Obtuse angle
$b_{12.3}$	: Partial regression coefficient of $x_1$on x_2
$b_{13.2}$	: Partial regression coefficient of $x_1$on x_3
$x_{1.23}$	: Estimate of x_1
$e_{1.23}$	: Error of estimate or residual
r_{ij}	: Correlation coefficient between x_i and x_j
σ_i^2	: Variance of x_i
Cov ($x_{i,}$ x_j)	: Covariance between x_i and x_j
W_{ij}	: Co-factor of the element in the i^{th} row and j^{th} column of matrix W
$\sigma_{1.23}^2$	: Variance of residual $e_{1.23}$
$R_{1.23}$	: Multiple correlation coefficient x_1 on x_2 and x_3
$r_{12.3}$	: Partial correlation coefficient between x_1 and x_2
$e_{1.3}$	: Residual for x_1 and x_3
$e_{2.3}$	: Residual for x_2 and x_3
$V(e_{1.3})=\sigma_{1.3}^2$	: Variance of residual $e_{1.3}$
$\mathrm{Cov}(e_{1.3}, e_{2.3})$	: Covariance between residual $e_{1.3}$ and $e_{2.3}$

UNIT 9 LINEAR REGRESSION

Structure

9.1 Introduction

Objectives

9.2 Concept of Linear Regression

9.3 Lines of Regression

9.4 Regression Coefficients

9.5 Properties of Regression Coefficients

9.6 Distinction between Correlation and Regression

9.7 Angle between Two Lines of Regression

9.8 Summary

9.9 Solutions / Answers

9.1 INTRODUCTION

In Block 2, you have studied the curve fitting, correlation, rank correlation and intra-class correlation. Correlation studies the linear relationship between two variables. Correlation coefficient measures the strength of linear relationship and direction of the correlation whether it is positive or negative. When one variable is considered as an independent variable and another as dependent variable, and if we are interested in estimating the value of dependent variable for a particular value of independent variable, we study regression analysis. For example we might be interested in estimation of production of a crop for particular amount of rainfall or in prediction of demand on the price or prediction of marks on the basis of study hours of students. In these types of cases, regression would be the choice of statisticians or researchers. In general sense, regression analysis means estimation or prediction of the unknown value of one variable from the other variable.

In this unit, you will study the concept of regression, linear regression, regression coefficient with its properties and angle between two linear regression lines. Since correlation coefficient plays very important role in understanding the regression coefficient and its properties, you are advised to see the correlation coefficient with its properties carefully. You also go through the principle of least squares which will be used in finding the regression lines. This unit will also clearly discriminate the correlation and regression.

Section 9.2 explains the concept of linear regression and Section 9.3 describes how to obtain the regression line of dependent variable y on independent variable x. Regression line considering the x as a dependent variable and y as an independent variable is also discussed. Regression coefficients of y on x and x on y are defined in Section 9.4 whereas Section 9.5 gives the properties of regression coefficients with their proofs.

Section 9.6 differentiate between correlation and regression. Angles between two linear regression lines are explained in Section 9.7.

Objectives

After reading this unit, you will be able to

- define independent variable and dependent variable;
- explain the concept of regression and linear regression;
- describe lines of regression of y on x and x on y;
- define the regression coefficients of y on x and x on y;
- explore the properties of regression coefficient;
- explain the distinction between correlation and regression; and
- define the acute angle and obtuse angle.

9.2 CONCEPT OF LINEAR REGRESSION

Prediction or estimation is one of the major problems in most of the human activities. Like prediction of future production of any crop, consumption, price of any good, sales, income, profit, etc. are very important in business world. Similarly, prediction of population, consumption of agricultural product, rainfall, revenue, etc. have great importance to the government of any country for effective planning.

If two variables are correlated significantly, then it is possible to predict or estimate the values of one variable from the other. This leads us to very important concept of regression analysis. In fact, regression analysis is a statistical technique which is used to investigate the relationship between variables. The effect of price increase on demand, the effect of change in the money supply on the increase rate, effect of change in expenditure on advertisement on sales and profit in business are such examples where investigators or researchers try to construct cause and affect relationship. To handle these type of situations, investigators collect data on variables of interest and apply regression method to estimate the quantitative effect of the causal variables upon the variable that they influence.

Regression analysis describes how the independent variable(s) is (are) related to the dependent variable i.e. regression analysis measures the average relationship between independent variables and dependent variable. The literal meaning of regression is “stepping back towards the average” which was used by British Biometrician Sir Francis Galton (1822-1911) regarding the height of parents and their offspring’s.

Regression analysis is a mathematical measure of the average relationship between two or more variables.

There are two types of variables in regression analysis:

1. Independent variable
2. Dependent variable

The variable which is used for prediction is called independent variable. It is also known as regressor or predictor or explanatory variable.

The variable whose value is predicted by the independent variable is called dependent variable. It is also known as regressed or explained variable.

If scatter diagram shows some relationship between independent variable X and dependent variable Y, then the scatter diagram will be more or less concentrated round a curve, which may be called the curve of regression.

When the curve is a straight line, it is known as line of regression and the regression is said to be linear regression.

If the relationship between dependent and independent variables is not a straight line but curve of any other type then regression is known as non-linear regression.

Regression can also be classified according to number of variables being used. If only two variables are being used this is considered as simple regression whereas the involvement of more than two variables in regression is categorized as multiple regression.

Let us solve some little exercises.

E1) What do you mean by independent and dependent variables?
E2) Define regression.

9.3 LINES OF REGRESSION

Regression lines are the lines of best fit which express the average relationship between variables. Here, the concept of lines of best fit is based on principle of least squares.

Let X be the independent variable and Y be the dependent variable and we have observations (x_i, y_i); $i = 1, 2, ..., n$; Let the equation of line of regression of y on x be

$$y = a + bx \qquad \dots (1)$$

In this case, the value of the dependent variable changes at a constant rate as a unit change in the independent variable or explanatory variable.

Let $Y_i = a + bx_i$ be the estimated value of y_i for the observed or given value of $x = x_i$. According to the principle of least squares, we have to determine a and b so that the sum of squares of deviations of observed values of y from expected values of y i.e.

$$U = \sum_{i=1}^{n} (y_i - Y_i)^2$$

$$\text{or,} \quad U = \sum_{i=1}^{n} (y_i - a - bx_i)^2 \qquad \dots (2)$$

is minimum. From the principle of maxima and minima, we take partial derivatives of U with respect to a and b and equating to zero, i.e.

$$\frac{\partial U}{\partial a} = 0$$

$$\Rightarrow \frac{\partial}{\partial a}\sum_{i=1}^{n}(y_i - a - bx_i)^2 = 0$$

$$\Rightarrow 2\sum_{i=1}^{n}(y_i - a - bx_i)(-1) = 0$$

$$\Rightarrow \sum_{i=1}^{n}(y_i - a - bx_i) = 0$$

$$\Rightarrow \sum_{i=1}^{n} y_i - na - b\sum_{i=1}^{n} x_i = 0$$

$$\Rightarrow \sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i \qquad \ldots (3)$$

and $\frac{\partial U}{\partial b} = 0$

$$\Rightarrow \frac{\partial}{\partial b}\sum_{i=1}^{n}(y_i - a - bx_i)^2 = 0$$

$$\Rightarrow 2\sum_{i=1}^{n}(y_i - a - bx_i)(-x_i) = 0$$

$$\Rightarrow \sum_{i=1}^{n}(y_i x_i - ax_i - bx_i^2) = 0$$

$$\Rightarrow \sum_{i=1}^{n} y_i x_i - a\sum_{i=1}^{n} x_i - b\sum_{i=1}^{n} x_i^2 = 0$$

$$\Rightarrow \sum_{i=1}^{n} y_i x_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 \qquad \ldots (4)$$

Equations (3) and (4) are known as normal equations for straight line (1). Dividing equation (3) by n, we get

$$\bar{y} = a + b\,\bar{x} \qquad \ldots (5)$$

This indicates that regression line of y on x passes through the point $(\bar{x}, \bar{y})$. By the definition of covariance, we know that

$$\mathrm{Cov}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$$\mathrm{Cov}(x, y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\,\bar{y}$$

$$\Rightarrow \frac{1}{n}\sum_{i=1}^{n} x_i y_i = \mathrm{Cov}(x, y) + \bar{x}\,\bar{y} \qquad \ldots (6)$$

The variance of variable x can be expressed as

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$$

$$\Rightarrow \frac{1}{n}\sum_{i=1}^{n} x_i^2 = \sigma_x^2 + \bar{x}^2 \qquad \ldots (7)$$

Dividing equation (4) by n we get

$$\frac{1}{n}\sum_{i=1}^{n} y_i x_i = a\frac{1}{n}\sum_{i=1}^{n} x_i + b\frac{1}{n}\sum_{i=1}^{n} x_i^2 \qquad \ldots (8)$$

Using equations (6) and (7) in equation (8) gives

$$\mathrm{Cov}(x, y) + \bar{x}\,\bar{y} = a\bar{x} + b(\sigma_x^2 + \bar{x}^2) \qquad \ldots (9)$$

Multiplying equation (5) by $\bar{x}$, we get

$$\bar{y}\,\bar{x} = a\bar{x} + b\bar{x}^2 \qquad \ldots (10)$$

Subtracting equation (10) from equation (9), we get

$$\mathrm{Cov}(x, y) = b\sigma_x^2$$

$$\Rightarrow b = \frac{\mathrm{Cov}(x, y)}{\sigma_x^2} \qquad \ldots (11)$$

Since b is the slope of the line of regression of y on x and the line of regression passes through the point $(\bar{x}, \bar{y})$, so the equation of line of regression of y on x is

$$(y - \bar{y}) = b(x - \bar{x})$$

$$= \frac{\mathrm{Cov}(x, y)}{\sigma_x^2}(x - \bar{x}) \qquad \left(\because b = \frac{\mathrm{Cov}(x, y)}{\sigma_x^2}\right)$$

$$= \frac{r\sigma_x\sigma_y}{\sigma_x^2}(x - \bar{x}) \qquad (\mathrm{Cov}(x, y) = r\sigma_x\sigma_y\ )$$

$$(y - \bar{y}) = \frac{r\sigma_y}{\sigma_x}(x - \bar{x}) \qquad \ldots (12)$$

This is known as regression line of y on x.

If we consider the straight line $x = c + d\,y$ and proceeding similarly as in case of equation (1), we get the line of regression of x on y as

$$(x - \bar{x}) = \frac{\mathrm{Cov}(x, y)}{\sigma_y^2}(y - \bar{y})$$

$$(x - \bar{x}) = \frac{r\sigma_x}{\sigma_y}(y - \bar{y}) \qquad \ldots (13)$$

Therefore, we have two lines of regression, one of y on x and other of x on y. In case of perfect correlation $(r = \pm 1)$, either perfect positive or perfect negative, both lines of regression coincide and we have single line.

Lines of regression provide the best estimate to the value of dependent variable for specific value of the independent variable.

Equation (12) shows that the regression line of y on x passes through $(\bar{x}, \bar{y})$ and equation (13) also implies that regression line of x on y passes through $(\bar{x}, \bar{y})$.

Passing of both lines through the points $(\bar{x}, \bar{y})$ indicates that the point $(\bar{x}, \bar{y})$ is the intersection point of both lines. We can also conclude that solution of both lines as simultaneous equations provide mean of both variables, i.e. $\bar{x}$ and $\bar{y}$.

Regression line of y on x is used to estimate or predict the value of dependent variable y for the given value of independent variable x. Estimate of y obtained by this line will be best because this line minimizes the sum of squares of the errors of the estimates in y. If x is considered as dependent variable and y as independent variable then regression line of x on y is used to estimate or predict the value of variable x for the given value of y. Estimate of x obtained by regression line of x on y will be best because it minimizes the sum of squares of the errors of the estimates in x.

It is important to know that these regression lines y on x and x on y are different. These lines can’t be interchanged. Regression line of y on x cannot be used for the prediction of independent variable x. Similarly regression line of x on y cannot be used for the prediction of independent variable y.

When two regression lines cut each other at right angle (at the angle of 90 degree), it shows no correlation between y and x.

Let us solve some exercises.

E3) Which line is used for the prediction of dependent variable x and why?

E4) How many regression lines exist when two variables are considered in regression analysis and why?

9.4 REGRESSION COEFFICIENTS

If regression line of y on x is

$$(y - \bar{y}) = \frac{r\sigma_y}{\sigma_x}(x - \bar{x})$$

Then $\frac{r\sigma_y}{\sigma_x}$ is called the regression coefficient of y on x and it is denoted by b_{yx}. Thus, Regression coefficient of y on x ,

$$b_{yx} = \frac{r\sigma_y}{\sigma_x}$$

It gives the change in the dependent variable y as a unit change in the independent variable x. If we are considering the production of a crop by y and amount of rain fall by x, then regression coefficient of y on x represents change in production of crop as a unit change in rainfall i.e. if rainfall increases or decreases by one cm or by one inch how much production of a crop increases or decreases, is given by regression coefficient b_{yx}. In another example of investment in advertisement (x) and sales revenue (y), regression coefficient b_{yx} gives the change in sales revenue when the investment in advertisement is changed by a unit (a unit may be one thousand or one lakh or any other convenient figure).

Similarly, the regression coefficient of x on y gives the change in the value of dependent variable x as a unit change in the value of independent variable y and it is defined as

Regression coefficient of x on y, $b_{xy} = \frac{r\sigma_x}{\sigma_y}$.

For calculation purpose we can use the formula of regression coefficient of x on y as

$$b_{xy} = \frac{r\sigma_x}{\sigma_y}$$

$$= \frac{\left(\sum(x-\overline{x})(y-\overline{y}) \Big/ \sqrt{\sum(x-\overline{x})^2 \sum(y-\overline{y})^2}\right)\sqrt{\frac{1}{n}\sum(x-\overline{x})^2}}{\sqrt{\frac{1}{n}\sum(y-\overline{y})^2}}$$

$$b_{xy} = \frac{\sum(x-\overline{x})(y-\overline{y})}{\sum(y-\overline{y})^2}$$

Similarly, $b_{yx} = \frac{\sum(x-\overline{x})(y-\overline{y})}{\sum(x-\overline{x})^2}$

9.5 PROPERTIES OF REGRESSION COEFFICIENTS

Property 1: Geometric mean of the regression coefficients is correlation coefficient.

Description: If regression coefficient of y on x is b_{yx} and regression coefficient of x on y is b_{xy} then geometric mean of b_{yx} and b_{xy} is correlation coefficient i.e. $\sqrt{b_{yx} \times b_{xy}} = r$.

Proof: If regression coefficient of y on x is b_{yx} and regression coefficient of x on y is b_{xy}, then

$$b_{yx} \times b_{xy} = r\frac{\sigma_y}{\sigma_x} \times r\frac{\sigma_x}{\sigma_y}$$

$$\Rightarrow b_{yx} \times b_{xy} = r^2$$

$$\Rightarrow \pm\sqrt{b_{yx} \times b_{xy}} = r$$

It shows that geometric mean of regression coefficients is correlation coefficient.

Property 2: If one of the regression coefficients is greater than one, then other must be less than one.

Description: If b_{yx} is greater than one then b_{xy} must be less than one.

Proof: Let b_{yx}, the regression coefficient of y on x is greater than one i.e.

$$b_{yx} > 1$$

or
$$\frac{1}{b_{yx}} < 1$$

We know that

$$r^2 \leq 1 \Rightarrow b_{yx} \times b_{xy} \leq 1 \qquad \text{(From Property 1)}$$

$$\Rightarrow b_{xy} \leq \frac{1}{b_{yx}} < 1.$$

Thus if b_{yx} is greater than one then b_{xy} is less than one.

Property 3: Arithmetic mean of the regression coefficients is greater than the correlation coefficient i.e. $\frac{1}{2}(b_{yx} + b_{xy}) \geq r$, subject to the condition $r > 0$.

Proof: Suppose that arithmetic mean of regression coefficients is greater than correlation coefficient thus,

$$\frac{1}{2}(b_{yx} + b_{xy}) \geq r$$

$$\Rightarrow (b_{yx} + b_{xy}) \geq 2r$$

$$\Rightarrow (b_{yx} + b_{xy}) \geq 2(\pm\sqrt{(b_{xy} \times b_{yx})}$$

$$\text{(From Property 1 } r = \pm\sqrt{b_{yx} \times b_{xy}}\text{)}$$

Therefore,

$$\Rightarrow (b_{yx} + b_{xy}) \mp 2\sqrt{b_{yx}} \times \sqrt{b_{xy}} \geq 0$$

$$\Rightarrow (\sqrt{b_{yx}} \mp \sqrt{b_{xy}})^2 \geq 0$$

which is always true since the square of a real quantity is always positive.

Thus, $\frac{1}{2}(b_{yx} + b_{xy}) \geq r$, i.e. arithmetic mean of regression coefficients is greater than correlation coefficient.

Let us do some problems related to regression coefficients.

Example 1: Height of fathers and sons in inches are given below:

Height of Father	65	66	67	67	68	69	70	71
Height of Son	66	68	65	69	74	73	72	70

Find two lines of regression and calculate the estimated average height of son when the height of father is 68.5 inches.

Solution: Let us denote the father's height by x and son's height by y then two lines of regression can be expressed as

$$(y-\bar{y})=\frac{r\sigma_y}{\sigma_x}(x-\bar{x}) \text{ and}$$

$$(x-\bar{x})=\frac{r\sigma_x}{\sigma_y}(y-\bar{y})$$

To find the regression lines we need $\bar{x}$, $\bar{y}$, σ_x, σ_y and r which will be obtained from the following table:

x	**y**	**x^2**	**y**	**xy**
65	66	4225	4356	4290
66	68	4356	4624	4488
67	65	4489	4225	4355
67	69	4489	4761	4623
68	74	4624	5476	5032
69	73	4761	5329	5037
70	72	4900	5184	5040
71	70	5041	4900	4970
$\sum x=543$	$\sum y=557$	$\sum x^2=36885$	$\sum y^2=38855$	$\sum xy=37835$

Now,

Mean of x = $\bar{x}=\frac{1}{n}\sum x=\frac{1}{8}543=67.88$

Mean of y = $\bar{y}=\frac{1}{n}\sum y=\frac{1}{8}557=69.62$

Standard Deviation of x = $\sigma_x=\sqrt{\frac{1}{n}\sum(x-\bar{x})^2}=\sqrt{\frac{1}{n}\sum x^2-\bar{x}^2}$

$$=\sqrt{\frac{1}{8}\times 36885-(67.88)^2}$$

$$=\sqrt{4610.62-4607.69}$$

$$=\sqrt{2.93}=1.71$$

Similarly,

Standard deviation of y = $\sigma_y=\sqrt{\frac{1}{n}\sum(y_i-\bar{y})^2}=\sqrt{\frac{1}{n}\sum y^2-\bar{y}^2}$

$$=\sqrt{\frac{1}{8}\times 38855-(69.62)^2}$$

$$=\sqrt{4856.88-4846.94}$$

$$=\sqrt{9.94}=3.15$$

Now, correlation coefficient

$$r = \text{Corr}(x, y) = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{\left\{n\sum x^2 - \left(\sum x\right)^2\right\}\left\{n\sum y^2 - \left(\sum y\right)^2\right\}}}$$

$$= \frac{8\times 37835 - (543)\times(557)}{\sqrt{\left\{(8\times 36885 - (543)^2\right\}\left\{(8\times 38855) - (557)^2\right\}}}$$

$$= \frac{302680 - 302451}{\sqrt{\left\{(295080 - 294849\right\}\left\{310840 - 310249\right\}}}$$

$$= \frac{229}{\sqrt{231\times 591}}$$

$$= \frac{229}{\sqrt{136521}} = \frac{229}{369.49}$$

$$r = 0.62$$

Substituting the value of $\bar{x}$, $\bar{y}$, σ_x, σ_y and r in regression equations, we get regression line of y on x as

$$(y - 69.62) = 0.62\times\frac{3.15}{1.71}(x - 67.88)$$

$$y = 1.14x - 77.38 + 69.62$$

$$y = 1.14x - 7.76$$

and regression line of x on y

$$(x - 67.88) = 0.62\times\frac{1.71}{3.15}(y - 69.62)$$

$$x = 0.34y - 23.67 + 67.88$$

$$x = 0.34y + 44.21$$

Estimate of height of son for the height of father = 68.5 inch is obtained by the regression line of y on x

$$y = 1.14x - 7.76$$

Putting x = 68.5 in above regression line

$$y = 1.14\times 68.50 - 7.76 = 78.09 - 7.76 = 70.33$$

Thus, the estimate of son's height for the father's height 68.5 inch is 70.33 inch.

Note: For the estimation of x for the given value of y, we use regression line of x on y whereas for the estimation of y for the given value of x we use regression line of y on x.

Example 2: Regression line of y on x and x on y respectively are

$$2x - 3y = -8$$

$$5x - y = 6$$

Then, find

(i) the mean values of x and y,

(ii) coefficient of correlation between x and y, and

(iii) the standard deviation of y for given variance of x = 5.

Solution:

(i) Since regression lines of y on x and x on y passes through $\bar{x}$ and $\bar{y}$ so $\bar{x}$ and $\bar{y}$ are the intersection points. Thus to get the mean values of variable x and y, we solve given simultaneous equations

$$2x - 3y = -8$$

$$5x - y = 6$$

By solving these equations as simultaneous equations we get x =2 and y = 4 which are means of x and y respectively.

Note: You have already solved simultaneous equations in Unit 1 of the Block-2 during the fitting of various curves for given data. If you face some problems in solving these equations go through problems given in Unit 5 of the Block 2.

(ii) To find the correlation coefficient, given equations are expressed as

$$y = 2.67 + 0.67x \qquad \ldots (14)$$

$$x = 1.20 + 0.20y \qquad \ldots (15)$$

Note: To find the regression coefficient of y on x, regression line of y on x is expressed in the form of $y = a + bx$, where b is the regression coefficient of y on x. Similarly, to find the regression coefficient of x on y, regression line of x on y is expressed in the form of $x = c + dy$, where d is the regression coefficient of x on y. In our problem, by dividing the first line by 3, i.e. by the coefficient of y gives equation in the form $y = a + bx$ which is y = 2.67 +0.67 x. Similarly, dividing the second regression equation by 5 gives equation (15).

From equations (14) and (15) we, find the regression coefficient of y on x and x on y respectively as

$$b_{yx} = \frac{r\sigma_y}{\sigma_x} = 0.67 \text{ and}$$

$$b_{xy} = \frac{r\sigma_x}{\sigma_y} = 0.20$$

By the property of regression coefficients

$$\pm\sqrt{b_{yx} \times b_{xy}} = r \Rightarrow r = \sqrt{0.67 \times 0.20}0 = 0.37$$

Thus, correlation coefficient r = 0.37

Note: We are taking $(+)$ sign because correlation coefficient and regression coefficients have same sign.

(iii) By the definition of regression coefficient of y on x i.e.

$$b_{yx} = \frac{r\sigma_y}{\sigma_x} = 0.67$$

Variance of x i.e., $\sigma_x^2 = 5 \Rightarrow \sigma_x = 2.24$

Now,

$$b_{yx} = \frac{r\sigma_y}{\sigma_x} \Rightarrow 0.67 = \frac{0.37\sigma_y}{2.24} \Rightarrow \sigma_y = 4.05,$$

$$\Rightarrow \sigma_y^2 = (4.05)^2 = 16.45$$

Thus, the variance of y is 16.45.

E5) What is the geometric mean of regression coefficients?

E6) Both regression coefficients can have same sign. Do you agree?

E7) Marks of 6 students of a class in paper I and paper II of Statistics are given below:

Paper I	45	55	66	75	85	100
Paper II	56	55	45	65	62	71

Find

(i) both regression coefficients,

(ii) both regression lines, and

(iii) correlation coefficient.

E8) We have data on variables x and y as

x	5	4	3	2	1
y	9	8	10	11	12

Calculate

(i) both regression coefficients,

(ii) correlation coefficient,

(iii) regression lines of y on x and x on y, and

(iv) estimate y for x = 4.5.

E9) If two regression lines are

$$6x + 15y = 27$$

$$6x + 3y = 15,$$

Then, calculate

(i) correlation coefficient, and

(ii) mean values of x and y.

9.6 DISTINCTION BETWEEN CORRELATION AND REGRESSION

Both correlation and regression have important role in relationship study but there are some distinctions between them which can be described as follow:

(i) Correlation studies the linear relationship between two variables while regression analysis is a mathematical measure of the average relationship between two or more variables.

(ii) Correlation has limited application because it gives the strength of linear relationship while the purpose of regression is to "predict" the value of the dependent variable for the given values of one or more independent variables.

(iii) Correlation makes no distinction between independent and dependent variables while linear regression does it, i.e. correlation does not consider the concept of dependent and independent variables while in regression analysis one variable is considered as dependent variable and other(s) is/are as independent variable(s).

9.7 ANGLE BETWEEN TWO LINES OF REGRESSION

When we consider y as dependent variable and x as independent variable than regression line of y on x is

$$(y-\overline{y}) = r\frac{\sigma_y}{\sigma_x}(x-\overline{x})$$

Similarly, when x is considered as dependent variable and y as an independent variable then regression line of x on y is

$$(x-\overline{x}) = r\frac{\sigma_x}{\sigma_y}(y-\overline{y})$$

If m_1 and m_2 are the slopes of two lines and θ be the angle between them therefore

$$\tan\theta = \left|\frac{m_1-m_2}{1+m_1m_2}\right|$$

For these regression lines it is observed that the slope of regression line of y on x is $\frac{r\sigma_y}{\sigma_x}$ and slope of regression line of x on y is $\frac{\sigma_y}{r\sigma_x}$. If the angle between the two lines of regression is denoted by θ, then

$$\theta = \tan^{-1}\left\{\frac{(1-r^2)}{r}\left(\frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}\right)\right\} \qquad \dots (16)$$

If $r = 0$ i.e. variables are uncorrelated then

$$\tan\theta = \infty \Rightarrow \theta = \frac{\pi}{2}$$

In this case lines of regression are perpendicular to each other.

If r = ±1 i.e. variables are perfect positive or negative correlated then

$$\tan\theta = 0 \Rightarrow \theta = 0 \text{ or } \pi$$

In this case, regression lines either coincide or parallel to each other.

There are two angles between regression lines whenever two lines intersect each other, one is acute angle and another is obtuse angle. The tan θ would be greater than zero if θ lies between 0 and $\frac{\pi}{2}$ then θ is called acute angle, which is obtained by

$$\theta_1 = \text{acute angle} = \tan^{-1}\left\{\frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}.\frac{1-r^2}{r}\right\}, \quad r>0, \quad \text{if } 0< \theta < \frac{\pi}{2}$$

The tan θ would be less than zero if θ lies between $\frac{\pi}{2}$ and π then θ is called obtuse angle, which is obtained by

$$\theta_2 = \text{obtuse angle} = \tan^{-1}\left\{\frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}.\frac{r^2-1}{r}\right\}, \quad r>0, \quad \text{if } \frac{\pi}{2} < \theta < \pi$$

9.8 SUMMARY

In this unit, we have discussed:

1. The concept of regression,
2. How to obtain lines of regression and regression coefficient,
3. Properties of regression coefficients,
4. How to use regression line for the prediction of dependent variable on the basis of value of independent variable,
5. The difference between correlation and regression, and
6. The angle between the two regression lines.

9.9 SOLUTIONS /ANSWERS

E1) The variable which is used for prediction is called independent variable. It is also known as regressor or predictor or explanatory variable.

The variable whose value is predicted by the independent variable is called dependent variable. It is also known as regressed or explained variable.

E2) If two variables are correlated significantly, then it is possible to predict or estimate the values of one variable from the other. This leads us to very important concept of regression analysis. In fact, regression analysis is a statistical technique which is used to investigate the relationship between variables. The effect of price increase on demand, the effect of change in the money supply on the inflation rate, effect of change in expenditure on advertisement on sales and profit in business are such examples where investigators or researchers try to construct cause and affect relationship. To handle these type of situations investigators collect data on variables of interest and apply regression method to estimate the quantitative effect of the causal variables upon the variable that they influence.

The literal meaning of regression is “stepping back towards the average”. It was first used by British Biometrician Sir Francis Galton (1822-1911) in connection with the height of parents and their offspring’s.

Regression analysis is a mathematical measure of the average relationship between two or more variables.

E3) If x is considered as dependent variable and y as independent variable then regression line of x on y is used to estimate or predict the value of variable x for the given value of y. Estimate of x obtained by regression line of x on y will be best because it minimizes the sum of squares of errors of estimates in x.

E4) When two variables are considered in regression analysis, There are two regression lines

(i) Regression line of y on x and

(ii) Regression line of x on y.

Regression line of y on x is used to estimate or predict the value of dependent variable y for the given value of independent variable x. Estimate of y obtained by this line will be best because this line minimizes the sum of squares of the errors of the estimates in y. If x is considered as dependent variable and y as independent variable then regression line of x on y is used to estimate or predict the value of variable x for the given value of y. Estimate of x obtained by regression line of x on y will be best because it minimizes the sum of squares of the errors of the estimates in x.

E5) Correlation coefficient is the geometric mean of the regression coefficients.

Description: If regression coefficient of y on x is b_{yx} in and regression coefficient of x on y is b_{xy} , then $\sqrt{b_{yx} \times b_{xy}} = r$.

Proof:

If regression coefficient of y on x is b_{yx} and regression coefficient of x on y is b_{xy}, then

$$b_{yx} \times b_{xy} = r\frac{\sigma_y}{\sigma_x} \times r\frac{\sigma_x}{\sigma_y}$$

$$\Rightarrow b_{yx} \times b_{xy} = r^2$$

$$\Rightarrow \pm\sqrt{b_{yx} \times b_{xy}} = r$$

It shows that geometric mean of regression coefficients is correlation coefficient.

E6) Yes, both regression coefficients have same sign (positive or negative).

E7) (i) Let us denote the marks in paper I by x and marks in paper II by y then, here we will use the direct formula of b_{yx} and b_{xy} which are

$$b_{xy} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(y-\bar{y})^2} \quad \text{and}$$

$$b_{yx} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}$$

Now, $\bar{x} = \frac{\sum x}{n} = \frac{426}{6} = 71$

and $\bar{y} = \frac{\sum y}{n} = \frac{354}{6} = 59$

S. No.	**x**	**y**	$(x-\bar{x})$	$(x-\bar{x})^2$	$(y-\bar{y})$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
1	45	56	-26	676	-3	9	78
2	55	55	-16	256	-4	16	64
3	66	45	-5	25	-14	196	70
4	75	65	4	16	6	36	24
5	85	62	14	196	3	9	42
6	100	71	29	841	12	144	348
Total	426	354	0	2010	0	410	626

Thus $b_{yx} = \frac{626}{2010} = 0.31$

$$b_{xy} = \frac{626}{410} = 1.53.$$

(ii) Regression line of y on x is $(y-\bar{y}) = b_{yx}(x-\bar{x})$ and the regression line of x on y is $(x-\bar{x}) = b_{xy}(y-\bar{y})$

So we need $\bar{y}$, $\bar{x}$, b_{yx} and b_{xy}. In (i) we have calculated b_{yx} and b_{xy}.

Thus, the Regression line of y on x is $(y-59) = 0.31\,(x-71)$

$$\Rightarrow y = 0.31x - 22.01 + 51$$

$$\Rightarrow y = 0.31x + 36.99$$

Regression line of x on y is $(x-71) = 1.53(y-59)$

$$\Rightarrow x = 1.53y - 19.27$$

(iii) By the first property of regression coefficients, we know that

$$r = \pm\sqrt{b_{yx}\times b_{xy}} \;\Rightarrow \sqrt{0.31\times 1.53} = 0.68$$

E8) Here, for the calculation in table

$$\bar{x} = \frac{\sum x}{n} = \frac{15}{3} = 3 \quad \text{and} \quad \bar{y} = \frac{\sum y}{n} = \frac{50}{5} = 10$$

S. No.	x	y	$(x-\bar{x})$	$(x-\bar{x})^2$	$(y-\bar{y})$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
1	5	9	2	4	-1	1	-2
2	4	8	1	1	-2	4	-2
3	3	10	0	0	0	0	0
4	2	11	-1	1	1	1	-1
5	1	12	-2	4	2	4	-4
Total	15	50	0	10	0	10	-9

(i) Regression coefficient of x on y $\quad b_{xy} = \dfrac{\sum(x-\bar{x})(y-\bar{y})}{\sum(y-\bar{y})^2}$

$$b_{xy} = \frac{-9}{10} = -0.9$$

Regression coefficient of y on x $\quad b_{yx} = \dfrac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}$

$$b_{yx} = \frac{-9}{10} = -0.9$$

(ii) $r = \pm\sqrt{b_{yx} \times b_{xy}} \Rightarrow \pm\sqrt{(-0.9)\times(-0.9)} = -\ 0.90$

Note: We are taking $(-)$ sign because correlation coefficient and regression coefficients have same sign.

(iii) To find regression lines we need $\bar{y}, \bar{x}, b_{yx}$ and b_{xy}. From the calculation in table

$$\bar{x} = \frac{\sum x}{n} = \frac{15}{5} = 3 \text{ and}$$

$$\bar{y} = \frac{\sum y}{n} = \frac{50}{5} = 10$$

Thus, regression line of y on x $(y-10) = -0.90(x-3)$

Regression line of x on y $(x-3) = -\ 0.90\ (y-10)$

(iv) To estimate y we use regression line of y on x which is

$$(y-10) = -\ 0.9\ (x-3)$$

putting x = 4.5 we get

$$(y-10) = -\ 0.9\ (4.5-3)$$

$$y = 8.65$$

E9) (i) To find correlation coefficient, we need regression coefficients which can be obtained from the given regression lines by presenting them in the following form

$$y = 1.80 - 0.40x \text{ which gives } b_{yx} = -\ 0.40$$

and

$$x = 2.50 - 0.50y \text{ which gives } b_{xy} = -0.50$$

$$r = \pm\sqrt{b_{yx} \times b_{xy}} = \pm\sqrt{(-\ 0.4)\times(-\ 0.5)} = -\ 0.44$$

(ii) Since both regression lines passes through means of y and x so the solution of these equation as a simultaneous equation gives mean values. Subtracting second equation from first equation we have

$$6x + 15y = 27$$

$$6x + 3y = 15$$

$$12\ y = 12 \quad \Rightarrow y = 1$$

Substituting the value of y in first equation, we get x = 2.

Thus x = 2 and y = 1 are mean values of variables x and y respectively i.e. $\bar{x} = 2, \bar{y} = 1$

UNIT10 PLANE OF REGRESSION

Structure

10.1 Introduction

Objectives

10.2 Yule's Notation

10.3 Plane of Regression for three Variables

10.4 Properties of Residuals

10.5 Variance of the Residuals

10.6 Summary

10.7 Solutions / Answers

10.1 INTRODUCTION

In Unit 9, you learnt the concept of regression, linear regression, lines of regression and regression coefficient with its properties. Unit 9 was based on linear regression in which we were considering only two variables. In Unit 6 of MST-002, you studied the correlation that measures the linear relationship between two variables. In many situations, we are interested in studying the relationship among more than two variables in which one variable is influenced by many others. For example, production of any crop depends upon soil fertility, fertilizers used, irrigation methods, weather conditions, etc. Similarly, marks obtained by students in exam may depend upon their IQ, attendance in class, study at home, etc. In these type of situations, we study the multiple and partial correlations. In this unit you will study the basics of multiple and partial correlations that includes Yule's notations, plane of regression, residuals, properties of residuals and variance of residuals.

For the study of more than two variables, there would be need of much more notations in comparison to the notations used in Unit 9. These notations were given by G.U. Yule (1897). Yule's notations and residuals are described in Section 10.2. Plane of regression and normal equations are given in Section 10.3. Properties of residuals are explored in Section 10.4 whereas Section 10.5 explains the variance of residuals.

Objectives

After reading this unit, you would be able to

- define the Yule's notation;
- describe the plane of regression for three variable;
- explain the properties of residuals;
- describe the variance of the residuals;
- find out the lines of regression; and
- find out the estimates of dependent variable from the regression lines.

10.2 YULE'S NOTATION

Karl Pearson developed the theory of multiple and partial correlation for three variables and it was generalized by G.U. Yule (1897). Let us consider three random variables $X_1,\ X_2$ and X_3 and $\overline{X}_1,\ \overline{X}_2$ and $\overline{X}_3$ are their respective means. Then regression equation of X_1 on X_2 and X_3 is defined as

$$X_1 = a + b_{12.3}X_2 + b_{13.2}X_3 \qquad \dots (1)$$

where, $b_{12.3}$ and $b_{13.2}$ are the partial regression coefficients of X_1 on X_2 and X_1 on X_3 keeping the effect of X_3 and X_2 fixed respectively.

Taking summation of equation (1) and dividing it by N, we get

$$\overline{X}_1 = a + b_{12.3}\overline{X}_2 + b_{13.2}\overline{X}_3 \qquad \dots (2)$$

On subtracting equation (2) from equation (1), we get

$$X_1 - \overline{X}_1 = b_{12.3}\left(X_2 - \overline{X}_2\right) + b_{13.2}\left(X_3 - \overline{X}_3\right) \qquad \dots (3)$$

Suppose $X_1 - \overline{X}_1 = x_1, X_2 - \overline{X}_2 = x_2$ and $X_3 - \overline{X}_3 = x_3$

Now, plane of regression x_1 on x_2 and x_3 (equation (3)) can be rewritten as

$$x_1 = b_{12.3}x_2 + b_{13.2}x_3 \qquad \dots (4)$$

Right hand side of equation (2) is called the estimate of x_1 which is denoted by $x_{1.23}$. Thus,

$$x_{1.23} = b_{12.3}x_2 + b_{13.2}x_3 \qquad \dots (5)$$

Error of estimate or residual is defined as

$$e_{1.23} = x_1 - x_{1.23}$$

$$e_{1.23} = x_1 - b_{12.3}x_2 - b_{13.2}x_3 \qquad \dots (6)$$

This residual is of order 2.

If we are considering n variables $x_1, x_2, ..., x_n$, the equation of the plane of regression of x_1 on $x_2, ..., x_n$ is

$$x_1 = b_{12.3456...n}x_2 + b_{13.24...n}x_3 + ... + b_{1n.23...(n-1)}x_n \qquad \dots (7)$$

and error of estimate or residual is

$$x_{1.23...n} = x_1 - (b_{12.34...n}x_2 + b_{13.24...n}x_3 + ... + b_{1n.23...(n-1)}x_n) \qquad \dots (8)$$

Note: In above expressions we have used subscripts involving digits 1, 2, 3,…, n and dot (.). Subscripts before the dot are known as the primary subscripts whereas the subscripts after the dot are called secondary subscripts.

The number of secondary subscripts decides the order of regression coefficient.

For example $b_{12.3}$ is the regression coefficient of order 1, $b_{12.34}$ is of order 2 and so on $b_{1n.23...(n-1)}$ of order $(n-2)$.

Order in which secondary subscripts ($b_{12\,.\,34}$ or $b_{12\,.\,43}$) is immaterial but the order of primary subscripts is very important and decides the dependent and independent variables. In $b_{12.34}$, x_1 is dependent variable and x_2 is independent variable whereas in $b_{21.34}$, x_2 is dependent variable and x_1 is independent variable.

Order of residuals is determined by the number of secondary subscripts. For example $x_{1.23}$ is residual of order 2 while as $x_{1.234}$ is of order 3. Similarly, $x_{1.234....n}$ is residual of order $(n-1)$.

10.3 PLANE OF REGRESSION FOR THREE VARIABLES

Let us consider three variables x_1, x_2 and x_3 measured from their respective means. The regression equation of x_1 depends upon x_2 and x_3 is given by (from equation (4)).

$$x_1 = b_{12.3}x_2 + b_{13.2}x_3 \qquad \ldots (9)$$

If x_3 is considered as a constant then the graph of x_1 and x_2 is a straight line with slope $b_{12.3}$, similarly the graph of the x_1 and x_3 would be the straight line with slope $b_{13.2}$, if x_2 is considered as a constant.

According to the principle of least squares, we have to determine constants $b_{12.3}$ and $b_{13.2}$ in such a way that sum of squares of residuals is minimum i.e.

$$U = \sum (x_1 - b_{12.3}x_2 - b_{13.2}x_3)^2 = \sum e_{1.23}^2 \text{ is minimum.}$$

here, $e_{1.23} = x_1 - b_{12.3}x_2 - b_{13.2}x_3$ … (10)

By the principle of maxima and minima, we take partial derivatives of U with respect to $b_{12.3}$ and $b_{13.2}$

Thus,

$$\frac{\partial U}{\partial b_{12.3}} = \frac{\partial U}{\partial b_{13.2}} = 0$$

Let us take

$$\frac{\partial U}{\partial b_{12.3}} = 0$$

$$\Rightarrow \sum 2(x_1 - b_{12.3}x_2 - b_{13.2}x_3)(-x_2) = 0$$

$$\Rightarrow \sum x_2(x_1 - b_{12.3}x_2 - b_{13.2}x_3) = 0 \qquad \ldots (11)$$

$$\Rightarrow \sum(x_2x_1 - b_{12.3}x_2^2 - b_{13.2}x_2x_3) = 0$$

$$\Rightarrow b_{12.3}\sum x_2^2 + b_{13.2}\sum x_2x_3 - \sum x_1x_2 = 0 \qquad \ldots (12)$$

Similarly,

$$\frac{\partial U}{\partial b_{13.2}} = 0 \Rightarrow \sum x_3(x_1 - b_{12.3}x_2 - b_{13.2}x_3) = 0 \qquad \ldots (13)$$

$$\Rightarrow b_{12.3}\sum x_2x_3 + b_{13.2}\sum x_3^2 - \sum x_1x_3 = 0 \qquad \ldots (14)$$

As we know that

$$\sigma_i^2 = \frac{1}{N}\sum(x_i - \overline{x}_i)^2 \qquad \text{(for } i = 1, 2 \text{ and } 3)$$

$$= \frac{1}{N}\sum x_i^2 - \overline{x}_i^{\,2}$$

Since, x_1 , x_2 and x_3 are measured from their means therefore $\overline{x}_1 = \overline{x}_2 = \overline{x}_3 = 0$ then

$$\sigma_i^2 = \frac{1}{N}\sum x_i^2 \qquad \ldots (15)$$

Similarly, we can write (for i ≠ j = 1, 2, 3)

$$\text{Cov}(x_i, x_j) = \frac{1}{N}\sum x_ix_j \qquad \ldots (16)$$

and consequently, using equations (15) and (16), correlation coefficient between x_i and x_j can be expressed as

$$r_{ij} = \frac{\text{Cov}(x_i, x_j)}{\sqrt{V(x_i)V(x_j)}} = \frac{\sum x_ix_j}{N\sigma_i\sigma_j} \Rightarrow \text{Cov}(x_i, x_j) = r_{ij}\sigma_i\sigma_j \qquad \ldots (17)$$

Dividing equations (12) and (14) by N provides

$$b_{12.3}\frac{1}{N}\sum x_2^2 + b_{13.2}\frac{1}{N}\sum x_2x_3 - \frac{1}{N}\sum x_1x_2 = 0 \quad \text{and} \qquad \ldots (18)$$

$$b_{12.3}\frac{1}{N}\sum x_2x_3 + b_{13.2}\frac{1}{N}\sum x_3^2 - \frac{1}{N}\sum x_1x_3 = 0 \qquad \ldots (19)$$

Using equations (15), (16) and (17) in equations (18) and (19), we have

$$b_{12.3}\sigma_2^2 + b_{13.2}\text{Cov}(x_2, x_3) - \text{Cov}(x_1, x_2) = 0 \quad \text{From equation (18)}$$

$$\Rightarrow b_{12.3}\sigma_2^2 + b_{13.2}r_{23}\sigma_2\sigma_3 - r_{12}\sigma_1\sigma_2 = 0$$

$$\Rightarrow \sigma_2(b_{12.3}\sigma_2 + b_{13.2}r_{23}\sigma_3 - r_{12}\sigma_1) = 0$$

$$\Rightarrow b_{12.3}\sigma_2 + b_{13.2}r_{23}\sigma_3 - r_{12}\sigma_1 = 0 \qquad \ldots (20)$$

Similarly,

$$b_{12.3}\text{Cov}(x_2, x_3) + b_{13.2}\sigma_3^2 - \text{Cov}(x_1, x_3) = 0 \quad \text{From equation (19)}$$

$$\Rightarrow b_{12.3}r_{23}\sigma_2\sigma_3 + b_{13.2}\sigma_3^2 - r_{13}\sigma_1\sigma_3 = 0$$

$$\Rightarrow \sigma_3(b_{12.3}r_{23}\sigma_2 + b_{13.2}\sigma_3 - r_{13}\sigma_1) = 0$$

$$\Rightarrow b_{12.3}r_{23}\sigma_2 + b_{13.2}\sigma_3 - r_{13}\sigma_1 = 0 \qquad \ldots (21)$$

where, r_{12} is the total correlation coefficient between x_1 and x_2, r_{13} is the total correlation coefficient between x_1 and x_3 and similarly, r_{23} is the total correlation coefficient between x_1 and x_3. Thus, we have two equations (20) and (21).

Solving the equations (20) and (21) for $b_{12.3}$ and $b_{13.2}$, we obtained

$$b_{12.3} = \frac{\begin{vmatrix} r_{12}\sigma_1 & r_{23}\sigma_3 \\ r_{13}\sigma_1 & \sigma_3 \end{vmatrix}}{\begin{vmatrix} \sigma_2 & r_{23}\sigma_3 \\ r_{23}\sigma_2 & \sigma_3 \end{vmatrix}}$$

$$b_{12.3} = \frac{\sigma_1}{\sigma_2}\frac{\begin{vmatrix} r_{12} & r_{23} \\ r_{13} & 1 \end{vmatrix}}{\begin{vmatrix} 1 & r_{23} \\ r_{23} & 1 \end{vmatrix}} = \frac{\sigma_1\left(r_{12} - r_{13}r_{23}\right)}{\sigma_2\left(1 - r_{23}^2\right)} \qquad \ldots (22)$$

Similarly,

$$b_{13.2} = \frac{\sigma_1}{\sigma_3}\frac{\begin{vmatrix} 1 & r_{12} \\ r_{23} & r_{13} \end{vmatrix}}{\begin{vmatrix} 1 & r_{23} \\ r_{23} & 1 \end{vmatrix}} = \frac{\sigma_1\left(r_{13} - r_{12}r_{23}\right)}{\sigma_3\left(1 - r_{23}^2\right)} \qquad \ldots (23)$$

If we write

$$t = \begin{vmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{vmatrix} \qquad \ldots (24)$$

$b_{12.3}$ and $b_{13.2}$ can be written as

$$b_{12.3} = -\frac{\sigma_1}{\sigma_2}\frac{t_{12}}{t_{11}}$$

and

$$b_{13.2} = -\frac{\sigma_1}{\sigma_3}\frac{t_{13}}{t_{11}}$$

where, t_{ij} is the cofactor of the element in the i^{th} row and j^{th} column of t.

Substituting the values of $b_{12.3}$ and $b_{13.2}$ in equation (9), we get the equation of the plane of regression of x_1 on x_2 and x_3 as

$$x_1 = -\frac{\sigma_1}{\sigma_2}\frac{t_{12}}{t_{11}}x_2 - \frac{\sigma_1}{\sigma_3}\frac{t_{13}}{t_{11}}x_3 \qquad \text{where, a = 0}$$

$$\Rightarrow \frac{x_1}{\sigma_1}t_{11} + \frac{x_2}{\sigma_2}t_{12} + \frac{x_3}{\sigma_3}t_{13} = 0 \qquad \dots (25)$$

Similarly, the plane of regression of x_2 on x_1 and x_3 is given by

$$\Rightarrow \frac{x_1}{\sigma_1}t_{21} + \frac{x_2}{\sigma_2}t_{22} + \frac{x_3}{\sigma_3}t_{23} = 0 \qquad \dots (26)$$

and the plane of regression of x_3 on x_1 and x_2 is

$$\Rightarrow \frac{x_1}{\sigma_1}t_{31} + \frac{x_2}{\sigma_2}t_{32} + \frac{x_3}{\sigma_3}t_{33} = 0 \qquad \dots (27)$$

In general the plane of regression of x_i on the remaining variable x_j $(j \neq i = 1, 2, ..., n)$ is given by

$$\Rightarrow \frac{x_1}{\sigma_1}t_{i1} + \frac{x_2}{\sigma_2}t_{i2} + \dots + \frac{x_i}{\sigma_i}t_{ii} + \dots + \frac{x_n}{\sigma_n}t_{in} = 0;\ i = 1,2,...,n \qquad \dots (28)$$

Example 1: From the given data in the following table find out

(i) Least square regression equation of X_1 on X_2 and X_3.

(ii) Estimate the value of X_1 for $X_2 = 45$ and $X_3 = 8$.

X_1	1	2	3	4	5
X_2	3	4	5	6	7
X_3	4	5	6	7	8

Solution: (i) Here X_1, X_2 and X_3 are three random variables with their respective means $\overline{X}_1, \overline{X}_2$ and $\overline{X}_3$.

Let $X_1 - \overline{X}_1 = x_1,\ X_2 - \overline{X}_2 = x_2$ and $X_3 - \overline{X}_3 = x_3$

Then linear regression equation of x_1 on x_2 and x_3 is

$$x_1 = b_{12.3}x_2 + b_{13.2}x_3$$

From equation (22) and (23), we have

$$b_{12.3} = \frac{\sigma_1 (r_{12} - r_{13}r_{23})}{\sigma_2 (1 - r_{23}^2)}$$

and,

$$b_{13.2} = \frac{\sigma_1 (r_{13} - r_{12}r_{23})}{\sigma_3 (1 - r_{23}^2)}$$

The value of σ_1, σ_2, σ_3, r_{12} r_{13} and r_{23} can be obtained through some calculations given in the following table:

S. No.	X_1	X_2	X_3	$x_1 = X_1 - 5$	$x_2 = X_2 - 6$	$x_3 = X_3 - 7$	$(x_1)^2$	$(x_2)^2$	$(x_3)^2$	x_1x_2	x_1x_3	x_2x_3
1	1	3	4	−4	−3	−3	16	9	9	12	12	9
2	3	5	5	−2	−1	−2	4	1	4	2	4	2
3	4	6	6	−1	0	−1	1	0	1	0	1	0
4	7	7	9	2	1	2	4	1	4	2	4	2
5	10	9	11	5	3	4	25	9	16	15	20	12
Total	25	30	35	0	0	0	50	20	34	31	41	25

$$\overline{X}_1 = \frac{\sum X_1}{N} = \frac{25}{5} = 5$$

$$\overline{X}_2 = \frac{\sum X_2}{N} = \frac{30}{5} = 6$$

$$\overline{X}_3 = \frac{\sum X_3}{N} = \frac{35}{5} = 7$$

$$\sigma_1^2 = \frac{1}{N}\sum x_1^2$$ from equation (15)

$$= \frac{1}{5}(50) = 10$$

$$\Rightarrow \sigma_1 = \sqrt{10}$$

$$= 3.162$$

$$\sigma_2^2 = \frac{1}{N}\sum x_2^2$$

$$= \frac{1}{5}(20) = 4$$

$$\Rightarrow \sigma_2 = \sqrt{4} = 2$$

$$\sigma_3^2 = \frac{1}{N}\sum x_3^2$$

$$= \frac{1}{5}(34) = 6.8$$

$$\Rightarrow \sigma_3 = \sqrt{6.8}$$

$$= 2.608$$

$$r_{12} = \frac{\sum x_1 x_2}{N\sigma_1\sigma_2}$$ from equation (17)

$$= \frac{31}{5 \times 3.162 \times 2} = 0.98$$

$$r_{13} = \frac{\sum x_1 x_3}{N\sigma_1\sigma_3}$$

$$= \frac{41}{5 \times 3.162 \times 2.608} = 0.994$$

$$r_{23} = \frac{\sum x_2 x_3}{N\sigma_2\sigma_3}$$

$$= \frac{25}{5 \times 2 \times 2.608} = 0.959$$

Now, we have

$$b_{12.3} = \frac{\sigma_1\left(r_{12} - r_{13}r_{23}\right)}{\sigma_2\left(1 - r_{23}^2\right)}$$

$$= \frac{3.162 \times \left(0.98 - 0.994 \times 0.959\right)}{2 \times \left(1 - (0.959)^2\right)} = 0.527$$

$$b_{13.2} = \frac{\sigma_1\left(r_{13} - r_{12}r_{23}\right)}{\sigma_3\left(1 - r_{23}^2\right)}$$

$$= \frac{3.162 \times \left(0.994 - 0.98 \times 0.959\right)}{2.608 \times \left(1 - (0.959)^2\right)} = 0.818$$

Thus, regression equation of x_1 on x_2 and x_3 is

$$x_1 = 5.276x_2 + 0.818x_3$$

After substituting the value of x_1, x_2 and x_3, we will get the following regression equation of X_1 on X_2 and X_3 is

$$\left(X_1 - 5\right) = 0.527\left(X_2 - 6\right) + 0.818\left(X_3 - 7\right)$$

$$\Rightarrow X_1 = -3.891 + 5.276\ X_2 + 0.818\ X_3$$

(ii) Substituting $X_2 = 45$ and $X_3 = 8$ in regression equation

$$\Rightarrow X_1 = -3.891 + 5.276 X_2 + 0.818\ X_3$$

we get estimated value of X_1 i.e. $X_1 = 26.38$

Let us solve some exercises.

E1) For the data given in the following table find out

(i) Regression equation of X_1 on X_2 and X_3.

(ii) Estimate the value of the value of X_1 for $X_2 = 6$ and $X_3 = 8$.

X_1	2	6	8	10
X_2	4	5	9	12
X_3	4	6	10	12

10.4 PROPERTIES OF RESIDUALS

Property 1: The sum of the product of a variate and a residual is zero if the subscript of the variate occurs among the secondary subscripts of the residual, i.e. $\sum x_2 e_{1.23} = 0$. Here, subscript of the variate x i.e. 2 is appearing in the second subscripts of the $e_{1.23}$.

Proof: If the regression equation of $x_1 \, \text{on} \, x_2 \, \text{and} \, x_3$ is

$$x_1 = b_{12.3} x_2 + b_{13.2} x_3$$

Here, x_1, x_2 and x_3 are measured from their respective means.

Using equation (10) in equations (11) and (13) we have following normal equations

$$\sum x_2 e_{1.23} = 0 \;\; = \sum x_3 e_{1.23}$$

Similarly, normal equation for regression lines of x_2 on x_1 and x_3 & x_3 on x_2 and x_1 are

$$\sum x_1 e_{2.13} = 0 = \sum x_3 e_{2.13}$$

$$\sum x_2 e_{3.12} = 0 = \sum x_1 e_{3.12}$$

Property 2: The sum of the product of two residuals is zero provided all the subscripts, primary as well as secondary, are appearing among the secondary subscripts of second residual i.e. $\sum x_{3.2} e_{1.23} = 0$, since primary as well as secondary subscripts (3 and 2) of the first residual is appearing among the secondary subscripts of the second residual.

Proof: We have

$$\sum x_{3.2} e_{1.23} = \sum (x_3 - b_{32} x_2) e_{1.23}$$

$$= \sum (x_3 e_{1.23} - b_{32} x_2 e_{1.23}) \; = 0$$

(From Property 1: $\sum x_3 e_{1.23} = 0$ and $\sum x_2 e_{1.23} = 0$)

Similarly,

$$\sum x_{2.3} e_{1.23} = 0$$

Property 3: The sum of the product of any two residuals is unaltered if all the secondary subscript of the first occur among the secondary subscripts of the second and we omit any or all of the secondary subscripts of the first.

Proof: We have

$$\sum x_{1.2}e_{1.23} = \sum x_1 e_{1.23},$$

i.e. Right hand side and left hand side are equal even

$$\sum x_{1.2}e_{1.23} = \sum (x_1 - b_{12}x_2)e_{1.23}$$

$$= \sum (x_1 e_{1.23} - b_{12}x_2 e_{1.23})$$

$$= \sum x_1 e_{1.23}$$

(From Property 1, $\sum x_2 e_{1.23} = 0$)

Now let us do some little exercises.

E2) Show that $\sum x_2 e_{1.23} = 0$.

E3) Show that $\sum x_3 e_{1.23} = 0$.

10.5 VARIANCE OF THE RESIDUALS

Let x_1, x_2 and x_3 be three random variables then plane of regression of x_1 on x_2 and x_3 is defined as

$$x_1 = a + b_{12\,.\,3}x_2 + b_{13\,.\,2}x_3$$

Since x_1 x_2 and x_3 are measured from their respective means so

$\sum x_1 = \sum x_2 = \sum x_3 = 0$ and we get a = 0 and regression equation becomes

$$x_1 = b_{12.3}x_2 + b_{13.2}x_3$$

and error of the estimate or residual is (See Section 10.2)

$$e_{1.23} = x_1 - b_{12.3}x_2 - b_{13.2}x_3$$

Now the variance of the residual is denoted by $\sigma^2_{1.23}$ and defined as

$$\sigma^2_{1.23} = \frac{1}{N}\sum (e_{1.23} - \bar{e}_{1.23})^2$$

$$\bar{e}_{1.23} = \bar{x}_1 - b_{12\,.\,3}\bar{x}_2 + b_{13\,.\,2}\bar{x}_3 = 0 \text{ because } \sum x_1 = \sum x_2 = \sum x_3 = 0$$

and

$$\sigma^2_{1.23} = \frac{1}{N}\sum e^2_{1.23} \qquad ...(29)$$

$$= \frac{1}{N}\sum (x_1 - b_{12.3}x_2 - b_{13.2}x_3)(x_1 - b_{12.3}x_2 - b_{13.2}x_3)$$

$$= \frac{1}{N}\sum x_1(x_1 - b_{12.3}x_2 - b_{13.2}x_3)$$

$$-\frac{1}{N}\sum b_{12.3}x_2(x_1 - b_{12.3}x_2 - b_{13.2}x_3)$$

$$-\frac{1}{N}\sum b_{13.2}x_3(x_1 - b_{12.3}x_2 - b_{13.2}x_3)$$

$$\sigma^2_{1.23} = \frac{1}{N}\sum (x_1^2 - b_{12.3}x_1x_2 - b_{13.2}x_1x_3)$$

$$- \frac{1}{N}\sum (b_{12.3}x_2x_1 - b^2_{12.3}x_2^2 - b_{12.3}b_{13.2}x_2x_3)$$

$$-\frac{1}{N}\sum (b_{13.2}x_3x_1 - b_{13.2}b_{12.3}x_3x_2 - b^2_{13.2}x_3^2)$$

We know that

$$b_{12.3}\sum x_2^2 + b_{13.2}\sum x_2x_3 - \sum x_1x_2 = 0$$

and

$$b_{12.3}\sum x_2x_3 + b_{13.2}\sum x_3^2 - \sum x_1x_3 = 0$$

(see equations (12) and (14) of Section 10.3)

Therefore,

$$\sigma^2_{1.23} = \frac{1}{N}\sum x_1^2 - b_{12.3}\frac{1}{N}\sum x_1x_2 - b_{13.2}\frac{1}{N}\sum x_1x_3$$

$$\sigma^2_{1.23} = \sigma_1^2 - b_{1.23}r_{12}\sigma_1\sigma_2 - b_{13.2}r_{13}\sigma_1\sigma_3$$

$$\sigma^2_{1.23} = \sigma_1^2 - \frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2}.\sigma_1^2 r_{12} - \frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2}.\sigma_1^2 r_{13}$$

$$\sigma^2_{1.23} = \frac{\sigma_1^2}{1 - r_{23}^2}(1 - r_{23}^2 - r_{12}^2 - r_{13}^2 + 2r_{12}r_{23}r_{13}) \qquad \ldots (30)$$

10.6 SUMMARY

In this unit, we have discussed:

1. The Yule's notation for trivariate distribution;
2. The plane of regression for trivariate distribution;
3. How to get normal equations for the regression equation of x_1 on x_2 and x_3 ;
4. The properties of residuals;
5. The variance of residuals; and
6. How to find the estimates of dependent variable of regression equations of three variables.

10.7 SOLUTIONS / ANSWERS

E1) (i) Here X_1, X_2 and X_3 are three random variables with their respective means $\overline{X}_1, \overline{X}_2$ and $\overline{X}_3$.

Let $X_1 - \overline{X}_1 = x_1, X_2 - \overline{X}_2 = x_2$ and $X_3 - \overline{X}_3 = x_3$

Then linear regression equation of x_1 on x_2 and x_3 is

$$x_1 = b_{12.3}x_2 + b_{13.2}x_3$$

From equation (22) and (23), we have

$$b_{12.3} = \frac{\sigma_1(r_{12} - r_{13}r_{23})}{\sigma_2(1 - r_{23}^2)}$$

and

$$b_{13.2} = \frac{\sigma_1(r_{13} - r_{12}r_{23})}{\sigma_3(1 - r_{23}^2)}$$

σ_1, σ_2, σ_3, r_{12} r_{13} and r_{23} can be obtained through the following table.

S. No.	X_1	X_2	X_3	$x_1 = X_1 - 6.5$	$x_2 = X_2 - 7.5$	$x_3 = X_3 - 8$	$(x_1)^2$	$(x_2)^2$	$(x_3)^2$	x_1x_2	x_1x_3	x_2x_3
1	2	4	4	−4.5	−3.5	−4	20.25	12.25	16	15.75	18	14
2	6	5	6	−0.5	−2.5	−2	0.25	6.25	4	1.25	1	5
3	8	9	10	1.5	1.5	2	2.25	2.25	4	2.25	3	3
4	10	12	12	3.5	4.5	4	12.25	20.25	16	15.75	14	18
Total	26	30	32	0	0	0	35	41	40	35	36	40

$$\overline{X}_1 = \frac{\sum X_1}{N} = \frac{26}{4} = 6.5$$

$$\overline{X}_2 = \frac{\sum X_2}{N} = \frac{30}{4} = 7.5$$

$$\overline{X}_3 = \frac{\sum X_3}{N} = \frac{32}{4} = 8$$

$$\sigma_1^2 = \frac{1}{N}\sum x_1^2 \quad \text{from equation (15)}$$

$$= \frac{1}{4}(35) = 8.75$$

$$\Rightarrow \quad \sigma_1 = \sqrt{35} = 2.958$$

$$\sigma_2^2 = \frac{1}{N}\sum x_2^2$$

$$= \frac{1}{4}(41) = 10.25$$

$$\Rightarrow \quad \sigma_2 = \sqrt{10.25} = 3.202$$

$$\sigma_3^2 = \frac{1}{N}\sum x_3^2$$

$$= \frac{1}{4}(40) = 10$$

$$\Rightarrow \quad \sigma_3 = \sqrt{10} = 3.162$$

$$r_{12} = \frac{\sum x_1 x_2}{N\sigma_1\sigma_2}$$ from equation (17)

$$= \frac{35}{4 \times 2.958 \times 3.202} = 0.924$$

$$r_{13} = \frac{\sum x_1 x_3}{N\sigma_1\sigma_3}$$

$$= \frac{36}{4 \times 2.958 \times 3.162} = 0.962$$

$$r_{23} = \frac{\sum x_2 x_3}{N\sigma_2\sigma_3}$$

$$= \frac{40}{4 \times 3.202 \times 3.162} = 0.988$$

Now, we have

$$b_{12.3} = \frac{\sigma_1 (r_{12} - r_{13} r_{23})}{\sigma_2 (1 - r_{23}^2)}$$

$$= \frac{2.958 \times (0.924 - 0.962 \times 0.988)}{3.202 \times \{1 - (0.988)^2\}} = -1$$

$$b_{13.2} = \frac{\sigma_1 (r_{13} - r_{12} r_{23})}{\sigma_3 (1 - r_{23}^2)}$$

$$= \frac{2.958 \times (0.962 - 0.924 \times 0.988)}{3.162 \times \{1 - (0.988)^2\}} = 1.9$$

Thus, regression equation of x_1 on x_2 and x_3 is

$$x_1 = -x_2 + 1.9\,x_3$$

After substituting the value of x_1, x_2 and x_3, we will get the following regression equation of X_1on X_2 and X_3 is

$$(X_1 - 6.5) = -(X_2 - 7.5) + 1.9(X_3 - 8)$$

$$\Rightarrow X_1 = -1.2 - X_2 + 1.9X_3$$

(ii) Substituting $X_2 = 6$ and $X_3 = 8$ in regression equation

$$\Rightarrow X_1 = -1.2 - X_2 + 1.9X_3$$

we get estimated value of X_1 i.e. $X_1 = 8$

E2) **Hint:** According to the property 1: $\sum x_2 e_{1.23} = 0$, since subscript of the variate x_2 i.e. 2 is appearing in the second subscript of $e_{1.23}$ i.e. in 23.

.

E3) **Hint:** According to the property 1: $\sum x_3 e_{1.23} = 0$, since subscript of the variate x_3 i.e. 3 is appearing in the second subscript of $e_{1.23}$ i.e. in 23.

UNIT 11 MULTIPLE CORRELATION

Structure

11.1 Introduction

Objectives

11.2 Coefficient of Multiple Correlation

11.3 Properties of Multiple Correlation Coefficient

11.4 Summary

11.5 Solutions / Answers

11.1 INTRODUCTION

In Unit 9, you have studied the concept of regression and linear regression. Regression coefficient was also discussed with its properties. You learned how to determine the relationship between two variables in regression and how to predict value of one variable from the given value of the other variable. Plane of regression for trivariate, properties of residuals and variance of the residuals were discussed in Unit 10 of this block, which are basis for multiple and partial correlation coefficients. In Block 2, you have studied the coefficient of correlation that provides the degree of linear relationship between the two variables.

If we have more than two variables which are interrelated in someway and our interest is to know the relationship between one variable and set of others. This leads us to multiple correlation study.

In this unit, you will study the multiple correlation and multiple correlation coefficient with its properties .To understand the concept of multiple correlation you must be well versed with correlation coefficient. Before starting this unit, you go through the correlation coefficient given in Unit 6 of the Block 2. You should also clear the basics given in Unit 10 of this block to understand the mathematical formulation of multiple correlation coefficients.

Section 11.2 discusses the concept of multiple correlation and multiple correlation coefficient. It gives the derivation of the multiple correlation coefficient formula. Properties of multiple correlation coefficients are described in Section 11.3

Objectives

After reading this unit, you would be able to

- describe the concept of multiple correlation;
- define multiple correlation coefficient;
- derive the multiple correlation coefficient formula; and
- explain the properties of multiple correlation coefficient.

11.2 COEFFICIENT OF MULTIPLE CORRELATION

If information on two variables like height and weight, income and expenditure, demand and supply, etc. are available and we want to study the linear relationship between two variables, correlation coefficient serves our purpose which provides the strength or degree of linear relationship with direction whether it is positive or negative. But in biological, physical and social sciences, often data are available on more than two variables and value of one variable seems to be influenced by two or more variables. For example, crimes in a city may be influenced by illiteracy, increased population and unemployment in the city, etc. The production of a crop may depend upon amount of rainfall, quality of seeds, quantity of fertilizers used and method of irrigation, etc. Similarly, performance of students in university exam may depend upon his/her IQ, mother's qualification, father's qualification, parents income, number of hours of studies, etc. Whenever we are interested in studying the joint effect of two or more variables on a single variable, multiple correlation gives the solution of our problem.

In fact, multiple correlation is the study of combined influence of two or more variables on a single variable.

Suppose, X_1, X_2 and X_3 are three variables having observations on N individuals or units. Then multiple correlation coefficient of X_1 on X_2 and X_3 is the simple correlation coefficient between X_1 and the joint effect of X_2 and X_3. It can also be defined as the correlation between X_1 and its estimate based on X_2 and X_3.

Multiple correlation coefficient is the simple correlation coefficient between a variable and its estimate.

Let us define a regression equation of X_1 on X_2 and X_3 as

$$X_1 = a + b_{12.3}X_2 + b_{13.2}X_3$$

Let us consider three variables x_1, x_2 and x_3 measured from their respective means. The regression equation of x_1 depends upon x_2 and x_3 is given by

$$x_1 = b_{12.3}x_2 + b_{13.2}x_3 \qquad \ldots (1)$$

Where $X_1 - \overline{X}_1 = x_1, X_2 - \overline{X}_2 = x_2$ and $X_3 - \overline{X}_3 = x_3$

$$\therefore \sum x_1 = \sum x_2 = \sum x_3 = 0$$

Right hand side of equation (1) can be considered as expected or estimated value of x_1 based on x_2 and x_3 which may be expressed as

$$x_{1.23} = b_{12.3}x_2 + b_{13.2}x_3 \qquad \ldots (2)$$

Residual $e_{1.23}$ (see definition of residual in Unit 5 of Block 2 of MST 002) is written as

$$e_{1.23} = x_1 - b_{12.3}x_2 - b_{13.2}x_3 = x_1 - x_{1.23}$$

$$\Rightarrow e_{1.23} = x_1 - x_{1.23}$$

$$\Rightarrow x_{1.23} = x_1 - e_{1.23} \qquad \ldots (3)$$

The multiple correlation coefficient can be defined as the simple correlation coefficient between x_1 and its estimate $e_{1.23}$. It is usually denoted by $R_{1.23}$ and defined as

$$R_{1.23} = \frac{Cov(x_1, x_{1.23})}{\sqrt{V(x_1)V(x_{1.23})}} \qquad \ldots (4)$$

Now,

$$Cov(x_1, x_{1.23}) = \frac{1}{N}\sum(x_1 - \overline{x}_1)(x_{1.23} - \overline{x}_{1.23})$$

(By the definition of covariance)

Since, x_1, x_2 and x_3 are measured from their respective means, so

$$\sum x_1 = \sum x_2 = \sum x_3 = 0 \Rightarrow \overline{x}_1 = \overline{x}_2 = \overline{x}_3 = 0$$

and consequently

$$\overline{x}_{1.23} = b_{12.3}\overline{x}_2 + b_{13.2}\overline{x}_3 = 0 \qquad \text{(From equation (2))}$$

Thus,

$$Cov(x_1, x_{1.23}) = \frac{1}{N}\sum x_1 x_{1.23}$$

$$= \frac{1}{N}\sum x_1(x_1 - e_{1.23}) \qquad \text{(From equation (3))}$$

$$= \frac{1}{N}\sum x_1^2 - \frac{1}{N}\sum x_1 e_{1.23} \qquad \text{(By third property of residuals)}$$

$$= \frac{1}{N}\sum x_1^2 - \frac{1}{N}\sum e_{1.23}^2$$

$$= \sigma_1^2 - \sigma_{1.23}^2 \qquad \text{(From equation (29) of Unit10)}$$

Now $V(x_{1.23}) = \frac{1}{N}\sum(x_{1.23} - \overline{x}_{1.23})^2$

$$= \frac{1}{N}\sum(x_{1.23})^2 \qquad \text{(Since } \overline{x}_{1.23} = 0\text{)}$$

$$= \frac{1}{N}\sum(x_1 - e_{1.23})^2 \qquad \text{(From equation (3))}$$

$$= \frac{1}{N}\sum(x_1^2 + e_{1.23}^2 - 2x_1 e_{1.23})$$

$$= \frac{1}{N}\sum x_1^2 + \frac{1}{N}\sum e_{1.23}^2 - 2\frac{1}{N}\sum x_1 e_{1.23}$$

$$= \frac{1}{N}\sum x_1^2 + \frac{1}{N}\sum e_{1.23}^2 - 2\frac{1}{N}\sum e_{1.23}^2$$

(By third property of residuals)

$$= \frac{1}{N}\sum x_1^2 - \frac{1}{N}\sum e_{1.23}^2$$

$$V(x_{1.23}) = \sigma_1^2 - \sigma_{1.23}^2 \qquad \text{(From equation (29) of Unit 10)}$$

Substituting the value of $Cov(x_1, x_{1.23})$ and $V(x_{1.23})$ in equation (4), we have

$$R_{1.23} = \frac{\sigma_1^2 - \sigma_{1.23}^2}{\sqrt{\sigma_1^2(\sigma_1^2 - \sigma_{1.23}^2)}}$$

$$R_{1.23}^2 = \frac{(\sigma_1^2 - \sigma_{1.23}^2)^2}{\sigma_1^2(\sigma_1^2 - \sigma_{1.23}^2)}$$

$$R_{1.23}^2 = \frac{\sigma_1^2 - \sigma_{1.23}^2}{\sigma_1^2} = 1 - \frac{\sigma_{1.23}^2}{\sigma_1^2}$$

here, $\sigma_{1.23}^2$ is the variance of residual, which is

$$\sigma_{1.23}^2 = \frac{\sigma_1^2}{1 - r_{23}^2}(1 - r_{23}^2 - r_{12}^2 - r_{13}^2 + 2r_{12}r_{23}r_{13})$$

(From equation (30) of unit 10)

Then,

$$R_{1.23}^2 = 1 - \frac{\sigma_1^2(1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2r_{12}r_{13}r_{23})}{\sigma_1^2(1 - r_{23}^2)}$$

$$R_{1.23}^2 = 1 - \frac{1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}$$

$$R_{1.23}^2 = \frac{1 - r_{23}^2 - 1 + r_{12}^2 + r_{13}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}$$

$$R_{1.23}^2 = \frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}$$

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}} \qquad \ldots (5)$$

which is required formula for multiple correlation coefficient.

where, r_{12} is the total correlation coefficient between variable X_1 and X_2,

r_{23} is the total correlation coefficient between variable X_2 and X_3,

r_{13} is the total correlation coefficient between variable X_1 and X_3.

Now let us solve a problem on multiple correlation coefficients.

Example 1: From the following data, obtain $R_{1.23}$ and $R_{2.13}$

X_1	65	72	54	68	55	59	78	58	57	51
X_2	56	58	48	61	50	51	55	48	52	42
X_3	9	11	8	13	10	8	11	10	11	7

Solution: To obtain multiple correlation coefficients $R_{1.23}$ and $R_{2.13}$, we use following formulae

$$R^2_{1.23} = \frac{r^2_{12} + r^2_{13} - 2r_{12}r_{13}r_{23}}{1 - r^2_{23}} \quad \text{and}$$

$$R^2_{2.13} = \frac{r^2_{12} + r^2_{23} - 2r_{12}r_{13}r_{23}}{1 - r^2_{13}}$$

We need r_{12}, r_{13} and r_{23} which are obtained from the following table:

S. No.	X_1	X_2	X_3	$(X_1)^2$	$(X_2)^2$	$(X_3)^2$	X_1X_2	X_1X_3	X_2X_3
1	65	56	9	4225	3136	81	3640	585	504
2	72	58	11	5184	3364	121	4176	792	638
3	54	48	8	2916	2304	64	2592	432	384
4	68	61	13	4624	3721	169	4148	884	793
5	55	50	10	3025	2500	100	2750	550	500
6	59	51	8	3481	2601	64	3009	472	408
7	78	55	11	6084	3025	121	4290	858	605
8	58	48	10	3364	2304	100	2784	580	480
9	57	52	11	3249	2704	121	2964	627	572
10	51	42	7	2601	1764	49	2142	357	294
Total	617	521	98	38753	27423	990	32495	6137	5178

Now we get the total correlation coefficient r_{12}, r_{13} and r_{23}

$$r_{12} = \frac{N(\sum X_1X_2) - (\sum X_1)(\sum X_2)}{\sqrt{\{N(\sum X_1^2) - (\sum X_1)^2\}\{N(\sum X_2^2) - (\sum X_2)^2\}}}$$

$$r_{12} = \frac{(10 \times 32495) - (617) \times (521)}{\sqrt{\{(10 \times 38753) - (617) \times (617)\}\{(10 \times 27423) - (521) \times (521)\}}}$$

$$r_{12} = \frac{3493}{\sqrt{\{6841\}\times\{2789\}}} = \frac{3493}{4368.01} = 0.80$$

$$r_{13} = \frac{N(\sum X_1X_3) - (\sum X_1)(\sum X_3)}{\sqrt{\{N(\sum X_1^2) - (\sum X_1)^2\}\{N(\sum X_3^2) - (\sum X_3)^2\}}}$$

$$r_{13} = \frac{(10\times 6137) - (617)\times(98)}{\sqrt{\{(10\times 38753) - (617\times 617)\}\{(10\times 990) - (98\times 98)\}}}$$

$$r_{13} = \frac{904}{\sqrt{\{6841\}\{296\}}} = \frac{904}{1423.00} = 0.64$$

and

$$r_{23} = \frac{N(\sum X_2X_3) - (\sum X_2)(\sum X_3)}{\sqrt{\{N(\sum X_2^2) - (\sum X_2)^2\}\{N(\sum X_3^2) - (\sum X_3)^2\}}}$$

$$r_{23} = \frac{(10\times 5178) - (521)\times(98)}{\sqrt{\{(10\times 27423) - (521\times 521)\}\{(10\times 990) - (98\times 98)\}}}$$

$$r_{23} = \frac{722}{\sqrt{\{2789\}\{296\}}} = \frac{722}{908.59} = 0.79$$

Now, we calculate $R_{1.23}$

We have, $r_{12} = 0.80$, $r_{13} = 0.64$ and $r_{23} = 0.79$, then

$$R^2_{1.23} = \frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}$$

$$= \frac{0.80^2 + 0.64^2 - 2\times 0.80\times 0.64\times 0.79}{1 - 0.79^2}$$

$$= \frac{0.64 + 0.41 - 0.81}{1 - 0.62}$$

$$R^2_{1.23} = \frac{0.24}{0.38} = 0.63$$

Then

$$R_{1.23} = 0.79.$$

$$R^2_{2.13} = \frac{r_{12}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{13}^2}$$

$$= \frac{0.80^2 + 0.79^2 - 2\times 0.80\times 0.64\times 0.79}{1 - 0.64^2}$$

$$= \frac{0.64 + 0.62 - 0.81}{1 - 0.49}$$

$$= \frac{0.45}{0.51} = 0.88$$

Thus,

$$R_{2.13} = 0.94$$

Example 2: From the following data, obtain $R_{1.23}$, $R_{2.13}$ and $R_{3.12}$

X_1	2	5	7	11
X_2	3	6	10	12
X_3	1	3	6	10

Solution: To obtain multiple correlation coefficients $R_{1.23}$ $R_{2.13}$ and $R_{3.12}$, we use following formulae

$$R^2_{1.23} = \frac{r^2_{12} + r^2_{13} - 2r_{12}r_{13}r_{23}}{1 - r^2_{23}} \quad ,$$

$$R^2_{2.13} = \frac{r^2_{12} + r^2_{23} - 2r_{12}r_{13}r_{23}}{1 - r^2_{13}} \quad \text{and}$$

$$R^2_{3.12} = \frac{r^2_{13} + r^2_{23} - 2r_{12}r_{13}r_{23}}{1 - r^2_{12}}$$

We need r_{12} , r_{13} and r_{23} which are obtained from the following table:

S. No.	**X_1**	**X_2**	**X_3**	**$(X_1)^2$**	**$(X_2)^2$**	**$(X_3)^2$**	**X_1X_2**	**X_1X_3**	**X_2X_3**
1	2	3	1	4	9	1	6	2	3
2	5	6	3	25	36	9	30	15	18
3	7	10	6	49	100	36	70	42	60
4	11	12	10	121	144	100	132	110	120
Total	25	31	20	199	289	146	238	169	201

Now we get the total correlation coefficient r_{12} , r_{13} and r_{23}

$$r_{12} = \frac{N(\sum X_1X_2) - (\sum X_1)(\sum X_2)}{\sqrt{\{N(\sum X_1^2) - (\sum X_1)^2\}\{N(\sum X_2^2) - (\sum X_2)^2\}}}$$

$$r_{12} = \frac{(4 \times 238) - (25) \times (31)}{\sqrt{\{(4 \times 199) - (25) \times (25)\}\{(4 \times 289) - (31) \times (31)\}}}$$

$$r_{12} = \frac{177}{\sqrt{\{171\}\{195\}}} = \frac{177}{182.61} = 0.0.97$$

$$r_{13} = \frac{N(\sum X_1X_3) - (\sum X_1)(\sum X_3)}{\sqrt{\{N(\sum X_1^2) - (\sum X_1)^2\}\{N(\sum X_3^2) - (\sum X_3)^2\}}}$$

$$r_{13} = \frac{(4\times 169)-(25)\times(20)}{\sqrt{\{(4\times 199)-(25\times 25)\}\{(4\times 146)-(20\times 20)\}}}$$

$$r_{13} = \frac{176}{\sqrt{\{171\}\{184\}}} = \frac{176}{177.38} = 0.99$$

and

$$r_{23} = \frac{N(\sum X_2X_3)-(\sum X_2)(\sum X_3)}{\sqrt{\{N(\sum X_2^2)-(\sum X_2)^2\}\{N(\sum X_3^2)-(\sum X_3)^2\}}}$$

$$r_{23} = \frac{(4\times 201)-(31)\times(20)}{\sqrt{\{4\times 289)-(31\times 31)\}\{(4\times 146)-(20\times 20)\}}}$$

$$r_{23} = \frac{184}{\sqrt{\{195\}\{184\}}} = \frac{184}{189.42} = 0.97$$

Now, we calculate $R_{1.23}$

We have, $r_{12} = 0.97$, $r_{13} = 0.99$ and $r_{23} = 0.97$, then

$$R^2_{1.23} = \frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1-r_{23}^2}$$

$$= \frac{0.97^2 + 0.99^2 - 2\times 0.97\times 0.99\times 0.97}{1-0.97^2}$$

$$= \frac{0.058}{0.059} = 0.98$$

Then

$$R_{1.23} = 0.99.$$

$$R^2_{2.13} = \frac{r_{12}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1-r_{13}^2}$$

$$= \frac{0.97^2 + 0.97^2 - 2\times 0.97\times 0.99\times 0.97}{1-0.99^2}$$

$$= \frac{0.19}{0.20} = 0.95$$

Thus,

$$R_{2.13} = 0.97$$

$$R^2_{3.12} = \frac{r_{13}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1-r_{12}^2}$$

$$= \frac{0.99^2 + 0.97^2 - 2\times 0.97\times 0.99\times 0.97}{1-0.97^2}$$

$$= \frac{0.58}{0.591}$$

$$= 0.981$$

Thus,

$$R_{3.12} = 0.99$$

Example 3: The following data is given:

X_1	60	68	50	66	60	55	72	60	62	51
X_2	42	56	45	64	50	55	57	48	56	42
X_3	74	71	78	80	72	62	70	70	76	65

Obtain $R_{1.23}$, $R_{2.13}$ and $R_{3.12}$

Solution: To obtain multiple correlation coefficients $R_{1.23}$, $R_{2.13}$ and $R_{3.12}$ we use following formulae:

$$R^2_{1.23} = \frac{r^2_{12} + r^2_{13} - 2r_{12}r_{13}r_{23}}{1 - r^2_{23}},$$

$$R^2_{2.13} = \frac{r^2_{12} + r^2_{23} - 2r_{12}r_{13}r_{23}}{1 - r^2_{13}}$$ and

$$R^2_{3.12} = \frac{r^2_{13} + r^2_{23} - 2r_{12}r_{13}r_{23}}{1 - r^2_{12}}$$

We need r_{12} , r_{13} and r_{23} which are obtained from the following table:

S. No.	**X_1**	**X_2**	**X_3**	**d_1= X_1 -60**	**d_2 = X_2 -50**	**d_3= X_3 -70**	**$(d_1)^2$**	**$(d_2)^2$**	**$(d_3)^2$**	**d_1d_2**	**d_1d_3**	**d_2d_3**
1	60	42	74	0	-8	4	0	64	16	0	0	-32
2	68	56	71	8	6	1	64	36	1	48	8	6
3	50	45	78	-10	-5	8	100	25	64	50	-80	-40
4	66	64	80	6	14	10	36	196	100	84	60	140
5	60	50	72	0	0	2	0	0	4	0	0	0
6	55	55	62	-5	5	-8	25	25	64	-25	40	-40
7	72	57	70	12	7	0	144	49	0	84	0	0
8	60	48	70	0	-2	0	0	4	0	0	0	0
9	62	56	76	2	6	6	4	36	36	12	12	36
10	51	42	65	-9	-8	-5	81	64	25	72	45	40
Total				4	15	18	454	499	310	325	85	110

Here, we can also use shortcut method to calculate r_{12}, r_{13} & r_{23},

Let $d_1 = X_1 - 60$

$d_2 = X_2 - 50$

$d_3 = X_1 - 70$

Now we get the total correlation coefficient r_{12}, r_{13} and r_{23}

$$r_{12} = \frac{N(\sum d_1 d_2) - (\sum d_1)(\sum d_2)}{\sqrt{\{N(\sum d_1^2) - (\sum d_1)^2\}\{N(\sum d_2^2) - (\sum d_2)^2\}}}$$

$$r_{12} = \frac{(10 \times 325) - (4) \times (15)}{\sqrt{\{(10 \times 454) - (4) \times (4)\}\{(10 \times 499) - (15) \times (15)\}}}$$

$$r_{12} = \frac{3190}{\sqrt{\{4524\}\{4765\}}} = \frac{3190}{4642.94} = 0.69$$

$$r_{13} = \frac{N(\sum d_1 d_3) - (\sum d_1)(\sum d_3)}{\sqrt{\{N(\sum d_1^2) - (\sum d_1)^2\}\{N(\sum d_3^2) - (\sum d_3)^2\}}}$$

$$r_{13} = \frac{(10 \times 85) - (4) \times (18)}{\sqrt{\{(10 \times 454) - (4 \times 4)\}\{(10 \times 310) - (18 \times 18)\}}}$$

$$r_{13} = \frac{778}{\sqrt{\{4524\}\{2776\}}} = \frac{778}{3543.81} = 0.22$$

and

$$r_{23} = \frac{N(\sum d_2 d_3) - (\sum d_2)(\sum d_3)}{\sqrt{\{N(\sum d_2^2) - (\sum d_2)^2\}\{N(\sum d_3^2) - (\sum d_3)^2\}}}$$

$$r_{23} = \frac{(10 \times 110) - (15) \times (18)}{\sqrt{\{(10 \times 499) - (15 \times 15)\}\{(10 \times 310) - (18 \times 18)\}}}$$

$$r_{23} = \frac{830}{\sqrt{\{4765\}\{2776\}}} = \frac{830}{3636.98} = 0.23$$

Now, we calculate $R_{1.23}$

We have, $r_{12} = 0.69$, $r_{13} = 0.22$ and $r_{23} = 0.23$, then

$$R_{1.23}^2 = \frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}$$

$$= \frac{0.69^2 + 0.22^2 - 2 \times 0.69 \times 0.22 \times 0.23}{1 - 0.23^2}$$

$$= \frac{0.4547}{0.9471} = 0.4801$$

Then

$$R_{1.23} = 0.69$$

$$R^2_{2.13} = \frac{r^2_{12} + r^2_{23} - 2r_{12}r_{13}r_{23}}{1 - r^2_{13}}$$

$$= \frac{0.69^2 + 0.23^2 - 2 \times 0.69 \times 0.22 \times 0.23}{1 - 0.22^2}$$

$$= \frac{0.4592}{0.9516} = 0.4825$$

Thus,

$$R_{2.13} = 0.69$$

$$R^2_{3.12} = \frac{r^2_{13} + r^2_{23} - 2r_{12}r_{13}r_{23}}{1 - r^2_{12}}$$

$$= \frac{0.22^2 + 0.23^2 - 2 \times 0.69 \times 0.22 \times 0.23}{1 - 0.69^2}$$

$$= \frac{0.0315}{0.5239} = 0.0601$$

Thus,

$$R_{3.12} = 0.25$$

Now let us solve some exercises.

E1) In bivariate distribution, $r_{12} = 0.6$, $r_{23} = r_{31} = 0.54$, then calculate $R_{1.23}$.

E2) If $r_{12} = 0.70$, $r_{13} = 0.74$ and $r_{23} = 0.54$, calculate multiple correlation coefficient $R_{2.13}$.

E3) Calculate multiple correlation coefficients $R_{1.23}$ and $R_{2.13}$ from the following information: $r_{12} = 0.82$, $r_{23} = -0.57$ and $r_{13} = -0.42$.

E4) From the following data,

X_1	22	15	27	28	30	42	40
X_2	12	15	17	15	42	15	28
X_3	13	16	12	18	22	20	12

Obtain $R_{1.23}$, $R_{2.13}$ and $R_{3.12}$

E5) The following data is given:

X_1	50	54	50	56	50	55	52	50	52	51
X_2	42	46	45	44	40	45	43	42	41	42
X_3	72	71	73	70	72	72	70	71	75	71

By using the short-cut method obtain $R_{1.23}$, $R_{2.13}$ and $R_{3.12}$

11.3 PROPERTIES OF MULTIPLE CORRELATION COEFFICIENT

The following are some of the properties of multiple correlation coefficients:

1. Multiple correlation coefficient is the degree of association between observed value of the dependent variable and its estimate obtained by multiple regression,
2. Multiple Correlation coefficient lies between 0 and 1,
3. If multiple correlation coefficient is 1, then association is perfect and multiple regression equation may said to be perfect prediction formula,
4. If multiple correlation coefficient is 0, dependent variable is uncorrelated with other independent variables. From this, it can be concluded that multiple regression equation fails to predict the value of dependent variable when values of independent variables are known,
5. Multiple correlation coefficient is always greater or equal than any total correlation coefficient. If $R_{1.23}$ is the multiple correlation coefficient than $R_{1.23} \geq r_{12}$ or r_{13} or r_{23}, and
6. Multiple correlation coefficient obtained by method of least squares would always be greater than the multiple correlation coefficient obtained by any other method.

11.4 SUMMARY

In this unit, we have discussed:

1. The multiple correlation, which is the study of joint effect of a group of two or more variables on a single variable which is not included in that group,
2. The estimate obtained by regression equation of that variable on other variables,
3. Limit of multiple correlation coefficient, which lies between 0 and +1,
4. The numerical problems of multiple correlation coefficient, and
5. The properties of multiple correlation coefficient.

11.5 SOLUTIONS / ANSWERS

E1) We have,

$$r_{12} = 0.6, \quad r_{23} = r_{31} = 0.54$$

$$R^2_{1.23} = \frac{r^2_{12} + r^2_{13} - 2r_{12}r_{13}r_{23}}{1 - r^2_{23}}$$

$$= \frac{0.36 + 0.29 - 0.35}{0.71}$$

$$= \frac{0.30}{0.71} = 0.42$$

Then

$$R_{1.23} = 0.65$$

E2) We have

$$R^2_{2.13} = \frac{r^2_{12} + r^2_{23} - 2r_{12}r_{13}r_{23}}{1 - r^2_{13}}$$

$$= \frac{0.49 + 0.29 - 0.56}{1 - 0.55}$$

$$= \frac{0.22}{0.45} = 0.49$$

Thus

$$R_{2.13} = 0.70.$$

E3) We have

$r_{12} = 0.82$, $r_{23} = -0.57$ $r_{13} = -0.42$.

$$R^2_{1.23} = \frac{r^2_{12} + r^2_{13} - 2r_{12}r_{13}r_{23}}{1 - r^2_{23}}$$

$$= \frac{0.67 + 0.18 - 0.39}{0.68}$$

$$= \frac{0.46}{0.68} = 0.68$$

Then

$$R_{1.23} = 0.82$$

$$R^2_{2.13} = \frac{r^2_{12} + r^2_{23} - 2r_{12}r_{13}r_{23}}{1 - r^2_{13}}$$

$$= \frac{0.67 + 0.32 - 0.39}{0.82}$$

$$= \frac{0.60}{0.82} = 0.73$$

Thus,

$$R_{2.13} = 0.85.$$

E4) To obtain multiple correlation coefficients $R_{1.23}$, $R_{2.13}$ and $R_{3.12}$ we use following formulae:

$$R^2_{1.23} = \frac{r^2_{12} + r^2_{13} - 2r_{12}r_{13}r_{23}}{1 - r^2_{23}} ,$$

$$R^2_{2.13} = \frac{r^2_{12} + r^2_{23} - 2r_{12}r_{13}r_{23}}{1 - r^2_{13}} \text{ and}$$

$$R^2_{3.12} = \frac{r^2_{13} + r^2_{23} - 2r_{12}r_{13}r_{23}}{1 - r^2_{12}}$$

We need r_{12}, r_{13} and r_{23} which are obtained from the following table:

S. No.	**X_1**	**X_2**	**X_3**	**$(X_1)^2$**	**$(X_2)^2$**	**$(X_3)^2$**	**X_1X_2**	**X_1X_3**	**X_2X_3**
1	22	12	13	484	144	169	264	286	156
2	15	15	16	225	225	256	225	240	240
3	27	17	12	729	289	144	459	324	204
4	28	15	18	784	225	324	420	504	270
5	30	42	22	900	1764	484	1260	660	924
6	42	15	20	1764	225	400	630	840	300
7	40	28	12	1600	784	144	1120	480	336
Total	204	144	113	6486	3656	1921	4378	3334	2430

Now, we get the total correlation coefficient r_{12}, r_{13} and r_{23}

$$r_{12} = \frac{N(\sum X_1X_2) - (\sum X_1)(\sum X_2)}{\sqrt{\left\{N(\sum X_1^2) - (\sum X_1)^2\right\}\left\{N(\sum X_2^2) - (\sum X_2)^2\right\}}}$$

$$r_{12} = \frac{(7 \times 4378) - (204) \times (144)}{\sqrt{\{(7 \times 6486) - (204) \times (204)\}\{(7 \times 3656) - (144) \times (144)\}}}$$

$$r_{12} = \frac{1270}{\sqrt{\{3786\}\{4856\}}}$$

$$= \frac{1270}{4287.75} = 0.30$$

$$r_{13} = \frac{N(\sum X_1X_3) - (\sum X_1)(\sum X_3)}{\sqrt{\left\{N(\sum X_1^2) - (\sum X_1)^2\right\}\left\{N(\sum X_3^2) - (\sum X_3)^2\right\}}}$$

$$r_{13} = \frac{(7 \times 3334) - (204) \times (113)}{\sqrt{\{(7 \times 6486) - (204 \times 204)\}\{(7 \times 1921) - (113 \times 113)\}}}$$

$$r_{13} = \frac{286}{\sqrt{3786 \times 678}}$$

$$= \frac{286}{1602.16} = 0.18$$

and

$$r_{23} = \frac{N(\sum X_2 X_3) - (\sum X_2)(\sum X_3)}{\sqrt{\left\{N(\sum X_2^2) - (\sum X_2)^2\right\}\left\{N(\sum X_3^2) - (\sum X_3)^2\right\}}}$$

$$r_{23} = \frac{(7 \times 2430) - (144) \times (113)}{\sqrt{\{7 \times 3656) - (144 \times 144)\}\ \{(7 \times 1921) - (113 \times 113)\}}}$$

$$r_{23} = \frac{738}{\sqrt{\{4856\}\{678\}}}$$

$$= \frac{738}{1814.49} = 0.41$$

Now, we calculate $R_{1.23}$

We have, $r_{12} = 0.30$, $r_{13} = 0.18$ and $r_{23} = 0.41$, then

$$R^2_{1.23} = \frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}$$

$$= \frac{0.30^2 + 0.18^2 - 2 \times .30 \times 0.18 \times 0.41}{1 - (0.41)^2}$$

$$= \frac{0.0781}{0.8319} = 0.9380$$

Then

$$R_{1.23} = 0.30.$$

$$R^2_{2.13} = \frac{r_{12}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{13}^2}$$

$$= \frac{0.30^2 + 0.41^2 - 2 \times .30 \times 0.18 \times 0.41}{1 - 0.18^2}$$

$$= \frac{0.2138}{0.9676} = 0.221$$

Thus,

$$R_{2.13} = 0.47$$

$$R^2_{3.12} = \frac{r_{13}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{12}^2}$$

$$= \frac{0.18^2 + 0.41^2 - 2\times 0.30\times 0.18\times 0.41}{1-0.30^2}$$

$$= \frac{0.1562}{0.9100} = 0.1717$$

Thus,

$$R_{3.12} = 0.41$$

E5) To obtain multiple correlation coefficients $R_{1.23}$, $R_{2.13}$ and $R_{3.12}$ we use following formulae

$$R^2_{1.23} = \frac{r^2_{12} + r^2_{13} - 2r_{12}r_{13}r_{23}}{1-r^2_{23}} \quad ,$$

$$R^2_{3.12} = \frac{r^2_{13} + r^2_{23} - 2r_{12}r_{13}r_{23}}{1-r^2_{12}}$$

and

$$R^2_{2.13} = \frac{r^2_{12} + r^2_{23} - 2r_{12}r_{13}r_{23}}{1-r^2_{13}}$$

We need r_{12} , r_{13} and r_{23} which are obtained from the following table:

S. No.	X_1	X_2	X_3	$d_1 = X_1 - 50$	$d_2 = X_2 - 40$	$d_3 = X_3 - 70$	$(d_1)^2$	$(d_2)^2$	$(d_3)^2$	d_1d_2	d_1d_3	d_2d_3
1	50	42	72	0	2	2	0	4	4	0	0	4
2	54	46	71	4	6	1	16	36	1	24	4	6
3	50	45	73	0	5	3	0	25	9	0	0	15
4	56	44	70	6	4	0	36	16	0	24	0	0
5	50	40	72	0	0	2	0	0	4	0	0	0
6	55	45	72	5	5	2	25	25	4	25	10	10
7	52	43	70	2	3	0	4	9	0	6	0	0
8	50	42	71	0	2	1	0	4	1	0	0	2
9	52	41	75	2	1	5	4	1	25	2	10	5
10	51	42	71	1	2	1	1	4	1	2	1	2
Total				20	30	17	86	124	49	83	25	44

Now, we get the total correlation coefficient r_{12} , r_{13} and r_{23}

$$r_{12} = \frac{N(\sum d_1d_2) - (\sum d_1)(\sum d_2)}{\sqrt{\left\{N(\sum d_1^2) - (\sum d_1)^2\right\}\left\{N(\sum d_2^2) - (\sum d_2)^2\right\}}}$$

$$r_{12} = \frac{(10\times 83) - (20)\times(30)}{\sqrt{\left\{(10\times 86) - (20)\times(20)\right\}\left\{(10\times 124) - (30)\times(30)\right\}}}$$

$$r_{12} = \frac{230}{\sqrt{460 \times 340}}$$

$$= \frac{230}{395.47} = 0.58$$

$$r_{13} = \frac{N(\sum d_1 d_3) - (\sum d_1)\ (\sum d_3)}{\sqrt{\left\{N(\sum d_1^2) - (\sum d_1)^2\right\}\left\{N(\sum d_3^2) - (\sum d_3)^2\right\}}}$$

$$r_{13} = \frac{(10 \times 25) - (20) \times (17)}{\sqrt{\{(10 \times 86) - (20 \times 20)\}\ \{(10 \times 49) - (17 \times 17)\}}}$$

$$r_{13} = \frac{-90}{\sqrt{\{460\}\{201\}}}$$

$$= \frac{-90}{304.07} = -\ 0.30$$

and

$$r_{23} = \frac{N(\sum X_2 X_3) - (\sum X_2)(\sum X_3)}{\sqrt{\left\{N(\sum X_2^2) - (\sum X_2)^2\right\}\left\{N(\sum X_3^2) - (\sum X_3)^2\right\}}}$$

$$r_{23} = \frac{(10 \times 44) - (30) \times (17)}{\sqrt{\{(10 \times 124) - (20 \times 20)\}\{(10 \times 49) - (17 \times 17)\}}}$$

$$r_{23} = \frac{-70}{\sqrt{\{340\}\{201\}}}$$

$$= \frac{-70}{261.42} = -\ 0.27$$

Now, we calculate $R_{1.23}$

We have, $r_{12} = 0.58$, $r_{13} = -\ 0.30$ and $r_{23} = -\ 0.27$, then

$$R_{1.23}^2 = \frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}$$

$$= \frac{0.58^2 + (-\ 0.30)^2 - 2 \times 0.58 \times (-\ 0.30) \times (-0.27)}{1 - (-\ 0.27)^2}$$

$$= \frac{0.3324}{0.9271} = 0.36$$

Then

$$R_{1.23} = 0.60\,.$$

$$R_{2.13}^2 = \frac{r_{12}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{13}^2}$$

$$= \frac{0.58^2 + (-0.27)^2 - 2 \times 0.58 \times (-0.30) \times (-0.27)}{1 - (-0.30)^2}$$

$$= \frac{0.3153}{0.9100} = 0.35$$

Thus,

$$R_{2.13} = 0.59$$

$$R^2_{3.12} = \frac{r^2_{13} + r^2_{23} - 2r_{12}r_{13}r_{23}}{1 - r^2_{12}}$$

$$= \frac{(-0.30)^2 + (-0.27)^2 - 2 \times 0.58 \times (-0.30) \times (-0.27)}{1 - (0.58)^2}$$

$$= \frac{0.0689}{0.6636} = 0.10$$

Thus,

$$R_{3.12} = 0.32$$

UNIT 12 PARTIALCORRELATION

Structure

12.1 Introduction

Objectives

12.2 Coefficient of Partial Correlation

12.3 Multiple Correlation Coefficient in Terms of Total and Partial Correlation Coefficients

12.4 Summary

12.5 Solutions / Answers

12.1 INTRODUCTION

In Unit 11 of this block, you studied the concept of multiple correlation and multiple correlation coefficient with its properties. In this unit, you will study the partial correlation. To understand the mathematical formulation of partial correlation coefficient, you go through the Unit 10 of this block. You will learn also how to derive the multiple correlation coefficients in terms of total and partial correlation coefficients.

Section 12.2 discusses the concept of partial correlation and derivation of partial correlation coefficient formula. Multiple correlation coefficients in terms of total and partial correlation coefficients are expressed in Section 12.3.

Objectives

After reading this unit, you will be able to

- describe the concept of partial correlation;
- derive the partial correlation coefficient formula; and
- derscribe multiple correlation coefficient in terms of total and partial correlation coefficients.

12.2 COEFFICIENT OF PARTIAL CORRELATION

As we have seen in Unit 11 of this block that multiple correlation studies the joint effect of group of variables on a single variable and multiple correlation coefficient provides a degree of association between a variable and its estimate. Many times correlation between two variables is partly due to the third variable. For example correlation between height and weight is due to age. In such situations, one may be interested to know the relationship between two variables ignoring the effect of third and fourth or more other variables. Partial correlation studies this type of situations.

In fact, partial correlation is the correlation between two variables, after removing the linear effects of other variables on them.

Let us consider the case of three variables X_1, X_2 and X_3. Sometimes the correlation between two variables X_1 and X_2 may be partly due to the correlation of a third variable X_3 with both X_1 and X_2. In this type of situation one may be interested to study the correlation between X_1 and X_2 when the effect of X_3 on each of X_1 and X_2 is eliminated. This correlation is known as partial correlation. The correlation coefficient between X_1 and X_2 after eliminating the linear effect of X_3 on X_1 and X_2 is called the partial correlation coefficient.

If we consider the regression equation of X_1 on X_3 i.e.

$$X_1 = a + b_{13} X_3$$

and suppose three variables x_1, x_2 and x_3 are measured from their respective means i.e.

$$X_1 - \overline{X}_1 = x_1, X_2 - \overline{X}_2 = x_2 \text{ and } X_3 - \overline{X}_3 = x_3$$

then the regression equation of x_1 on x_3 is given by $x_1 = b_{13} x_3$.

The residual $e_{1.3}$ for x_1 can be expressed as

$$e_{1.3} = x_1 - b_{13} x_3 \qquad \dots (1)$$

Equation (1) may be considered as a part of the dependent variable x_1 which remains when the linear effect of x_3 on x_1 is eliminated.

Similarly, the regression equation x_2 on x_3 i.e. $x_2 = b_{23} x_3$ then the residual $e_{2.3}$ is expressed as

$$e_{2.3} = x_2 - b_{23} x_3 \qquad \dots (2)$$

which may be considered as a part of the dependent variable x_2, which remains when the linear effect of x_3 on x_2 is eliminated. Thus, the correlation between $e_{1.3}$ and $e_{2.3}$ is considered as the partial correlation coefficient.

12.2.1 Derivation of Partial Correlation Coefficient Formula

Partial correlation coefficient is the correlation coefficient between two variables after removing the linear effect of other variables on them.

If there are three variables x_1, x_2 and x_3 then partial correlation coefficient between x_1 and x_2 is denoted by $r_{12.3}$ and defined by

$$r_{12.3} = \frac{Cov(e_{1.3}, e_{2.3})}{\sqrt{V(e_{1.3)})V(e_{2.3})}} \qquad \dots (3)$$

We know that

$$Cov(e_{1.3}, e_{2.3}) = \frac{1}{N} \sum (e_{1.3} - \overline{e}_{1.3})(e_{2.3} - \overline{e}_{2.3})$$

Since x_1, x_2 and x_3 are measured from their respective means so

$\sum x_1 = \sum x_2 = \sum x_3 = 0 \Rightarrow \bar{e}_{1.3} = 0 = \bar{e}_{2.3}$ (See equations (1) and (2))

So,

$$\begin{aligned}
\text{Cov}(e_{1.3}, e_{2.3}) &= \frac{1}{N}\sum e_{1.3}e_{2.3} \\
&= \frac{1}{N}\sum (x_1 - b_{13}x_3)(x_2 - b_{23}x_3)
\end{aligned}$$

From equations (1) and (2)

$$\begin{aligned}
&= \frac{1}{N}\sum x_1x_2 - b_{13}\frac{1}{N}\sum x_2x_3 \\
&\quad - b_{23}\frac{1}{N}\sum x_1x_3 + b_{13}b_{23}\frac{1}{N}\sum x_3^2 \\
\text{Cov}(e_{1.3}, e_{2.3}) &= r_{12}\sigma_1\sigma_2 - b_{13}r_{23}\sigma_2\sigma_3 - b_{23}r_{13}\sigma_1\sigma_3 + b_{13}b_{23}\sigma_3^2
\end{aligned}$$

Explanatory Note:

$\text{Cov}(x_1, x_2) = \frac{1}{N}\sum (x_1 - \bar{x}_1)(x_2 - \bar{x}_2)$

Since x_1, x_2 and x_3 are measured from their respective means so

$\sum x_1 = \sum x_2 = 0 \Rightarrow \bar{x}_1 = \bar{x}_2 = 0$

$\text{Cov}(x_1, x_2) = \frac{1}{N}\sum x_1x_2$

So

$\Rightarrow \frac{1}{N}\sum x_1x_2 = \text{Cov}(x_1, x_2) = r_{12}\sigma_1\sigma_2$

Similarly $\Rightarrow \frac{1}{N}\sum x_2x_3 = \text{Cov}(x_2, x_3) = r_{23}\sigma_2\sigma_3$

We know that $\sigma_3^2 = \frac{1}{N}\sum (x_3 - \bar{x}_3)^2 = \frac{1}{N}\sum x_3^2$. Similarly, other expressions can be obtained.

We know that the simple regression coefficient of x_1 on x_3 is $b_{13} = r_{13}\frac{\sigma_1}{\sigma_3}$

similarly, the regression coefficient of x_2 on x_3 is

$$b_{23} = r_{23}\frac{\sigma_2}{\sigma_3}$$

So

$$\begin{aligned}
\text{Cov}(e_{1.3}, e_{2.3}) &= r_{12}\sigma_1\sigma_2 - r_{13}\frac{\sigma_1}{\sigma_3}r_{23}\sigma_2\sigma_3 - r_{23}\frac{\sigma_2}{\sigma_3}r_{13}\sigma_1\sigma_3 + r_{13}\frac{\sigma_1}{\sigma_3}r_{23}\frac{\sigma_2}{\sigma_3}\sigma_3^2 \\
\text{Cov}(e_{1.3}, e_{2.3}) &= \sigma_1\sigma_2(r_{12} - r_{13}r_{23})
\end{aligned}$$

and

$$\begin{aligned}
V(e_{1.3}) &= \frac{1}{N}\sum (e_{1.3} - \bar{e}_{1.3})^2 \\
V(e_{1.3}) &= \frac{1}{N}\sum e_{1.3}^2 \qquad \left[\because \bar{e}_{1.3} = 0\right]
\end{aligned}$$

$$= \frac{1}{N}\sum e_{1.3}e_{1.3}$$

$$= \frac{1}{N}\sum x_1 e_{1.3} \quad \text{(By the third property of residuals)}$$

$$= \frac{1}{N}\sum x_1(x_1 - b_{13}x_3) \qquad \text{From equation (1)}$$

$$= \frac{1}{N}\sum x_1^2 - b_{13}\frac{1}{N}\sum x_1 x_3$$

$$= \sigma_1^2 - b_{13}r_{13}\sigma_1\sigma_3$$

$$= \sigma_1^2 - r_{13}\frac{\sigma_1}{\sigma_3}r_{13}\sigma_1\sigma_3$$

Since, $\quad b_{13} = r_{13}\frac{\sigma_1}{\sigma_3} = \sigma_1^2 - r_{13}^2\sigma_1^2$

$$V(e_{1.3}) = \sigma_1^2(1 - r_{13}^2)$$

Similarly $\quad V(e_{2.3}) = \sigma_2^2(1 - r_{23}^2) \qquad \ldots (4)$

Substituting the value of $Cov(e_{1.3}, e_{2.3}), V(e_{1.3)})$ and $V(e_{2.3})$ in equation (3), we have

$$r_{12.3} = \frac{\sigma_1\sigma_2(r_{12} - r_{13}r_{23})}{\sqrt{\sigma_1^2(1 - r_{13}^2)\sigma_2^2(1 - r_{23}^2)}}$$

$$r_{12.3} = \frac{(r_{12} - r_{13}r_{23})}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}} \qquad \ldots (5)$$

Similarly, expression for $r_{13.2}$ may be obtained as

$$r_{13.2} = \frac{r_{13} - r_{12}r_{32}}{\sqrt{(1 - r_{12}^2)(1 - r_{32}^2)}} \qquad \ldots (6)$$

and

$$r_{23.1} = \frac{r_{23} - r_{21}r_{31}}{\sqrt{(1 - r_{21}^2)(1 - r_{31}^2)}}. \qquad \ldots (7)$$

If $r_{12.3} = 0$, i.e. partial correlation coefficient is zero but $r_{12.3} = 0 \Rightarrow r_{12} = r_{13}r_{23}$ it means correlation coefficient between X_1 and X_2 is not zero if X_3 is correlated with X_1 and X_2.

12.3 MULTIPLE CORRELATION COEFFICIENT IN TERMS OF TOTAL AND PARTIAL CORRELATION COEFFCIENTS

If three variables X_1, X_2 and X_3 are considered then multiple correlation coefficient between X_1 and joint effect of X_2 and X_3 on X_1 is

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}} \qquad \ldots (8)$$

$$R^2_{1.23} = \frac{r^2_{12} + r^2_{13} - 2r_{12}r_{13}r_{23}}{1 - r^2_{23}} \qquad \ldots (9)$$

$$\Rightarrow 1 - R^2_{1.23} = 1 - \frac{r^2_{12} + r^2_{13} - 2r_{12}r_{13}r_{23}}{1 - r^2_{23}}$$

$$\Rightarrow 1 - R^2_{1.23} = \frac{1 - r^2_{23} - r^2_{12} - r^2_{13} + 2r_{12}r_{13}r_{23}}{1 - r^2_{23}} \qquad \ldots (10)$$

We know that the partial correlation coefficient between x_1 and x_3 when the effect of x_2 on each of x_1 and x_3 are eliminated is

$$r_{13.2} = \frac{(r_{13} - r_{12}r_{23})}{\sqrt{(1 - r^2_{12})(1 - r^2_{23})}} \qquad \ldots (11)$$

Squaring equation (11), we get

$$r^2_{13.2} = \frac{(r_{13} - r_{12}r_{23})^2}{(1 - r^2_{12})(1 - r^2_{23})}$$

$$\Rightarrow 1 - r^2_{13.2} = 1 - \frac{(r_{13} - r_{12}r_{23})^2}{(1 - r^2_{12})(1 - r^2_{23})}$$

$$\Rightarrow 1 - r^2_{13.2} = \frac{(1 - r^2_{12})(1 - r^2_{23}) - (r_{13} - r_{12}r_{23})^2}{(1 - r^2_{12})(1 - r^2_{23})}$$

$$\Rightarrow 1 - r^2_{13.2} = \frac{1 - r^2_{23} - r^2_{12} + r^2_{23}r^2_{12} - r^2_{13} - r^2_{23}r^2_{12} + 2r_{12}r_{13}r_{23}}{(1 - r^2_{12})(1 - r^2_{23})}$$

$$\Rightarrow 1 - r^2_{13.2} = \frac{1 - r^2_{23} - r^2_{12} - r^2_{13} + 2r_{12}r_{13}r_{23}}{(1 - r^2_{12})(1 - r^2_{23})}$$

$$\Rightarrow (1 - r^2_{13.2})(1 - r^2_{12}) = \frac{1 - r^2_{23} - r^2_{12} - r^2_{13} + 2r_{12}r_{13}r_{23}}{(1 - r^2_{23})} \qquad \ldots (12)$$

From equations (9) and (12)

$$\Rightarrow (1 - r^2_{13.2})(1 - r^2_{12}) = 1 - R^2_{1.23}$$

$$\Rightarrow (1 - r^2_{13.2})(1 - r^2_{12}) = 1 - R^2_{1.23}$$

$$\Rightarrow R_{1.23} = \sqrt{1 - (1 - r^2_{12})(1 - r^2_{13.2})}$$

It is the required formula and similarly, we may obtain

$$R_{2.13} = \sqrt{1 - (1 - r^2_{12})(1 - r^2_{23.1})}$$

and $$R_{3.12} = \sqrt{1 - (1 - r^2_{13})(1 - r^2_{32.1})}$$

Let us solve some problem on partial correlation coefficient.

Example 1: If $r_{12} = 0.60$, $r_{13} = 0.50$ and $r_{23} = 0.45$ then calculate $r_{12.3}$, $r_{13.2}$ and $r_{23.1}$.

Solution: We have

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}}$$

$$= \frac{0.60-(0.50)(0.45)}{\sqrt{\{1-(0.50)^2\}\{1-(0.45)^2\}}}$$

$$= \frac{0.60-0.23}{\sqrt{\{1-0.25\}\{1-0.20\}}}$$

$$= \frac{0.37}{\sqrt{0.75\times 0.80}}$$

$$= \frac{0.37}{\sqrt{0.60}}$$

$$r_{12.3} = 0.48$$

$$r_{13.2} = \frac{r_{13} - r_{12}r_{32}}{\sqrt{(1-r_{12}^2)(1-r_{23}^2)}}$$

$$= \frac{0.50-(0.60)\times(0.45)}{\sqrt{\{1-(0.60)^2\}\{1-(0.45)^2\}}}$$

$$= \frac{0.50-0.27}{\sqrt{\{1-0.36\{1-0.20\}}}$$

$$= \frac{0.23}{\sqrt{0.64\times 0..80}}$$

$$= \frac{0.23}{\sqrt{0..51`}}$$

$$= \frac{0.23}{0.71}$$

$$r_{13.2} = 0.32$$

Now,

$$r_{23.1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{(1-r_{12}^2)(1-r_{13}^2)}}$$

$$= \frac{0.45-(0.60)(0.50)}{\sqrt{\{1-(0.60)^2\}\{1-(0.50)^2\}}}$$

$$= \frac{0.45-0.30}{\sqrt{\{1-0.36\}\{1-0.25\}}}$$

$$= \frac{0.15}{\sqrt{0.64 \times 0.75}}$$

$$= \frac{0.15}{\sqrt{0.48}}$$

$$= \frac{0.15}{0.69}$$

$$r_{23.1} = 0.22$$

Example 2: From the following data, obtain $r_{12.3}$, $r_{13.2}$ and $r_{23.1}$.

X_1	20	15	25	26	28	40	38
X_2	12	13	16	15	23	15	28
X_3	13	15	12	16	14	18	14

Solution: To obtain partial correlation coefficients $r_{12.3}$, $r_{13.2}$ and $r_{23.1}$ we use following formulae:

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}},$$

$$r_{13.2} = \frac{r_{13} - r_{12}r_{32}}{\sqrt{(1-r_{12}^2)(1-r_{23}^2)}} \quad \text{and}$$

$$r_{23.1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{(1-r_{12}^2)(1-r_{13}^2)}}$$

We need r_{12} , r_{13} and r_{23} which are obtained from the following table:

S. No.	X_1	X_2	X_3	$(X_1)^2$	$(X_2)^2$	$(X_3)^2$	X_1X_2	X_1X_3	X_2X_3
1	20	12	13	400	144	169	240	260	156
2	15	13	15	225	169	225	195	225	195
3	25	16	12	625	256	144	400	300	192
4	26	15	16	676	225	256	390	416	240
5	28	23	14	784	529	196	644	392	322
6	40	15	28	1600	225	784	600	1120	420
7	38	28	14	1444	784	196	1064	532	392
Total	192	122	112	5754	2332	1970	3533	3245	1917

Now, we get the total correlation coefficient r_{12} , r_{13} and r_{23}

$$r_{12} = \frac{N(\sum X_1X_2) - (\sum X_1)\,(\sum X_2)}{\sqrt{\left\{N(\sum X_1^2) - (\sum X_1)^2\right\}\left\{N(\sum X_2^2) - (\sum X_2)^2\right\}}}$$

$$r_{12} = \frac{(7 \times 3533) - (192) \times (122)}{\sqrt{\left\{(7 \times 5754) - (192) \times (192)\right\}\left\{(7 \times 2332) - (122) \times (122)\right\}}}$$

$$r_{12} = \frac{1307}{\sqrt{\{3414\}\{1440\}}} = \frac{1307}{2217.24} = 0.59$$

$$r_{13} = \frac{N(\sum X_1X_3) - (\sum X_1)\,(\sum X_3)}{\sqrt{\left\{N(\sum X_1^2) - (\sum X_1)^2\right\}\left\{N(\sum X_3^2) - (\sum X_3)^2\right\}}}$$

$$r_{13} = \frac{(7 \times 3245) - (192) \times (112)}{\sqrt{\left\{(7 \times 5754) - (192 \times 192)\right\}\left\{(7 \times 1970) - (112 \times 112)\right\}}}$$

$$r_{13} = \frac{1211}{\sqrt{\{3414\}\{1246\}}} = \frac{1211}{2062 \cdot 48} = 0.59$$

and

$$r_{23} = \frac{N(\sum X_2X_3) - (\sum X_2)(\sum X_3)}{\sqrt{\left\{N(\sum X_2^2) - (\sum X_2)^2\right\}\left\{N(\sum X_3^2) - (\sum X_3)^2\right\}}}$$

$$r_{23} = \frac{(7 \times 1917) - (122) \times (112)}{\sqrt{\left\{7 \times 2332) - (122 \times 122)\right\}\left\{(7 \times 1970) - (112 \times 112)\right\}}}$$

$$r_{23} = \frac{-245}{\sqrt{1440 \times 1246}} = \frac{-245}{1339.50} = -0.18$$

Now, we calculate $r_{12.3}$

We have, $r_{12} = 0.59$, $r_{13} = 0.59$ and $r_{23} = -0.18$, then

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}}$$

$$= \frac{0.59 - (0.59)(-0.18)}{\sqrt{\{1 - (0.59)^2\}\{1 - (-0.18)^2\}}}$$

$$= \frac{0.6962}{\sqrt{0.6519 \times 0.9676}}$$

$$= \frac{0.6962}{0.7942}$$

$$r_{12.3} = 0.88$$

$$r_{13.2} = \frac{r_{13} - r_{12}r_{32}}{\sqrt{(1 - r_{12}^2)(1 - r_{23}^2)}}$$

$$= \frac{0.59-(0.59)\times(-0.18)}{\sqrt{\{1-(.59)^2\}\{1-(-0.18^2\}}}$$

$$= \frac{0.6962}{\sqrt{0.6519\times 0.9676}}$$

$$= \frac{0.6962}{0.7942}$$

Thus,

$$r_{13.2} = 0.88$$

$$r_{23.1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{(1-r_{12}^2)(1-r_{13}^2)}}$$

$$= \frac{-0.18-(0.59)(0.59)}{\sqrt{\{1-(0.59)^2\}\{1-(0.59)^2\}}}$$

$$= \frac{-0.5281}{\sqrt{0.6519\times 0.6519}}$$

$$= \frac{-\ 0.5281}{0.6519}$$

$$r_{23.1} = -0.81$$

Example 3: From the following data, obtain $r_{12.3}$, $r_{13.2}$ and $r_{23.1}$:

X_1	40	44	42	45	40	45	40	40	42	41
X_2	18	20	26	24	20	25	23	19	18	16
X_3	52	51	50	48	47	52	50	51	49	50

Solution: To obtain partial correlation coefficients $r_{12.3}$, $r_{13.2}$ and $r_{23.1}$ we use following formulae

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}},$$

$$r_{13.2} = \frac{r_{13} - r_{12}r_{32}}{\sqrt{(1-r_{12}^2)(1-r_{23}^2)}} \quad \text{and}$$

$$r_{23.1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{(1-r_{12}^2)(1-r_{13}^2)}}$$

We need r_{12} , r_{13} and r_{23} which are obtained from the following table:

Here we are using shortcut method to find correlation coefficient.

Let $d_1 = X_1 - 60$, $d_2 = X_2 - 50$ and $d_3 = X_3 - 70$

S. No.	X_1	X_2	X_3	$d_1 = X_1 - 60$	$d_2 = X_2 - 50$	$d_3 = X_3 - 70$	$(d_1)^2$	$(d_2)^2$	$(d_3)^2$	d_1d_2	d_1d_3	d_2d_3
1	40	18	52	0	-2	2	0	4	4	0	0	-4
2	44	20	51	4	0	1	16	0	1	0	4	0
3	42	26	50	2	6	0	4	36	0	12	0	0
4	45	24	48	5	4	-2	25	16	4	20	-10	-8
5	40	20	47	0	0	-3	0	0	9	0	0	0
6	45	25	52	5	5	2	25	25	4	25	10	10
7	40	23	50	0	3	0	0	9	0	0	0	0
8	40	19	51	0	-1	1	0	1	1	0	0	-1
9	42	18	49	2	-2	-1	4	4	1	-4	-2	2
10	41	16	50	1	-4	0	1	16	0	-4	0	0
Total				19	9	0	75	111	24	49	2	-1

Now we get the total correlation coefficient r_{12} , r_{13} and r_{23}

$$r_{12} = \frac{N(\sum d_1d_2) - (\sum d_1)(\sum d_2)}{\sqrt{\{N(\sum d_1^2) - (\sum d_1)^2\}\{N(\sum d_2^2) - (\sum d_2)^2\}}}$$

$$r_{12} = \frac{(10 \times 49) - (19) \times (9)}{\sqrt{\{(10 \times 75) - (19) \times (19)\}\{(10 \times 111) - (9) \times (9)\}}}$$

$$r_{12} = \frac{319}{\sqrt{\{389\}\{1029\}}} = \frac{319}{632.68} = 0.50$$

$$r_{13} = \frac{N(\sum d_1d_3) - (\sum d_1)(\sum d_3)}{\sqrt{\{N(\sum d_1^2) - (\sum d_1)^2\}\{N(\sum d_3^2) - (\sum d_3)^2\}}}$$

$$r_{13} = \frac{(10 \times 2) - (19) \times (0)}{\sqrt{\{(10 \times 75) - (19 \times 19)\}\{(10 \times 24) - (0 \times 0)\}}}$$

$$r_{13} = \frac{20}{\sqrt{\{389\}\{240\}}} = \frac{20}{305.55} = 0.07$$

and

$$r_{23} = \frac{N(\sum d_2d_3) - (\sum d_2)(\sum d_3)}{\sqrt{\{N(\sum d_2^2) - (\sum d_2)^2\}\{N(\sum d_3^2) - (\sum d_3)^2\}}}$$

$$r_{23} = \frac{(10 \times -1) - (9) \times (0)}{\sqrt{\{(10 \times 111) - (9 \times 9)\}\{(10 \times 24) - (0 \times 0)\}}}$$

$$r_{23} = \frac{-10}{\sqrt{\{1029\}\{240\}}} = \frac{-10}{496.95} = -\ 0.02$$

Now, we calculate $r_{12.3}$

We have, $r_{12} = 0.50$, $r_{13} = 0.07$ and $r_{23} = -0.02$, then

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}}$$

$$= \frac{0.50-(0.07)(-\ 0.02)}{\sqrt{\{1-(0.07)^2\}\{1-(-\ 0.02)^2\}}}$$

$$= \frac{0.5014}{\sqrt{0.9951 \times 0.9996}}$$

$$= \frac{0.5014}{0.9933}$$

$$r_{12.3} = 0.50$$

$$r_{13.2} = \frac{r_{13} - r_{12}r_{32}}{\sqrt{(1-r_{12}^2)(1-r_{23}^2)}}$$

$$= \frac{0.20-(0.50)\times(-\ 0.02)}{\sqrt{\{1-(0.50)^2\}\{1-(-\ 0.02)^2\}}}$$

$$= \frac{0.0800}{\sqrt{0.7500 \times 0.9996}}$$

$$= \frac{0.0800}{0.8659}$$

Now, $r_{13.2} = 0.09$

$$r_{23.1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{(1-r_{12}^2)(1-r_{13}^2)}}$$

$$= \frac{-\ 0.02-(0.50)(0.07)}{\sqrt{\{1-(0.50)^2\}\{1-(0.07)^2\}}}$$

$$= \frac{-0.0550}{\sqrt{0.7500 \times 0.9951}}$$

$$= \frac{-\ 0.0550}{0.8639}$$

$$r_{23.1} = -\ 0.06$$

Now let us solve some exercises

E1) If $r_{12} = 0.87$, $r_{13} = 0.82$ and $r_{23} = 0.62$, compute partial correlation coefficient $r_{12.3}$.

E2) In trivariate distribution $r_{12} = 0.8$, $r_{23} = 0.6$, $r_{13} = 0.6$

Compute (a) $r_{12.3}$ (b) $r_{23.1}$

E3) From the following data, obtain $r_{12.3}$, $r_{13.2}$ and $r_{23.1}$.

X_1	12	7	9	15	14	18	18
X_2	10	7	16	15	8	12	10
X_3	7	9	4	8	10	12	8

E4) From the following data, obtain $r_{12.3}$, $r_{13.2}$ and $r_{23.1}$ by using shortcut method

X_1	200	204	202	205	199	200	198	200	202	201
X_2	180	185	179	180	175	184	180	181	178	181
X_3	152	150	149	148	152	150	150	148	153	150

12.4 SUMMARY

In this unit, we have discussed:

1. The correlation coefficient between X_1 and X_2 after eliminating the linear effect of X_3 on X_1 and X_2 is called the partial correlation coefficient,
2. How to derive the formula of partial correlation coefficient, and
3. Multiple correlation coefficient can be expressed in terms of total and partial correlation coefficients as $R_{1.23} = \sqrt{1-(1-r_{12}^2)(1-r_{13.2}^2)}$.

12.5 SOLUTIONS /ANSWERS

E1) We have

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}}$$

$$= \frac{0.87-(0.82)(0.62)}{\sqrt{\{1-(0.82)^2\}\{1-(0.62)^2\}}}$$

$$= \frac{0.87-0.51}{\sqrt{\{1-0.67\}\{1-0.38\}}}$$

$$= \frac{0.36}{\sqrt{0.33\times 0.62}} = \frac{0.36}{\sqrt{0.20}} = \frac{0.36}{0.45} = 0.80$$

E2) We have

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}}$$

$$= \frac{0.8-(0.6)(0.6)}{\sqrt{\{1-(0.6)^2\}\{1-(0.6)^2\}}}$$

$$= \frac{0.44}{\sqrt{0.64\times 0.64}} = \frac{0.44}{0.64} = 0.69$$

$$r_{23.1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{(1-r_{12}^2)(1-r_{13}^2)}}$$

$$= \frac{0.6-(0.8)(0.6)}{\sqrt{\{1-(0.8)^2\}\{1-(0.6)^2\}}}$$

$$= \frac{0.6-0.48}{\sqrt{0.36\times 0.64}} = \frac{0.12}{\sqrt{0.23}} = \frac{0.12}{0.48} = 0.25$$

E3) To obtain partial correlation coefficients $r_{12.3}$, $r_{13.2}$ and $r_{23.1}$ we use following formulae:

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}},$$

$$r_{13.2} = \frac{r_{13} - r_{12}r_{32}}{\sqrt{(1-r_{12}^2)(1-r_{23}^2)}} \quad \text{and}$$

$$r_{23.1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{(1-r_{12}^2)(1-r_{13}^2)}}$$

We need r_{12} , r_{13} and r_{23} which are obtained from the following table:

S. No.	**X_1**	**X_2**	**X_3**	**$(X_1)^2$**	**$(X_2)^2$**	**$(X_3)^2$**	**X_1X_2**	**X_1X_3**	**X_2X_3**
1	12	10	7	144	100	49	120	84	70
2	7	7	9	49	49	81	49	63	63
3	9	16	4	81	256	16	144	36	64
4	15	15	8	225	225	64	225	120	120
5	14	8	10	196	64	100	112	140	80
6	18	12	12	324	144	144	216	216	144
7	18	10	8	324	100	64	180	144	80
Total	93	78	58	1343	938	518	1046	803	621

Now we get the total correlation coefficient r_{12} , r_{13} and r_{23}

$$r_{12} = \frac{N(\sum X_1X_2) - (\sum X_1)(\sum X_2)}{\sqrt{\left\{N(\sum X_1^2) - (\sum X_1)^2\right\}\left\{N(\sum X_2^2) - (\sum X_2)^2\right\}}}$$

$$r_{12} = \frac{(7\times 1046) - (93)\times(78)}{\sqrt{\{(7\times 1343) - (93)\times(93)\}\{(7\times 938) - (78)\times(78)\}}}$$

$$r_{12} = \frac{68}{\sqrt{\{752\}\{482\}}} = \frac{68}{602.05} = 0.11$$

$$r_{13} = \frac{N(\sum X_1X_3) - (\sum X_1)(\sum X_3)}{\sqrt{\left\{N(\sum X_1^2) - (\sum X_1)^2\right\}\left\{N(\sum X_3^2) - (\sum X_3)^2\right\}}}$$

$$r_{13} = \frac{(7\times 803) - (93)\times(58)}{\sqrt{\{(7\times 1343) - (93\times 93)\}\{(7\times 518) - (58\times 58)\}}}$$

$$r_{13} = \frac{227}{\sqrt{\{752\}\{262\}}} = \frac{227}{443.87} = 0.51$$

and

$$r_{23} = \frac{N(\sum X_2X_3) - (\sum X_2)(\sum X_3)}{\sqrt{\left\{N(\sum X_2^2) - (\sum X_2)^2\right\}\left\{N(\sum X_3^2) - (\sum X_3)^2\right\}}}$$

$$r_{23} = \frac{(7\times 621) - (78)\times(58)}{\sqrt{\{7\times 938) - (78\times 78)\}\ \{(7\times 518) - (58\times 58)\}}}$$

$$r_{23} = \frac{-177}{\sqrt{\{482\}\{262\}}} = \frac{-177}{355.36} = -0.50$$

Now, we calculate $r_{12.3}$

We have, $r_{12} = 0.11$, $r_{13} = 0.51$ and $r_{23} = -0.50$, then

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}}$$

$$= \frac{0.11 - (0.51)(-0.50)}{\sqrt{\{1 - (0.51)^2\}\{1 - (-0.50)^2\}}}$$

$$= \frac{0.3650}{\sqrt{0.7399\times 0.7500}} = \frac{0.3650}{0.7449}$$

$$r_{12.3} = 0.49$$

$$r_{13.2} = \frac{r_{13} - r_{12}r_{32}}{\sqrt{(1 - r_{12}^2)(1 - r_{23}^2)}}$$

$$= \frac{0.51 - (0.11)\times(-0.50)}{\sqrt{\{1 - (0.11)^2\}\{1 - (-0.50)^2\}}}$$

$$= \frac{0.5650}{\sqrt{0.9879 \times 0.7500}} = \frac{0.5650}{0.8608}$$

$$r_{13.2} = 0.66$$

Now,

$$r_{23.1} = \frac{r_{23} - r_{12} r_{13}}{\sqrt{(1 - r_{12}^2)(1 - r_{13}^2)}}$$

$$= \frac{-0.50 - (0.11)(0.51)}{\sqrt{\{1 - (0.11)^2\}\{1 - (0.51)^2\}}}$$

$$= \frac{-0.5561}{\sqrt{0.9879 \times 0.7399}} = \frac{-0.5561}{0.8550}$$

$$r_{23.1} = -0.65$$

E4) To obtain partial correlation coefficients $r_{12.3}$, $r_{13.2}$ and $r_{23.1}$ we use following formulae

$$r_{12.3} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}},$$

$$r_{13.2} = \frac{r_{13} - r_{12} r_{32}}{\sqrt{(1 - r_{12}^2)(1 - r_{23}^2)}} \quad \text{and}$$

$$r_{23.1} = \frac{r_{23} - r_{12} r_{13}}{\sqrt{(1 - r_{12}^2)(1 - r_{13}^2)}}$$

We need r_{12} , r_{13} and r_{23} which are obtained from the following table:

S. No.	X_1	X_2	X_3	$d_1 = X_1 - 60$	$d_2 = X_2 - 50$	$d_3 = X_3 - 70$	$(d_1)^2$	$(d_2)^2$	$(d_3)^2$	$d_1 d_2$	$d_1 d_3$	$d_2 d_3$
1	201	180	152	1	0	2	1	0	4	0	2	0
2	204	185	150	4	5	0	16	25	0	20	0	0
3	202	179	149	2	-1	-1	4	1	1	-2	-2	1
4	205	180	148	5	0	-2	25	0	4	0	-10	0
5	199	175	152	-1	-5	2	1	25	4	5	-2	-10
6	200	184	150	0	4	0	0	16	0	0	0	0
7	198	180	150	-2	0	0	4	0	0	0	0	0
8	200	181	148	0	1	-2	0	1	4	0	0	-2
9	202	178	153	2	-2	3	4	4	9	-4	6	-6
10	201	181	150	1	1	0	1	1	0	1	0	0
Total				12	3	2	56	73	26	20	-6	-17

Now we get the total correlation coefficient r_{12} , r_{13} and r_{23}

$$r_{12} = \frac{N(\sum d_1 d_2) - (\sum d_1)(\sum d_2)}{\sqrt{\left\{N(\sum d_1^2) - (\sum d_1)^2\right\}\left\{N(\sum d_2^2) - (\sum d_2)^2\right\}}}$$

$$r_{12} = \frac{(10 \times 20) - (12) \times (3)}{\sqrt{\{(10 \times 56) - (12) \times (12)\}\{(10 \times 73) - (3) \times (3)\}}}$$

$$r_{12} = \frac{164}{\sqrt{\{416\}\{721\}}} = \frac{164}{547.66} = 0.30$$

$$r_{13} = \frac{N(\sum d_1 d_3) - (\sum d_1)(\sum d_3)}{\sqrt{\left\{N(\sum d_1^2) - (\sum d_1)^2\right\}\left\{N(\sum d_3^2) - (\sum d_3)^2\right\}}}$$

$$r_{13} = \frac{(10 \times -6) - (12) \times (3)}{\sqrt{\{(10 \times 56) - (12 \times 12)\}\{(10 \times 26) - (3 \times 3)\}}}$$

$$r_{13} = \frac{-84}{\sqrt{\{416\}\{256\}}} = \frac{-84}{326.34} = -0.26$$

and

$$r_{23} = \frac{N(\sum d_2 d_3) - (\sum d_2)(\sum d_3)}{\sqrt{\left\{N(\sum d_2^2) - (\sum d_2)^2\right\}\left\{N(\sum d_3^2) - (\sum d_3)^2\right\}}}$$

$$r_{23} = \frac{(10 \times -17) - (3) \times (2)}{\sqrt{\{(10 \times 73) - (3 \times 3)\}\{(10 \times 26) - (2 \times 2)\}}}$$

$$r_{23} = \frac{-176}{\sqrt{\{721\}\{256\}}} = \frac{-176}{429.62} = -0.41$$

Now, we calculate $r_{12.3}$

We have, $r_{12} = 0.30$, $r_{13} = -0.26$ and $r_{23} = -0.41$, then

$$r_{12.3} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}}$$

$$= \frac{0.30 - (-0.26)(-0.41)}{\sqrt{\{1 - (-0.26)^2\}\{1 - (-0.41)^2\}}}$$

$$= \frac{0.1934}{\sqrt{0.9324 \times 0.8319}} = \frac{0.1934}{0.8807}$$

$$r_{12.3} = 0.22$$

$$r_{13.2} = \frac{r_{13} - r_{12} r_{32}}{\sqrt{(1 - r_{12}^2)(1 - r_{23}^2)}}$$

$$= \frac{-0.26 - (0.30) \times (-0.41)}{\sqrt{\{1 - (0.30)^2\}\{1 - (-0.41)^2\}}}$$

$$= \frac{-\ 0.1370}{\sqrt{0.9100 \times 0.8300}} = \frac{-0.1370}{0.8701}$$

$$r_{13.2} = 0.16$$

Now,

$$r_{23.1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{(1 - r_{12}^2)(1 - r_{13}^2)}}$$

$$= \frac{-\ 0.41 - (0.30)(-0.26)}{\sqrt{\{1 - (0.30)^2\}\{1 - (-0.26)^2\}}}$$

$$= \frac{-\ 0.3320}{\sqrt{0.9100 \times 0.9324}} = \frac{-0.3320}{0.9211}$$

$$r_{23.1} = -\ 0.36$$

Indira Gandhi
National Open University
School of Sciences

MST-002
DESCRIPTIVE STATISTICS

Block

4

THEORY OF ATTRIBUTES

UNIT 13
Classification of Attributes 5

UNIT 14
Independence of Attributes 15

UNIT 15
Association of Attributes 25

UNIT 16
Association of Attributes for r × s Contingency Table 41

Curriculum and Course Design Committee

Faculty members of School of Sciences, IGNOU

Block Preparation Team

Content Editor
Dr. Soubhik Chakraborty
Department of Applied Mathematics
Birla Institute of Technology Mesra, Ranchi

Language Editor
Dr. Nandini Sahu
School of Humanities, IGNOU

Secretarial Support
Mr. Deepak Singh

Course Writer
Dr. Meenakshi Srivastava
Institute of Social Sciences
Dr. B. R. Ambedkar University, Agra

Formatted By
Dr. Manish Trivedi
Mr. Prabhat Kumar Sangal
School of Sciences, IGNOU

Programme and Course Coordinator: Dr. Manish Trivedi

Block Production

Mr. Y. N. Sharma, SO (P.)
School of Sciences, IGNOU

Acknowledgement: We gratefully acknowledge to Prof. Geeta Kaicker Director, School of Sciences for her great support and guidance.

December, 2011

ISBN-978-81-266-

Further information on the Indira Gandhi National Open University may be obtained from University's Office at Maidan Garhi, New Delhi-110068 or visit University's website http://www.ignou.ac.in

Printed and published on behalf of the Indira Gandhi National Open University, New Delhi by the Director, School of Sciences.

Laser Typeset by : Tessa Media & Computers, C-206, A.F.E.-II, Okhla, New Delhi

Printed at :

THEORY OF ATTRIBUTES

You have studied quantitative techniques in Block 1. The purpose of those techniques is to make you aware of the measures of Central Tendency, measures of Dispersion and measures of Skewness and Kurtosis which describe a set of quantitative data. The concept of statistical relationship between two variables is discussed in Block 2. The concepts of regression analysis are elaborated in Block 3.

The statistical methods discussed in these three blocks are based on the data whose actual magnitude can be measured. However, in some situations, data might be such that it may not be possible to measure their actual magnitude. One can only study the presence or absence of a particular quality or attribute. The statistical methodology for the analysis of such type of data will be slightly different. The present block is mainly concerned with the qualitative characteristics and analysis of qualitative data. Such type of data arises when a sample from some population is classified with respect to two or more qualitative variables. We may then “count” the number of individuals in each category.

This block contains four units. In Unit 13, we shall commence by defining various terms, introducing nomenclature and describing how such kind of data arise. The consistency of the data, independence of the attributes and the condition of independence are discussed in Unit 14. Unit 15 deals with the association of attributes, types of association and the methods to measure the association of attributes. Unit 16 is primarily concerned with the concept of the contingency tables and general notations for higher dimensional contingency tables. This unit also introduces Chi-Square Test for investigating the degree of association between two qualitative variables.

Suggested Readings:

1. Agrawal, B. L.; Basic Statistics, New Age International (P) Ltd. Publishers, New Delhi, 3rd edn., 1996
2. Agrawal, B. L.; Programmed Statistics, New Age International (P) Ltd. Publishers, New Delhi, 2nd edn., 2003
3. Ansari, M. A., Gupta, O. P. and Chaudhari S. S.; Applied Statistics, Kedar Nath Ram Nath & Co., Meerut 1979.
4. Arora, S. and Bansi Lal; New Mathematical Statistics, Satya Prakashan, New Delhi, 1989.
5. Chaturvedi, J. C.; Elementary Statistics, Prakash Brothers, Agra, 1963
6. Elhance, D. N.; Fundamentals of Statistics, Kitab Mahal, Allahabad, 1987.
7. Everitt, B. S.; The Analysis of Contingency Tables, Chapman and Hall Ltd. London, 1st edn. 1977.
8. Garg, N. L.; Practical Problems in Statistics, Ramesh Book Depot, Jaipur 1978.
9. Goodman, L. A. and Kruskal,W. H.; Measures of Association for Cross Classification, Springer – Verlag, Berlin, 1979.
10. Gupta, S. C. and Kapoor, V. K.; Fundamentals of Mathematical Statistics, Sultan Chand & Sons, New Delhi, 11th edn. 2002.

Notations and Symbols

A	:	Presence of attribute A
B	:	Presence of attribute B
C	:	Presence of attribute C
AB	:	Presence of attributes A & B
ABC	:	Presence of attributes A, B & C
α	:	Absence of attribute A
β	:	Absence of attribute B
γ	:	Absence of attribute C
(A)	:	Positive class frequency of attribute A
(α)	:	Negative class frequency of attribute A
$(A\alpha)$	:	Contrary class frequency of attributes A and α
(AB)	:	Positive class frequency of attributes A and B
$(AB)_0$	:	Association of attributes A and B
Q	:	Yule's coefficient of association
γ	:	Coefficient of colligation
(A_i)	:	Number of persons possessing the attribute A_i
$\sum A_i = N$	:	Total frequency
χ^2	:	Chi-square
ϕ^2	:	Mean square contingency
C	:	Karl Pearson's coefficient of mean square contingency

UNIT 13 CLASSIFICATION OF ATTRIBUTES

Structure

13.1 Introduction
Objectives
13.2 Notations
13.3 Classes and Class Frequencies
13.4 Order of Classes and Class Frequencies
13.5 Relation between Class Frequencies
13.6 Class Symbols as Operators
13.7 Summary
13.8 Solutions / Answers
Glossary

13.1 INTRODUCTION

A characteristic that varies from one person or thing to another is called a variable. Income, height, weight, blindness, honesty, sickness are a few examples of variables among humans. The first three of these variables yield numerical information and an investigator can measure their actual magnitude. They are thus termed as quantitative variables. The last three yield non-numerical information and an investigator may not be able to measure their actual magnitude numerically. Hence they are called qualitative variables. They can be arranged in order or in rank. Sometimes they are also referred to as categorical variables.

In the earlier blocks you must have gone through various statistical methods viz. measures of central tendency, measures of dispersion, skewness, correlation, etc. These are important statistical methods that can be used for analysis of the data, which is quantitative in nature i.e. for the first case. However, in the second case when we are dealing with the qualitative variables, which are usually referred to as attributes, the aforesaid methods cannot be used as such, as we cannot measure these qualitative characteristics numerically. All that we can do is to count the number of persons possessing a particular attribute or quality or the number of persons not possessing a particular attribute or quality. Different statistical treatment is required to numerically measure qualitative characteristics. However, they can be related in an indirect manner to a numerical data after assigning particular quantitative indicator. For example, the presence of an attribute can be represented by numeral 1 and the absence of an attribute by numeral 0. Thus, methods of statistics dealing with quantitative variables can also be used for analysing and interpreting the qualitative variables, i.e. attributes. But to have a clear insight on available data through study, analysis and interpretation of attributes there are independently developed statistical methods in the theory of attributes. This unit forms the base for understanding the theory of attributes.

Attributes are Qualitative Variables and cannot be measured directly as they do not yield any numerical information. They can be arranged in order or in rank

In Section 13.2 the basic notations regarding the presence and absence of the attributes are explained. The dichotomy of data is defined in Sub-section 13.2.2. Concepts of classes and class frequencies are described in Section 13.4 whereas order of classes and class frequencies are elaborated in Section 13.5.

The relation between the class frequencies is described in Section 13.6 and class symbols as operators are defined in Section 13.7.

Objectives

After studying this unit, you will be able to

- describe the difference between quantitative and qualitative variables;
- explore the notations and terminology used in the classification of attributes;
- describe the classes and class frequencies;
- define the order of classes and class frequencies;
- explain the basic concepts of relation between the class frequencies; and
- describe the class symbols as operators.

13.2 NOTATIONS

Capital A, B and C denote the presence of attributes and Greek letters α, β and γ denote the absence of attributes.

For convenience in analysis it is necessary to use certain symbols for different classes and for the number of observations assigned to each class. Usually the capital letters A, B, C, etc. are used to denote the presence of attributes and Greek letters α, β, γ, etc. are used to denote the absence of these attributes respectively. Thus, if A represents the attribute of being wealthy, α would represent the attribute of being poor i.e. not wealthy. If B represents blindness, β, would represent absence of blindness. The two classes viz. A (presence of attribute) and α (absence of attribute) are called complementary classes and the attribute α is called complimentary attribute to A. Similar interpretation is for β, γ i.e. they are complementary attributes to B and C respectively. Two or more attributes present in an individual or individuals are indicated by combination of capital Latin letters AB, AC, BC, ABC, etc. Also the presence of one attribute and the absence of other will be represented by $A\beta$, αB, $AB\gamma$.

$A\beta$ represents presence of A and absence of B in an individual. Similarly, the combination $AB\gamma$ represents the absence of the attribute C and the presence of A and B. Likewise any combination of letters can be interpreted.

For example, if A denotes attribute of being wealthy and B denotes the attribute of honesty, then

AB stands for wealthy and honest

$A\beta$ stands for wealthy and dishonest

αB stands for poor and honest

$\alpha\beta$ stands for poor and dishonest.

If a third attribute is noted e.g. sickness then ABC includes those who are wealthy, honest and sick. $A\beta C$ stands for those individuals who are wealthy, dishonest and sick. Similar interpretations can be given to other combination of letters.

Dichotomous classification has only 2 sub-classes

13.2.1 Dichotomy of Data

When the individuals are divided only into two sub-classes or complementary classes and no more, with respect to each of the attributes A, B, C, etc., it is called dichotomous classification. If an attribute has many classes it is called manifold classification.

13.3 CLASSES AND CLASS FREQUENCIES

Different attributes are themselves called classes. For example, if only one attribute, say, tallness is being studied the population would be divided into two classes-one consisting of those possessing this attribute and the other consisting of persons not possessing this attribute i.e. one class would be of tall persons and the other class would be of dwarfs. If more than one attribute were taken into account then the number of classes would be more than two. If, for example, the attribute of sex were also studied along with tallness, then there would be more than two classes in which the total population would be divided. There would be “tall”, “dwarf”, “male”, “female”, “tall female”, “tall male”, “dwarf male” and “dwarf female”.

The number of observations assigned to any class is termed as class frequency. Putting a letter or letters within brackets generally denotes the class frequencies. Thus, (A) stands for the frequency of the attribute A. ($A\beta$) denotes the number possessing the attribute A but not B. Class frequencies (A) (AB) (ABC), etc. are known as positive frequencies and class frequencies of type (α) ($\alpha\beta$) ($\alpha\beta\gamma$) are called negative frequencies whereas the frequencies of the type (αB) ($A\beta$) ($A\beta C$), etc. are called contrary frequencies.

13.4 ORDER OF CLASSES AND CLASS FREQUENCIES

The total of all frequencies denoted by N is called the class of zero order. The classes A, α , B, β, etc. are called the class of first order whereas the combination of any two letters showing the presence or absence of attributes are called class of second order e.g. AB, $A\beta$, αB, $\alpha\beta$, etc. Similarly, combinations like ABC, $AB\gamma$, $A\beta\gamma$, $\alpha\beta\gamma$, etc. are known as class of third order and so on. The frequencies of these classes are known as frequencies of zero order, first order, second order, third order respectively.

13.4.1 Total Number of Class Frequencies

Only one class with N number of members is known as a class of order 0. In a class of order 1 there are 2n number of classes, because there are n attributes, each of them contributing two symbols i.e. one of type A and other of type α.

By binomial theorem

$(1 + 2)^n = 1 + {}^nC_1\, 2 + {}^nC_2\, 2^2 + \ldots + {}^nC_n 2^n$

Similarly, a class of order 2 has ${}^nC_2 \times 2^2$ classes. Since each class contains two symbols, two attributes can be chosen from n in ${}^nC_2$ ways, and each pair give rise to 2^2 different frequencies of the types (AB), ($A\beta$), (αB) and ($\alpha\beta$).

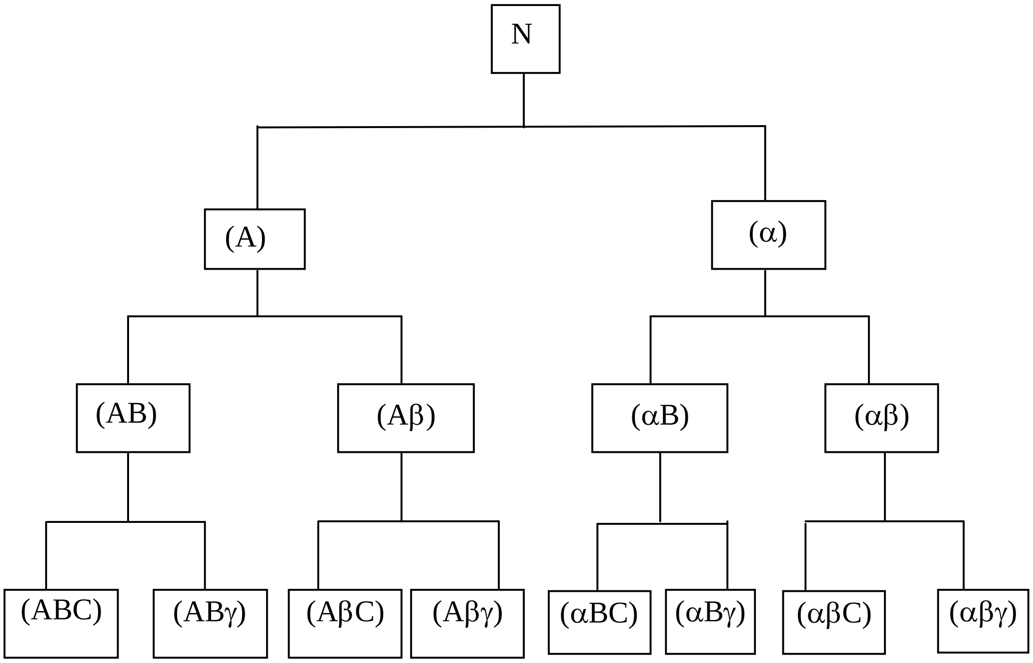
In the same way, it can be shown that total number of classes of order r, there are ${}^nC_r \times 2^r$ classes. Thus, total number of class frequencies of all orders, for n attributes

$$= 1 + {}^nC_1\, 2 + {}^nC_2\, 2^2 + \cdots + {}^nC_r\, 2^r + \cdots + {}^nC_n\, 2^n$$

$$= (1 + 2)^n = 3^n$$

13.5 RELATION BETWEEN CLASS FREQUENCIES

The frequencies of lower order class can always be expressed in terms of higher order class frequencies. For three factors A, B and C, all possible twenty seven combinations of attributes belonging to different classes in the form of a pedigree can be displayed in the following manner.



Similar relations can be given by taking N, B, γ and N, C, γ.

From the above relations we find that

$$(A) = (AB) + (A\beta); (\alpha) = (\alpha B) + (\alpha\beta)$$

$$(AB) = (ABC) + (AB\gamma); (A\beta) = (A\beta C) + (A\beta\gamma)$$

$$(\alpha B) = (\alpha BC) + (\alpha B\gamma); (\alpha\beta) = (\alpha\beta C) + (\alpha\beta\gamma)$$

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

$$N = (AB) + (A\beta) + (\alpha B) + (\alpha\beta)$$

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

Similarly, other relations can be given. The classes of higher order are known as ultimate classes and their frequencies as the ultimate class frequencies. In case of n attributes the ultimate class frequencies will be of n order. For example, if there are three attributes A, B and C the ultimate frequencies will be (ABC), (ABγ), (AβC), (Aβγ), (αBC), (αBγ), (αβC) and (αβγ).

13.6 CLASS SYMBOLS AS OPERATORS

Symbol AN is taken for the operation of dichotomizing N according to the attribute A and is written as AN = (A)
Similarly, we can write $\alpha N = (\alpha)$

Adding these two expressions we get

$$AN + \alpha N = (A) + (\alpha)$$

$$\Rightarrow N(A+\alpha) = N$$

$$\therefore A + \alpha = 1$$

Thus, A can be replaced by $(1-\alpha)$ and α can be replaced by $(1-A)$ respectively. Similarly, B can be replaced by $(1-\beta)$ and β by $(1-B)$ and so on.

Similarly, we may write $ABN = (AB)$ or $\alpha\beta N = (\alpha\beta)$

Thus, $(\alpha\beta) = (1-A)(1-B)N$

$$= (1-A- B + AB)N$$

$$= N-(A)-(B)+(AB)$$

Again

$$(\alpha\beta\gamma) = (1-A)(1-B)(1-C)N$$

$$= (1- A - B - C + AB + BC + AC - ABC) N$$

$$= N-(A)-(B)-(C)+(AB)+(BC)+(AC)-(ABC)$$

Let us consider some problems.

Example 1: Given that $(AB) = 150$, $(A\beta) = 230$, $(\alpha B) = 260$, $(\alpha\beta) = 2340$. Find other frequencies and the value of N.

Solution: We have

$$(A) = (AB) + (A\beta) = 150 + 230 = 380$$

$$(\alpha) = (\alpha B) + (\alpha\beta) = 260 + 2340 = 2600$$

$$(B) = (AB) + (\alpha B) = 150 + 260 = 410$$

$$(\beta) = (A\beta) + (\alpha\beta) = 230 + 2340 = 2570$$

$$N = (A) + (\alpha) = 380 + 2600 = 2980$$

$$N = (B) + (\beta) = 410 + 2570 = 2980$$

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

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

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

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

$(AB) = (ABC) + (AB\gamma)$;

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

$(\alpha B) = (\alpha BC) + (\alpha B\gamma)$;

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

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

Example 2: A number of school children were examined for the presence or absence of certain defects of which three chief descriptions were noted. Let A development defects; B nerve sign; C low nutrition. Given the following ultimate frequencies, find the frequencies of the class defined by the presence of the defects.

$$(ABC) = 60,\ (\alpha BC) = 75,\ (AB\gamma) = 250,\ (\alpha B\gamma) = 650,$$

$$(A\beta C) = 80,\ (\alpha\beta C) = 55,\ (A\beta\gamma) = 350,\ (\alpha\beta\gamma) = 8200$$

Solution: We have

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

$$= (ABC) + (AB\gamma) + (A\beta C) + (A\beta\gamma)$$

$$= 60 + 250 + 80 + 350$$

$$= 740$$

Similarly,

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

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

$= 60+250+75+650$

$= 1035$

$(B) = (AC) + (\alpha C)$

$= (ABC) + (A\beta C) + (\alpha BC) + (\alpha\beta C)$

$= 60 + 80 + 75 + 55$

$= 270$

Again

$(AB) = (ABC) + (AB\gamma) = 60+250 = 310$

$(AC) = (ABC) + (A\beta C) = 60 + 80 = 140$

$(BC) = (ABC) + (\alpha BC) = 60 + 75 = 135$

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

$= 60 + 250 + 80 + 350 + 75 + 650 + 55 + 8200$

$= 9720$

Example 3: Measurements are made on a thousand husbands and a thousand wives. If the measurement of husbands exceeds the measurement of wives in 600 cases for one measurement, in 500 cases for another and in 480 cases for both measurements, then in how many cases would both measurements on wives exceed the measurement on husbands?

Solution: Let (A) denotes the husbands exceeding wives in first measurement and (B) denotes husbands exceeding wives in second measurement. Then given N = 1000, (A) = 600, (B) = 500, (AB) = 480

We have to find $(\alpha\beta)$

$(\alpha\beta) = N – (A) – (B) + (AB)$

$= 1000 – 600 – 500 + 480$

$= 380$

Now, let us solve following exercise:

E1) Given, the following frequencies of the positive classes. Find the frequencies of the rest of the classes:

(A) = 975, (AB) = 455, (ABC) = 125, (B) = 1,187, (AC) = 290,

N= 12,000, (C) = 585 and (BC) = 250

E2) In an observation of 100 cases it was found that the number of unmarried students was 40, number failing the examination was 55 and the number of married who failed was 30. From the information given above find out:

1. The number of married students,
2. The number of students passing the examination,
3. The number of married students who passed,
4. The number of unmarried students who passed,
5. The number of unmarried students who failed.

E3) In a Girls' High school there were 200 students. Their results in the quarterly, half yearly and annual examinations were as follows

85 passed the quarterly examination.

80 passed the half yearly examination.

94 passed the annual examination.

28 passed all the three and 40 failed all the three.

25 passed the first two and failed in the annual examination.

43 failed the first two but passed the annual examination.

Find how many students passed the annual examination.

13.7 SUMMARY

In this unit, we have discussed:

1. Variables can be classified as quantitative and qualitative. The magnitude of quantitative variables can be measured whereas qualitative variables are non, numerical in nature and their magnitude cannot be measured numerically. They are called attributes. The qualitative data can be quantified by assigning number 1 to a person possessing the particular attribute and number 0 to a person not possessing that attribute. Thus, total number of ones would denote total number of persons possessing that attribute. Similarly, total number of zeros would denote total number of persons not possessing that attribute;

2. The statistical methods used to study the qualitative variables are different from the methods to study quantitative variables;

3. The data for the analysis of qualitative variables can be dichotomous where the individuals are classified only into two sub-classes or complementary classes with respect to each attribute A, B, C, etc. On the other hand if the individuals are classified into many classes on the basis of an attribute, the classification is manifold;

4. The number of observations belonging to each class is called class frequency of that class. Putting the letter or letters within brackets generally denotes the class frequencies;

5. The total number of observation denoted by N is called the class of zero order. The class denoted by single letter e.g. A, B, C, α, β, γ, etc. are called class of first order, whereas class represented by combination of two letters e.g. AB, Aβ, αB, etc. are classes of second order. Similarly combinations of three letters indicating presence or absence of attributes are class of third order and so on. The frequencies of these classes are known as frequencies of zero, first, second and third order respectively; and

6. For order r, there are ${}^{n}C_{r} \times 2^{r}$ classes. If there were n attributes, total number of class frequencies would be 3^{n}. The frequencies of lower order class can be expressed in terms of higher order class frequencies.

13.8 SOLUTIONS /ANSWERS

E1) We have

$(\alpha) = N - (A) = 12{,}000 - 975 = 11025$

$(\gamma) = N - (C) = 12000 - 585 = 11415$

$(\beta) = N - (B) = 12000 - 1187 = 10815$

$(AB\gamma) = (AB) - (ABC) = 455 - 125 = 330$

$(A\beta C) = (AC) - (ABC) = 290 - 125 = 165$

$(A\beta\gamma) = (A) - (AB) - (AC) + (ABC)$

$= 975 - 455 - 290 + 125 = 355$

$(\alpha BC) = (BC) - (ABC)$

$= 250 - 125 = 125$

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

$= 1187 - 455 - 250 + 125 = 607$

$(\alpha\beta C) = (C) - (AC) - (BC) + (ABC)$

$= 585 - 290 - 250 + 125 = 170$

$(\alpha\beta\gamma) = N - (A) - (B) - (C) + (AB) + (AC) + (BC) - (ABC)$

$= 12000 - 975 - 1187 - 585 + 455 + 290 + 250 - 125 = 10123$

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

$= 165 + 355 = 520$

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

$= 125 + 607 = 732$

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

$= 170 + 10123 = 10293$

$(A\gamma) = (A) - (AC)$

$= 975 - 290 = 685$

$(\alpha C) = (C) - (AC)$

$= 585 - 290 = 295$

$(\alpha\gamma) = (\alpha) - (\alpha C)$

$= 11025 - 295 = 10730$

$(B\gamma) = (B) - (BC)$

$= 1187 - 250 = 937$

$(\beta C) = (C) - (BC)$

$= 585 - 250 = 335$

$(\beta\gamma) = (\beta) - (\beta C)$

$= 10815 - 335 = 10480$

E2) Let A represents married

Then, α represents unmarried

Let B represents passed

Then, β represents failed

The given data are

$N = 100, (\alpha) = 40, (\beta) = 55, (A\beta) = 30$

(i) Number of married students

$(A) = N - (\alpha) = 100 - 40 = 60$

(ii) Number of students passing the examination

$(B) = N - (\beta) = 100 - 55 = 45$

(iii) Number of married students who passed

$(AB) = (A) - (A\beta) = 60 - 30 = 30$

(iv) Number of unmarried students who passed

$(\alpha B) = (B) - (AB) = 45 - 30 = 15$

(v) Number of unmarried students who failed

$(\alpha\beta) = (\alpha) - (\alpha B) = 40 - 15 = 25$

E3) Let us denote

Success in quarterly examination by A and failure by α

Success in half yearly examination by B and failure by β

Success in annual examination by C and failure by γ

Thus, we have data in the question as

$N = 200, (A) = 85, (B) = 80, (C) = 94, (ABC) = 28,$

$(\alpha\beta\gamma) = 40, (AB\gamma) = 25, (\alpha\beta C) = 43$

Now we have to find the value of

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

We have the relation

$(\alpha BC) + (A\beta C) + (ABC) + (\alpha\beta C) = (C)$

$\therefore\ (\alpha BC) + (A\beta C) = (C) - (ABC) - (\alpha\beta C)$

$= 94 - 28 - 43 = 23$

$\therefore\ (\alpha BC) + (A\beta C) + (AB\gamma) + (ABC)$

$= 23 + 25 + 28 = 76$

Thus, the number of students who passed at least two examinations is 76.

GLOSSARY

Attribute	**:**	A qualitative measurement assigned to objects or individuals.
Dichotomy	**:**	A sharp division into two opposed groups or classes.
Frequency	**:**	The number of observations in some given category.
Manifold classification	**:**	Numerous or various classes.
Variable	**:**	A single one-dimensional property, which may vary along a scale.

UNIT 14 INDEPENDENCE OF ATTRIBUTES

Structure

14.1 Introduction

Objectives

14.2 Consistency of Data

14.3 Conditions of Consistency of the Data

14.4 Independence of Attributes

Criterion of Independence

14.5 Summary

14.6 Solutions / Answers

Glossary

14.1 INTRODUCTION

In Unit 13, you have seen that statistics which deals with the measurement of variables that can be broadly classified as quantitative and qualitative. Quantitative variables are those whose magnitude can be measured numerically. For example, income, height, weight of a group of individuals or number of laborers getting particular amount of wage, etc. Qualitative variables are those whose magnitude cannot be directly measured. An investigator can only study the presence or absence of particular quality in a group. Examples of such variables are sickness, insanity, extravagance, etc. We have also discussed the statistical methodology used for the analysis of quantitative data. You must have also noted that these qualitative variables are called attributes and theory of attributes deals with the measurement of data whose magnitude cannot be directly measured numerically. Though, the qualitative data can be quantified but for the sake of clear understanding and convenience, the statistical methodologies for the analysis of qualitative data have been separately developed. By reading Unit 13, you must now be familiar with the notations, terminology and concepts that are pre–requisites for proceeding any further.

In the present unit we will discuss consistency of the data, the conditions for consistency and the independence of attributes. Section 14.2 deals with the idea of consistency of data. A data is said to be consistent if no class frequency turns out to be negative. Section 14.3 discusses the conditions for consistency of the data. The conditions will be obtained in terms of ultimate class frequencies (already discussed in Unit 13). Section 14.4 illustrates the independence of the attributes i.e. we will study whether or not there is relationship of any kind between two attributes say A and B.

In consistent data no class frequency is negative

Objectives

After reading this unit, you should be able to

- check whether the data is consistent or not;
- describe the conditions for consistency of the data;
- explain the independence of the attributes; and

- test if there exists any relationship of any kind between two attributes or they are independent.

14.2 CONSISTENCY OF DATA

It is a well known fact that no frequency can be negative. If the frequencies of various classes are counted and any class frequency obtained comes out to be negative, then the data is said to be inconsistent. Such inconsistency arises due to wrong counting, or inaccurate addition or subtraction or sometimes due to error in printing. In order to test whether the data is consistent, all the class frequencies are calculated and if none of them is found to be negative, the data is consistent. It should be noted that if the data is consistent it does not mean that the counting is correct or calculations are accurate. But if the data is inconsistent, it means that there is either mistake or misprint in figures.

Data is inconsistent if frequency of attribute or combination is greater than total frequency N.

In order to test the consistency of data, obtain the ultimate class frequencies. If any of them is negative, the data is inconsistent. It would also be seen that no higher order class could have a greater frequency than the lower order class frequency. If any frequency of an attribute or combination of attributes is greater than the total frequency N (frequency of zero order), the data is inconsistent. The easy way to check whether the ultimate class frequencies are negative or not (i.e. checking the data for consistency), is to enter the class frequencies in the chart given in the Section 13.6 of the Unit 13. This will present an overall picture of all the ultimate class frequencies.

It is also possible to lay down conditions for consistency of data.The following section deals with the rules for testing the consistency of the data.

14.3 CONDITIONS FOR CONSISTENCY OF THE DATA

Condition 1: If there is only one attribute A

(i) $(A) \geq 0$

(ii) $(\alpha) \geq 0 \quad \Rightarrow \quad (A) \leq N$ since $N = (A) + (\alpha)$

Condition 2: If there are two attributes A and B then

(i) $(AB) \geq 0$ otherwise (AB) would be negative

(ii) $(AB) \geq (A) + (B) - N$ otherwise $(\alpha\beta)$ would be negative

Proof: We have

$$\begin{aligned}(\alpha\beta) &= (\alpha) - (\alpha B)\\ &= N - (A) - [(B) - (AB)]\\ (\alpha\beta) &= N - (A) - (B) + (AB)\end{aligned}$$

$$\therefore (AB) = (A) + (B) - N + (\alpha\beta)$$

Now if (AB) is less than $(A) + (B) - N$, then $(\alpha\beta)$ would be negative.

(iii) $(AB) \leq (A)$ otherwise $(A\beta)$ would be negative since

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

(iv) $(AB) \leq (B)$ otherwise (αB) would be negative since

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

Condition 3: If there are three attributes A, B and C then

(i) $(ABC) \geq 0$ otherwise (ABC) would be negative.

(ii) $(ABC) \geq (AB) + (AC) - (A)$ otherwise $(AB\gamma)$ would be negative.

(iii) $(ABC) \geq (AB) + (BC) - (B)$ otherwise (αBC) would be negative.

(iv) $(ABC) \geq (AC) + (BC) - (C)$ otherwise $(A\beta C)$ would be negative.

(v) $(ABC) \leq (AB)$

(vi) $(ABC) \leq (AC)$

(vii) $(ABC) \leq (BC)$

(viii) $(ABC) \leq (AB) + (AC) + (BC) - (A) - (B) - (C) + N$ otherwise $(\alpha\beta\gamma)$ would be negative.

Proof: Relation (ii) of (3) is obtained in the following way

Since, $(AB\gamma) \leq (A\gamma)$

i.e. $(AB) - (ABC) \leq (A) - (AC)$

i.e. $(ABC) \geq (AB) + (AC) - (A)$

Similarly, other relations can be computed. Now (i) and (viii) give

$$(AB) + (AC) + (BC) \geq (A) + (B) + (C) - N \qquad \text{... (1)}$$

(ii) and (vii) give

$$(AB) + (AC) - (BC) \leq (A) \qquad \text{... (2)}$$

(iii) and (vi) give

$$(AB) - (AC) + (BC) \leq (B) \qquad \text{... (3)}$$

(iv) and (v) give

$$(AB) + (AC) + (BC) \leq (C) \qquad \text{... (4)}$$

Expressions given by equations (1), (2), (3), (4) are the conditions of consistency which are of course obtained from (i) - (viii) conditions of non negativity of class frequencies.

Example 1: Examine the consistency of the following data N = 1000, (A)= 800, (B) = 400, (AB) = 80, the symbols having their usual meaning.

Solution: We have

$$\begin{aligned}(\alpha\beta) &= N - (A) - (B) + (AB)\\ &= 1000 - 800 - 400 + 80 = -120\end{aligned}$$

Since $(\alpha\beta) < 0$, the data is inconsistent.

Example 2: In a locality having a population of 1000 persons, 750 were males out of whom 530 were married. Among females the number of married ones were 350. Check the consistency of the data.

Solution: Let A represent Males, α represent Females, B represent Married and β represent Unmarried

Given N = 1000, (A) = 750, (AB) = 530 and $(\alpha\beta) = 350$

$$(\alpha) = N - (A) = 1000 - 750 = 250$$

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

$$(\alpha\beta) = (\alpha) - (\alpha B) = 250 - 350 = -100$$

Since $(\alpha\beta) < 0$, the data are inconsistent

Example 3: If all A's are B's and all B's are C's show that all A's are C's.

Solution: Given (AB) = (A), (BC) = (B)

We have to prove (AC) = (A)

We have the relation

$$(AB) + (BC) - (AC) \leq (B)$$

$$(A) + (B) - (AC) \leq (B)$$

$$\Rightarrow \quad (A) \leq (AC)$$

But we know that (A) cannot be less than (AC), hence (A) = (AC)

Example 4: Among the adult population of a certain town 50% of the population is male, 60% are wage earners and 50% are 45 years of age or over. 10% of the males are not wage earners and 40% of the males are under 45. Can we infer anything about what percentage of the population of 45 or over are wage earners?

Solution: Let A, B, C denote the attributes male, wage earners and 45 years old respectively.

Then, N = 100, (A) = 50, (B) = 60, (C) = 50

$$(A\beta) = \frac{10}{100} \times 50 = 5,$$

$$(A\gamma) = \frac{40}{100} \times 50 = 20$$

We are to find out the limits of (BC)

$$(AB) = (A) - (A\beta) = 45,$$

$$(AC) = (A) - (A\gamma) = 30$$

Applying the conditions of consistency

(i) $(AB) + (AC) + (BC) \geq (A) + (B) + (C) - N$

$$\Rightarrow (BC) \geq -15$$

(ii) $(AB) + (AC) - (BC) \leq (A)$

$$\Rightarrow (BC) \geq 25$$

(iii) $(AB) - (AC) + (BC) \leq (B)$

$$\Rightarrow (BC) \leq 45$$

(iv) $(ABC) - (AB) + (AC) + (BC) \leq (C)$

$$\Rightarrow (BC) \leq 65$$

(ii) to (iv) $\Rightarrow 25 \leq (BC) \leq 45$

Hence the percentage of wage earning population of 45 years or over must be between 25 and 45.

14.4 INDEPENDENCE OF ATTRIBUTES

Two attributes A and B are said to be independent if there does not exist any kind of relationship between them. Thus, if A and B are independent we may expect (i) the same proportion of A's in B's as in β's and (ii) same proportion of B's in A's as in α's or we can say two attributes A and B are independent if A is equally popular in B's and in β's and B is equally popular in A's and in α's. For example, intelligence and honesty are independent the proportion of intelligent persons among honest and dishonest person must be equal. If proportion of intelligent persons among honest persons is more than the proportion of intelligent persons among dishonest persons, then obviously intelligence and honesty are not independent. There exists an association between them.

14.4.1 Criterion of Independence

If two attributes are independent then (i) in Section 14.4 gives

$$\frac{(AB)}{(B)} = \frac{(A\beta)}{(\beta)} \qquad \dots (5)$$

$$\Rightarrow 1 - \frac{(AB)}{(B)} = 1 - \frac{(A\beta)}{(\beta)}$$

$$\frac{(\alpha B)}{(B)} = \frac{(\alpha\beta)}{(\beta)} \qquad \dots (6)$$

Similarly, condition (ii) in Section 14.4 gives

$$\frac{(AB)}{(A)} = \frac{(\alpha B)}{(\alpha)} \qquad \dots (7)$$

$$\Rightarrow 1 - \frac{(AB)}{(A)} = 1 - \frac{(\alpha B)}{(\alpha)}$$

$$\therefore \frac{(A\beta)}{(A)} = \frac{(\alpha\beta)}{(\alpha)} \qquad \dots (8)$$

Infact, equation (5) $\Leftrightarrow$ equation (7) i.e. equation (5) implies equation (7) and equation (7) implies equation (5).

Now, we know that for independence

$$\frac{(AB)}{(B)} = \frac{(A\beta)}{(\beta)} = \frac{(AB) + (A\beta)}{(B) + (\beta)} = \frac{(A)}{N}$$

Since (AB) + (Aβ) = (A) and (B) + (β) = N

$$\therefore (AB) = (A).\frac{(B)}{N} \qquad \dots (9)$$

Relation (9) and the expressions, which can be derived like this, give the condition to test the independence of two attributes A and B. The relation (9) can also be written as

$$\frac{(AB)}{N} = \frac{(A)}{N} \cdot \frac{(B)}{N}$$

Thus, an important rule for judging the independence between two attributes A and B can be formulated in terms of proportion. We may say that for independence, the proportion of AB's in the population should be equal to the product of the proportions of A's and B's in the population. The criteria of the independence between two attributes would be more comprehensible and easily understood with the help of following table. In the Table 1, the class frequencies are displayed in the relevant cells.

Table 1

Attributes	**A**	**α**	**Total**
B	(AB)	(αB)	(B)
β	(Aβ)	(αβ)	β
Total	(A)	(α)	(N)

Observing the above table,we may obtain the condition of independence as

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

or

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

Now, let us solve the following exercise:

E1) Given the following class frequencies, do you find any inconsistency in the data?
(A) = 300; (B) = 150; (αβ) = 110; N = 500.

E2) In a survey of 1000 children, 811 liked pizza; 752 liked chowmein and 418 liked burger; 570 liked pizza and chowmein; 356 liked pizza and burger; 348 liked chowmein and burger; 297 liked all the three. Test the consistency of the data.

E3) In a competitive examination 200 graduates appeared. Following facts were noted.

No. of boys =139

No. of Science graduate girls who failed to qualify for interview = 25

No. of Arts graduate girls who qualified for interview = 30

No. of Arts graduate girls who failed to qualify for interview = 18.

Test the consistency of the data.

E4) If report gives the following frequencies as actually observed, show that there is misprint or mistake of some sort.
N=1000; (A) = 525; (B) = 485; (C) = 427; (AB) = 189; (AC) = 140; (BC) = 85.

E5) A study was made about the studying habits of the students of certain university and the following facts were observed. Of the student surveyed, 75% were from well to do families, 55% were boys and 60% were irregular in their studies out of irregular ones 50% were boys and 2/3 were from well to do families. The percentage of irregular boys from well to do families was 8. Is there any consistency in the data?

Before ending this unit let us go over its main points.

14.5 SUMMARY

In this unit, we have discussed:

1. The data is consistent if none of the class frequency is negative. Consistency of the data does not imply that counting of the frequencies or calculations are correct. But the inconsistency in the data means that there is somewhere error or misprint in figures;
2. There are certain conditions laid down to check the consistency of data. These must be applied at the very outset of analysis to get correct and measurable results from data; and
3. Two attributes A and B are independent if

$$(AB) = \frac{(A).(B)}{N}$$

$$\text{or } \frac{(AB)}{N} = \frac{(A)}{N}.\frac{(B)}{N}$$

14.6 SOLUTIONS / ANSWERS

E1) First find out (AB)

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

$$350 = (A\beta) + 110$$

$$\therefore (A\beta) = 350 - 110 = 240$$

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

$$300 = (AB) + 240$$

$$\therefore (AB) = 60$$

For two attributes A and B conditions for consistency are

(i) $(AB) \geq 0$

$60 > 0$

(ii) $(AB) \leq (A)$

$60 < 300$

(iii) $(AB) \leq (B)$

$60 < 150$

(iv) $(AB) \geq (A) + (B) - N$

$60 \geq 300 + 150 - 500 = -50$

Since, all the conditions are satisfied

$\therefore$ the data are consistent

E2) Let A, B, C represent liking of pizza, chowmein and burger respectively.

The given data are

$N = 1000; (A) = 811; (B) = 752; (C) = 418;$

$(AB) = 570; (AC) = 356; (BC) = 348; (ABC) = 297$

Appling the conditions for testing consistency of three attributes we find that the condition (viii) is not satisfied i.e.

$(ABC) \leq (AB) + (AC) + (BC) - (A) - (B) - (C) + N$

Otherwise, $(\alpha\beta\gamma)$ would be negative.

$$297 \leq 570 + 356 + 348 - 811 - 752 - 418 + 1000$$

$$= 2274 - 1981 = 293$$

But $(ABC) > 293$

Thus, the data are inconsistent as $(\alpha\beta\gamma)$ would be negative.

E3) Let A represents boys

α represents girls

B represents science graduates

β represents arts graduates

C those who qualified for interviews

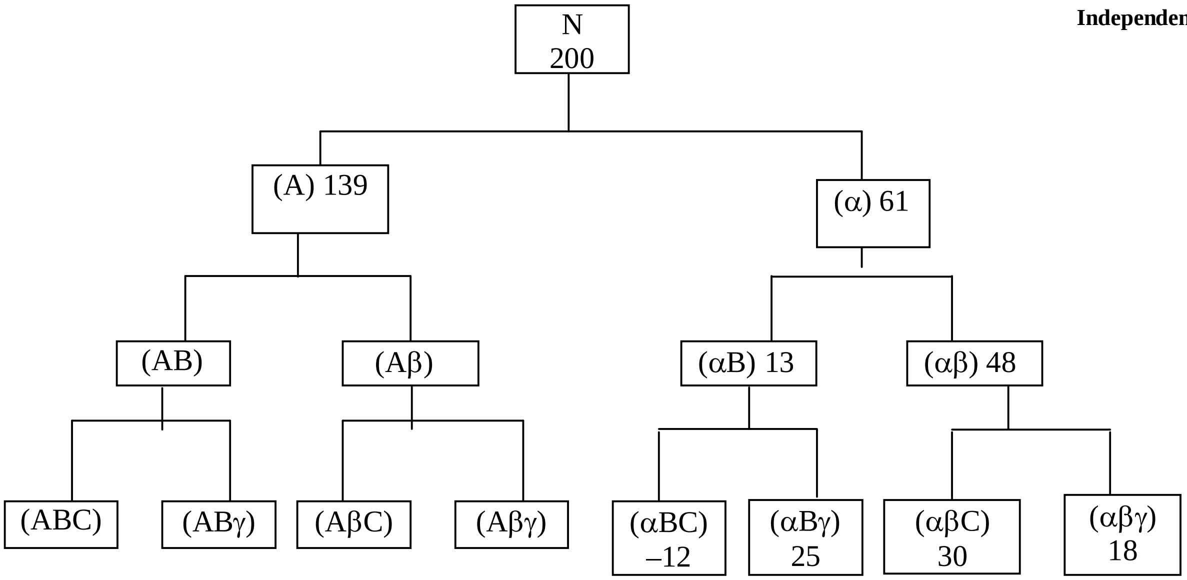
γ those who failed to qualify for interview

Hence, the given data are

$N= 200; (A) = 139; (\alpha B\gamma) = 25; (\alpha\beta\gamma) = 18; (\alpha\beta C) = 30$

In order to check the consistency of the data we have to find whether any ultimate class frequency is negative or not.

The easiest way is to enter the class frequencies in the chart given in the Section 13.6 of Unit 13, i.e.



∵ (αBC) is negative,

∴ The data are inconsistent.

E4) The condition of consistency when positive class frequencies are given (see equation (1) of Section 14.3)

$$(AB) + (AC) + (BC) \geq (A) + (B) + (C) - N \qquad \text{or}$$

$$189 + 140 + 85 \geq 525 + 485 + 427 - 1000$$

$$414 \geq 437$$

which is not true.

Therefore, there is misprint or mistake of any sort in the report.

E5) Let A represent well to do families

B represents boys and C represents irregulars. The data given then are

$$N = 100; (A) = 75; (B) = 55; (C) = 60$$

$$(BC) = \frac{60 \times 50}{100} = 30 \;; (AC) = \frac{60 \times 2}{3} = 40 \;; (ABC) = 8$$

∵ (AB) is not given

∴ applying the (iv) condition of consistency of three attributes

$$(ABC) \geq (AC) + (BC) - (C)$$

$$8 \geq 40 + 30 - 60 = 10$$

which is not true.

Hence, data is inconsistent.

GLOSSARY

Consistency : Degree of firmness, reliably unchanging in deed, compatible.

Independence : No relationship of any kind.

UNIT 15 ASSOCIATION OF ATTRIBUTES

Structure

15.1 Introduction

Objectives

15.2 Association of Attributes

Types of Association

The Symbols $(AB)_0$ and δ

15.3 Methods of Measures of Association

Yule's Coefficient of Association

Coefficient of Colligation

15.4 Summary

15.5 Solutions / Answers

Glossary

15.1 INTRODUCTION

You have seen in Units 13 and 14 that statistical methods deal with the quantitative measurements and the quantitative data can be obtained in two ways:

1. Where the actual magnitude of the variables can be measured for individuals or items for example, income of the people.
2. Where we can count the number of people possessing a particular attribute or the number of people not possessing this particular attribute. For example, number of persons having dark brown eyes. The quantitative character, in this case arises only by counting the number of persons with the presence or absence of certain quality or attribute.

You have also seen that the statistical methodologies for studying the nature of the relationship between two variables for the two aforesaid cases are different. For case (1), relationship between two variables can be measured by correlation coefficient, which not only gives the magnitude of the relationship but also the direction of the relationship i.e. the two variables are positively or negatively related. However, for the second case, coefficient correlation cannot be used, as the data is not numerically expressed, only we know the number possessing particular attribute or not. We have seen in Unit14, the methods used to study the relationship are different. They are covered under theory of attributes.

We have also seen that if an attribute has only two classes, it is said to be dichotomous and if it has many classes, it is called manifold classification. Hence, the need is often felt as to whether there is some association between attributes or not. For this, some measures are developed specifically known as Measures of Association of Attributes. In this unit, we will discuss the measures of association which tell us if there is any association between attributes or not and also whether association is positive or negative.

In this unit, we will focus on association of attributes i.e. the relationship between attributes and how to measure such relationship.

In Section 15.2 we shall discuss association of attributes while in Section 15.3 we shall deal with some of the measures of association commonly known as coefficient of association.

Objectives

After studying this unit, you would be able to

- define the association of attributes;
- differentiate between coefficient of correlation and coefficient of association;
- assess what kind of associations among attributes are likely to occur; and
- distinguish between different methods of measures of association.

15.2 ASSOCIATION OF ATTRIBUTES

The meaning of association in statistical language is quiet different from the common meaning of association. Commonly, if two attributes A and B appear together number of times then they can be said to be as associated. But according to Yule and Kendall, “In Statistics A and B are associated only if they appear together in a greater number of cases than is to be expected, if they are independent.”

Yule
Statisti
associa
appear
numbe
be ex
indepe

Correlation coefficient is a measure of degree or extent of linear relationship between two variables, whereas the coefficient of association indicates association between two attributes and also whether the association is positive or negative.

Methods used to measure the association of attributes refer to those techniques, which are used to measure the relationship between two such phenomena, whose size cannot be measured and where we can only find the presence or absence of an attribute.

In the case of correlation analysis, we study the relationship between two variables, which we can measure quantitatively. Similarly, in the case of association we study the relationship between two attributes, which are not quantitatively measurable. For example, level of education and crime. In association no variables are involved. As it has been stated earlier an attribute divides the universe into two classes, one possessing the attribute and another not possessing the attribute whereas the variable can divide the universe into any number of classes. Correlation coefficient is a measure of degree or extent of linear relationship between two variables, whereas the coefficient of association indicates association between two attributes and also whether the association is positive or negative. But with the help of coefficient of association we cannot find expected change in A for a given change in B and vice-versa, as possible by regression coefficient, which is derived from correlation coefficient.

15.2.1 Types of Association

Two attributes A and B are said to be associated if they are not independent but are related in some way or the other. There are three kinds of associations, which possibly occur between attributes.

1. Positive association
2. Negative association or disassociation
3. No association or independence.

In positive association, the presence of one attribute is accompanied by the presence of other attribute. For example, health and hygiene are positively associated.

or if $(AB) > \frac{(A)(B)}{N}$

Then attributes A and B are positively associated.

In negative association, the presence of one attribute say A ensures the absence of another attribute say B or vice versa. For example, vaccination and occurrence of disease for which vaccine is meant are negatively associated.

or if $(AB) < \frac{(A)(B)}{N}$

Then attributes A and B are negatively associated.

If two attributes are such that presence or absence of one attribute has nothing to do with the absence or presence of another, they are said to independent or not associated. For example, Honesty and Boldness

or if $(AB) = \frac{(A)(B)}{N}$

Then attributes A and B are independent.

Note:

1. Two attributes A and B are said to be completely associated if A cannot occur without B, though B may occur without A and vice-versa. In other words, A and B are completely associated if all A's are B's i.e. (AB) = (A) or all B's are A's i.e. (AB) = (B), according as whether either A's or B's are in a minority.
2. Complete disassociation means that no A's are B's i.e. (AB) = 0 or no α's are β's i.e. $(\alpha\beta) = 0$.

15.2.2 The Symbols $(AB)_0$ and δ

In this unit, following symbols will be used

δ is a Greek letter called Delta.

$$(AB)_0 = \frac{(A)(B)}{N}, \quad (\alpha\beta)_0 = \frac{(\alpha)(\beta)}{N}$$

$$(\alpha B)_0 = \frac{(\alpha)(B)}{N}, \quad (A\beta)_0 = \frac{(A)(\beta)}{N}$$

$$\delta = (AB) - (AB)_0 = (AB) - \frac{(A)(B)}{N}$$

If $\delta = 0$, then $(AB) = \frac{(A)(B)}{N}$

$\Rightarrow$ A and B are independent.

If $\delta > 0$ then attributes A and B are positively associated and if $\delta < 0$ then attributes A and B are negatively associated.

Remark: It is to be noted that if $\delta \neq 0$ and its value is very small then it is possible that this association (either positive or negative) is just by chance and not really significant of any real association between the attributes. This difference is significant or not should be tested by the test statistic (χ^2: Chi-square).

Example 1: Show whether A and B are independent, positively associated or negatively associated in each of the following cases:

(i) N = 1000; (A) = 450; (B) = 600; (AB) = 340

(ii) (A) = 480; (AB) = 290; (α) = 585; (αB) = 383

(iii) N = 1000; (A) = 500; (B) = 400; (AB) = 200

Solution: We have given

(i) $\frac{(A)(B)}{N} = \frac{450 \times 600}{1000} = 270 = (AB)_0$

Thus, $(AB) = 340 > \frac{(A)(B)}{N}$

Since $(AB) > (AB)_0$ hence they are positively associated.

(ii) $\because\ (B) = (AB) + (\alpha B) = 290 + 383 = 673$

$$N = (A) + (\alpha) = 480 + 585 = 1065$$

$$\therefore \frac{(A)(B)}{N} = \frac{480 \times 673}{1065} = 303.32 = (AB)_0$$

Thus, $(AB) = 290 < 303.32$

$\because (AB) < (AB)_0$

$\therefore$ A and B are negatively associated.

(iii) $\frac{(A)(B)}{N} = \frac{500 \times 400}{1000} = 200 = (AB)_0$

Thus, we find $(AB) = (AB)_0$

Hence, A and B are independent, i.e. $\delta = 0$

Example 2: The male population of certain state is 250 lakhs. The number of literate males is 26 lakhs and the total number of male criminals is 32 thousand. The number of literate male criminal is 3000. Do you find any association between literacy and criminality?

Solution: Let literate males be denoted by A so illiterate males would be denoted by α.

Let B represents male criminal so that males who are not criminal would be denoted by β.

Then in lakhs

(A)= 26; (B) = 0.32 ;(AB) = 0.03; N = 250

To study association between A and B let us compute

$$(AB)_0 = \frac{(A)(B)}{N} = \frac{26 \times 0.32}{250} = 0.0332$$

Since $(AB) = 0.03 < (AB)_0$. Hence literacy and criminality are negatively associated.

Example 3: 1660 candidates appeared for competitive examination 425 were successful, 252 had attended a coaching class and of these 150 came successful. Is there any association between success and utility of coaching class?

Solution: Let A denotes successful candidates and B denotes candidates attending coaching class

Given N = 1660; (A) = 425; (B) = 252; (AB) = 150

$$(AB)_0 = \frac{(A)(B)}{N} = \frac{425 \times 252}{1660} = 64.52$$

Since, $(AB) = 150 > (AB)_0$ therefore, there is a positive association between success and utility of coaching class.

Now, let us solve the following exercise:

E1) Find if A and B are independent, positively associated or negatively associated in each of the following cases:

(i) N = 100; (A) = 47; (B) = 62 and (AB) = 32

(ii) (A) = 495; (AB) = 292; (α) = 572 and $(\alpha\beta)$ = 380

(iii) (AB) = 2560; (αB) = 7680;$(A\beta)$ = 480 and $(\alpha\beta)$ = 1440

E2) Out of total population of 1000 the number of vaccinated persons was 600. In all 200 had an attack of smallpox and out of these 30 were those who were vaccinated. Do you find any association between vaccination and freedom from attack?

E3) In an area with a total population of 7000 adults, 3400 are males and out of a total 600 graduates, 70 are females. Out of 120 graduate employees, 20 are females.

(i) Is there any association between sex and education?

(ii) Is their any association between appointment and sex?

15.3 METHODS OF MEASURES OF ASSOCIATION

The methods we have discussed so far can give you an idea whether two attributes are positively associated, negatively associated or independent. Sometimes, this is enough for taking decisions for practical purposes. But most of the times, it is not sufficient as we are always interested in the extent of association, so that we can measure the degree of association mathematically. In the present section, we shall discuss the possibility of obtaining coefficient of association, which can give some idea about the extent of association between two attributes. It would be easy for taking decision if the coefficient of association is such that its value is 0 when two attributes are independent; +1 when they are perfectly associated and –1 when they are perfectly dissociated. In between –1 to +1 lie different levels of association.

Many authors have developed many such coefficients of association, but we will be discussing the one given by Yule.

15.3.1 Yule’s Coefficient of Association

Yule’s coefficient of association is named after its inventor G. Udny Yule. For two attributes A and B, the coefficient of association is given as

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

Value of Q lies between –1 and +1.

1. If Q = 1, A and B has perfect positive association. It can be verified that under perfect positive association

   $(AB) = (A) \Rightarrow (A\beta) = 0$

   $(AB) = (B) \Rightarrow (\alpha B) = 0$

2. If Q = –1, A and B possess perfect negative association. This leads to following relationship:

   $$(AB) = 0 \text{ or } (\alpha\beta) = 0$$

3. If Q = 0, A and B are independent. Here, we have following relation

   $(AB)(\alpha\beta) = (A\beta)(\alpha B)$

4. Any value between –1 to +1 tells us the degree of relationship between two attributes A and B. Conventionally, if Q > 0.5 the association between two attributes is considered to be of high order and the value of Q less than 0.5 shows low degree of association between two attributes.

Q =
asso
Q = -
asso
Q = (
Q > (
high
Q <
low

Remarks: It is to be noted that Q is independent of the relative preposition of A’s or α’s in the data. This property of Q is useful when the prepositions are arbitrary.

15.3.2 Coefficient of Colligation

This is another important coefficient of association given by Yule. It is defined as

$$\gamma = \frac{1 - \sqrt{\dfrac{(A\beta)(\alpha B)}{(AB)(\alpha\beta)}}}{1 + \sqrt{\dfrac{(A\beta)(\alpha B)}{(AB)(\alpha\beta)}}}$$

It can be shown that

$$Q = \frac{2\gamma}{1 + \gamma^2}$$

The range of γ is from –1 to +1. It can be interpreted in the same manner as Q.

Following exercises will illustrate the calculation and interpretation of above two coefficients of association.

Example 4: Calculate Yule’s coefficient of association for the following data:

(i) (A) = 600; (B) = 800 ;(AB) = 480; N = 1000

(ii) (A) = 600; (B) = 800; (AB) = 600; N = 1000

(iii) (A) = 600; (B) = 800; (AB) = 400; N = 1000

(iv) (A) = 600; (B) = 800; (AB) = 500; N = 1000

Solution: We have

(i) (AB) = 480

$(A\beta) = (A) – (AB) = 600 – 480 = 120$

$(\alpha B) = (B) – (AB) = 800 – 480 = 320$

$(\alpha\beta) = (\alpha) – (\alpha B) = 400 – 320 = 80$

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

$$= \frac{(480 \times 80) - (120 \times 320)}{(480 \times 80) + (120 \times 320)}$$

$$= \frac{0}{38400 + 38400} = 0$$

Thus, two attributes are independent

(ii) Here

$(AB) = 600$

$(A\beta) = (A) – (AB) = 600 – 600 = 0$

$(\alpha B) = (B) – (AB) = 800 – 600 = 200$

$(\alpha\beta) = (\alpha) – (\alpha B) = 400 –200 = 200$

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

$$= \frac{(600 \times 200) - (0 \times 200)}{(600 \times 200) + (0 \times 200)}$$

$$= \frac{120000}{120000} = +1$$

Thus, there is a perfect positive association between attributes A and B.

(iii) In this case

$(AB) = 400$

$(A\beta) = (A) – (AB) = 600 – 400 = 200$

$(\alpha B) = (B) – (AB) = 800 – 400 = 400$

$(\alpha\beta) = (\alpha) – (\alpha B) = 400 – 400 = 0$

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

$$= \frac{(400 \times 0) - (200 \times 400)}{(400 \times 0) + (200 \times 400)}$$

$$= \frac{-80000}{80000} = -1$$

(iv) Here, we have

(AB) = 500

(Aβ) = (A) – (AB) = 600 – 500 = 100

(αB) = (B) – (AB) = 800 – 500 = 300

(αβ) = (α) – (αB) = 400 – 300 = 100

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

$$= \frac{(500 \times 100) - (100 \times 300)}{(500 \times 100) + (100 \times 300)}$$

$$= \frac{20000}{80000} = +0.25$$

Thus, there is very nominal association between A and B.

Example 5: In a sample of 1000 children, 400 came from higher income group and rest from lower income group. The number of delinquent children in these groups was 50 and 200 respectively. Calculate the coefficient of association between delinquency and income groups.

Solution: Let A denotes the higher income group, then α would denote lower income group. Let B denotes delinquent children then β would denote non-delinquent children. To get the frequencies of second order we form following nine square table (or 2× 2 table):

Attributes	**A**	**α**	**Total**
B	AB 50	αB 200	B 250
β	Aβ 350	αβ 400	β 750
Total	A 400	α 600	N 1000

From the table

(α) = N – (A) = 1000 – 400 = 600

(B) = (AB) + (αB) = 50 + 200 = 250

(Aβ) = (A) – (AB) = 400 – 50 = 350

(αB) = (B) – (AB) = 250 – 50 = 200

(αβ) = (α) – (αB) = 600 –200 = 400

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

$$= \frac{(50 \times 400) - (350 \times 200)}{(50 \times 400) + (350 \times 200)}$$

$$= \frac{-50000}{90000} = -\ 0.55$$

Thus, there is a negative association between income and delinquency.

Example 6: Investigate if there is any association between extravagance in father and son from the following:

Extravagant sons with extravagant fathers (AB) = 450

Miser sons with extravagant fathers (αB) = 155

Extravagant sons with miser fathers (Aβ) = 175

Miser sons with miser fathers ($\alpha\beta$) = 1150

Solution: For association between extravagance in father and son we calculate coefficient of association as

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

$$= \frac{(450 \times 1150) - (175 \times 155)}{(450 \times 1150) + (175 \times 155)}$$

$$= \frac{490375}{544625} = 0.90$$

Thus, there is very high degree of positive association between extravagance in father and son. An extravagant father in general has an extravagant son.

Example 7: In an examination, at which 600 candidates appeared, boys out numbered girls by 16% of all candidates. Number of passed exceeded the number of failed candidates by 310. Boys failing in the examination numbered 88. Find the Yule's coefficient of association between male sex and success in examination.

Solution: Let boys be denoted by A so girls are denoted by α.

Let the success in examination be denoted by B so that failure in examination is denoted by β.

Data given are

(i) Boys outnumber girls by 16% of the total is

$$= \frac{16 \times 600}{100} = 96$$

$\therefore$ by condition we have (A) – (α) = 96

Also we have (A) + (α) = 600

$\therefore$ 2 (A) = 696

$\Rightarrow$ (A) = 348

which is the number of boys so the number of girls (α) would be

(A) – 96 = (α)

$\therefore$ 348 – 96 = (α)

$\Rightarrow$ (α) = 252

(ii) Number of passed candidates exceeded number that failed by 310

i.e. (B) – (β) = 310

also we have (B) + (β) = 600

$\therefore$ 2 (B) = 910

$\Rightarrow$ (B) = 455

Thus, number of passed candidates is 455, so the number of failures would be

(β) = (B) – 310 = 455 –310 =145

(iii) Boys failing in the examination (Aβ) = 88

Other values can be obtained from the 2× 2 table

Attributes	**A**	**α**	**Total**
B	AB 260	α B 195	B 455
β	Aβ 88	α β 57	β 145
Total	A 348	α 252	N 600

Yule's coefficient of association is

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

$$= \frac{(260 \times 57) - (88 \times 195)}{(260 \times 57) + (88 \times 195)}$$

$$= \frac{-2340}{31980} = -0.07$$

Thus, there is insignificant association between male sex and success.

Example 8: Given

(AB) = 35 (αβ) = 7

(Aβ) = 8 (αB) = 6

calculate the coefficient of colligation.

Solution: Coefficient of colligation

$$\gamma = \frac{1 - \sqrt{\frac{(A\beta)(\alpha B)}{(AB)(\alpha\beta)}}}{1 + \sqrt{\frac{(A\beta)(\alpha B)}{(AB)(\alpha\beta)}}} = \frac{1 - \sqrt{\frac{8 \times 6}{35 \times 7}}}{1 + \sqrt{\frac{8 \times 6}{35 \times 7}}}$$

$$= \frac{1-\sqrt{0.196}}{1+\sqrt{0.196}} = \frac{1-0.44}{1+0.44} = 0.39$$

Thus, two attributes A and B are positively associated.

Now, let us solve the following exercise:

E4) The following table is reproduced from a memoir written by Karl Pearson

	Eye colour in son	
Eye colour in father	Not light	Light
Not light	230	148
Light	151	471

Discuss whether the colour of the son's eye is associated with that of father.

E5) Can vaccination be regarded as a preventive measure for small pox from the data given below:

'Of 1482 persons in a locality exposed to small pox 368 in all were attacked'

'Of 1482 persons, 343 had been vaccinated and of these only 35 were attacked'.

E6) From the following data prepare 2 × 2 table and using Yule's coefficient discuss whether there is any association between literacy a unemployment.

Illiterate Unemployed	250 persons
Literate Employed	25 persons
Illiterate Employed	180 persons
Total number of persons	500 persons

E7) From the data given below calculate Yule's coefficient of association between weight of the children and then economic condition, and interpret it.

	Poor Children	Rich Children
Below Normal Weight	85	20
Above Normal Weight	6	47

This brings us end of this unit. In the next unit we will take up the study of some more aspects of theory of attributes. But before that let us briefly recall what we have studied in this unit.

15.4 SUMMARY

In this unit, we have discussed:

1. Meaning of association in Statistics is different from the general meaning of association. In Statistics, attributes A and B are associated only if they appear together in greater number of cases than is to be expected if they are independent. In common language association means if A and B occur together a number of times then A and B are associated;
2. In association, we study the relationship between two attributes, which are not quantitatively measured;
3. Correlation coefficient measures the extent of relationship between two quantitative variables, whereas coefficient of association only suggests that the association is positive or negative;
5. If there exist no relationship of any kind between two attributes then they are said to be independent otherwise are said to be associated. Attributes A and B are said to be

   Positively associated if $(AB) > \frac{(A)(B)}{N}$

   Negatively associated if $(AB) < \frac{(A)(B)}{N}$

   Independent if $(AB) = \frac{(A)(B)}{N}$
6. Some times only the knowledge of the association (whether positive or negative) or independence between attributes is not sufficient. We are interested in finding the extent or degree of association between attributes, so that we can take decision more precisely and easily. In this regard, we have discussed Yule's coefficient of association in this unit. The value of Yule's coefficient of association lies between –1 to +1. If Q = +1, A and B are perfectly associated. In between –1 to +1, are lying different degrees of association;
7. Another important coefficient of association is coefficient of colligation. Q and γ are related by the following expression

   $Q = \frac{2\gamma}{1+\gamma^2}$; and
8. γ also lies between –1 to +1 and have the interpretation as that of Q.

15.5 SOLUTIONS / ANSWERS

E1) $(AB)_0 = \frac{(A)(B)}{N} = \frac{47 \times 62}{100} = 29.14$

$\because (AB) = 32 > (AB)_0$

$\therefore$ A and B are positively related.

(ii) We have

$N = (A) + (\alpha) = 495 + 572 = 1067$

$(B) = (AB) + (\alpha B) = 292 + 380 = 672$

$$\therefore (AB)_0 = \frac{(A)(B)}{N} = \frac{495 \times 672}{1067} = 31.75$$

$$\because (AB) = 292 < (AB)_0$$

$\therefore$ A and B are negatively related.

(iii) (A) = (AB) + (Aβ) = 2560 + 480 = 3040

(B) = (AB) + (αB) = 2560 + 7680 = 10240

N = (AB) + (Aβ) + (αB) + ($\alpha\beta$)

= 2560 + 480 + 7680 + 1440

= 12160

$$\therefore (AB)_0 = \frac{(A)(B)}{N} = \frac{3040 \times 10240}{12160} = 2560$$

$$\because (AB) = 2560 = (AB)_0$$

$\therefore$ A and B are independent.

$\therefore$ i.e. $\delta = 0$

E2) Let A represents vaccinated and B freedom from attack.

The given data are

N = 1000; (A) = 600; (β) = 200; (Aβ) = 30

We have (AB) + (Aβ) = N

$\therefore$ (AB) = N – (Aβ) = 1000 – 30 = 970

Again, we have

(B) + (β) = N

$\therefore$ (B) = N – (β) = 1000 –200 = 800

$$\text{Thus } (AB)_0 = \frac{(A)(B)}{N} = \frac{600 \times 800}{1000} = 480$$

$$\text{Since } (AB) = 970 > (AB)_0$$

Hence, A and B are positively associated.

E3) (i) Let A represents males; α will be females

Let B represents graduates; β will be non graduates

The given data are

N = 7000, (A) = 3400, (B) = 600, (αB) = 70

(AB) = (αB) + (B) = 600 + 70 = 670

$$(AB)_0 = \frac{(A)(B)}{N} = \frac{3400 \times 600}{7000} = 291.43$$

$$\text{Since } (AB) = 670 > (AB)_0$$

Hence, A and B are positively associated.

(ii) Let A represents male graduates then α will be female graduates.

Let B represents employed then β will be unemployed.

Given (A) = 530; (β) = 70; (B) = 120; (αB) = 20

$N = (\beta) + (B) = 70 + 120 = 190$

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

$$\Rightarrow \quad (AB) = 120 - 20 = 100$$

Now, $$(AB)_0 = \frac{(A)(B)}{N} = \frac{530 \times 120}{100} = 636$$

$$\because \quad (AB) = 100 < 636$$

$\therefore$ A and B are negatively related.

E4) Let A represents the light eye colour of father an B represents the light eyecolour of son. Then α represents not light eye colour of father and β represents not light eye colour of son. Then the given data is

$(\alpha\beta)$ = 230, (αB) = 148

$(A\beta)$ = 151, (AB) = 471

Coefficient of Association is

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

$$= \frac{(471 \times 230) - (151 \times 148)}{(471 \times 230) + (151 \times 148)}$$

$$= \frac{85982}{130678} = +0.657$$

This shows that there is fairly high degree of positive association between eye colour of father and son.

E5) Let A denotes attribute of vaccination, and B that of attack

Then the given data are N = 1482; (B) = 368; (A) = 343; (AB) = 35

Now,

$(\alpha\beta) = N - (A) - (B) + (AB) = 1482 - 343 - 368 + 35 = 806$

$(A\beta) = (A) - (AB) = 343 - 35 = 308$

$(\alpha B) = (B) - (AB) = 368 - 35 = 333$

Yule’s coefficient of association

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

$$= \frac{(35 \times 806) - (308 \times 333)}{(35 \times 806) + (308 \times 333)}$$

$$= \frac{-74354}{130774} = -0.57$$

Thus, we find that vaccination and small pox are in a high degree of negative association. Hence, vaccination can be regarded as a preventive measure of small pox.

E6) Let A denotes literacy and B denotes unemployment so that α denotes illiteracy and β denotes employment.

Now we have

$(\alpha B) = 250$; $(A\beta) = 25$; $(\alpha\beta) = 180$; $N = 500$

We put these figures in 2 × 2 table and get the frequencies of the remaining class

Attributes	**A**	**α**	**Total**
B	AB 45	αB 250	B 295
β	Aβ 25	α β 180	β 205
Total	A 70	α 430	N 500

Yule's Coefficient of Association is

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

$$= \frac{(45 \times 180) - (25 \times 250)}{(45 \times 180) + (25 \times 250)}$$

$$= \frac{1850}{14350} = +0.13$$

This shows that there is only marginal positive association between literacy and unemployment.

E7) Let A denotes poor children and B denotes children below normal weight. Then α would denote rich children and β would denote children above normal weight.

The data given are

$(AB) = 85$; $(\alpha B) = 20$; $(A\beta) = 6$; $(\alpha\beta) = 47$;

Yule's coefficient of association is

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

$$= \frac{(85 \times 47) - (6 \times 20)}{(85 \times 47) + (6 \times 20)}$$

$$= \frac{3875}{4115} = +0.94$$

There is high degree of association between poor children and children below normal weight.

This means chances of poor children being below normal weight are very high. Rich children will generally be above normal weight.

GLOSSARY

Correlation : Study of relationship between two variables say x and y.

Correlation Coefficient : A measure that gives the degree of in lines relationship between x and y.

Regression Coefficient : Measure of change in variable y corresponding to unit change in variable x and vice versa.

UNIT 16 ASSOCIATION OF ATTRIBUTES FOR r × s CONTINGENCY TABLE

Structure

16.1 Introduction
Objectives
16.2 Contingency Table: Manifold Classification
16.3 Chi-square and Coefficient of Contingency
16.4 Summary
16.5 Solutions/Answers
Glossary

16.1 INTRODUCTION

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

table
ells.

Table 1

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

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

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

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

Objectives

After reading this unit, you should be able to

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

16.2 CONTINGENCY TABLE: MANIFOLD CLASSIFICATION

We have already learnt that if an attribute is divided into more than two parts or groups, we have manifold classification. For example, instead of dividing the universe into two parts-heavy and not heavy, we may sub-divide it in a large number of parts very heavy, heavy, normal, light and very light. This type of subdivision can be done for both the attributes of the universe. Thus, attribute A can be divided into a number of groups A_1, A_2,…, A_r. Similarly, the attribute B can be subdivided into B_1, B_2,…, B_r. When the observations are classified according to two attributes and arranged in a table, the display is called contingency table. This table can be 3×3, 4×4, etc. In 3×3 table both of the attributes A and B have three subdivisions. Similarly, in 4×4 table, each of the attributes A and B is divided into four parts, viz. $A_{1,}$ $A_{2,}$ $A_{3,}$ A_4 and $B_{1,}$ $B_{2,}$ $B_{3,}$ B_4.

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

If or
attri
class
has
class
the c
man

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

Following is the layout of r × s contingency table:

Table 2: r × s Contingency Table

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

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

In the table

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

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

etc. Similarly,

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

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

etc. And

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

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

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

16.3 CHI - SQUARE AND COEFFICIENT OF CONTINGENCY

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

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

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

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

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

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

Table 3: Expected Frequency Table

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

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

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

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

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

Thus,

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

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

The above expression can also be written as

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

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

E - is the expected frequency of that class.

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

Therefore, Square contingency = χ^2

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

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

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

The coefficient of mean square contingency is defined as

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

If we calculate χ^2 by the formula

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

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

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

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

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

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

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

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

Table 4

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

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

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

Similarly, the expected frequency of class

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

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

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

(i) O- the observed frequency

(ii) E -the expected frequency.

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

Table 5

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

Thus, the value of chi-square is

$$\chi^2 = 27.94$$

and

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

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

$$= 0.0736$$

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

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

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

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

Table 6: Incidence of Cerebral Tumour

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

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

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

Similarly, expected frequency

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

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

Table 7

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

Hence, the value of χ^2 = 6.72

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

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

$$= 0.21$$

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

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

Now, let us solve the following exercise:

E1) 1000 students at college level were graded according to their IQ level and the economic condition of their parents.

Economic Condition	IQ level		
	High	Low	Total
Rich	460	140	600
Poor	240	160	400
Total	700	300	1000

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

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

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

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

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

16.4 SUMMARY

In this unit, we have discussed:

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

	Boys	Girls	Total
Right handed			
Left handed			
Total			

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

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

16.5 SOLUTIONS /ANSWERS

E1) Calculations for expected frequencies are below:

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

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

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

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

Enter the expected frequencies in the table given below

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

Therefore, $\chi^2 = 31.746$

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

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

$$= 0.175$$

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

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

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

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

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

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

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

Hence $\chi^2 = 7.291$

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

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

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

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

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

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

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

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

Therefore, $\chi^2 = 8.18$

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

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

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

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

GLOSSARY

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

Expected frequency : Frequencies expected under certain assumptions.

Observed frequency : Actually recorded frequency.