MAEND-407

Educational Statistics

MA EDUCATION

2nd Semester

Rajiv Gandhi University

www.ide.rgu.ac.in

EDUCATIONAL STATISTICS

MA [Education]
Second
Semester

MAEDN - 407

RAJIV GANDHI UNIVERSITY

Authors

Prof J. Sahoo
Dr. C Siva Sankar
Moyir Riba

About the University

Rajiv Gandhi University (formerly Arunachal University) is a premier institution for higher education in the state of Arunachal Pradesh and has completed twenty-five year of its existence. Late Smt. Indira Gandhi, the then Prime Minister of India, laid the foundation stone of the university on 4th February, 1984 at Rono Hills, where the present campus is located.

Ever since its inception, the university has been trying to achieve excellence and fulfill the objectives as envisaged in the University Act. The University received academic recognition under Section 2(f) from the University Grants Comission on 28th March, 1985 and started functioning from 1st April, 1985. It got financial recognition under section 12-B of the UGC on 25th March, 1994. Since then Rajiv Gandhi University, (then Arunachal University) has carved a niche for itself in the educational scenario of the country following its selection as a University with potential for excellence by a high-level expert committee of the University Grants Commission from among universities in India.

The University was converted into a Central University with effect from 9th April, 2007 as per notification of the Ministry of Human Resource Development, Government of India.

The University is located atop Rono Hills on a picturesque tableland of 302 acres overlooking the river Dikrong. It is 6.5 km from the National Highway by the Dikrong Bridge.

The teaching and research programmes of the University are designed with a view to play a positive role in the socio-economic and cultural development of the State. The University offers Undergraduate, Post-graduate, M.Phil and Ph.D programmes. The Department of Education also offers the B.Ed Programme.

There are fifteen colleges affiliated to the University. The University has been extending educational facilities to students from the neighbouring states, particularly Assam. The Strength of students in different departments of the University and in affiliated colleges has been steadily increasing.

The faculty members have been actively engaged in research activities with financial support from UGC and other funding agencies. Since inception, a number of proposals on research projects have been sanctioned by various funding agencies to the University. Various departments have organized numerous seminars, workshops and conferences. Many faculty members have participated in national and international conferences and seminars held within the country and abroad. Eminent scholars and distinguished personalities have visited the University and delivered lectures on various disciplines.

The academic year 2000-2001 was a year of consolidation for the University. The switch over from the annual to the semester system took off smoothly and the performance of the students registered a marked improvements. Various syllabi designed by Boards of Post-graduate Studies (BPGS) have been implemented. VSAT facility installed by the ERNET India, New Delhi under the UGC-Infonet program, provides Internet access.

In spite of infrastructural constraints, the University has been maintaining its Academic excellence. The University has strictly adhered to the academic calendar, conducted the examinations and declared the results on time.The students from the University have found placements not only in State and Central Government Services, but also in various institutions, industries and organizations. Many students have emerged successful in the National Eligibility Test (NET).

Since inception, the University has made significant progress in teaching, research, innovations in curriculum development and developing infrastructure.

About IDE

The formal system of higher education in our country is facing the problems of access, limitation of seats, lack of facilities and infrastructure. Academicians from various disciplines opine that it is learning which is more important and not the channel of education. The education through distance mode is an alternative mode of imparting instruction to overcome the problems of access, infrastructure and socio-economic barriers. This will meet the demand for qualitative higher education of millions of people who cannot get admission in the regular system and wish to pursue their education. It also helps interested employed and unemployed men and women to continue with their higher education. Distance education is a distinct approach to impart education to learners who remained away in the space and/or time from the teachers and teaching institutions on account of economic, social and other considerations. Our main aim is to provide higher education opportunities to those who are unable to join regular academic and vocational education programmes in the affiliated colleges of the University and make higher education reach to the doorsteps in rural and geographically remote areas of Arunachal Pradesh in particular and North-eastern part of India in general. In 2008, the Centre for Distance Education has been renamed as "Institute of Distance Education (IDE)."

Continuing the endeavor to expand the learning opportunities for distant learners, IDE has introduced Post-Graduate Courses in 5 subjects (Education, English, Hindi, History and Political Science) from the Academy Session 2013-14.

The Institute of Distance Education is housed in the Physical Sciences Faculty Building(First floor) next to the University Library. The University campus is 6 kms from NERIST point on National Highway 52A. The University buses ply to NERIST point regularly.

Outstanding Features of Institute of Distance Education :

(i) At per with Regular Mode.

Eligibility requirements, curricular content, mode of examination and the award of degrees are on par with the colleges affiliated to the Rajiv Gandhi University and the Department(s) of the University

(ii) Self-Instructional Study Material (SISM)

The students are provided SISM prepared by the Institute and approved by Distance Education Council (DEC), New Delhi. This will be provided at the time of admission at the IDE or its Study Centres.SISM is provided only in English except Hindi subject.

(iii) Contact and Counselling Programme (CCP)

The course curriculum of every programme involves counsellig in the form of personal contact programmes of duration of approximately 7-15 days. The CCP shall not be compulsory for BA. However for professional courses and MA the attendance in CCP will be mandatory.

(iv) Field Training and Project

For professional course(s) there shall be provision of field training and project writing in the concerned subject.

(v) Medium of Instructions and Examination

The medium of instruction and examination will be English for all the subjects except for those subjects where the learners will need to write in the respective languages.

(vi) Subject /Counselling Coordinators

For developing study material, the IDE appoints subject coordinators from within and outside the University. In order to run the PCCP effectively Counselling Coordinators are engaged from the Departments of the University, The counseling-Coordinators do necessary coordination for involving resource persons in contact and counseling programme and assignemt evaluation.The learners can also contact them for clarifying their difficulties in then respective subjects.

SYLLABUS

Objectives:

1. To make the students understand the role of statistics in educational research and compute measures of central tendency and variability
2. To develop the skill of using the statistical techniques appropriately.
3. To enable the students how to test hypotheses using appropriate Statistics

Course Content:

UNIT-I. Measures of central tendency and variability:

- Measures of Central Tendency and their computation and uses
- Measures of Variability and their computation and uses

UNIT II. Correlation

- Correlation: Concept and its applications:
- Methods of computing coefficient of correlation:
  Rank difference and Pearson's coefficient of correlation.

UNIT-III. Normal probability curve and tests of significance :

- Properties and applications
- The concept of statistical significance
- Testing the significance of mean, proportion and correlation

UNIT IV. Hypothesis Testing

- Testing the significance of difference between means, proportion and correlation
- Chi-square ($x^{2)}$, Types of errors, one-tailed and two tailed tests(ANOVA-One way)

Practicum :

1. Construction of attitude scale using appropriate Statistics
2. Construction of test using appropriate Statistics

UNIT 1 MEASURES OF CENTRAL TENDENCY AND VARIABILITY

Structure

Introduction
Unit Objectives
Measures of Central Tendency
Weighted Arithmetic Mean
Different Positional Numbers
Geometric Mean
Harmonic Mean
Measures of Variation
Types of Measures
Range
Skewness
Kurtosis
Comparison of Various Measures of Dispersion
Coefficient of Variation
Summary
Key Terms
Answers to 'Check Your Progress'
Questions and Exercises
Further Reading

INTRODUCTION

In this unit, you will learn about the measures of central tendency and dispersion. There are several commonly used measures of central tendency, such as arithmetic mean, mode and median. These values are very useful not only in presenting the overall picture of the entire data but also for the purpose of making comparisons among two or more sets of data. In addition, you will learn about the geometric mean and harmonic mean. If a, b, c are in GP, then b is called a geometric mean between a and c, written as GM. If a, b, c are in HP, then b is called a Harmonic Mean between a and c, written as HM. Moreover, you will also learn about the measures of dispersion. A measure of dispersion or simply dispersion may be defined as statistics signifying the extent of the scatteredness of items around a measure of central tendency. Finally, you will learn about the coefficient of variation.

UNIT OBJECTIVES

After going through this unit, you will be able to:

- Explain how to measure central tendency
- Learn about the geometric mean and harmonic mean
- Discuss the various types of measures of dispersion

- Understand about the coefficient of variation

MEASURES OF CENTRAL TENDENCY

There are several commonly used measures of central tendency such as arithmetic mean, mode and median. These values are very useful not only in presenting the overall picture of the entire data but also for the purpose of making comparisons among two or more sets of data.

As an example, questions like 'How hot is the month of June in Delhi?' can be answered, generally by a single figure of the average for that month. Similarly, suppose we want to find out if boys and girls at age 10 years differ in height for the purpose of making comparisons. Then, by taking the average height of boys of that age and average height of girls of the same age, we can compare and record the differences.

While arithmetic mean is the most commonly used measure of central location, mode and median are more suitable measures under certain set of conditions and for certain types of data. However, each measure of central tendency should meet the following requisites.

1. It should be easy to calculate and understand.
2. It should be rigidly defined. It should have only one interpretation so that the personal prejudice or bias of the investigator does not affect its usefulness.
3. It should be representative of the data. If it is calculated from a sample, then the sample should be random enough to be accurately representing the population.
4. It should have sampling stability. It should not be affected by sampling fluctuations. This means that if we pick 10 different groups of college students at random and compute the average of each group, then we should expect to get approximately the same value from each of these groups.
5. It should not be affected much by extreme values. If few very small or very large items are present in the data, they will unduly influence the value of the average by shifting it to one side or other, so that the average would not be really typical of the entire series. Hence, the average chosen should be such that it is not unduly affected by such extreme values.

Let us consider the three measures of central tendency.

(a) Arithmetic Mean: This is also commonly known as simply the mean. Even though average, in general, means any measure of central location, when we use the word average in our daily routine, we always mean the arithmetic average. The term is widely used by almost every one in daily communication. We speak of an individual being an average student or of average intelligence. We always talk about average family size or average family income or grade point average (GPA) for students and so on.

Combined Mean: If the arithmetic averages and the number of items in two or more related groups are known, the combined (or composite) mean of the entire group can be obtained by the following formula:

$$\overline{\overline{X}} = \left[\frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} \right]$$

The advantage of combined arithmetic mean is that, one can determine the over, all mean of the combined data without having to going back to the original data.

An example:

Find the combined mean for the data given below

$n_1 = 10, x_1 = 2, n_2 = 15, x_2 = 3$

Solution:

$$\overline{X} = \left[\frac{n_1x_1 + n_2x_2}{n_1 + n_2}\right]$$

$$= \left[\frac{10 \times 2 + 15 \times 3}{10 + 15}\right]$$

$$= \frac{20 + 45}{25}$$

$$= 2.6$$

For discussion purposes, let us assume a variable *X* which stands for some scores such as the ages of students. Let the ages of 5 students be 19, 20, 22, 22 and 17 years. Then variable *X* would represent these ages as follows:

X: 19, 20, 22, 22, 17

Placing the Greek symbol S(Sigma) before *X* would indicate a command that all values of *X* are to be added together. Thus:

S*X* = 19 + 20 + 22 + 22 + 17

The mean is computed by adding all the data values and dividing it by the number of such values. The symbol used for sample average is $\overline{X}$ so that:

$$\overline{X} = \frac{19 + 20 + 22 + 22 + 17}{5}$$

In general, if there are *n* values in the sample, then

$$\overline{X} = \frac{X_1 + X_2 + \qquad + X_n}{n}$$

In other words,

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}, \qquad i = 1,\ 2\ ...\ n.$$

The above formula states, add up all the values of X_i where the value of *i* starts at 1 and ends at n with unit increments so that *i* = 1, 2, 3, ... n.

If instead of taking a sample, we take the entire population in our calculations of the mean, then the symbol for the mean of the population is μ (mu) and the size of the population is *N*, so that:

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}, \qquad i = 1,\ 2\ ...N.$$

NOTES

If we have the data in grouped discrete form with frequencies, then the sample mean is given by:

$$\overline{X} = \frac{\Sigma f\,(X)}{\Sigma f}$$

where Σf = Summation of all frequencies' n

$\Sigma f(X)$ = Summation of each value of X multiplied by its corresponding frequency (f).

Example 1: Let us take the ages of 10 students as follows:

19, 20, 22, 22, 17, 22, 20, 23, 17, 18

This data can be arranged in a frequency distribution as follows:

(X)	(f)	$f(X)$
17	2	34
18	1	18
19	1	19
20	2	40
22	3	66
23	1	23
	Total = 10	200

In the above case we have Sf = 10 and S$f(X)$ = 200, so that:

$$\overline{X} = \frac{\Sigma f\,(X)}{\Sigma f}$$

$$= 200/10 = 20$$

Characteristics of the Mean

The arithmetic mean has three interesting properties. These are:

1. The sum of the deviations of individual values of X from the mean will always add up to zero. This means that if we subtract all the individual values from their mean, then some values will be negative and some will be positive, but if all these differences are added together then the total sum will be zero. In other words, the positive deviations must balance the negative deviations. Or symbolically:

$$\sum_{i=1}^{n}(X_i - \overline{X}) = 0, i = 1, 2, ... n.$$

2. The second important characteristic of the mean is that it is very sensitive to extreme values. Since the computation of the mean is based upon inclusion of all values in the data, an extreme value in the data would shift the mean towards it, thus making the mean unrepresentative of the data.
3. The third property of the mean is that the sum of squares of the deviations about the mean is minimum. This means that if we take differences between individual values and the mean and square these differences individually and then add these squared differences, then the final figure will be less than the sum of the squared deviations around any other number other than the mean. Symbolically, it means that:

$$\sum_{i=1}^{n}(X_i - \overline{X})^2 = \text{Minimum}, i = 1, 2, ... n.$$

(b) Mode: The mode is another form of average and can be defined as the most frequently occurring value in the data. The mode is not affected by extreme values in the data and can easily be obtained from an ordered set of data. It can be useful and more representative of the data under certain conditions and is the only measure of central tendency that can be used for qualitative data. For instance, when a researcher quotes the opinion of an average person, he is probably referring to the most frequently expressed opinion which is the modal opinion. In our example of ages of 10 students as:

19, 20, 22, 22, 17, 22, 20, 23, 17 and 18

The mode is 22, since it occurs more often than any other value in this data.

(c) Median: The median is a measure of central tendency and it appears in the centre of an ordered data. It divides the list of ordered values in the data into two equal parts so that half of the data will have values less than the median and half will have values greater than the median.

If the total number of values is odd, then we simply take the middle value as the median. For instance, if there are 5 numbers arranged in order such as 2, 3, 3, 5, 7, then 3 is the middle number and this will be the median. However, if the total number of values in the data is even, then we take the average of the middle two values. For instance, let there be 6 numbers in the ordered data such as 2, 3, 3, 5, 7, 8, then the average of middle two numbers which are 3 and 5 would be the median, which is

$$\text{Median} = \frac{(3+5)}{2} = 4$$

In general, the median is $\frac{n+1}{2}$th observation in the ordered data.

The median is a useful measure in the sense that it is not unduly affected by extreme values and is specially useful in open ended frequencies.

Advantages of Mean

The following are the various advantages of mean:

1. Its concept is familiar to most people and is intuitively clear.
2. Every data set has a mean, which is unique and describes the entire data to some degree. For instance, when we say that the average salary of a professor is ` 25,000 per month, it gives us a reasonable idea about the salaries of professors.
3. It is a measure that can be easily calculated.
4. It includes all values of the data set in its calculation.
5. Its value varies very little from sample to sample taken from the same population.
6. It is useful for performing statistical procedures such as computing and comparing the means of several data sets.

Disadvantages of Mean

The following are the various disadvantages of mean:

1. It is affected by extreme values, and hence, not very reliable when the data set has extreme values especially when these extreme values are on one side of the ordered data. Thus, a mean of such data is not truly a representative of such data. For instance, the average age of three persons of ages 4, 6 and 80 years gives us an average of 30.

2. It is tedious to compute for a large data set as every point in the data set is to be used in computations.
3. We are unable to compute the mean for a data set that has open-ended classes either at the high or at the low end of the scale.
4. The mean cannot be calculated for qualitative characteristics such as beauty or intelligence, unless these can be converted into quantitative figures such as intelligence into IQs.

Advantages of Median

The following are the advantages of median:

1. Median is a positional average and hence the extreme values in the data set do not affect it as much as they do to the mean.
2. Median is easy to understand and can be calculated from any kind of data, even for grouped data with open-ended classes.
3. We can find the median even when our data set is qualitative and can be arranged in the ascending or the descending order, such as average beauty or average intelligence.
4. Similar to mean, median is also unique meaning that there is only one median in a given set of data.
5. Median can be located visually when the data is in the form of ordered data.
6. The sum of absolute differences of all values in the data set from the median value is minimum meaning that it is less than any other value of central tendency in the data set, which makes it more central in certain situations.

Disadvantages of Median

The following are the disadvantages of median:

1. The data must be arranged in order to find the median. This can be very time consuming for a large number of elements in the data set.
2. The value of the median is affected more by sampling variations. Different samples from the same population may give significantly different values of the median.
3. The calculation of median in case of grouped data is based on the assumption that the values of observations are evenly spaced over the entire class interval and this is usually not so.
4. Median is comparatively less stable than the mean, particularly for small samples, due to fluctuations in sampling.
5. Median is not suitable for further mathematical treatment. For instance, we cannot compute the median of the combined group from the median values of different groups.

Advantages of Mode

The following are the advantages of mode:

1. Similar to median, the mode is not affected by extreme values in the data.
2. Its value can be obtained in open-ended distributions without ascertaining the class limits.
3. It can be easily used to describe qualitative phenomenon. For instance, if most people prefer a certain brand of tea then this will become the modal point.
4. Mode is easy to calculate and understand. In some cases it can be located simply by observation or inspection.

Disadvantages of Mode

The following are the disadvantages of mode:

1. Quite often, there is no modal value.
2. It can be bi-modal or multi-modal or it can have all modal values making its significance more difficult to measure.
3. If there is more than one modal value, the data is difficult to interpret.
4. A mode is not suitable for algebraic manipulations.
5. Since the mode is the value of maximum frequency in the data set, it cannot be rigidly defined if such frequency occurs at the beginning or at the end of the distribution.
6. It does not include all observations in the data set, and hence, less reliable in most of the situations.

Weighted Arithmetic Mean

In the computation of arithmetic mean we had given equal importance to each observation in the series. This equal importance may be misleading if the individual values constituting the series have different importance as in the following example:

The Raja Toy shop sells

Toy Cars at	` 3 each
Toy Locomotives at	` 5 each
Toy Aeroplanes at	` 7 each
Toy Double Decker at	` 9 each

What shall be the average price of the toys sold, if the shop sells 4 toys, one of each kind?

Mean Price, i.e., $\bar{x} = \frac{\sum x}{4} = \text{Rs}\,\frac{24}{4} = \text{`}\ 6$

In this case the importance of each observation (Price quotation) is equal in as much as one toy of each variety has been sold. In the above computation of the arithmetic mean this fact has been taken care of by including 'once only' the price of each toy.

But if the shop sells 100 toys: 50 cars, 25 locomotives, 15 aeroplanes and 10 double deckers, the importance of the four price quotations to the dealer is **not equal** as a source of earning revenue. In fact their respective importance is equal to the number of units of each toy sold, i.e.,

The importance of Toy Car	50
The importance of Locomotive	25
The importance of Aeroplane	15
The importance of Double Decker	10

It may be noted that 50, 25, 15, 10 are the quantities of the various classes of toys sold. It is for these quantities that the term 'weights' is used in statistical language. Weight is represented by symbol '*w*', and Σ*w* represents the sum of weights.

While determining the 'average price of toy sold' these weights are of great importance and are taken into account in the manner illustrated below:

$$\bar{x} = \frac{w_1x_1 \square w_2x_2 \square w_3x_3 \square w_4x_4}{w_1 \square w_2 \square w_3 \square w_4} = \frac{\sum wx}{\sum w}$$

When w_1, w_2, w_3, w_4 are the respective weights of x_1, x_2, x_3, x_4 which in turn represent the price of four varieties of toys, viz., car, locomotive, aeroplane and double decker, respectively.

$$\bar{x} = \frac{(50 ¥ 3) \square (25 ¥ 5) \square (15 ¥ 7) \square (10 ¥ 9)}{50 \square 25 \square 15 \square 10}$$

$$= \frac{(150)+(125)+(105)+(90)}{100} = \frac{470}{100} = \text{`}\ 4.70$$

The table below summarizes the steps taken in the computation of the weighted arithmetic mean.

$$\Sigma w = 100; \quad \Sigma wx = 470$$

$$\bar{x} = \frac{\sum wx}{\sum w} = \frac{470}{100} = 4.70$$

The weighted arithmetic mean is particularly useful where we have to compute the *mean of means*. If we are given two arithmetic means, one for each of two different series, in respect of the *same variable*, and are required to find the arithmetic mean of the combined series, the weighted arithmetic mean is the only suitable method of its determination.

Weighted Arithmetic Mean of Toys Sold by the Raja Toy Shop

Toys	*Price per Toy* ` *x*	*Number Sold* *w*	*Price × Weight* *xw*
Car	3	50	150
Locomotive	5	25	125
Aeroplane	7	15	105
Double Decker	9	10	90
		Σ*w* = 100	Σ*xw* = 470

Example 2: The arithmetic mean of daily wages of two manufacturing concerns A Ltd. and B Ltd. is ` 5 and ` 7, respectively. Determine the average daily wages of both concerns if the number of workers employed were 2,000 and 4,000, respectively.

Solution: (*a*) Multiply each average (viz. 5 and 7) by the number of workers in the concern it represents.

(b) Add up the two products obtained in (*a*) above and

(c) Divide the total obtained in (*b*) by the total number of workers.

Weighted Mean of Mean Wages of A Ltd. and B Ltd.

Manufacturing Concern	*Mean Wages* *x*	*Workers Employed* *w*	*Mean Wages × Workers Employed* *wx*
A Ltd.	5	2,000	10,000
B Ltd.	7	4,000	28,000
		$\sum w$ = 6,000	$\sum wx$ = 38,000

$$\bar{x} = \frac{\hat{A}\, wx}{\hat{A}\, w}$$

$$= \frac{38{,}000}{6{,}000}$$

$$= \text{`}\ 6.33$$

The above mentioned examples explain that 'Arithmetic Means and Percentage' are not original data. They are derived figures and their importance is relative to the original data from which they are obtained. This relative importance must be taken into account by weighting while averaging them (means and percentage).

Different Positional Numbers

The position of value in statistics is determined using specific methods for a given set of data or observations. The following are the popular common measures of positions:

- **Percentiles:** Percentiles are those values which divide a given data set into hundred equal parts. It is the value of a variable below which certain per cent of observations fall. For instance, the 25th percentile is the value below which 25 per cent of the observations occur. The 25th percentile is also referred as the first quartile, the 50th percentile as the median or second quartile and the 75th percentile as the third quartile.
- **Quartiles:** It segments the data in four regions and is commonly used to measure the position of value in statistics. It is a number and not a range of values.
- **Standard Scores:** It is also termed as *Z*-values, *Z*-scores, normal scores and standardized variables. It is a dimensionless quantity and can be calculated using the following formula:

$$Z = (X - \mu)/\sigma$$

Measures of Position Values

We have defined the median as the value of the item which is located at the centre of the array, we can define other measures which are located at other specidied points. Thus, the *N*th *percentile* of an array is the value of the item such that *N* per cent items lie *below* it. Clearly then the *N*th percentile P_n of grouped data is given by,

$$P_n = l + \frac{\frac{nN}{100} - C}{f} \times i$$

where, *l* is the lower limit of the class in which *nN*/100th item lies, *i* its width, *f* its frequency, *C* the cumulative frequency upto (but not including) this class, and *N* is the total number of items.

We similarly define the *N*th *decile* as the value of the item below which (*nN*/10) items of the array lie. Clearly,

$$D_n = P_{10n} = l + \frac{\frac{nN}{10} - C}{f} \times i \qquad (5.1)$$

The other most commonly referred to measures of location are the quartiles. Thus, *n*th quartile is the value of the item which lie at the *n*(*N*/4)th item. Clearly Q_2, the second quartile is the median. For grouped data,

$$Q_n = P_{25n} = l + \frac{\frac{nN}{4} - C}{f} \times i \qquad (5.2)$$

Some measures other than measures of central tendency are often employed when summarizing or describing a set of data where it is necessary to divide the data into equal parts. These are positional measures and are called quantiles and consist of quartiles, deciles and percentiles. The quartiles divide the data into four equal parts. The deciles divide the total ordered data into ten equal parts and percentiles divide the data into 100 equal parts. Consequently, there are three quartiles, nine deciles and 99 percentiles. The quartiles are denoted by the symbol *Q* so that Q_1 will be such point in the ordered data which has 25 per cent of the data below and 75 per cent of the data above it. In other words, Q_1 is the value corresponding to $\left(\frac{n+1}{4}\right)$ th ordered observation. Similarly, Q_2 divides the data in the middle, and is also equal to the median and its value Q_2 is given by:

NOTES

$$Q_2 = \text{The value of } 2\left(\frac{n+1}{4}\right) \text{th ordered observation in the data.}$$

Similarly, we can calculate the values of various deciles. For instance,

$$D_1 = \left(\frac{n+1}{10}\right) \text{th observaton in the data, and}$$

$$D_7 = 7\left(\frac{n+1}{10}\right) \text{ th observation in the ordered data.}$$

Percentiles are generally used in the research area of education where people are given standard tests and it is desirable to compare the relative position of the subject's performance on the test. Percentiles are similarly calculated as:

$$P_7 = 7\left(\frac{n+1}{100}\right) \text{ th observation in the ordered data.}$$

and,

$$P_{69} = 69\left(\frac{n+1}{100}\right) \text{ th observation in the ordered data.}$$

Quartiles

The formula for calculating the values of quartiles for grouped data is given as follows.

$$Q = L + (j/f)C$$

Where,

Q = The quartile under consideration.

L = Lower limit of the class interval which contains the value of Q.

j = The number of units we lack from the class interval which contains the value of Q, in reaching the value of Q.

f = Frequency of the class interval containing Q.

C = Size of the class interval.

Let us assume we took the data of the ages of 100 students and a frequency distribution for this data has been constructed as shown.

The frequency distribution is as follows:

Ages (*CI*)	*Mid-point* (*X*)	(*f*)	*f*(*X*)	*f*(*X*)^2
16 and upto 17	16.5	4	66	1089.0
17 and upto 18	17.5	14	245	4287.5
18 and upto 19	18.5	18	333	6160.5
19 and upto 20	19.5	28	546	10647.0
20 and upto 21	20.5	20	410	8405.0
21 and upto 22	21.5	12	258	5547.0
22 and upto 23	22.5	4	90	2025.0
		Totals = 100	1948	38161

In our case, in order to find Q_1, where Q_1 is the cut-off point so that 25 per cent of the data is below this point and 75 per cent of the data is above, we see that the first group has 4 students and the second group has 14 students making a total of 18 students. Since Q_1 cuts off at 25 students, it is the third class interval which contains Q_1. This means that the value of L in our formula is 18.

Since we already have 18 students in the first two groups, we need 7 more students from the third group to make it a total of 25 students, which is the value of Q_1. Hence, the value of (j) is 7. Also, since the frequency of this third class interval which contains Q_1 is 18, the value of (f) in our formula is 18. The size of the class interval C is given as 1. Substituting these values in the formula for Q, we get

$$\begin{aligned} Q_1 &= 18 + (7/18)1 \\ &= 18 + 0.38 = 18.38 \end{aligned}$$

This means that 25 per cent of the students are below 18.38 years of age and 75 per cent are above this age.

Similarly, we can calculate the value of Q_2, using the same formula. Hence,

$$\begin{aligned} Q_2 &= L + (j/f)C \\ &= 19 + (14/28)1 \\ &= 19.5 \end{aligned}$$

This also happens to be the median.

By using the same formula and the same logic we can calculate the values of all deciles as well as percentiles.

We have defined the median as the value of the item which is located at the centre of the array. We can define other measures which are located at other specified points. Thus, the Nth percentile of an array is the value of the item such that N per cent items lie below it. Clearly then, the Nth percentile Pn of grouped data is given by

$$P_n = l + \frac{\frac{nN}{100} - C}{f} \times i$$

Here, l is the lower limit of the class in which nN/100th item lies, i its width, f its frequency, C the cumulative frequency upto (but not including) this class, and N is the total number of items.

We can similarly define the Nth decile as the value of the item below which (nN/10) items of the array lie. Clearly,

$$D_n = P_{10n} = l + \frac{\frac{nN}{10} - C}{f} \times i$$

where the symbols have the obvious meanings.

The other most commonly referred to measures of location are the quartiles. Thus, nth quartile is the value of the item which lies at the n(N/5)th item. Clearly, Q_2, the second quartile is the median for grouped data.

$$Q_n = P_{25n} = l + \frac{\frac{nN}{4} - C}{f} \times i$$

Geometric Mean

If α, β, γ are in GP, then β is called a *geometric mean* between α and γ, written as GM.

If $a_1, a_2, ...,a_n$ are in GP, then $a_2, ..., a_{n-1}$ are called *geometric means* between a_1 and a_n.

Thus, 3, 9, 27 are three geometric means between 1 and 81.

Non-zero quantities $a_1, a_2, a_3, ..., a_n,....$, each term of which is equal to the product of preceding term and a constant number, form a *Geometrical Progression* (written as G.P.).

Thus, all the following quantities are in G.P.

(*a*) 1, 2, 4, 8, 16,...

(*b*) 3, –1, $\frac{1}{3}$, $\frac{-1}{9}$, $\frac{1}{27}$, ...

(*c*) 1, $\sqrt{2}$, 2, 2$\sqrt{2}$,....

(d) a, $\frac{a}{b}$, $\frac{a}{b^2}$, $\frac{a}{b^3}$,..., where $a \neq 0$, $b \neq 0$.

(*e*) 1, $\frac{1}{5}$, $\frac{1}{25}$, $\frac{1}{125}$,...

The constant number is termed as the *common ratio* of the G.P.

The *n*th Term of a G.P.

Let first term be a and r, the common ratio, By definition the G.P. is a, ar, ar^2,...

$$\text{1st term} = a = ar^0 = ar^{1-1}$$

$$\text{2nd term} = ar = ar^1 = ar^{2-1}$$

... ...

In general, nth term = ar^{n-1}.

In examples of the preceding section, we compute 5th, 7th, 3rd, 11th and 8th term of (*a*), (*b*), (*c*), (*d*) and (*e*) respectively.

In (*a*) Ist term is 1 and common ratio = 2.

Hence, 5th term = $ar^4 = 1.2^4 = 16$.

In (*b*) $a = 3$, $r = \frac{-1}{3}$, hence, 7th term = $ar^6 = 3\left(\frac{-1}{3}\right)^6 = \frac{1}{243}$.

In (*c*) $a = 1$, $r = \sqrt{2}$, hence, 3rd term = $ar^2 = 2$.

In (*d*) Ist term = a, $r = \frac{1}{b}$, hence, 11th term = $ar^{10} = \frac{a}{b^{10}}$.

In (*e*) $a = 1$, $r = \frac{1}{5}$, hence, 8th term = $ar^2 = \frac{1}{5^7} = \frac{1}{78125}$.

Sum of First *n* Terms of a G.P.

Let a, ar, ar^2,... be a given G.P. and let S_n be the sum of its first n terms.

Then, $S_n = a + ar + ar^2 + ... + ar^{n-1}$.

This gives that $rS_n = ar + ar^2 + ... + ar^{n-1} + ar^n$

Subtracting, we get, $S_n - r\,S_n = a - ar^n = a\,(1 - r^n)$

In case $r \neq 1$, $S_n = \frac{a(1-r^n)}{(1-r)}$

In case $r = 1$, $S_n = a + a + a + ... + a$ (n times)

$= na.$

Thus, sum of n terms of a G.P. is $\frac{a(1-r^n)}{1-r}$ provided $r \neq 1$.

In case $r = 1$, sum of G.P. is na.

Example 3: Find the sum of the first 14 terms of a G.P.

3, 9, 27, 81, 243, 729,...

Solution: In this case $a = 3$, $r = 3$, $n = 14$.

$$\text{So,} \quad S_n = \frac{a\left(1-r^n\right)}{1-r} = \frac{3\left(1-3^{14}\right)}{1-3}$$

$$= \frac{3}{2}\left(3^{14}-1\right).$$

NOTES

Example 4: Find the sum of first 11 terms of a G.P. given by

$$1,\ -\frac{1}{2},\ \frac{1}{4},\ -\frac{1}{8}\ \ldots,\ldots$$

Solution: Here, $a = 1$, $r = -\frac{1}{2}$, $n = 11$.

$$\text{So,} \quad S_n = \frac{a\left(1-r^n\right)}{1-r} = \frac{1\left[1-\left(-\frac{1}{2}\right)^{11}\right]}{1+\frac{1}{2}}$$

$$= \frac{2^{11}+1}{3\times 2^{10}} = \frac{683}{1024}.$$

To Insert *n* Geometric Means between Two given Numbers *a* and *b*

Let $G_1, G_3, \ldots G_n$ be n geometric means between a and b. Thus, $a, G_1, G_2, \ldots G_n, b$ is a GP, b being $(n+2)$th term $= ar^{n+1}$, where r is the common ratio of GP

$$\text{Thus,} \quad b = ar^{n+1} \Rightarrow r = \left(\frac{b}{a}\right)^{\frac{1}{n+1}}$$

$$\text{So,} \quad G_1 = ar = a\left(\frac{b}{a}\right)^{\frac{1}{n+1}} = \left(a^n b\right)^{\frac{1}{n+1}}$$

$$G_2 = ar^2 = a\left(\frac{b}{a}\right)^{\frac{2}{n+1}} = \left(a^{n-1}b^2\right)^{\frac{1}{n+1}}$$

$$\ldots\ \ldots.\ldots \qquad \ldots \qquad \ldots\ \ldots\ \ldots.\ \ldots\ \ldots\ \ldots\ \ldots.$$

$$G_n = ar^{n-1} = a\left(\frac{b}{a}\right)^{\frac{n-1}{n+1}} = \left(a^2 b^{n-1}\right)^{\frac{1}{n+1}}$$

Example 5: Find 7 GM's between 1 and 256.

Solution: Let $G_1, G_2, \ldots G_7$, be 7 GM's between 1 and 256

Then, 256= 9th term of GP,

$= 1.\,r^8$, where r is the common ratio of the GP

This gives that, $r^8 = 256 \Rightarrow r = 2$

Thus,

$$G_1 = ar = 1.2 = 2$$
$$G_2 = ar^2 = 1.4 = 4$$
$$G_3 = ar^3 = 1.8 = 8$$
$$G_4 = ar^4 = 1.16 = 16$$
$$G_5 = ar^5 = 1.32 = 32$$
$$G_6 = ar^6 = 1.64 = 64$$
$$G_7 = ar^7 = 1.128 = 128$$

Hence, required GM's are 2, 4, 8, 16, 32, 64, 128.

NOTES

Example 6: Sum the series $1 + 3x + 5x^2 + 7x^3 + ...$ up to n terms, $x \neq 1$.

Solution: Note that nth term of this series = $(2n - 1)\, x^{n-1}$

Let $S_n = 1 + 3x + 5x^2 + ... + (2n - 1)\, x^{n-1}$

Then, $xS_n = x + 3x^2 + ... + (2n - 3)\, x^{n-1} + (2n - 1)\, x^n$

Subtracing, we get

$$S_n(1 - x) = 1 + 2x + 2x^2 + ... + 2x^{n-1} - (2n - 1)\, x^n$$

$$= 1 + 2x \left(\frac{1 - x^{n-1}}{1 - x}\right) - (2n - 1)\, x^n$$

$$= \frac{1 - x + 2x - 2x^n - (2n - 1)\, x^n (1 - x)}{1 - x}$$

$$= \frac{1 + x - 2x^n - (2n - 1)\, x^n + (2n - 1)\, x^{n+1}}{1 - x}$$

$$= \frac{1 + x - (2n + 1)\, x^n + (2n - 1)\, x^{n+1}}{1 - x}$$

Hence, $$S = \frac{1 + x - (2n + 1)\, x^n + (2n - 1)\, x^{n+1}}{(1 - x)^2}$$

Example 7: If in a GP $(p + q)$th term = m and $(p - q)$th term = n, then find its pth and qth terms.

Solution: Suppose that the given GP be $a, ar, ar^2, ar^3, . . .$

By hypothesis, $(p + q)$th term $= m = ar^{\,p+q-1}$

$(p - q)$th term $= n = ar^{p-q-1}$

Then, $$\frac{m}{n} = r^{2q} \Rightarrow r = \left(\frac{m}{n}\right)^{1/2q}$$

Hence, $$m = a\left(\frac{m}{n}\right)^{(p+q-1)/2q} \Rightarrow a = m^{(q-p+1)/2q}\; n^{(p+q-1)/2q}$$

Thus, pth term $= ar^{p-1} = m^{1/2}\, n^{1/2} = \sqrt{mn}$

qth term $= ar^{q-1} = m^{\frac{2q-p}{2q}}\; n^{\frac{p}{2q}}$

Example 8: Sum the series $5 + 55 + 555 + ...$ up to n terms.

Solution: Let $S_n = 5 + 55 + 555 + ...$

$S_n = 5\,(1 + 11 + 111 + ...)$

$= \frac{5}{9}\,(9 + 99 + 999 + . . .)$

$= \frac{5}{9}\,[(10 - 1) + (100 - 1) + (1000 - 1) + ...]$

$= \frac{5}{9}\,[(10 + 10^2 + 10^3 + ... + 10^n) - (1 + 1 + ...n \text{ terms})]$

$= \frac{5}{9}\,[(10 + 10^2 + 10^3 +.....+ 10^n) - n]$

$$= \frac{5}{9}\left[\frac{10(1-10^n)}{1-10} - n\right]$$

$$= \frac{5}{9}\left[\frac{10(10^n - 1)}{9} - n\right]$$

$$= \frac{50}{81}(10^n - 1) - \frac{5n}{9}$$

Example 9: If a, b, c, d are in GP, prove that $a^2 - b^2, b^2 - c^2$ and $c^2 - d^2$ are also in GP.

Solution: Since, $\frac{b}{a} = \frac{c}{b} = \frac{d}{c} = k$ (say)

we have, $b = ak, c = bk, d = ck$

i.e., $b = ak, c = ak^2, d = ak^3$

Now, $(b^2 - c^2)^2 = (a^2k^2 - a^2k^4)^2$

$= a^4k^4(1 - k^2)^2$

Also, $(a^2 - b^2)(c^2 - d^2) = (a^2 - a^2k^2)(a^2k^4 - a^2k^6)$

$= a^4(1 - k^2)(k^4 - k^6)$

$= a^4k^4(1 - k^2)^2$

Hence, $(b^2 - c^2) = (a^2 - b^2)(c^2 - d^2)$

This gives that, $a^2 - b^2, b^2 - c^2, c^2 - d^2$ are in GP.

Example 10: Three numbers are in GP. Their product is 64 and sum is $\frac{124}{5}$. Find them.

Solution: Let the numbers be $\frac{a}{r}, a, ar$

Since, $\frac{a}{r} + a + ar = \frac{124}{5}$ and $\frac{a}{r} \times a \times ar = 64,$

we have, $a^3 = 64 \Rightarrow a = 4$

This gives, $\frac{4}{r} + 4 + 4r = \frac{124}{5}$

$\Rightarrow \frac{1}{r} + 1 + r = \frac{31}{5}$

$\Rightarrow \frac{r^2 + 1}{r} = \frac{26}{5}$

$\Rightarrow 5r^2 + 5 = 26r$

$\Rightarrow 5r^2 - 26r + 5 = 0$

$\Rightarrow 5r^2 - 25r - r + 5 = 0$

$\Rightarrow 5r(r - 5) - 1(r - 5) = 0$

$\Rightarrow (r - 5)(5r - 1) = 0$

$\Rightarrow r = \frac{1}{5}$ or 5

In either case, numbers are $\frac{4}{5}$, 4 and 20.

NOTES

Example 11: If a, b, c are in GP and $a^x = b^y = c^z$, prove that

$$\frac{1}{x}+\frac{1}{z} = \frac{2}{y}$$

Solution: a, b, c are in GP, $b^2 = ac$

But, $b^y = a^x \Rightarrow a = b^{y/x}$

and, $b^y = c^z \Rightarrow c = b^{y/z}$

So, we get $b^2 = b^{y/x}.\, b^{y/z}$

$$= b^{y\left(\frac{1}{x}+\frac{1}{z}\right)}$$

$$\Rightarrow \quad 2 = y\left(\frac{1}{x}+\frac{1}{z}\right)$$

$$\Rightarrow \quad \frac{1}{x}+\frac{1}{z} = \frac{2}{y}$$

Example 12: Sum to n terms the series

$$0.7 + 0.77 + 0.777 + \ldots$$

Solution: Given series,

$$= 0.7 + 0.77 + 0.777 + \ldots \text{ up to } n \text{ terms}$$

$$= 7\,(0.1 + 0.11 + 0.111 + \ldots \text{ up to } n \text{ terms})$$

$$= \frac{7}{9}(0.9 + 0.99 + 0.999 + \ldots \text{ up to } n \text{ terms})$$

$$= \frac{7}{9}\left[\left(1-\frac{1}{10}\right)+\left(1-\frac{1}{10^2}\right)+\left(1-\frac{1}{10^3}\right)+\ldots\right]$$

$$= \frac{7}{9}\left[n-\left(\frac{1}{10}+\frac{1}{10^2}+\ldots \text{ up to } n \text{ terms}\right)\right]$$

$$= \frac{7}{9}\left[n-\frac{\frac{1}{10}(1-1/10^n)}{1-\frac{1}{10}}\right]$$

$$= \frac{7}{9}\left[n-\frac{1}{9}\left(1-\frac{1}{10^n}\right)\right]$$

$$= \frac{7}{9}\left[n-\frac{1}{9}\left(1-\frac{1}{10^n}\right)\right]$$

Example 13: The sum of three numbers in GP is 35 and their product is 1000. Find the numbers.

Solution: Let the numbers be $\frac{\alpha}{r}$, α, αr

The product of $\frac{\alpha}{r} \times \alpha \times \alpha r = 1000$

$$\alpha^3 = 1000$$

$$\Rightarrow \quad \alpha = 10$$

So, the numbers are $\frac{10}{r}$, 10, $10r$

The sum of these numbers = 35

$\Rightarrow \quad \frac{10}{r} + 10 + 10r = 35$

$\Rightarrow \quad \frac{2}{r} + 2r = 5$

$\Rightarrow \quad 2r^2 - 5r + 2 = 0$

$\Rightarrow \quad (2r - 1)(r - 2) = 0$

$\Rightarrow \quad r = 2 \quad \text{or} \quad \frac{1}{2}$

$r = 2$ gives the numbers as 5, 10, 20

$r = \frac{1}{2}$, gives the numbers as 20, 10, 5, the same as the first set.

Hence, the required numbers are 5, 10 and 20.

Example 14: The sum of the first eight terms of a GP (of real terms) is five times the sum of the first four terms. Find the common ratio.

Solution: Let the GP be $a, ar, ar^2, \ldots$

$$S_8 = \text{Sum of first eight terms} = \frac{a(1 \square r^8)}{1 \square r}$$

$$S_4 = \text{Sum of first four terms} = \frac{a(1 \square r^4)}{1 \square r}$$

By hypothesis, $\quad S_8 = 5S_4 \Rightarrow \frac{a(1 \square r^8)}{1 \square r} = \frac{5a(1 \square r^4)}{1 \square r}$

$\Rightarrow \quad 1 - r^8 = 5(1 - r^4)$

$\Rightarrow \quad (1 - r^4)(1 + r^4) = 5(1 - r^4)$

In case, $\quad r^4 - 1 = 0 \quad$ we get, $r^2 - 1 = 0 \Rightarrow r = \pm 1$

(Note that $\quad r^2 + 1 = 0 \quad \Rightarrow \quad r$ is imaginary)

Now, $\quad r = 1 \quad \Rightarrow \quad$ The given series is $a + a + a + \ldots$

but, $\quad S_8 = 8a \quad$ and $\quad S_4 = 4a$

So, $S_8 \neq 5S_4$

In case $r = -1$, we get, $S_8 = 0$ and $S_4 = 0$, hence the hypothesis is satisfied.

Suppose now, $\quad r^4 - 1 \neq 0, \quad$ then $1 + r^4 = 5$

$\Rightarrow \quad r^4 = 4 \quad \Rightarrow \quad r^2 = 2 \quad (r^2 \neq -2)$

$\Rightarrow \quad r = \pm\sqrt{2}$

Hence, $\quad r = -1 \quad \text{or} \quad \pm\sqrt{2}$

Example 15: If S is the sum, P the product of n term of G.P. and R the sum of reciprocals of n terms in GP, then prove that

$$P^2R^n = S^n.$$

Solution: Let $\quad a, ar, ar^2, \ldots$ be the given GP

Then, $\quad S = a + ar + ar^2 + \ldots \quad$ up to n terms

$$= \frac{a(1 \square r^n)}{1 \square r} \qquad \ldots(1)$$

NOTES

$$P = a \cdot ar \cdot ar^2 \cdots ar^{n-1}$$

$$= a^n r^{1+2+3+\ldots+(n-1)}$$

$$= a^n r^{\frac{(n \square 1)}{2}(2 \square n \square 2)}$$

$$= a^n r^{n \square \frac{n \square 1}{2} \square n} \qquad \ldots(2)$$

$$R = \frac{1}{a} \square \frac{1}{ar} \square \frac{1}{ar^2} \square \ldots \quad \text{up to } n \text{ terms}$$

$$= \frac{\frac{1}{a} \square 1 \square \frac{1}{r^n} \square}{1 \square \frac{1}{r}} = \frac{r}{a} \frac{(r^n \square 1)}{(r \square 1)\, r^n}$$

$$= \frac{(1 \square r^n)}{a\,(1 \square r)\, r^{n \square 1}} \qquad \ldots(3)$$

By Equations (2) and (3),

$$P^2R^n = a^{2n} r^{n(n-1)} \frac{(1 \square r^n)^n}{a^n\,(1 \square r)^n\, r^{n(n \square 1)}}$$

$$= \frac{a^n (1 \square r^n)^n}{(1 \square r)^n} = S^n, \text{ by (1)}$$

Example 16: The ratio of the 4th to the 12th term of a GP with positive common ratio is $\frac{1}{256}$. If the sum of the two terms is 61.68, find the sum of series to 8 terms.

Solution: Let the series be $a, ar, ar^2, \ldots,$

$$T_4 = \text{4th term} = ar^3$$

$$T_{12} = \text{12th term} = ar^{11}$$

By hypothesis, $\frac{T_4}{T_{12}} = \frac{1}{256}$

i.e., $\frac{ar^3}{ar^{11}} = \frac{1}{256}$

$$\frac{1}{r^8} = \frac{1}{256}$$

$\Rightarrow$ $r^8 = 256$

$\Rightarrow$ $r = \pm 2$

Since r is given to be positive, we reject negative sign.

Again, it is given that

$$T_4 + T_{12} = 61.68$$

i.e., $a\,(r^3 + r^{11}) = 61.68$

$$a\,(8 + 2048) = 61.68$$

$$a = \frac{61.68}{2056} = 0.03$$

Hence, S_8 = Sum to eight terms

$$= \frac{a\,(1 \square r^8)}{1 \square r} \square \frac{a\,(r^8 \square 1)}{r \square 1}$$

$$= \frac{(0.03)\,(256 - 1)}{(2 - 1)} = 0.03 \times 255 = 7.65$$

NOTES

Example 17: A manufacturer reckons that the value of a machine which costs him Rs 18750 will depreciate each year by 20%. Find the estimated value at the end of 5 years.

Solution: At the end of first year the value of machine is

$$= 18750 \times \frac{80}{100} = \frac{4}{5}\,(18750)$$

At the end of 2nd year it is equal to $\left(\frac{4}{5}\right)^2$ (18750); proceeding in this manner, the estimated value of machine at the end of 5 years is $\left(\frac{4}{5}\right)^5$ (18750)

$$= \frac{64 \times 16}{125 \times 25} \times 18750$$

$$= \frac{1024}{125} \times 750 = 1024 \times 6$$

$$= 6144 \text{ rupees}$$

Example 18: Show that a given sum of money accumulated at 20 % per annum, more than doubles itself in 4 years at compound interest.

Solution: Let the given sum be *a* rupees. After 1 year it becomes $\frac{6a}{5}$ (it is increased by $\frac{a}{5}$).

At the end of two years it becomes $\frac{6}{5}\left(\frac{6a}{5}\right) = \left(\frac{6}{5}\right)^2 a$

Proceeding in this manner, we get that at the end of 4th year, the amount will be $\left(\frac{6}{5}\right)^4 a = \frac{1296}{625}\, a$

Now, $\frac{1296}{625}\, a - 2a = \frac{46}{625}\, a$, since *a* is a + ve quantity, so the amount after 4 years is more than double of the original amount.

Example 19: If

$$x = a + \frac{a}{r} + \frac{a}{r^2} + \dots \infty$$

$$y = b - \frac{b}{r} + \frac{b}{r^2} + \dots \infty$$

and

$$z = c + \frac{c}{r^2} + \frac{c}{r^4} + \dots \infty$$

Show that

$$\frac{xy}{z} = \frac{ab}{c}$$

Solution: Clearly,

$$x = \frac{a}{1 - \frac{1}{r}} = \frac{ar}{r - 1},$$

$$y = \frac{b}{1 - (-1/r)} = \frac{br}{r + 1}$$

NOTES

and, $$z = \frac{c}{1 \square \frac{1}{r^2}} \square \frac{cr^2}{r^2 \square 1}$$

Now, $$\frac{xy}{z} = \frac{abr^2}{(r^2-1)} \Big/ \left(\frac{cr^2}{r^2-1} \right) = \frac{ab}{c}$$

Example 20: If $a^2 + b^2$, $ab + bc$ and $b^2 + c^2$ are in GP, prove that a, b, c are also in GP.

Solution: Since $a^2 + b^2$, $ab + bc$ and $b^2 + c^2$ are in GP, we get,

$$(ab + bc)^2 = (a^2 + b^2)(b^2 + c^2)$$

$$b^2(a^2 + 2ac + c^2) = a^2b^2 + a^2c^2 + b^4 + b^2c^2$$

$$\Rightarrow \quad 2ab^2c = a^2c^2 + b^4$$

$$\Rightarrow \quad a^2c^2 - 2ab^2c + b^4 = 0$$

$$\Rightarrow \quad (ac - b^2)^2 = 0$$

$$\Rightarrow \quad ac = b^2$$

$\Rightarrow$ a, b, c are in GP.

Harmonic Mean

If a, b, c are in HP, then b is called a *Harmonic Mean* between a and c, written as HM.

Harmonical Progression

Non zero quantities whose reciprocals are in AP, are said to be in *Harmonical Progression*, written as HP

Consider the following examples:

(*a*) $1, \frac{1}{3}, \frac{1}{5}, \frac{1}{7}, \ldots\ldots$

(*b*) $\frac{1}{2}, \frac{1}{5}, \frac{1}{8}, \frac{1}{11}, \ldots\ldots$

(*c*) $2, \frac{5}{2}, \frac{10}{3}, \ldots$

(*d*) $\frac{1}{a}, \frac{1}{a \square b}, \frac{1}{a \square 2b}, \ldots\ldots\ldots a, b \square 0$

(e) $5, \frac{55}{9}, \frac{55}{7}, 11, \ldots\ldots$

It can be easily checked that in each case, the series obtained by taking reciprocal of each of the term is an AP.

To Insert *n* Harmonic Means between *a* and *b*

Let $H_1, H_2, H_3, \ldots, H_n$ be the required Harmonic Means. Then $a, H_1, H_2, \ldots, H_n, b$ are in HP i.e., $\frac{1}{a}, \frac{1}{H_1}, \frac{1}{H_2}, ..., \frac{1}{H_n}, \frac{1}{b}$ are in AP

Then, $\frac{1}{b} = (n + 2)$th term of an AP

$= \frac{1}{a} + (n + 1)d$

Where d is the common difference of AP

This gives, $d = \frac{a \square b}{(n \square 1)ab}$

Now, $\frac{1}{H_1} = \frac{1}{a} \square d \square \frac{1}{a} \square \frac{a \square b}{(n \square 1)\, ab}$

$= \frac{nb \square b \square a \square b}{(n \square 1)\, ab} \square \frac{a \square nb}{(n \square 1)\, ab}$

So, $\frac{1}{H_1} = \frac{a \square nb}{(n \square 1)\, ab}$

⇒ $H_1 = \frac{(n \square 1)\, ab}{a \square nb}$

Again, $\frac{1}{H_2} = \frac{1}{a} \square 2d \square \frac{1}{a} \square \frac{2\,(a \square b)}{(n \square 1)\, ab}$

$= \frac{nb \square b \square 2a \square 2b}{(n \square 1)\, ab} \square \frac{2a \square b \square nb}{(n \square 1)\, ab}$

⇒ $H_2 = \frac{(n \square 1)\, ab}{2a \square b \square nb}$

Similarly, $\frac{1}{H_3} = \frac{1}{a} \square 3d \square \frac{3a \square 2b \square nb}{(n \square 1)\, ab}$

⇒ $H_3 = \frac{(n \square 1)\, ab}{3a \square 2b \square nb}$ and so on,

$\frac{1}{H_n} = \frac{1}{a} \square nd \square \frac{1}{a} \square \frac{n\,(a \square b)}{(n \square 1)\, ab}$

$= \frac{nb \square b \square na \square nb}{(n \square 1)\, ab}$

$= \frac{na \square b}{(n \square 1)\, ab} \Rightarrow H_n = \frac{(n \square 1)\, ab}{na \square b}$

Example 21: Find the 5th term of 2, $2\frac{1}{2}$, $3\frac{1}{3}$, … …

Solution: Let 5th term be x. Then, $\frac{1}{x}$ is 5th term of corresponding AP $\frac{1}{2}, \frac{2}{5}, \frac{3}{10}, \ldots\ldots$

Then, $\frac{1}{x} = \frac{1}{2} \pm 4\left(\frac{2}{5} - \frac{1}{2}\right) = \frac{1}{2} + 4\begin{pmatrix}-1\\ 10\end{pmatrix}$

$$\Rightarrow \qquad \frac{1}{x} = \frac{1}{2} \square \frac{2}{5} \square \frac{1}{10} \Rightarrow x = 10$$

Example 22: Insert two harmonic means between $\frac{1}{2}$ and $\frac{4}{17}$.

Solution: Let H_1, H_2 be two harmonic means between $\frac{1}{2}$ and $\frac{4}{17}$

Thus, 2, $\frac{1}{H_1}, \frac{1}{H_2}, \frac{17}{4}$ are in AP Let d be their common difference

Then, $$\frac{17}{4} = 2 + 3d$$

$$\Rightarrow \qquad 3d = \frac{9}{4} \quad \Rightarrow \quad d = \frac{3}{4}$$

Thus, $$\frac{1}{H_1} = 2 + \frac{3}{4} \square \frac{11}{4} \quad \Rightarrow \quad H_1 = \frac{4}{11}$$

$$\frac{1}{H_2} = 2 + 2 \times \frac{3}{4} \square \frac{7}{2} \quad \Rightarrow \quad H_2 = \frac{2}{7}$$

Required harmonic means are $\frac{4}{11}, \frac{2}{7}$.

Another important feature of a frequency distribution is its **variability**. The simplest measure of variability is the **range**, which is the difference between the highest and the lowest scores. In most applications, it is best to report the actual highest and lowest scores, as opposed to just the range.

MEASURES OF VARIATION

A measure of dispersion, or simply dispersion may be defined as statistics signifying the extent of the scatteredness of items around a measure of central tendency.

A measure of dispersion may be expressed in an 'absolute form', or in a 'relative form'. It is said to be in an absolute form when it states the actual amount by which the value of an item on an average deviates from a measure of central tendency. Absolute measures are expressed in concrete units, i.e., units in terms of which the data have been expressed, e.g., rupees, centimetres, kilograms, etc., and are used to describe frequency distribution.

A relative measure of dispersion computed is a quotient obtained by dividing the absolute measures by a quantity in respect to which absolute deviation has been computed. It is as such a pure number and is usually expressed in a percentage form. Relative measures are used for making comparisons between two or more distributions.

A measure of dispersion should possess all those characteristics which are considered essential for a measure of central tendency, which are as follows:

- It should be based on all observations.
- It should be readily comprehensible.

- It should be fairly easily calculated.
- It should be affected as little as possible by fluctuations of sampling.
- It should be amenable to algebraic treatment.

Some common measures of dispersion are (*a*) The range, (*b*) the semi-interquartile range or the quartile deviation, (*c*) the mean deviation, and (*d*) the standard deviation. Of these, the standard deviation is the best measure. We describe these measures in the following sections.

Types of Measures

The following are the various types of measures:

Quartile Deviation

There are many types of measures of dispersion, one of this is the semi-interquartile range, usually termed as 'quartile deviation'. Quartiles are the points which divide the array into four equal parts. More precisely, Q_1 gives the value of the item 1/4th the way up the distribution and Q_3 the value of the item 3/4th the way up the distribution. Between Q_1 and Q_3 are included half the total number of items. The difference between Q_1 and Q_3 includes only the central items but excludes the extremes. Since under most circumstances, the central half of the series tends to be fairly typical of all the items, the interquartile range ($Q_3 – Q_1$) affords a convenient and often a good indicator of the absolute variability. The larger the interquartile range, the larger the variability.

Usually, one-half of the difference between Q_3 and Q_1 is used and it is given the name of quartile deviation or semi-interquartile range. The interquartile range is divided by 2 for the reason that half of the interquartile range will, in a normal distribution, be equal to the difference between the median and any quartile. This means that 50 per cent items of a normal distribution will lie within the interval defined by the median plus and minus the semi-interquartile range.

Symbolically,

$$\text{Q.D.} = \frac{Q_3 - Q_1}{2} \qquad \text{...(5.3)}$$

Let us find quartile deviations for the weekly earnings of labour in the four workshops whose data is given in Table 5.1. The computations are as shown in Table 5.1.

Table 5.1 *Weekly Earnings of Labourers in Four Workshops of the Same Type*

Weekly earnings `	*No. of workers*			
	Workshop A	*Workshop B*	*Workshop C*	*Workshop D*
15–16	...	...	2	...
17–18	...	2	4	...
19–20	...	4	4	4
21–22	10	10	10	14
23–24	22	14	16	16
25–26	20	18	14	16
27–28	14	16	12	12
29–30	14	10	6	12
31–32	...	6	6	4
33–34	...	...	2	2
35–36	...	...	...	...
37–38	...	...	4	...

Total	80	80	80	80
Mean	25.5	25.5	25.5	25.5

Workshop	*Range*
A	9
B	15
C	23
D	15

As shown in Table 5.2, Q.D. of workshop *A* is ` 2.12 and median value in 25.3. This means that if the distribution is symmetrical, the number of workers whose wages vary between (25.3–2.1) = ` 23.2 and (25.3 + 2.1) = ` 27.4, shall be just half of the total cases. The other half of the workers will be more than ` 2.1 removed from the median wage. As this distribution is not symmetrical, the distance between Q_1 and the median Q_2 is not the same as between Q_3 and the median. Hence, the interval defined by median plus and minus semi inter-quartile range will not be exactly the same as given by the value of the two quartiles. Under such conditions the range between ` 23.2 and ` 27.4 will not include precisely 50 per cent of the workers.

If quartile deviation is to be used for comparing the variability of any two series, it is necessary to convert the absolute measure to a coefficient of quartile deviation. To do this the absolute measure is divided by the average size of the two quartiles.

Symbolically,

$$\text{Coefficient of quartile deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1} \qquad ...(5.4)$$

Applying this to our illustration of four workshops in Table 5.1 the coefficients of Q.D. are as given in Table 5.2.

Table 5.2 *Calculation of Quartile Deviation*

		Workshop	*Workshop*	*Workshop*	*Workshop*
		A	*B*	*C*	*D*
Location of Q_2	$\frac{N}{2}$	$\frac{80}{2}=40$	$\frac{80}{2}=40$	$\frac{80}{2}=40$	$\frac{80}{2}=40$
	Q_2	$24.5+\frac{40-30}{22}\times 2$ = 24.5 + 0.9 = 25.4	$24.5+\frac{40-30}{18}\times 2$ = 24.5 + 1.1 = 25.61	$24.5+\frac{40-30}{16}\times 2$ = 24.5 + 0.75 = 25.25	$24.5+\frac{40-30}{16}\times 2$ = 24.5 + 0.75 = 25.25
Location of Q_1	$\frac{N}{4}$	$\frac{80}{4}=20$	$\frac{80}{4}=20$	$\frac{80}{4}=20$	$\frac{80}{4}=20$
	Q_1	$22.5+\frac{20-10}{22}\times 2$ = 22.5 + .91 = 23.41	$22.5+\frac{20-16}{14}\times 2$ = 22.5 + .57 = 23.07	$20.5+\frac{20-10}{10}\times 2$ = 20.5 + 2 = 22.5	$22.5+\frac{20-18}{16}\times 2$ = 22.5 + .25 = 22.75
Location of Q_3	$\frac{3N}{4}$	$3\times\frac{80}{4}=60$	60	60	60
	Q_3	$26.5+\frac{60-52}{14}\times 2$ = 26.5 + 1.14 = 27.64	$26.5+\frac{60-48}{16}\times 2$ = 26.5 + 1.5 = 28.0	$26.5+\frac{60-50}{12}\times 2$ = 26.5 + 1.67 = 28.17	$26.5+\frac{60-50}{12}\times 2$ = 26.5 + 1.67 = 28.17

Quartile Deviation $\frac{Q_3-Q_1}{2}$	$\frac{27.64-23.41}{2}$ $=\frac{4.23}{2}$ = ` 2.12	$\frac{28-23.07}{2}$ $=\frac{4.93}{2}$ = ` 2.46	$\frac{28.17-22.5}{2}$ $=\frac{5.67}{2}$ = ` 2.83	$\frac{28.17-22.75}{2}$ $=\frac{5.42}{2}$ = `. 2.71
Coefficient of quartile deviation = $\frac{Q_3-Q_1}{Q_3+Q_1}$	$\frac{27.64-23.41}{27.64+23.41}$ = 0.083	$\frac{28-23.07}{28+23.07}$ = 0.097	$\frac{28.17-22.5}{28.17+22.5}$ = 0.112	$\frac{28.17-22.75}{28.17+22.75}$ = 0.106

Characteristics of Quartile Deviation

The following are the characteristics of quartile deviation:

(a) The size of the quartile deviation gives an indication about the uniformity or otherwise of the size of the items of a distribution. If the quartile deviation is small, it denotes large uniformity. Thus, a coefficient of quartile deviation may be used for comparing uniformity or variation in different distributions.

(b) Quartile deviation is not a measure of dispersion in the sense that it does not show the scatter around an average, but only a distance on scale. Consequently, quartile deviation is regarded as a measure of partition.

(c) It can be computed when the distribution has open-end classes.

Limitations of Quartile Deviation

Except for the fact that its computation is simple and it is easy to understand, a quartile deviation does not satisfy any other test of a good measure of variation.

Mean Deviation

In the following section you will study that a weakness of the measures of dispersion, based upon the range or a portion thereof, is that the precise size of most of the variants has no effect on the result. As an illustration, the quartile deviation will be the same whether the variates between Q_1 and Q_3 are concentrated just above Q_1 or they are spread uniformly from Q_1 to Q_3. This is an important defect from the viewpoint of measuring the divergence of the distribution from its typical value. The mean deviation is employed to answer the objection.

Mean deviation, also called average deviation, of a frequency distribution is the mean of the absolute values of the deviation from some measure of central tendency. In other words, mean deviation is the arithmetic average of the variations (deviations) of the individual items of the series from a measure of their central tendency.

We can measure the deviations from any measure of central tendency, but the most commonly employed ones are the median and the mean. The median is preferred because it has the important property that the average deviation from it is the least.

Calculation of mean deviation then involves the following steps:

(a) Calculate the median (or the mean) Me (or $\overline{X}$).

(b) Record the deviations $|d| = |x - Me|$ of each of the items, ignoring the sign.

(c) Find the average value of deviations.

$$\text{Mean Deviation} = \frac{\sum |d|}{N} \qquad ...(5.5)$$

Example 23 explains it better.

Example 23: Calculate the mean deviation from the following data giving marks obtained by 11 students in a class test.

14, 15, 23, 20, 10, 30, 19, 18, 16, 25, 12.

Solution:

Median = Size of $\frac{11+1}{2}$ th item

= size of 6th item = 18.

Serial No.	*Marks*	*| x – Median |* *| d |*
1	10	8
2	12	6
3	14	4
4	15	3
5	16	2
6	18	0
7	19	1
8	20	2
9	23	5
10	25	7
11	30	12
		$\sum \lvert d \rvert = 50$

Mean deviation from median $= \frac{\sum |d|}{N}$

$= \frac{50}{11} = 4.5$ marks

For grouped data, it is easy to see that the mean deviation is given by

$$\text{Mean deviation} = \frac{\sum f\,|d|}{\sum f} \qquad ...(5.5)$$

where,

$|\,d\,| = |\,x - \text{median}\,|$ for grouped discrete data

$|\,d\,| = M - \text{median}\,|$ for grouped continuous data with M as the mid-value of a particular group.

Examples 2.24 and 25 illustrate the use of this formula.

Example 24: Calculate the mean deviation from the following data:

Size of item	6	7	8	9	10	11	12
Frequency	3	6	9	13	8	5	4

Solution:

Size	*Frequency* *f*	*Cumulative frequency*	*Deviations from median (9)* | *d* |	*f* | *d* |
6	3	3	3	9
7	6	9	2	12
8	9	18	1	9
9	13	31	0	0
10	8	39	1	8
11	5	44	2	10
12	4	48	3	12
	48			60

$$\text{Median} = \text{the size of } \frac{48+1}{2} = 24.5\text{th item which is } 9.$$

Therefore, deviations *d* are calculated from 9, i.e., $|d| = |x - 9|$.

$$\text{Mean deviation} = \frac{\sum f|d|}{\sum f} = \frac{60}{48} = 1.25$$

Example 25: Calculate the mean deviation from the following data:

x	0–10	10–20	20–30	30–40	40–50	50–60	60–70	70–80
f	18	16	15	12	10	5	2	2

Solution:

This is a frequency distribution with continuous variable. Thus, deviations are calculated from mid-values.

x	*Mid-value*	*f*	*Less than c.f.*	*Deviation from median* | *d* |	*f* | *d* |
0–10	5	18	18	19	342
10–20	15	16	34	9	144
20–30	25	15	49	1	15
30–40	35	12	61	11	132
40–50	45	10	71	21	210
50–60	55	5	76	31	155
60–70	65	2	78	41	82
70–80	75	2	80	51	102
		80			1182

$$\text{Median} = \text{The size of } \frac{80}{2} \text{ th item}$$

$$= 20 + \frac{6}{15} \times 10 = 24$$

NOTES

and then, mean deviation $= \frac{\Sigma f\,|d|}{\Sigma f}$

$$= \frac{1182}{80} = 14.775.$$

Merits and Demerits of the Mean Deviation

Merits

The merits are as follows:

1. It is easy to understand.
2. As compared to standard deviation (discussed later), its computation is simple.
3. As compared to standard deviation, it is less affected by extreme values.
4. Since it is based on all values in the distribution, it is better than range or quartile deviation.

Demerits

The demerits are as follows:

1. It lacks those algebraic properties which would facilitate its computation and establish its relation to other measures.
2. Due to this, it is not suitable for further mathematical processing.

Coefficient of Mean Deviation

The coefficient or relative dispersion is found by dividing the mean deviations recorded. Thus,

$$\text{Coefficient of MD} = \frac{\text{Mean Deviation}}{\text{Mean}} \qquad ...(5.6)$$

(when deviations were recorded from the mean)

$$= \frac{\text{MD}}{\text{Median}} \qquad ...(5.7)$$

(when deviations were recorded from the median)

Applying the above formula to Example 25.

$$\text{Coefficient of MD} = \frac{14.775}{24}$$

$$= 0.616$$

Standard Deviation

By far the most universally used and the most useful measure of dispersion is the standard deviation or root mean square deviation about the mean. We have seen that all the methods of measuring dispersion so far discussed are not universally adopted for want of adequacy and accuracy. The range is not satisfactory as its magnitude is determined by most extreme cases in the entire group. Further, the range is notable because it is dependent on the item whose size is largely a matter of chance. Mean deviation method is also an unsatisfactory measure of scatter, as it ignores the algebraic signs of deviation. We desire a measure of scatter which is free from these shortcomings. To some extent standard deviation is one such measure.

The calculation of standard deviation differs in the following respects from that of mean deviation. First, in calculating standard deviation, the deviations are squared. This is done so as to get rid of negative signs without committing algebraic violence. Further, the squaring of deviations provides added weight to the extreme items, a desirable feature for certain types of series.

Second, the deviations are always recorded from the arithmetic mean, because although the sum of deviations is the minimum from the median, the sum of squares of deviations is minimum when deviations are measured from the arithmetic average. The deviation from $\bar{x}$ is represented by σ.

Thus, standard deviation, σ (sigma) is defined as the square root of the mean of the squares of the deviations of individual items from their arithmetic mean.

$$\sigma = \sqrt{\frac{\sum (x - \bar{x})^2}{N}} \qquad ...(5.8)$$

For grouped data (discrete variables),

$$\sigma = \sqrt{\frac{\sum f\,(x - \bar{x})^2}{\sum f}} \qquad ...(5.9)$$

and, for grouped data (continuous variables),

$$\sigma = \sqrt{\frac{\sum f\,(M - \bar{x})}{\sum f}} \qquad ...(5.10)$$

where *M* is the mid-value of the group.

The use of these formulae is illustrated by Examples 26 and 27.

Example 26: Compute the standard deviation for the following data:

11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21.

Solution:

Here, Formula (2.8) is appropriate. We first calculate the mean as $\bar{x} = \sum x/N = 176/11 = 16$, and then calculate the deviation as follows:

x	$(x - \bar{x})$	$(x - \bar{x})^2$
11	–5	25
12	–4	16
13	–3	9
14	–2	4
15	–1	1
16	0	0
17	+1	1
18	+2	4
19	+3	9
20	+4	16
21	+5	25
176		110

Thus by Formula (5.8),

$$\sigma = \sqrt{\frac{110}{11}} = \sqrt{10} = 3.16$$

Example 27: Find the standard deviation of the data in the following distributions:

x	12	13	14	15	16	17	18	20
f	4	11	32	21	15	8	6	4

Solution:

For this discrete variable grouped data, we use Formula (5.8). Since for calculation of $\bar{x}$, we need $\sum fx$ and then for σ we need $\sum f(x-\bar{x})^2$, the calculations are conveniently made in the following format.

x	f	fx	$d = x - \bar{x}$	d^2	fd^2
12	4	48	–3	9	36
13	11	143	–2	4	44
14	32	448	–1	1	32
15	21	315	0	0	0
16	15	240	1	1	15
17	8	136	2	4	32
18	5	90	3	9	45
20	4	80	5	25	100
	100	1500			304

Here, $\bar{x} = \sum fx/\sum f = 1500/100 = 15$

and $$\sigma = \sqrt{\frac{\sum fd^2}{\sum f}}$$

$$= \sqrt{\frac{304}{100}} = \sqrt{3.04} = 1.74$$

Calculation of Standard Deviation by Short-Cut Method

In most cases, it is very unlikely that $\bar{x}$ will turn out to be an integer simplifying problems. In such cases, the calculation of σ and σ^2 becomes quite time-consuming. Short-cut methods have consequently been developed. These are on the same lines asthose for calculation of mean itself.

In the short-cut method, we calculate deviations x' from an assumed mean A. Then for ungrouped data,

$$\sigma = \sqrt{\frac{\sum x'^2}{N} - \left(\frac{\sum x'}{N}\right)^2} \qquad ...(5.11)$$

and for grouped data

$$\sigma = \sqrt{\frac{\sum fx'^2}{\sum f} - \left(\frac{fx'}{\sum f}\right)^2} \qquad ...(5.12)$$

This formula is valid for both discrete and continuous variables. In case of continuous variables, x in the equation $x' = x - A$ stands for the mid-value of the class in question.

Note that the second term in each of the formulae is a correction term because of the difference in the values of A and $\bar{x}$. When A is taken as $\bar{x}$ itself, this correction is automatically reduced to zero. Examples 28 to 30 explain the use of these formulae.

Example 28: Compute the standard deviation by the short-cut method for the following data:

11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21

Solution: Let us assume that A = 15.

	$x' = (x - 15)$	x'^2
11	–4	16
12	–3	9
13	–2	4
14	–1	1
15	0	0
16	1	1
17	2	4
18	3	9
19	4	16
20	5	25
21	6	36
$N = 11$	$\sum x' = 11$	$\sum x'^2 = 121$

$$\sigma = \sqrt{\frac{\sum x'^2}{N} - \left(\frac{\sum x'}{N}\right)^2}$$

$$= \sqrt{\frac{121}{11} - \left(\frac{11}{11}\right)^2}$$

$$= \sqrt{11 - 1}$$

$$= \sqrt{10}$$

$$= 3.16$$

Another Method

If we assume A as zero, then the deviation of each item from the assumed mean is the same as the value of item itself. Thus, 11 deviates from the assumed mean of zero by 11, 12 deviates by 12, and so on. As such, we work with deviations without having to compute them, and the formula takes the following shape:

x	x^2
11	121
12	144
13	169
14	196
15	225
16	256
17	289
18	324
19	361
20	400
21	441
176	2,926

$$\sigma = \sqrt{\frac{\sum x^2}{N} - \left(\frac{\sum x}{N}\right)^2}$$

$$= \sqrt{\frac{2926}{11} - \left(\frac{176}{11}\right)^2} = \sqrt{266 - 256} = 3.16$$

Combining Standard Deviations of Two Distributions

If we were given two sets of data of N_1 and N_2 items with means $\bar{x}_1$ and $\bar{x}_2$ and standard deviations σ_1 and σ_2 respectively, we can obtain the mean and standard deviation $\bar{x}$ and σ of the combined distribution by the following formulae:

$$\bar{x} = \frac{N_1\bar{x}_1 + N_2\bar{x}_2}{N_1 + N_2} \qquad ...(5.13)$$

and $$\sigma = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_1(\bar{x} - \bar{x}_1)^2 + N_2(\bar{x} - \bar{x}_2)^2}{N_1 + N_2}} \qquad ...(5.14)$$

Example 29: The mean and standard deviations of two distributions of 100 and 150 items are 50, 5 and 40, 6 respectively. Find the standard deviation of all taken together.

Solution:

Combined mean,

$$\bar{x} = \frac{N_1\bar{x}_1 + N_2\bar{x}_2}{N_1 + N_2} = \frac{100 \times 50 + 150 \times 40}{100 + 150}$$

$$= 44$$

Combined standard deviation,

$$\sigma = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_1(\bar{x} - \bar{x}_1)^2 + N_2(\bar{x} - \bar{x}_2)^2}{N_1 + N_2}}$$

$$= \sqrt{\frac{100 \times (5)^2 + 150\,(6)^2 + 100\,(44 - 50)^2 + 150\,(44 - 40)^2}{100 + 150}}$$

$$= 7.46.$$

Example 30: A distribution consists of three components with 200, 250, 300 items having mean 25, 10 and 15 and standard deviation 3, 4 and 5, respectively. Find the standard deviation of the combined distribution.

Solution:

In the usual notations, we are given here

$N_1 = 200, N_2 = 250, N_3 = 300$

$\bar{x}_1 = 25,\ \bar{x}_2 = 10,\ \bar{x}_3 = 15$

The formulae (5.13) and (5.14) can easily be extended for combination of three series as

$$\bar{x} = \frac{N_1\bar{x}_1 + N_2\bar{x}_2 + N_3\bar{x}_3}{N_1 + N_2 + N_3}$$

$$= \frac{200 \times 25 + 250 \times 10 + 300 \times 15}{200 + 250 + 300}$$

$$= \frac{12000}{750} = 16$$

and,

NOTES

$$\sigma = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_3\sigma_3^2 + N_1(\bar{x} - \bar{x}_1)^2 + N_2(\bar{x} - \bar{x}_2)^2 + N_3(\bar{x} - \bar{x}_3)^2}{N_1 + N_2 + N_3}}$$

$$= \sqrt{\frac{200 \times 9 + 250 \times 16 + 300 \times 25 + 200 \times 81 + 250 \times 36 + 300 \times 1}{200 + 250 + 300}}$$

$$= \sqrt{51.73} = 7.19$$

Range

The crudest measure of dispersion is the range of the distribution. The range of any series is the difference between the highest and the lowest values in the series. If the marks received in an examination taken by 248 students are arranged in ascending order, then the range will be equal to the difference between the highest and the lowest marks.

In a frequency distribution, the range is taken to be the difference between the lower limit of the class at the lower extreme of the distribution and the upper limit of the class at the upper extreme.

Consider the data on weekly earnings of worker on four workshops given in Table 5.1.

From these figure in Table 5.1, it is clear that the greater the range, the greater is the variation of the values in the group.

The range is a measure of absolute dispersion and as such cannot be usefully employed for comparing the variability of two distributions expressed in different units. The amount of dispersion measured, say, in pounds, is not comparable with dispersion measured in inches. Thus, the need of measuring relative dispersion arises.

An absolute measure can be converted into a relative measure if we divide it by some other value regarded as standard for the purpose. We may use the mean of the distribution or any other positional average as the standard.

For Table 5.1, the relative dispersion would be,

$$\text{Workshop } A = \frac{9}{25.5} \qquad \text{Workshop } C = \frac{23}{25.5}$$

$$\text{Workshop } B = \frac{15}{25.5} \qquad \text{Workshop } D = \frac{15}{25.5}$$

An alternate method of converting an absolute variation into a relative one would be to use the total of the extremes as the standard. This will be equal to dividing the difference of the extreme items by the total of the extreme items. Thus,

$$\text{Relative Dispersion} = \frac{\text{Difference of extreme items, i.e, Range}}{\text{Sum of extreme items}}$$

NOTES

The relative dispersion of the series is called the coefficient or ratio of dispersion. In our example of weekly earnings of workers considered earlier, the coefficients would be,

$$\text{Workshop } A = \frac{9}{21+30} = \frac{9}{51} \qquad \text{Workshop } B = \frac{15}{17+32} = \frac{15}{49}$$

$$\text{Workshop } C = \frac{23}{15+38} = \frac{23}{53} \qquad \text{Workshop } D = \frac{15}{19+34} = \frac{15}{53}$$

Merits and Limitations of Range

Merits

Of the various characteristics that a good measure of dispersion should possess, the range has only two, which are as follows:

1. It is easy to understand.
2. Its computation is simple.

Limitations

Besides the aforesaid two qualities, the range does not satisfy the other test of a good measure and hence it is often termed as a crude measure of dispersion.

The following are the limitations that are inherent in the range as a concept of variability:

1. Since it is based upon two extreme cases in the entire distribution, the range may be considerably changed if either of the extreme cases happens to drop out, while the removal of any other case would not affect it at all.
2. It does not tell anything about the distribution of values in the series relative to a measure of central tendency.
3. It cannot be computed when distribution has open-end classes.
4. It does not take into account the entire data. These can be illustrated by the following illustration. Consider the data given in Table 5.3.

***Table 5.3** Distribution with the Same Number of Cases, but Different Variability*

Class	*No. of students*		
	Section A	*Section B*	*Section C*
0–10	...	...	...
10–20	1	...	...
20–30	12	12	19
30–40	17	20	18
40–50	29	35	16
50–60	18	25	18
60–70	16	10	18
70–80	6	8	21
80–90	11	...	...
90–100	...	...	...
Total	110	110	110
Range	80	60	60

The table is designed to illustrate three distributions with the same number of cases but different variability. The removal of two extreme students from Section *A* would make its range equal to that of *B* or *C*.

NOTES

The greater range of *A* is not a description of the entire group of 110 students, but of the two most extreme students only. Further, though sections *B* and *C* have the same range, the students in Section B cluster more closely around the central tendency of the group than they do in Section *C*. Thus, the range fails to reveal the greater homogeneity of *B* or the greater dispersion of *C*. Due to this defect, it is seldom used as a measure of dispersion.

Specific Uses of Range

In spite of the numerous limitations of the range as a measure of dispersion, it is the most appropriate under the following circumstances:

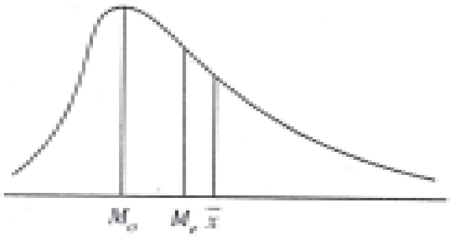
(a) In situations where the extremes involve some hazard for which preparation should be made, it may be more important to know the most extreme cases to be encountered than to know anything else about the distribution. For example, an explorer, would like to know the lowest and the highest temperatures on record in the region he is about to enter; or an engineer would like to know the maximum rainfall during 24 hours for the construction of a storm water drain.

(b) In the study of prices of securities, range has a special field of activity. Thus to highlight fluctuations in the prices of shares or bullion, it is a common practice to indicate the range over which the prices have moved during a certain period of time. This information, besides being of use to the operators, gives an indication of the stability of the bullion market, or that of the investment climate.

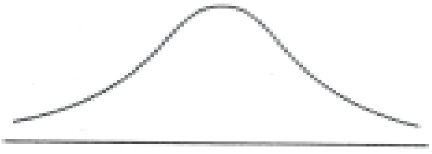
(c) In statistical quality control, range is used as a measure of variation. We, for example, determine the range over which variations in quality are due to random causes, which is made the basis for the fixation of control limits.

Skewness

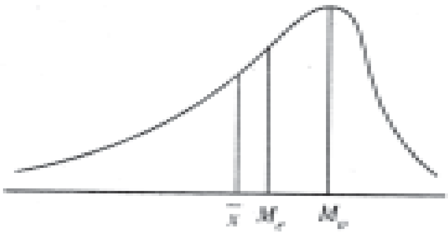
Skewness refers to lack of symmetry in a distribution. In a symmetrical distribution, the mean, median and mode coincide.



Positive Skewness



Symmetry



Negative Skewness

In a positively skewed distribution, the longer tail is on the right side and the mean is on the right of the median.

In a negatively skewed distribution, the longer tail is on the left and the mean is on the left of the median.

In a skewed distribution, the distance between the mean and the median is nearly one-third of that between the mean and the mode.

How to Check the Presence of Skewness in a Distribution

In the following cases skewness is present in the data:

(a) The graph is not symmetrical.

(b) The mean, median and mode do not coincide.

(c) The quartiles are not equidistant from the mean.

(d) The sum of positive and negative deviations from the median is not zero.

(e) Frequencies are not similarly distributed on either side of the mode.

Measure of Skewness

A measure of skewness gives a numerical expression and the direction of asymmetry in a distribution. It gives information about the shape of the distribution and the degree of variation on either side of the central value.

We consider some relative measures of skewness that are as follows:

(a) Pearson's Coefficient of Skewness

$$PSk = \frac{\bar{x} - Mo}{s} = \frac{3(\bar{x} - Me)}{s}$$

It may have any value, but usually it lies between –1 and +1.

Illustration 1**:** If for a given data it is found that

$\bar{x} = 10,$ Mode = 8, s = 4, we have

$$PSk = \frac{\bar{x} - Mo}{s} = \frac{10-8}{4} = 0.5$$

(b) Bowley's Coefficient of Skewness

$$BSk = \frac{Q_3 - Q_1 - 2\,Me}{Q_3 - Q_1}$$

Its value lies between –1 and +1.

Illustration 2: If for a given data $Q_1 = 2,\ Q_3 = 8,\ Me = 5$

$$BSk = \frac{Q_3 + Q_1 - 2\,Me}{Q_3 - Q_1} = \frac{8+2-5}{8-2} = 0.83$$

(c) Kelley's Coefficient of Skewness

$$KSk = P_{50} - \frac{1}{2}(P_{10} + P_{90})$$

where P_{10}, P_{50} and P_{90} are the 10th, 50th, and 90th percentiles of the data.

(d) Method of Moments

If μ_2, μ_3 are moments about the mean we have the coefficient of skewness

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \mu_3^2 / \sigma^6$$

Sometimes, we define the coefficient of skewness as,

$$\gamma_1 = \sqrt{\beta_1} = \sqrt{\frac{\mu_3^2}{\mu_2^3}} = \frac{\mu_3}{\sigma^3}$$

Kurtosis

Kurtosis is a measure of peakedness of a distribution. It shows the degree of convexity of a frequency curve.

If the normal curve is taken as the standard, symmetrical, bell-shaped curve, kurtosis gives a measure of departure from the normal convexity of a distribution. The normal curve is mesokurtic. It is of intermediate peakedness. The flat-topped curve, broader than the normal, is termed platykurtic. The slender, highly peaked curve is termed leptokurtic.

Measures of Kurtosis

(a) Moment Coefficient of Kurtosis : $\beta_2 = \frac{\mu_4}{\mu_2^2}$

Instead of β_2 we often use $\gamma_2 = \beta_2 - 3$ which is positive for a leptokurtic distribution, negative for a platykurtic distribution and zero for the normal distribution.

(b) Percentile Coefficient of Kurtosis $k = \frac{Q}{P_{90} - P_{10}}$

where $Q = \frac{1}{2}(Q_3 - Q_1)$ is the semi-interquartile range.

Comparison of Various Measures of Dispersion

The range is the easiest to calculate the measure of dispersion, but since it depends on extreme values, it is extremely sensitive to the size of the sample, and to the sample variability. In fact, as the sample size increases the range increases dramatically, because the more the items one considers, the more likely it is that one item will turn up which is larger than the previous maximum or smaller than the previous minimum. So, it is, in general, impossible to interpret properly the significance of a given range unless the sample size is constant. It is for this reason that there appears to be only one valid application of the range, namely in statistical quality control where the same sample size is repeatedly used, so that comparison of ranges is not distorted by differences in sample size.

The quartile deviations and other such positional measures of dispersions are also easy to calculate, but suffer from the disadvantage that they are not amenable to algebraic treatment. Similarly, the mean deviation is not suitable because we cannot obtain the mean deviation of a combined series from the deviations of component series. However, it is easy to interpret and easier to calculate than the standard deviation.

NOTES

The standard deviation of a set of data, on the other hand, is one of the most important statistics describing it. It lends itself to rigorous algebraic treatment, is rigidly defined and is based on all observations. It is, therefore, quite insensitive to sample size (provided the size is 'large enough') and is least affected by sampling variations.

It is used extensively in testing of hypothesis about population parameters based on sampling statistics.

In fact, the standard deviations has such stable mathematical properties that it is used as a standard scale for measuring deviations from the mean. If we are told that the performance of an individual is 10 points better than the mean, it really does not tell us enough, for 10 points may or may not be a large enough difference to be of significance. However, if we know that the *s* for the score is only 4 points, so that on this scale, the performance is 2.5*s* better than the mean, the statement becomes meaningful. This indicates an extremely good performance. This sigma scale is a very commonly used scale for measuring and specifying deviations which immediately suggest the significance of the deviation.

The only disadvantages of the standard deviation lies in the amount of work involved in its calculation, and the large weight it attaches to extreme values because of the process of squaring involved in its calculations.

COEFFICIENT OF VARIATION

The square of standard deviation, namely σ^2, is termed as variance and is more often specified than the standard deviation. Clearly, it has the same properties as standard deviation.

As is clear, the standard deviation σ or its square, the variance, cannot be very useful in comparing two series where either the units are different or the mean values are different. Thus, a σ of 5 on an examination where the mean score is 30 has an altogether different meaning than on an examination where the mean score is 90. Clearly, the variability in the second examination is much less. To take care of this problem, we define and use a coefficient of variation, *V*,

$$V = \frac{\sigma}{\bar{x}} \times 100$$

expressed as percentage.

Example 31: The following are the scores of two batsmen A and B in a series of innings:

A	12	115	6	73	7	19	119	36	84	29
B	47	12	76	42	4	51	37	48	13	0

Who is the better run-getter? Who is more consistent?

Solution: In order to decide as to which of the two batsmen, *A* and *B*, is the better run-getter, we should find their batting averages. The one whose average is higher will be considered as a better batsman.

To determine the consistency in batting we should determine the coefficient of variation. The less this coefficient the more consistent will be the player.

A			B		
Score *x*	*x*	*x*^{2}	*Scores* *x*	*x*	*x*^{2}
12	–38	1,444	47	14	196
115	+65	4,225	12	–21	441
6	–44	1,936	76	43	1,849
73	+23	529	42	9	81
7	–43	1,849	–4	– 29	841
19	–31	961	51	18	324
119	+69	4,761	37	4	16
36	–14	196	48	15	225
84	+34	1,156	13	–20	400
29	–21	441	0	–33	1,089
$\sum x$ = 500		17,498	$\sum x$ = 330		5,462

Batsman *A*:

$$\bar{x} = \frac{500}{10} = 50$$

$$\sigma = \sqrt{\frac{17{,}498}{10}} = 41.83$$

$$V = \frac{41.83 \times 100}{50}$$

= 83.66 per cent

Batsman *B:*

$$\bar{x} = \frac{330}{10} = 33$$

$$\sigma = \sqrt{\frac{5{,}462}{10}} = 23.37$$

$$V = \frac{23.37}{33} \times 100$$

= 70.8 per cent

A is a better batsman since his average is 50 as compared to 33 of *B*, but *B* is more consistent since the variation in his case is 70.8 as compared to 83.66 of *A*.

Example 32. The following table gives the age distribution of students admitted to a college in the years 1914 and 1918. Find which of the two groups is more variable in age.

Age	*Number of students in*	
	1914	*1918*
15	–	1
16	1	6
17	3	34
18	8	22
19	12	35
20	14	20
21	13	7
22	5	19
23	2	3
24	3	–
25	1	–
26	–	–
27	1	–

NOTES

Solution:

Age	Assumed Mean–2l 1914				Assumed Mean–19 1918			
	f	*x′*	*fx′*	*fx′*^{2}	*f*	*x′*	*fx*	*fx′*^{2}
15	0	–6	0	0	1	–4	–4	16
16	1	–5	–5	25	6	–3	–18	54
17	3	–4	–12	48	34	–2	–68	136
18	8	–3	–24	72	22	–1	–22	22
19	12	–2	–24	48			–112	
20	14	–1	–14	14				
			–79		35	0	0	0
21	13	0	0	0	20	1	20	20
22	5	1	5	5	7	2	14	28
23	2	2	4	8	19	3	57	171
24	3	3	9	27	3	4	12	48
25	1	4	4	16	147		+103	495
26	0	5	0	0			–9	
27	1	6	6	36				
	63		+28	299				
			–51					

1914 Group:

$$\sigma = \sqrt{\frac{\Sigma fx'^2}{N} - \left(\frac{\Sigma (fx')}{N}\right)^2}$$

$$= \sqrt{\frac{299}{63} - \left(\frac{-51}{63}\right)^2}$$

$$= \sqrt{4.476 - 0.655} = \sqrt{4.091}$$

$$= 2.02.$$

$$\bar{x} = 21 + \left(\frac{-51}{63}\right) = 21 - 8 = 20.2$$

$$V = \frac{2.02}{20.2} \times 100$$

$$= \frac{202}{20.2} = 10$$

1918 Group:

$$\sigma = \sqrt{\frac{495}{147} - \left(\frac{-9}{147}\right)^2} = \sqrt{3.3673 - 0.0037}$$

$$= \sqrt{3.3636} = 1.834$$

$$\bar{x} = 19 + \left(\frac{-9}{147}\right)$$

$$= 19 - .06 = 18.94$$

Check Your Progress

1. Name three measures of central tendency.
2. What is a mode?
3. What do you mean by median?
4. What is a geometric mean?
5. What do you understand by harmonic mean?
6. What is range?
7. What is a quartile?
8. What is mean deviation?
9. Define the term skewness?
10. What is kurtosis?

$$V = \frac{1.834}{18.94} \times 100$$
$$= 9.68$$

The coefficient of variation of the 1914 group is 10 and that of the 1918 group 9.68. This means that the 1914 group is more variable, but only barely so.

NOTES

ACTIVITY

1. 1. Determine the median and the value corresponding to the first and third quartiles in the following data set:

   46, 47, 49, 49, 51, 53, 54, 54, 55, 55, 59
2. Calculate mean deviation and its coefficient about median, arithmetic mean and mode for the following figures, and show that the mean deviation about the median is least.

   103, 50, 68, 110, 108, 105, 174, 103, 150, 200, 225, 350, 103

DID YOU KNOW

The choice of a particular measure of central tendency of location depends on the purpose of investigation. It should be noted that the Arithmetic Mean (AM) is quite precisely defined and is therefore more amenable to further mathematical manipulation. On the other hand, the mode can be located by inspection but cannot be manipulated easily. There are cases when we have the following condition:
1. Mean < Median < Mode
2. Mean > Median > Mode
3. Mean = Median = Mode

SUMMARY

- There are several commonly used measures of central tendency such as arithmetic mean, mode and median.
- Arithmetic mean is also commonly known as simply the mean.
- The advantage of combined arithmetic mean is that, one can determine the over, all mean of the combined data without having to going back to the original data.
- The sum of the deviations of individual values of X from the mean will always add up to zero. This means that if we subtract all the individual values from their mean, then some values will be negative and some will be positive, but if all these differences are added together then the total sum will be zero
- The second important characteristic of the mean is that it is very sensitive to extreme values.
- The third property of the mean is that the sum of squares of the deviations about the mean is minimum.
- The mode is another form of average and can be defined as the most frequently occurring value in the data.

- The mode is not affected by extreme values in the data and can easily be obtained from an ordered set of data.
- The median is a measure of central tendency and it appears in the centre of an ordered data.
- Median is a positional average and hence the extreme values in the data set do not affect it as much as they do to the mean.
- Median can be located visually when the data is in the form of ordered data.
- Median is comparatively less stable than the mean, particularly for small samples, due to fluctuations in sampling.
- A mode is not suitable for algebraic manipulations.
- The weighted arithmetic mean is particularly useful where we have to compute the mean of means.
- The position of value in statistics is determined using specific methods for a given set of data or observations.
- If α, β, γ are in GP, then β is called a geometric mean between α and γ, written as GM.
- If a, b, c are in HP, then b is called a Harmonic Mean between a and c, written as HM.
- A measure of dispersion or simply dispersion may be defined as statistics signifying the extent of the scatteredness of items around a measure of central tendency.
- A measure of dispersion may be expressed in an absolute form, or in a relative form.
- A measure of dispersion is said to be in an absolute form when it states the actual amount by which the value of an item on an average deviates from a measure of central tendency.
- A relative measure of dispersion computed is a quotient obtained by dividing the absolute measures by a quantity in respect to which absolute deviation has been computed.
- Quartiles are the points which divide the array into four equal parts.
- Standard deviation, σ (sigma) is defined as the square root of the mean of the squares of the deviations of individual items from their arithmetic mean.
- The crudest measure of dispersion is the range of the distribution.
- The range of any series is the difference between the highest and the lowest values in the series.
- Skewness refers to lack of symmetry in a distribution. In a symmetrical distribution, the mean, median and mode coincide.
- A measure of skewness gives a numerical expression and the direction of asymmetry in a distribution. It gives information about the shape of the distribution and the degree of variation on either side of the central value.
- Kurtosis is a measure of peakedness of a distribution. It shows the degree of convexity of a frequency curve.
- The square of standard deviation, namely σ^2, is termed as variance and is more often specified than the standard deviation.

- The term moment is obtained from mechanics where the moment of a force describes the tendency or capacity of a force to turn a pivoted lever.

KEY TERMS

- **Arithmetic mean:** The most commonly used measure of central location, otherwise referred to with the term 'average'
- **Mode:** The most frequently occurring value in a data
- **Median:** A measure that divides values in a data into two equal parts
- **Weighted arithmetic mean:** A measurement of a mean of means
- **Geometric mean:** The *n*th root of the product of *n* values
- **Harmonic mean:** An average of different rates
- **Dispersion:** The extent of scatteredness of items around a measure of central tendency.
- **Skewness:** Lack of symmetry in a distribution
- **Kurtosis:** A measure of peakedness of a distribution

ANSWERS TO 'CHECK YOUR PROGRESS'

1. The three measures of dispersion are arithmetic mean, median and mode.
2. The mode is another form of average and can be defined as the most frequently occurring value in the data.
3. The median is a measure of central tendency and it appears in the centre of an ordered data.
4. If α, β, γ are in GP, then β is called a geometric mean between α and γ, written as GM.
5. If a, b, c are in HP, then b is called a Harmonic Mean between a and c, written as HM.
6. Range is the difference between the highest and lowest values in a series.
7. Quartiles are points that divide an array into four equal parts.
8. Mean deviation is the arithmetic average of the variations of the individual items in a series from a measure of their central tendency.
9. Skewness refers to lack of symmetry in a distribution.
10. Kurtosis is a measure of peakedness of a distribution.

QUESTIONS AND EXERCISES

Short-Answer Questions

1. How is central tendency measured?
2. Define the term arithmetic mean.
3. Write three characteristics of mean.
4. What is the importance of arithmetic mean in statistics?

5. Define the term median with example.
6. How is location of median calculated using graphic analysis?
7. Define the terms quartiles, deciles and percentiles with suitable examples.
8. What is mode? How is it calculated?
9. Differentiate between a mean and a mode.
10. What is geometric mean? How is it calculated?
11. When a measure of dispersion is expressed in an absolute form and in relative form?
12. What is coefficient of mean deviation?
13. How is the calculation of standard deviation different from that of mean deviation?
14. What are the merits and demerits of range?
15. How will you measure skewness?
16. How will you measure the degree of convexity of a frequency curve using kurtosis?
17. The price of a commodity was four times higher in 1970 than what it was a decade back. Find the average rate of growth of price of the commodity.
18. Arithmetic mean of a group of 100 items is 50 and of another group of 150 items is 100. What will be the mean of all the items?
19. Arithmetic mean of 98 items is 50. Two items 60 and 70 were left out at the time of calculation. What is the correct mean of all the items?

Long-Answer Questions

1. Discuss the various measures of central tendency. What purposes do their measurement serve?
2. Explain geometric and harmonic mean and their uses.
3. Eight coins were tossed together and the number of heads resulting was observed. The operation was performed 256 times and the frequencies that were obtained for the different values of *x*, the number of heads, are shown in the following table. Calculate mean, median and quartiles of the distribution of *x*.

x	0	1	2	3	4	5	6	7	8
Frequency	1	9	26	59	72	52	29	7	1

4. Find the mean of the following distribution:

Breadth in mm	19–21	22–24	25–27	28–30	31–33	34–36	37–39
	6	13	19	23	18	12	9

5. The following table shows the number of persons employed in certain units of an industry. Find the average number of persons employed.

No. of Persons: below 20 20–30 30–50 50–100 100–200 200
and above 2 5 6 3 2 2

6. Calculate arithmetic mean for the following data:

Class Interval 5–10 10–15 15–20 20–25 25–30 30–35 35–40
40–45

Frequency	6	5	15	10	5	4	3	2

7. From the following table, calculate mean and median.

Crop Cutting Experimental Data on Plot Yields of Wheat

Yield (in lbs)	*No. of Plots*	*Yield (in lbs.)*	*No. of Plots*
Over 0	216	Over 300	31
Over 60	210	Over 360	13
Over 120	156	Over 420	7
Over 180	98	Over 480	2
Over 240	57	Up to 540	216

8. Find arithmetic mean, median and mode from the following:

Marks below	10	20	30	40	50	60	70	80
No. of Students	15	35	60	84	96	127	198	250

9. The wages of 1060 employees range from ` 300 to ` 450. They are grouped in 15 classes with a common class interval of ` 10. Class frequencies from lowest to the highest are 6, 17, 35, 48, 65, 90, 131, 173, 155, 177, 75, 52, 9, 6. Tabulate the data and calculate the mean wage.

10. From the table given below, find the mean.

Salary Per Day	*No. of Persons*	*Salary Per Day*	*No. of Persons*
1–5	7	26–30	18
6–10	10	31–35	10
11–15	16	36–40	5
16–20	32	41–45	1
21–25	24		

11. (a) From the data given below, find the mode.

Ages	20–25	25–30	30–35	35–40	40–45	45–50	50–55	55–60
No. of Persons	50	70	80	180	150	120	70	50

(b) If the mode and mean of a moderately asymmetrical series are respectively 16 inches and 20.2 inches, compute the most probable median.

12. Following is the distribution of the size of certain farms selected at random from a district. Calculate the mode of distribution.

Central Size of the Farm in Acres	10	20	30	40	50	60	70
No. of Farms	7	12	17	29	31	5	3

13. Draw a histogram from the following data and measure the modal value:

Class Size	*Frequency*	*Class Size*	*Frequency*
0–10	5	50–60	10
10–20	11	60–70	8
20–30	19	70–80	6
30–40	21	80–90	3
40–50	16	90–100	1

14. Monthly incomes of the families are given below in rupees:

    2000, 35, 400, 15, 40, 1500, 300, 6, 90, 250, 20, 12, 450, 10, 150, 8, 25, 30, 1200, 60.

    Calculate the geometric mean and harmonic mean of the above series.

15. The following table gives weights of 31 persons in a sample inquiry. Calculate mean weight using (a) Geometric mean and (b) Harmonic mean.

Weight in lbs.	130	135	140	145	146	148	149	150	157
No. of Persons	3	4	6	6	3	5	2	1	1

16. Peter travelled by car for 4 days. He drove 10 hours each day. He drove: first day at the rate of 45 km per hour and fourth day at the rate of 37 km per hour. What was his average speed?

17. The price of certain articles becomes $1\frac{1}{2}$ times in first year, $1\frac{5}{8}$ times in the second year and $\frac{7}{9}$ times in the third year. What is the average change per year?

18. You take a trip which entails travelling 900 miles by train at an average speed of 60 mph, 3000 miles by boat at an average of 25 mph, 400 miles by plane at 350 mph, and finally 15 miles by taxi at 25 mph. What is your average speed for the entire distance?

19. Calculate the simple average and weighted average of the following items:

Items	68	85	101	102	108	110	112	113	124	128	143	146	151	153	172
Weights	1	46	31	1	11	7	23	17	9	14	2	4	6	5	2

    Account for the difference in the two averages.

20. Find the QD from the mean of the series 5, 7, 10, 12, 6.

21. Calculate the mean deviation from the mean and median and their coefficients of the following data:

Size of Shoes:	3	6	11	2	4	10	5	7	8	9
Pairs sold:	10	15	25	6	4	3	2	8	9	4

22. Explain standard deviation in detail. Also, discuss the short-cut method of calculating standard deviation.

23. The following are the frequencies, means and standard deviations of two series. Find the SD of the combined series.

	Series I	Series II
Frequency	35	45
Mean	30	20
Standard Deviation	10	5

24. Discuss the instances when range can specifically be used.

25. Explain the different types of skewness.

26. Find SD and CV of the sample observations: 2, 5, 7, 6.

27. Find out the range and its coefficient in the following items:

    110, 117, 129, 300, 357, 100, 500, 630, 750

28. Find out the range and its coefficient in the following series :

Size	3	4	5	6	7	8	9	10
Frequency	35	30	20	10	6	3	2	1

29. Find out the range and its coefficient in the following series :

Size	10–60	60–120	120–180	180–240	240–300
Frequency	3	5	6	3	2

30. Calculate the quartile deviation and its coefficient from the following data :

Size	6	7	8	9	10	11	12
Frequency	3	6	9	13	8	5	4

31. The following table gives monthly wages (in hundreds of Rs.) of 72 workers in a factory. Compute the quartile deviation.

Wages (Rs.)	*No. of workers*	*Wages (Rs.)*	*No. of workers*
12.5 – 17.5	2	37.5 – 42.5	4
17.5 – 22.5	22	42.5 – 47.5	6
22.5 – 27.5	19	47.5 – 52.5	1
27.5 – 32.5	14	52.5 – 57.5	1
32.5 – 37.5	3		

32. Find the mean deviation of the set of numbers 3, 10, 9, 9, 4, 7, 14 (a) from mean (b) from median.

33. Calculate the mean deviation from the median of the following discrete series and find its coefficient:

Size	2	3	4	5	6	7
Frequency	5	4	10	8	3	2

34. Find the mean deviation from the mean of the following series :

x	56	63	70	77	84	91	98
f	3	6	14	16	13	6	2

35. Calculate the mean deviation from arithmetic mean of the following grouped data :

Class	0–10	10–20	20–30	30–40	40–50
Frequency	2	8	10	3	4

36. Calculate the mean deviation from the median of the following data :

Class	140–150	150–160	160 –170	170–180	180–190	190–200
Frequency	4	6	10	18	9	3

37. Find the standard deviation of the set of numbers : 3, 10, 9, 9, 4, 7, 14

38. Calculate the standard deviation the following distribution :

x	25	35	45	55	65	75	85
f	3	61	132	153	140	51	2

39. Calculate the standard deviation and its coefficient from the following table by short-cut method:

Class	5–15	15–25	25–35	35–45	45–55	55–65
Frequency	15	32	51	78	97	109

40. Calculate the standard deviation and its coefficient from the following table by step deviation method:

Class	0–10	10–20	20–30	30–40	40–50	50–60
Frequency	15	17	19	27	19	12

41. The following are some of the particulars of the distribution of weights of boys and girls in a class:

	Boys	*Girls*
Number	100	50
Mean weight	60 kg	45 kg
Variance	9	4

42. The no. of workers employed, the mean wages (in Rs.) per month and standard deviation (in Rs.) in each section of a factory are given below. Calculate the mean wages and standard deviation of all the workers taken together.

Section	*No. of workers employed*	*Mean wages (in Rs.)*	*Standard Deviation (in Rs.)*
A	50	1113	60
B	60	1120	70
C	90	1115	80

43. The mean and standard deviation of 200 items are found to be 60 and 20 respectively. If at the time of calculations, two items were wrongly taken as 3 and 67 instead of 13 and 17, find the correct mean and standard deviation.What is the correct coefficient of variation?

5.11 FURTHER READING

Best, John W. and James V. Kahn. 2005. *Research in Education*, 10th edition. New Jersey: Pearson Education.

Butcher, Harold John. 1966. *Sampling in Educational Research*, 3rd edition. United Kingdom: Manchester University Press.

Edwards, Allen Louis. 2006. *Experimental Design in Psychological Research*, 3rd edition. United States: The University of Michigan.

Garrett, Henry Edward. 1926. *Statistics in Psychology and Education*, New Jersey: Longmans, Green and Company.

Guilford, Joy Paul. 1977. *Fundamental Statistics in Psychology and Education*, 6th edition. New York: McGraw Hill.

Kerlinger, Fred Nichols and Howard Bing Lee. 2000. *Foundations of Behavioral Research*, 4th edition. United States: Harcourt College Publishers.

UNIT 2 CORRELATION

NOTES

Structure

Introduction
Unit Objectives
Correlation Analysis
Methods of Studying Simple Correlation
Properties of Correlation Coefficient
Method of Least Squares
Simple Regression Coefficients
Karl Pearson's Coefficient
Other Measures
Rank Correlation
Product Moment Correlation
Regression Analysis
Summary
Key Terms
Answers to 'Check Your Progress'
Questions and Exercises
Further Reading

INTRODUCTION

In this unit, you will learn about correlation analysis. This technique looks at indirect relationships and establishes the variables that are most closely associated with a given data or mindset. It is the process of finding how accurately the line fits using the observations. Correlation analysis can be referred to as the statistical tool used to describe the degree to which one variable is related to another. The relationship, if any, is assumed to be a linear one. In fact, the word 'correlation' refers to the relationship or the interdependence between two variables. There are various phenomena that are related to each other. The theory by means of which quantitative connections between two sets of phenomena are determined is called the 'Theory of Correlation'. On the basis of this theory, you can study the comparative changes occurring in two related phenomena and their cause–effect relation can also be examined. Thus, correlation is concerned with the relationship between two related and quantifiable variables and can be positive or negative.

In this unit, you will also learn about regression analysis, which is the mathematical process of using observations to find the line of best fit through the data in order to make estimates and predictions about the behaviour of variables. This technique is used to determine the statistical relationship between two or more variables and to make prediction of one variable on the basis of one or more other variables.

UNIT OBJECTIVES

After going through this unit, you will be able to:

- Explain the types of correlations
- Explain the various methods of studying simple correlation

- Describe the properties of coefficient
- Discuss the significance of regression analysis

CORRELATION ANALYSIS

Correlation analysis is the statistical tool generally used to describe the degree to which one variable is related to another. The relationship, if any, is usually assumed to be a linear one. This analysis is used quite frequently in conjunction with regression analysis to measure how well the regression line explains the variations of the dependent variable. In fact, the word correlation refers to the relationship or interdependence between two variables. There are various phenomena which are related to each other. For instance, when demand of a certain commodity increases, its price goes up and when its demand decreases, its price comes down. Similarly, with age the height of children, with height the weight of children, with money the supply and the general level of prices go up. Such sort of relationships can as well be noticed for several other phenomena. The theory by means of which quantitative connections between two sets of phenomena are determined is called the 'Theory of Correlation'.

On the basis of the theory of correlation, one can study the comparative changes occurring in two related phenomena and their cause–effect relation can be examined. It should, however, be borne in mind that relationships like 'black cat causes bad luck', 'filled up pitchers result in good fortune' and similar other beliefs of the people cannot be explained by the theory of correlation, since they are all imaginary and are incapable of being justified mathematically. Thus, correlation is concerned with relationship between two related and quantifiable variables. If two quantities vary in sympathy, so that a movement (an increase or decrease) in one tends to be accompanied by a movement in the same or opposite direction in the other and the greater the change in one, the greater is the change in the other, the quantities are said to be correlated. This type of relationship is known as correlation or what is sometimes called, in statistics, as covariation.

For correlation, it is essential that the two phenomena should have cause–effect relationship. If such relationship does not exist then one should not talk of correlation. For example, if the height of the students as well as the height of the trees increases, then one should not call it a case of correlation because the two phenomena, viz., the height of students and the height of trees are not even casually related. However, the relationship between the price of a commodity and its demand, the price of a commodity and its supply, the rate of interest and savings, etc. are examples of correlation, since in all such cases the change in one phenomenon is explained by a change in another phenomenon.

It is appropriate here to mention that correlation in case of phenomena pertaining to natural sciences can be reduced to absolute mathematical term, e.g., heat always increases with light. However, in phenomena pertaining to social sciences, it is often difficult to establish any absolute relationship between two phenomena. Hence, in social sciences, we must take the fact of correlation being established if in a large number of cases, two variables always tend to move in the same or opposite direction.

Correlation can either be positive or it can be negative. Whether correlation is positive or negative would depend upon the direction in which the variables are moving. If both variables are changing in the same direction, then correlation is said to be positive, but when the variations in the two variables take place in opposite direction, the correlation is termed as negative. This can be explained as follows:

Changes in Independent Variable	*Changes in Dependent Variable*	*Nature of Correlation*
Increase (+)↑	Increase (+)↑	Positive (+)
Decrease (–)↓	Decrease (–)↓	Positive (+)
Increase (+)↑	Decrease (–)↓	Negative (–)
Decrease (–)↓	Increase (+)↑	Negative (–)

Statisticians have developed two measures for describing the correlation between two variables, viz., the coefficient of determination and the coefficient of correlation. These two methods are explained in detail in the following sections.

METHODS OF STUDYING SIMPLE CORRELATION

Coefficient of Determination

The **coefficient of determination** (symbolically indicated as r^2, though some people would prefer to put it as R^2) is a measure of the degree of linear association or correlation between two variables, say X and Y, one of which happens to be an independent variable and the other being a dependent variable. This coefficient is based on the following two types of variations:

(a) The variation of the Y values around the fitted regression line, viz., $\sum\left(Y-\hat{Y}\right)^2$, technically known as the unexplained variation.

(b) The variation of the Y values around their own mean, viz., $\sum\left(Y-\overline{Y}\right)^2$, technically known as the total variation.

If we subtract the unexplained variation from the total variation, we obtain what is known as the explained variation, i.e., the variation explained by the line of regression. Thus, Explained Variation = (Total variation) – (Unexplained variation)

$$=\sum\left(Y-\overline{Y}\right)^2-\sum\left(Y-\hat{Y}\right)^2$$

$$=\sum\left(\hat{Y}-\overline{Y}\right)^2$$

The Total and Explained as well as Unexplained variations can be shown as given in Figure. 6.1.

NOTES

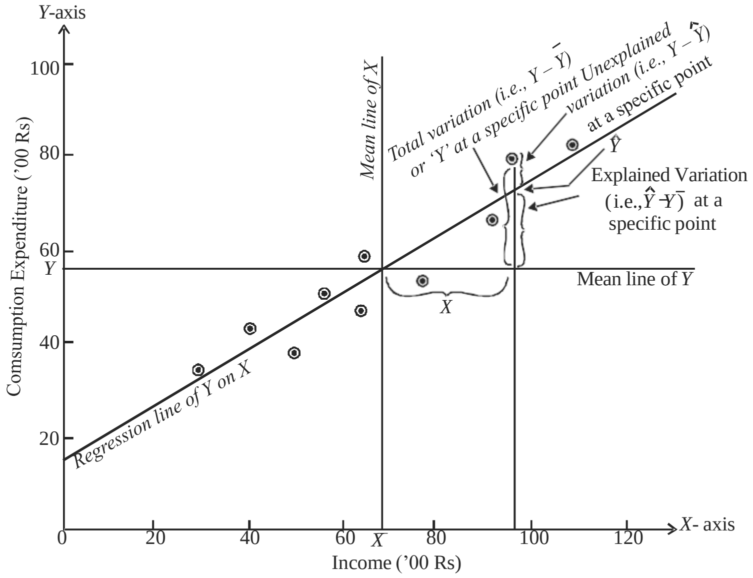


Fig. 6.1 *Diagram Showing Total, Explained and Unexplained Variations*

Coefficients of determination is that fraction of the total variation of Y which is explained by the regression line. In other words, coefficient of determination is the ratio of explained variation to total variation in the Y variable related to the X variable. Coefficient of determination can be algebraically stated as,

$$r^2 = \frac{\text{Explained variation}}{\text{Total variation}}$$

$$= \frac{\Sigma\left(\hat{Y} - \overline{Y}\right)^2}{\Sigma\left(Y - \overline{Y}\right)^2}$$

Alternatively, r^2 can also be stated as,

$$r^2 = 1 - \frac{\text{Explained variation}}{\text{Total variation}}$$

$$= 1 - \frac{\Sigma\left(\hat{Y} - \overline{Y}\right)^2}{\Sigma\left(Y - \overline{Y}\right)^2}$$

Interpreting r^2

The coefficient of determination can have a value ranging from 0 – 1. The value of 1 can occur only if the unexplained variation is 0, which simply means that all the data points in the Scatter diagram fall exactly on the regression line. For a 0 value to occur, $\Sigma(Y - \overline{Y})^2 = \Sigma(Y - \hat{Y})^2$, which simply means that X tells us nothing about Y and hence there is no regression relationship between X and Y variables. Values between 0 and 1 indicate the 'Goodness of fit' of the regression line to the sample data. The higher the value of r^2, the better the fit. In other words, the value of r^2 will lie somewhere between 0 and 1. If r^2 has a 0 value then it indicates no correlation, but if it has a value equal to 1 then it indicates that there is perfect correlation and as such the regression line is a perfect estimator. However, in most cases, the value of r^2 will lie somewhere between these two extremes of 1 and 0. One should remember that r^2 close to 1 indicates a strong correlation between X and Y, while an r^2 near 0 means there is little correlation between these two variables. r^2 value can as well be interpreted by looking at the amount of the variation in Y, the dependant variable, that is explained by the regression line. Supposing, we get a value of $r^2 = 0.925$ then this would mean that the variations in independent variable (say X) would explain 92.5 per cent of the variation in the dependent variable (say Y). If r^2 is close to 1 then it indicates that the regression equation explains most of the variations in the dependent variable (see Example 6.1).

NOTES

Example 6.1: Calculate the coefficient of determination (r^2) using the provided data. Calculate and analyse the result.

Observations	1	2	3	4	5	6	7	8	9	10
Income (X) (‘00 `)	41	65	50	57	96	94	110	30	79	65
Consumption Expenditure (Y) (‘00 `)	44	60	39	51	80	68	84	34	55	48

Solution:

r^2 can be worked out as follows:

Since,
$$r^2 = 1 - \frac{\text{Unexplained variation}}{\text{Total variation}} = 1 - \frac{\sum\left(Y - \hat{Y}\right)^2}{\sum\left(Y - \overline{Y}\right)^2}$$

As, $\sum\left(Y - \overline{Y}\right)^2 = \sum Y^2 = \left(\sum Y^2 - n\overline{Y}^2\right)$, we can write,

$$r^2 = 1 - \frac{\sum\left(Y - \hat{Y}\right)^2}{\sum Y^2 - n\overline{Y}^2}$$

Calculating and putting the various values, we have the following equation:

$$r^2 = 1 - \frac{260.54}{34223 - 10\left(56.3\right)^2} = 1 - \frac{260.54}{2526.10} = 0.897$$

Analysis of Result: The regression equation used to calculate the value of the coefficient of determination (r^2) from the sample data shows that, about 90 per cent of the variations in consumption expenditure can be explained. In other words, it means that the variations in income explain about 90 per cent of variations in consumption expenditure.

Observation	1	2	3	4	5	6	7	8	9	10
Income (X) (’00 `)	41	65	50	57	96	94	110	30	79	65
Consumption Expenditure (Y) (’00 `)	44	60	39	51	80	68	84	34	55	48

PROPERTIES OF CORRELATION COEFFICIENT

The coefficient of correlation, symbolically denoted by ‘r’, is another important measure to describe how well one variable is explained by another. It measures the degree of relationship between the two casually related variables. The value of this coefficient can never be more than +1 or less than –1. Thus, +1 and –1 are the limits of this coefficient. For a unit change in independent variable, if there happens to be a constant change in the dependent variable in the same direction, then the value of the coefficient will be +1 indicative of the perfect positive correlation; but if such a change occurs in the opposite direction, the value of the coefficient will be –1, indicating the perfect negative correlation. In practical life, the possibility of obtaining either a perfect positive or perfect negative correlation is very remote particularly in respect of phenomena concerning social sciences. If the coefficient of correlation has a zero value then it means that there exists no correlation between the variables under study.

There are several methods of finding the coefficient of correlation, but the following ones are considered important:

(a) Coefficient of correlation by the method of least squares

(b) Coefficient of correlation using simple regression coefficients

(c) Coefficient of correlation through product moment method or Karl Pearson's coefficient of correlation

Whichever of these three methods we adopt, we get the same value of *r*.

Method of Least Squares

Under this method, first, the estimating equation is obtained using the least square method of simple regression analysis. The equation is worked out as,

$$\hat{Y} = a + bX_i$$

Total variation $= \sum\left(Y - \overline{Y}\right)^2$

Unexplained variation $= \sum\left(Y - \hat{Y}\right)^2$

Explained variation $= \sum\left(\hat{Y} - \overline{Y}\right)^2$

Then, by applying the following formulae, we can find the value of the coefficient of correlation as,

$$r = \sqrt{r^2} = \sqrt{\frac{\text{Explained variation}}{\text{Total variation}}}$$

$$= \sqrt{1 - \frac{\text{Unexplained variation}}{\text{Total variation}}}$$

$$= \sqrt{1 - \frac{\sum\left(Y - \hat{Y}\right)^2}{\sum\left(Y - \overline{Y}\right)^2}}$$

This clearly shows that the coefficient of correlation happens to be the square root of the coefficient of determination.

Short-cut formula for finding the value of '*r*' by the method of least squares can be repeated and readily written as,

$$r = \sqrt{\frac{a\Sigma Y + b\Sigma XY - n\overline{Y}^2}{\Sigma Y^2 - n\overline{Y}^2}}$$

Where, a = Y-intercept

b = Slope of the estimating equation

X = Values of the independent variable

Y = Values of dependent variable

$\overline{Y}$ = Mean of the observed values of Y

n = Number of items in the sample (i.e., pairs of observed data)

The plus (+) or the minus (–) sign of the coefficient of correlation worked out by the method of least squares, is related to the sign of '*b*' in the estimating equation, viz., $\hat{Y} = a + bX_i$. If '*b*' has a minus sign, the sign of '*r*' will also be minus, but if '*b*' has a plus sign, then the sign of '*r*' will also be plus. The value of '*r*' indicates the degree along with the direction of the relationship between the two variables *X* and *Y*.

Simple Regression Coefficients

Under this method, the estimating equation of *Y* and the estimating equation of *X* is worked out using the method of least squares. From these estimating equations we find the regression coefficient of *X* on *Y*, i.e., the slope of the estimating equation of *X* (symbolically written as b_{XY}) and this happens to be equal to $r\frac{\sigma_X}{\sigma_Y}$ and similarly, we find the regression coefficient of *Y* on *X*, i.e., the slope of the estimating equation of *Y* (symbolically written as b_{YX}) and this happens to be equal to $r\frac{\sigma_Y}{\sigma_X}$. For finding '*r*', the square root of the product of these two regression coefficients are worked out as[1]

$$r = \sqrt{b_{XY}.b_{YX}}$$

$$= \sqrt{r\frac{\sigma_X}{\sigma_Y}.r\frac{\sigma_Y}{\sigma_X}}$$

$$= \sqrt{r^2} = r$$

As stated earlier, the sign of '*r*' will depend upon the sign of the regression coefficients. If they have minus sign, then '*r*' will take minus sign but the sign of '*r*' will be plus if regression coefficients have plus sign.

Karl Pearson's Coefficient

Karl Pearson's method is the most widely used method of measuring the relationship between two variables. This coefficient is based on the following assumptions:

(a) There is a linear relationship between the two variables, which means that a straight line would be obtained if the observed data is plotted on a graph.

(b) The two variables are casually related, which means that one of the variables is independent and the other one is dependent.

(c) A large number of independent causes operate on both the variables so as to produce a normal distribution.

According to Karl Pearson, '*r*' can be worked out as,

$$r = \frac{\sum XY}{n\sigma_X\sigma_Y}$$

[1] The short-cut formulae to workout b_{XY} and b_{YX}:

$$b_{XY} = \frac{\sum XY - n\bar{X}\bar{Y}}{\sum Y^2 - n\bar{Y}^2}$$

and
$$b_{YX} = \frac{\sum XY - nXY}{\sum X^2 - n\bar{X}^2}$$

Where,

$$X = (X - \overline{X})$$

$$Y = (Y - \overline{Y})$$

σ_X = Standard deviation of X series and is equal to $\sqrt{\dfrac{\sum X^2}{n}}$

σ_Y = Standard deviation of Y series and is equal to $\sqrt{\dfrac{\sum Y^2}{n}}$

n = Number of pairs of X and Y observed

A short-cut formula, known as the Product Moment Formula, can be derived from the above stated formula as,

$$r = \frac{\sum XY}{n\sigma_X\sigma_Y}$$

$$= \frac{\sum XY}{\sqrt{\dfrac{\sum X^2}{n} \cdot \dfrac{\sum Y^2}{n}}}$$

$$n = \frac{\sum XY}{\sqrt{\sum X^2 \sum Y^2}}$$

The above formulae are based on obtaining true means (viz. $\overline{X}$ and $\overline{Y}$) first and then doing all other calculations. This happens to be a tedious task, particularly if the true means are in fractions. To avoid difficult calculations, we make use of the assumed means in taking out deviations and doing the related calculations. In such a situation, we can use the following formula for finding the value of 'r':[2]

a. In Case of Ungrouped Data:

$$r = \frac{\dfrac{\sum dX.dY}{n} - \left(\dfrac{\sum dX}{n} \cdot \dfrac{\sum dY}{n}\right)}{\sqrt{\dfrac{\sum dX^2}{n} - \left(\dfrac{\sum dX}{n}\right)^2} \cdot \sqrt{\dfrac{\sum dY^2}{n} - \left(\dfrac{\sum dY}{n}\right)^2}}$$

[2] In case we take assumed mean to be zero for X variable as for Y variable, then our formula will be as,

$$r = \frac{\dfrac{\sum XY}{n} - \left(\dfrac{\sum X}{n}\right)\left(\dfrac{\sum Y}{n}\right)}{\sqrt{\dfrac{\sum X^2}{n} - \left(\dfrac{\sum X}{n}\right)^2}\sqrt{\dfrac{\sum Y^2}{n} - \left(\dfrac{\sum Y}{n}\right)^2}}$$

or

$$r = \frac{\dfrac{\sum XY}{n} - \overline{X}\overline{Y}}{\sqrt{\dfrac{\sum X^2}{n} - \overline{X}^2}\sqrt{\dfrac{\sum Y^2}{n} - \overline{Y}^2}}$$

$$r = \frac{\sum XY - n\overline{X}\overline{Y}}{\sqrt{\sum X^2 - n\overline{X}^2}\sqrt{\sum Y^2 - n\overline{Y}^2}}$$

NOTES

$$= \frac{\sum dX.dY - \left(\frac{\sum dX \times \sum dY}{n}\right)}{\sqrt{\sum dX^2 - \frac{(\sum dX)^2}{n}}\sqrt{\sum dY^2 - \frac{(\sum dY)^2}{n}}}$$

Where, $\sum dX = \sum(X - X_A)$ X_A = Assumed average of X

$\sum dY = \sum(Y - Y_A)$ Y_A = Assumed average of Y

$\sum dX^2 = \sum(X - X_A)^2$

$\sum dY^2 = \sum(Y - Y_A)^2$

$\sum dX \,.\, dY = \sum(X - X_A)\,(Y - Y_A)$

n = Number of pairs of observations of X and Y

b. In Case of Grouped Data:

$$r = \frac{\frac{\sum fdX.dY}{n} - \left(\frac{\sum fdX}{n} . \frac{\sum fdY}{n}\right)}{\sqrt{\frac{\sum fdX^2}{n} - \left(\frac{\sum fdX}{n}\right)^2}\sqrt{\frac{\sum fdY^2}{n} - \left(\frac{\sum fdY}{n}\right)^2}}$$

or $$r = \frac{\sum fdX.dY - \left(\frac{\sum fdX.\sum fdY}{n}\right)}{\sqrt{\sum fdX^2 - \left(\frac{\sum fdX}{n}\right)^2}\sqrt{\sum fdY^2 - \left(\frac{\sum fdY}{n}\right)^2}}$$

Where, $\sum fdX.dY = 0\sum f\,(X - X_A)\,(Y - Y_A)$

$\sum fdX = \sum f\,(X - X_A)$

$\sum fdY = \sum f\,(Y - Y_A)$

$\sum fdY^2 = \sum f\,(Y - Y_A)^2$

$\sum fdX^2 = \sum f\,(X - X_A)^2$

n = Number of pairs of observations of X and Y

Probable Error (P.E.) of the Coefficient of Correlation

Probable Error (P.E.) of r is very useful in interpreting the value of r and is worked out for Karl Pearson's coefficient of correlation as,

$$\text{P.E.} = 0.6745\frac{1 - r^2}{\sqrt{n}}$$

If r is less than its P.E., it is not at all significant. If r is more than P.E., there is correlation. If r is more than six times its P.E. and greater than ± 0.5, then it is considered significant. Let us consider Example 6.2.

Example 6.2: From the following data calculate 'r' between X and Y applying the following three methods:

(a) The method of least squares.

(b) The method based on regression coefficients.

(c) The product moment method of Karl Pearson.

NOTES

Verify the obtained result of any one method with that of another.

X	1	2	3	4	5	6	7	8	9
Y	9	8	10	12	11	13	14	16	15

Solution:

Let us develop the following table for calculating the value of '*r*':

X	Y	X^2	Y^2	XY
1	9	1	81	9
2	8	4	64	16
3	10	9	100	30
4	12	16	144	48
5	11	25	121	55
6	13	36	169	78
7	14	49	196	98
8	16	64	256	128
9	15	81	225	135
$n = 9$ $\sum X = 45$	$\sum Y = 108$	$\sum X^2 = 285$	$\sum Y^2 = 1356$	$\sum XY = 597$

$\therefore \quad \overline{X} = 5; \qquad \overline{Y} = 12$

(a) Coefficient of correlation by the method of least squares is worked out as follows: First find out the estimating equation,

$$\hat{Y} = a + bX_i$$

Where,
$$b = \frac{\sum XY - n\overline{X}\overline{Y}}{\sum X^2 - n\overline{X}^2}$$

$$= \frac{597 - 9(5)(12)}{285 - 9(25)} = \frac{597 - 540}{285 - 225} = \frac{57}{60} = 0.95$$

and
$$a = \overline{Y} - b\overline{X}$$

$$= 12 - 0.95(5) = 12 - 4.75 = 7.25$$

Hence,
$$\hat{Y} = 7.25 + 0.95X_i$$

Now '*r*' can be worked out as follows by the method of least squares,

$$r = \sqrt{1 - \frac{\text{Unexplained variation}}{\text{Total variation}}}$$

$$= \sqrt{1 - \frac{\Sigma\left(Y - \hat{Y}\right)^2}{\Sigma\left(Y - \bar{Y}\right)^2}} \quad = \sqrt{\frac{\Sigma\left(\hat{Y} - \bar{Y}\right)^2}{\Sigma\left(Y - \bar{Y}\right)^2}}$$

$$= \sqrt{\frac{a\sum Y + b\sum XY - n\bar{Y}^2}{\sum Y^2 - n\bar{Y}^2}}$$

This is as per short-cut formula,

$$r = \sqrt{\frac{7.25(108) + 0.95(597) - 9(12)^2}{1356 - 9(12)^2}}$$

$$= \sqrt{\frac{783 + 567.15 - 1296}{1356 - 1296}}$$

$$= \sqrt{\frac{54.15}{60}} \quad = \sqrt{0.9025} \; = 0.95$$

(b) Coefficient of correlation by the method based on regression coefficients is worked out as,

□ Regression coefficients of Y on X,

i.e.,
$$b_{YX} = \frac{\sum XY - n\bar{X}\bar{Y}}{\sum X^2 - n\bar{X}^2}$$

$$= \frac{597 - 9 \times 5 \times 12}{285 - 9(5)^2} = \frac{597 - 540}{285 - 225} = \frac{57}{60}$$

Regression coefficient of X on Y,

i.e.,
$$b_{XY} = \frac{\sum XY - n\bar{X}\bar{Y}}{\sum Y^2 - n\bar{Y}^2}$$

$$= \frac{597 - 9 \times 5 \times 12}{1356 - 9(12)^2} = \frac{597 - 540}{1356 - 1296} = \frac{57}{60}$$

Hence,
$$r = \sqrt{b_{YX} \, . \, b_{XY}}$$

$$= \sqrt{\frac{57}{60} \times \frac{57}{60}} = \frac{57}{60} = 0.95$$

(c) Coefficient of correlation by the product moment method of Karl Pearson is worked out as,

$$r = \frac{\sum XY - n\bar{X}\bar{Y}}{\sqrt{\sum X^2 - n\bar{X}^2}\sqrt{\sum Y^2 - n\bar{Y}^2}}$$

$$= \frac{597 - 9(5)(12)}{\sqrt{285 - 9(5)^2}\sqrt{1356 - 9(12)^2}}$$

$$= \frac{597 - 540}{\sqrt{285 - 225}\sqrt{1356 - 1296}} = \frac{57}{\sqrt{60}\sqrt{60}} = \frac{57}{60} = 0.95$$

Hence, we get the value of $r = 0.95$. We get the same value by applying the other two methods also. Therefore, whichever method we apply, the results will be the same.

Other Measures

Two other measures are often talked about along with the coefficients of determinations and that of correlation. These are as follows:

(a) Coefficient of Non-Determination: Instead of using coefficient of determination, sometimes coefficient of nondetermination is used. Coefficient of non-determination (denoted by k^2) is the ratio of unexplained variation to total variation in the Y variable related to the X variable. Algebrically,

$$k^2 = \frac{\text{Unexplained variation}}{\text{Total variation}} = \frac{\Sigma\left(Y-\hat{Y}\right)^2}{\Sigma\left(Y-\overline{Y}\right)^2}$$

Concerning the data of Example 6.1, coefficient of nondetermination will be calculated as follows:

$$k^2 = \frac{260.54}{2526.10} = 0.103$$

The value of k^2 shows that about 10 per cent of the variation in consumption expenditure remains unexplained by the regression equation we had worked out, viz., $\hat{Y} = 14.000 + 0.616X_i$. In simple terms, this means that variable other than X is responsible for 10 per cent of the variations in the dependent variable Y in the given case.

Coefficient of non-determination can as well be worked out as,

$$k^2 = 1 - r^2$$

Accordingly for Example 6.1, it will be equal to 1–0.897 = 0.103.

Note: Always remember that $r^2 + k^2 = 1$.

(b) Coefficient of Alienation: Based on k^2, we can work out one more measure, namely the coefficient of alienation, symbolically written as 'k'. Thus, coefficient of alienation, i.e., 'k' $= \sqrt{k^2}$.

Unlike $r + k^2 = 1$, the sum of 'r' and 'k' will not be equal to 1 unless one of the two coefficients is 1 and in this case the remaining coefficients must be zero. In all other cases, 'r' + 'k' > 1. Coefficient of alienation is not a popular measure from a practical point of view and is used very rarely.

RANK CORRELATION

If observations on two variables are given in the form of ranks and not as numerical values, it is possible to compute what is known as rank correlation between the two series.

The **rank correlation**, written as ρ, is a descriptive index of agreement between ranks over individuals. It is the same as the ordinary coefficient of correlation computed on ranks, but its formula is simpler.

$$\rho = 1 - \frac{6\Sigma D_i^2}{n(n^2-1)}$$

Here, n is the number of observations and D_i, the positive difference between ranks associated with the individuals i.

Like *r*, the rank correlation lies between –1 and +1. Consider Examples 6.3 and 6.4 for better understanding

Example 6.3: The ranks given by two judges to 10 individuals are as follows:

	Rank given by			
Individual	*Judge I*	*Judge II*	*D*	D^2
	x	*y*	$= x - y$	
1	1	7	6	36
2	2	5	3	9
3	7	8	1	1
4	9	10	1	1
5	8	9	1	1
6	6	4	2	4
7	4	1	3	9
8	3	6	3	9
9	10	3	7	49
10	5	2	3	9
				$\Sigma D^2 = 128$

Solution:

The rank correlation is given by,

$$\rho = 1 - \frac{6\Sigma D^2}{n^3 - n} = 1 - \frac{6 \times 128}{10^3 - 10} = 1 - 0.776 = 0.224$$

The value of ρ = 0.224 shows that the agreement between the judges is not high.

Example 6.4: Consider Example 6.3 and compute *r* and compare.

Solution:

The simple coefficient of correlation *r* for the previous data is calculated as follows:

x	*y*	x^2	y^2	*xy*
1	7	1	49	7
2	5	4	25	10
7	8	49	64	56
9	10	81	100	90
8	9	64	81	72
6	4	36	16	24
4	1	16	1	4
3	6	9	36	18
10	3	100	9	30
5	2	25	4	10
$\Sigma x = 55$	$\Sigma y = 55$	$\Sigma x^2 = 385$	$\Sigma y^2 = 385$	$\Sigma xy = 321$

$$r = \frac{321 - 10 \times \frac{55}{10} \times \frac{55}{10}}{\sqrt{385 - 10 \times \left(\frac{55}{10}\right)^2}\sqrt{385 - 10 \times \left(\frac{55}{10}\right)^2}} = \frac{18.5}{\sqrt{82.5 \times 82.5}} = \frac{18.5}{82.5} = 0.224$$

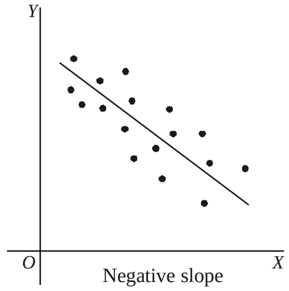
This shows that the Spearman ρ for any two sets of ranks is the same as the Pearson *r* for the set of ranks. However, it is much easier to compute ρ.

Often, the ranks are not given. Instead, the numerical values of observations are given. In such a case, we must attach the ranks to these values to calculate ρ.

Example 6.5: Show by means of diagrams various cases of scatter expressing correlation between *x*, *y*.

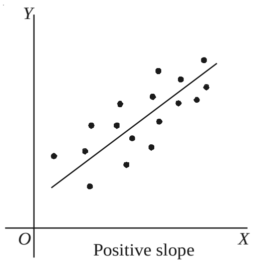
Solution:

(*a*)



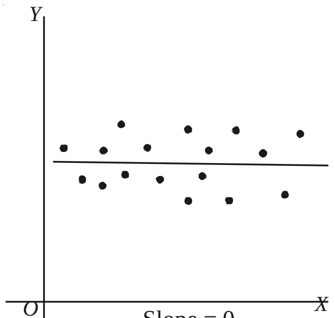
Negative slope
Inverse linear relationship
High scatter *r* low, negative

(*b*)



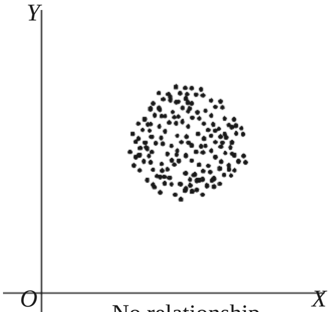
Positive slope
Direct linear relationship
High scatter *r* low, positive

(c)



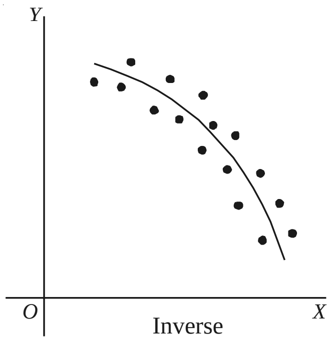
Slope = 0
r = 0

(d)



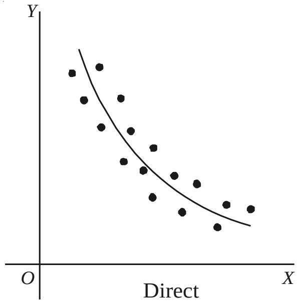
No relationship
$r \simeq 0$

(e)



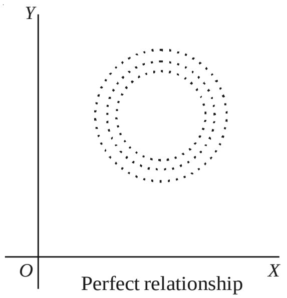
Inverse
curvilinear relationship

(f)



Direct
curvilinear relationship

(g)



Perfect relationship
But, *r* = 0 because of
non-linear relation

Correlation analysis helps us in determining the degree to which two or more variables are related to each other.

When there are only two variables, we can determine the degree to which one variable is linearly related to the other. Regression analysis helps in determining the pattern of relationship between one or more independent variables and a dependent variable. This is done by an equation estimated with the help of data.

PRODUCT MOMENT CORRELATION

In statistics, the Pearson Product Moment Correlation Coefficient is a measure of the correlation (linear dependence) between two variables *X* and *Y*, giving a value between +1 and –1 inclusive. It is sometimes referred the PPMCC or PCC or Pearson's *r*. It is widely used in the sciences as a measure of the strength of linear dependence between two variables. It was developed by Karl Pearson from a related idea introduced by Francis Galton in the 1880s.

Pearson's correlation coefficient between two variables is defined as the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a 'product moment', i.e., the mean (the first moment about the origin) of the product of the mean adjusted random variables; hence the modifier *product moment* in the name.

For a Population

Pearson's correlation coefficient when applied to a population is commonly represented by the Greek letter ρ (rho) and may be referred to as the *population correlation coefficient* or the *population Pearson correlation coefficient*. The formula for ρ is:

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E\left[(X-\mu_X)(Y-\mu_Y)\right]}{\sigma_X \sigma_Y}$$

For a Sample

Pearson's correlation coefficient when applied to a sample is commonly represented by the letter *r* and may be referred to as the *sample correlation coefficient* or the *sample Pearson correlation coefficient*. We can obtain a formula for *r* by substituting estimates of the covariances and variances based on a sample into the above formula. That formula for *r* is:

$$r = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}\sqrt{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}}$$

An equivalent expression gives the correlation coefficient as the mean of the products of the standard scores. Based on a sample of paired data (X_i, Y_i), the sample Pearson correlation coefficient is as follows:

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{s_X}\right)\left(\frac{Y_i - \bar{Y}}{s_Y}\right)$$

Where,

$\frac{X_i - \bar{X}}{s_X}$, $\bar{X}$, and s_X are the standard score, sample mean, and sample standard deviation, respectively.

How to Calculate Product Moment Correlation Coefficient?

The product moment correlation coefficient allows you to work out the linear dependence of two variables (referred to as *X* and *Y*). Let us consider an example, suppose you are the owner of a restaurant. You record the time of every 10th customer stayed in your restaurant (*X* in minutes) and the amount spend (*Y*, in rupees). If it is considered that the longer time the customer stayed the bigger is the amount spend, then this would be a positive correlation. Or it can also be considered in the other way, i.e., the richer the client the lesser time he takes for lunch in restaurant, then this would be a negative correlation. Pearson Product-Moment Correlation Coefficient or PMCC can be calculated to find the correlation in a situation.

Step 1: Remove Incomplete Pairs: After removing incomplete pairs, use only those observations where *both X* and *Y* are known. However, do not exclude observations just because one of the values equals zero.

Step 2: Summarize the Data into the Values needed for the Calculation: These are:

- *n*: The number of data.
- ΣX: The sum of all the *X* values.
- ΣX^2: The sum of the squares of the *X* values.
- ΣY: The sum of all the *Y* values.
- ΣY^2: The sum of the squares of the *Y* values.
- ΣXY: The sum of each *X* value multiplied by its corresponding *Y* value.

Step 3: Calculate S_{XY}, S_{XX} and S_{YY} using these values.

- $\mathbf{S}_{XY} = \Sigma XY - (\Sigma XY \div n)$
- $\mathbf{S}_{XX} = \Sigma X^2 - (\Sigma X\ \Sigma X \div n)$
- $\mathbf{S}_{XY} = \Sigma Y^2 - (\Sigma Y\ \Sigma Y \div n)$

Step 4: Insert these Values into the Equation below to Calculate the Product Moment Correlation Coefficient (*r*). The value should be between 1 and –1.

$$r = \frac{S_{xy}}{\sqrt{S_{xy} S_{yy}}}$$

- If a value is close to 1 implies strong positive correlation. The higher the *X*, the higher the *Y*.
- If a value close to 0 implies little or no correlation.
- If a value close to –1 implies strong negative correlation. The higher the *X*, the lower the *Y*.

REGRESSION ANALYSIS

The term 'regression' was first used in 1877 by Sir Francis Galton who made a study that showed that the height of children born to tall parents will tend to move back or 'regress' toward the mean height of the population. He designated the word regression as the name of the process of predicting one variable from another variable. He coined the term multiple regression to describe the process by which several variables are used to predict another. Thus, when there is a well-established relationship between variables, it is possible to make use of this relationship in making estimates and to forecast the value of one variable (the unknown or the dependent variable) on the basis of the other variable/s (the known or the independent variable/s). A banker, for example, could predict deposits on the basis of per capita income in the trading area of bank. A marketing manager, may plan his advertising expenditures on the basis of the expected effect on total sales revenue of a change in the level of advertising expenditure. Similarly, a hospital superintendent could project his need for beds on the basis of total population. Such predictions may be made by using regression analysis. An investigator may employ regression analysis to test his theory having the cause and effect relationship. All these explain that regression analysis is an extremely useful tool especially in problems of business and industry involving predictions.

NOTES

Assumptions in Regression Analysis

While making use of the regression techniques for making predictions, the following are always assumed:

(a) There is an actual relationship between the dependent and independent variables.

(b) The values of the dependent variable are random but the values of the independent variable are fixed quantities without error and are chosen by the experimentor.

(c) There is a clear indication of direction of the relationship. This means that dependent variable is a function of independent variable. (For example, when we say that advertising has an effect on sales, then we are saying that sales has an effect on advertising).

(d) The conditions (that existed when the relationship between the dependent and independent variable was estimated by the regression) are the same when the regression model is being used. In other words, it simply means that the relationship has not changed since the regression equation was computed.

(e) The analysis is to be used to predict values within the range (and not for values outside the range) for which it is valid.

Simple Linear Regression Model

In case of simple linear regression analysis, a single variable is used to predict another variable on the assumption of linear relationship (i.e., relationship of the type defined by $Y = a + bX$) between the given variables. The variable to be predicted is called the dependent variable and the variable on which the prediction is based is called the independent variable.

Simple linear regression model[3] (or the Regression Line) is stated as,

$$Y_i = a + bX_i + e_i$$

Where, Y_i = The dependent variable

X_i = The independent variable

e_i = Unpredictable random element (usually called residual or error term)

[3] Usually, the estimate of Y denoted by $\hat{Y}$ is written as,

$$\hat{Y} = a + bX_i$$

on the assumption that the random disturbance to the system averages out or has an expected value of zero (i.e., $e = 0$) for any single observation. This regression model is known as the Regression line of Y on X from which the value of Y can be estimated for the given value of X.

NOTES

(a) *a* represents the *Y*-intercept, i.e., the intercept specifies the value of the dependent variable when the independent variable has a value of zero. (However, this term has practical meaning only if a zero value for the independent variable is possible).

(b) *b* is a constant, indicating the slope of the regression line. Slope of the line indicates the amount of change in the value of the dependent variable for a unit change in the independent variable.

If the two constants (viz., *a* and *b*) are known, the accuracy of our prediction of *Y* (denoted by $\hat{Y}$ and read as *Y*-hat) depends on the magnitude of the values of e_i. If in the model, all the e_i tend to have very large values then the estimates will not be very good, but if these values are relatively small, then the predicted values ($\hat{Y}$) will tend to be close to the true values (Y_i).

Estimating the intercept and slope of the regression model (or estimating the regression equation)

The two constants or the parameters viz., '*a*' and '*b*' in the regression model for the entire population or universe are generally unknown and as such are estimated from sample information. The following are the two methods used for estimation:

(a) Scatter diagram method

(b) Least squares method

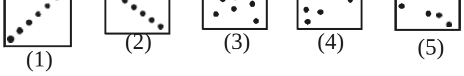
1. Scatter Diagram Method

This method makes use of the Scatter diagram also known as Dot diagram. *Scatter diagram*[4] is a diagram representing two series with the known variable, i.e., independent variable plotted on the *X*-axis and the variable to be estimated, i.e., dependent variable to be plotted on the *Y*-axis on a graph paper (see Figure 6.2) to get the following information illustrated in Table 6.1:

Table 6.1 *Table Derived from Scatter Diagram*

Income *X* *(Hundreds of Rupees)*	*Consumption Expenditure* *Y* *(Hundreds of Rupees)*
41	44
65	60
50	39
57	51
96	80
94	68
110	84
30	34
79	55
65	48

[4]



Five possible forms, which Scatter diagram may assume has been depicted in the above five diagrams. Diagram (1) is indicative of perfect positive relationship. Diagram (2) shows perfect negative relationship. Diagram (3) shows no relationship, Diagram (4) shows positive relationship and *Diagram* (5) shows negative relationship between the two variables under consideration.

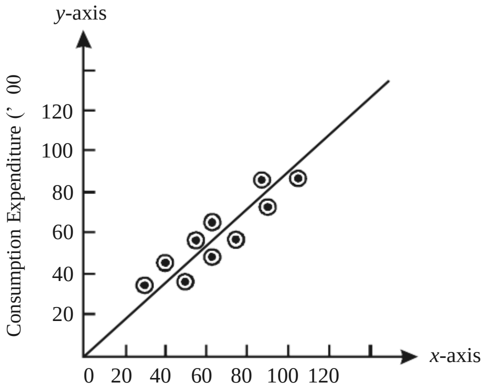


Fig. 6.2 *Scatter Diagram*

The scatter diagram by itself is not sufficient for predicting values of the dependent variable. Some formal expression of the relationship between the two variables is necessary for predictive purposes. For the purpose, one may simply take a ruler and draw a straight line through the points in the scatter diagram and this way can determine the intercept and the slope of the said line and then the line can be defined as $\hat{Y} = a + bX_i$, with the help of which we can predict *Y* for a given value of *X*. However, there are shortcomings in this approach. For example, if five different persons draw such a straight line in the same scatter diagram, it is possible that there may be five different estimates of *a* and *b*, especially when the dots are more dispersed in the diagram. Hence, the estimates cannot be worked out only through this approach. A more systematic and statistical method is required to estimate the constants of the predictive equation. The least squares method is used to draw the best fit line.

2. Least Square Method

The least squares method of fitting a line (the line of best fit or the regression line) through the scatter diagram is a method which minimizes the sum of the squared vertical deviations from the fitted line. In other words, the line to be fitted will pass through the points of the scatter diagram in such a way that the sum of the squares of the vertical deviations of these points from the line will be a minimum.

The meaning of the least squares criterion can be easily understood through Figure 6.3, where the earlier Figure 6.2 in scatter diagram has been reproduced along with a line which represents the least squares line to fit the data.

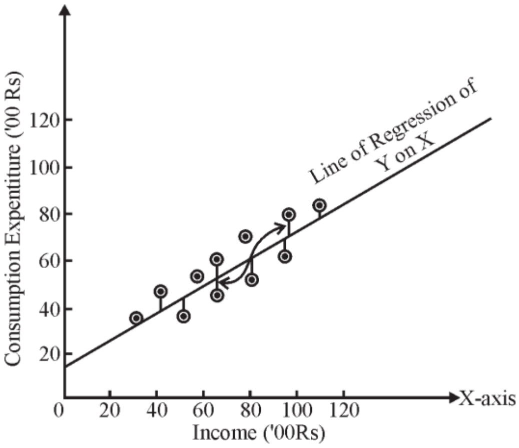


Fig. 6.3 *Scatter Diagram, Regression Line and Short Vertical Lines Representing 'e'*

In Figure 6.3, the vertical deviations of the individual points from the line are shown as the short vertical lines joining the points to the least squares line. These deviations will be denoted by the symbol '*e*'. The value of '*e*' varies from one point to another. In some cases it is positive, while in others it is negative. If the line drawn happens to be the least squares line, then the values of $\sum e_i$ is the least possible. It is because of this feature the method is known as Least Squares Method.

Why we insist on minimizing the sum of squared deviations is a question that needs explanation. If we denote the deviations from the actual value *Y* to the estimated value $\hat{Y}$ as $(Y - \hat{Y})$ or e_i, it is logical that we want the $\Sigma(Y - \hat{Y})$ or $\sum_{i=1}^{n} e_i$, to be as small as possible. However, mere examining $\Sigma(Y - \hat{Y})$ or $\sum_{i=1}^{n} e_i$, is inappropriate, since any e_i can be positive or negative. Large positive values and large negative values could cancel one another. However, large values of e_i regardless of their sign, indicate a poor prediction. Even if we ignore the signs while working out $\sum_{i=1}^{n} | e_i |$, the difficulties may continue. Hence, the standard procedure is to eliminate the effect of signs by squaring each observation. Squaring each term accomplishes two purposes, viz., (i) it magnifies (or penalizes) the larger errors, and (ii) it cancels the effect of the positive and negative values (since a negative error when squared becomes positive). The choice of minimizing the squared sum of errors rather than the sum of the absolute values implies that there are many small errors rather than a few large errors. Hence, in obtaining the regression line, we follow the approach that the sum of the squared deviations be minimum and on this basis work out the values of its constants viz., '*a*' and '*b*' also known as the intercept and the slope of the line. This is done with the help of the following two normal equations:[5]

$$\Sigma Y = na + b\Sigma X$$

$$\Sigma XY = a\Sigma X + \mathrm{b}\Sigma X^2$$

In these two equations, '*a*' and '*b*' are unknowns and all other values viz., $\sum X$, $\sum Y$, $\sum \mathrm{X}^2$, $\sum \mathrm{XY}$, are the sum of the products and cross products to be calculated from the sample data, and '*n*' means the number of observations in the sample.

Example 6.6 explains the Least squares method.

Example 6.6: Fit a regression line $\hat{Y} = a + bX_i$ by the method of Least squares to the following sample information.

[5] If we proceed centering each variable, i.e., setting its origin at its mean, then the two equations will be as under:

$$\sum Y = na + b\sum X$$
$$\sum XY = a\sum X + b\sum X^2$$

But since $\sum Y$ and $\sum X$ will be zero, the first equation and the first term of the second equation will disappear and we shall simply have the following equations:

$$\sum XY = b\sum X^2$$
$$b = \sum XY/\sum X^2$$

The value of '*a*' can then be worked out as:

$$a = \overline{Y} - b\overline{X}$$

Observations	1	2	3	4	5	6	7	8	9	10
Income (*X*) ('00 `)	41	65	50	57	96	94	110	30	79	65
Consumption Expenditure (*Y*) ('00 `)	44	60	39	51	80	68	84	34	55	48

Solution:

We are to fit a regression line $\hat{Y} = a + bX_i$ to the given data by the method of Least Squares. Accordingly, we work out the '*a*' and '*b*' values with the help of the normal equations as stated above and also for the purpose, work out $\sum X$, $\sum Y$, $\sum XY$, $\sum X^2$ values from the given sample information table on summations for regression equation.

Summations for Regression Equation

Observations	*Income* *X* ('00 `)	*Consumption Expenditure* *Y* ('00 `)	*XY*	X^2	Y^2
1	41	44	1804	1681	1936
2	65	60	3900	4225	3600
3	50	39	1950	2500	1521
4	57	51	2907	3249	2601
5	96	80	7680	9216	6400
6	94	68	6392	8836	4624
7	110	84	9240	12100	7056
8	30	34	1020	900	1156
9	79	55	4345	6241	3025
10	65	48	3120	4225	2304
$n = 10$	$\sum X = 687$	$\sum Y = 563$	$\sum XY = 42358$	$\sum X^2 = 53173$	$\sum Y^2 = 34223$

Putting the values in the required normal equations we have,

$$563 = 10a + 687b$$

$$42358 = 687a + 53173b$$

Solving these two equations for *a* and *b* we obtain,

$$a = 14.000 \quad \text{and} \quad b = 0.616$$

Hence, the equation for the required regression line is,

$$\hat{Y} = a + bX_i$$

or,

$$\hat{Y} = 14.000 + 0.616X_i$$

This equation is known as the regression equation of *Y* on *X* from which *Y* values can be estimated for given values of *X* variable.[6]

[6] It should be pointed out that the equation used to estimate the *Y* variable values from values of *X* should not be used to estimate the values of *X* variable from given values of *Y* variable. Another regression equation (known as the regression equation of *X* on *Y* of the type $X = a + bY$) that reverses the two value should be used if it is desired to estimate *X* from value of *Y*.

NOTES

Checking the Accuracy of Equation

After finding the regression line, one can check its accuracy also. The method to be used for the purpose follows from the mathematical property of a line fitted by the method of least squares, viz., the individual positive and negative errors must sum to zero. In other words, using the estimating equation one must find out whether the term $\Sigma\left(Y - \hat{Y}\right)$ is zero and if this is so, then one can reasonably be sure that he has not committed any mistake in determining the estimating equation.

The Problem of Prediction

When we talk about prediction or estimation, we usually imply that if the relationship $Y_i = a + bX_i + e_i$ exists, then the regression equation, $\hat{Y} = a + bX_i$ provides a base for making estimates of the value of *Y* which will be associated with particular values of *X*. In Example 6.6, we worked out the regression equation for the income and consumption data as,

$$\hat{Y} = 14.000 + 0.616X_i$$

On the basis of this equation, we can make a point estimate of *Y* for any given value of *X*. Suppose we wish to estimate the consumption expenditure of individuals with income of ` 10,000. We substitute *X* = 100 for the same in our equation and get an estimate of consumption expenditure as,

$$\hat{Y} = 14.000 + 0.616(100) = 75.60$$

Thus, the regression relationship indicates that individuals with ` 10,000 of income may be expected to spend approximately ` 7,560 on consumption. However, this is only an expected or an estimated value and it is possible that actual consumption expenditure of the individual with that income may deviate from this amount and if so, then our estimate will be an error, the likelihood of which will be high if the estimate is applied to any one individual. The interval estimate method is considered better and it states an interval in which the expected consumption expenditure may fall. Remember that the wider the interval, the greater the level of confidence we can have, but the width of the interval (or what is technically known as the precision of the estimate) is associated with a specified level of confidence and is dependent on the variability (consumption expenditure in our case) found in the sample. This variability is measured by the standard deviation of the error term, '*e*', and is popularly known as the standard error of the estimate.

Standard Error of the Estimate

Standard error of estimate is a measure developed by statisticians for measuring the reliability of the estimating equation. Like the standard deviation, the Standard Error (S.E.) of $\hat{Y}$ measures the variability or scatter of the observed values of *Y* around the regression line. Standard Error of Estimate (S.E. of $\hat{Y}$) is worked out as,

$$\text{S.E. of } \hat{Y} \text{ (or } S_e) = \sqrt{\frac{\sum (Y - \hat{Y})^2}{n-2}} = \sqrt{\frac{\sum e^2}{n-2}}$$

where, S.E. of $\hat{Y}$ (or S_e) = Standard error of the estimate

Y = Observed value of Y

$\hat{Y}$ = Estimated value of Y

e = The error term = $(Y - \hat{Y})$

n = Number of observations in the sample

Note: In the above formula, $n - 2$ is used instead of n because of the fact that two degrees of freedom are lost in basing the estimate on the variability of the sample observations about the line with two constants viz., 'a' and 'b' whose position is determined by those same sample observations.

The square of the S_e, also known as the variance of the error term, is the basic measure of reliability. The larger the variance, the more significant are the magnitudes of the e's and the less reliable is the regression analysis in predicting the data.

Interpreting the Standard Error of Estimate and Finding the Confidence Limits for the Estimate in Large and Small Samples

The larger the S.E. of estimate (SE$_e$), the greater happens to be the dispersion, or scattering, of given observations around the regression line. However, if the S.E. of estimate happens to be zero, then the estimating equation is a 'perfect' estimator (i.e., cent per cent correct estimator) of the dependent variable.

(a) In case of large samples, i.e., where $n > 30$ in a sample, it is assumed that the observed points are normally distributed around the regression line and we may find that,

- 68 per cent of all points lie within $\hat{Y} \pm 1$ SE$_e$ limits.
- 95.5 per cent of all points lie within $\hat{Y} \pm 2$ SE$_e$ limits.
- 99.7 per cent of all points lie within $\hat{Y} \pm 3$ SE$_e$ limits.

This can be stated as,

- The observed values of Y are normally distributed around each estimated value of $\hat{Y}$ and;
- The variance of the distributions around each possible value of $\hat{Y}$ is the same.

(b) In case of small samples, i.e., where $n \leq 30$ in a sample the 't' distribution is used for finding the two limits more appropriately.

This is done as follows:

Upper limit = $\hat{Y}$ + 't' (SE$_e$)

Lower limit = $\hat{Y}$ – 't' (SE$_e$)

Where, $\hat{Y}$ = The estimated value of Y for a given value of X.

SE_e = The standard error of estimate.

't' = Table value of 't' for given degrees of freedom for a specified confidence level.

Some other Details Concerning Simple Regression

Sometimes, the estimating equation of Y also known as the regression equation of Y on X, is written as,

$$(\hat{Y} - \overline{Y}) = r\frac{\sigma_Y}{\sigma_X}(X_i - \overline{X})$$

or,

$$\hat{Y} = r\frac{\sigma_Y}{\sigma_X}(X_i - \overline{X}) + \overline{Y}$$

NOTES

Where, r = Coefficient of simple correlation between X and Y

σ_Y = Standard deviation of Y

σ_X = Standard deviation of X

$\overline{X}$ = Mean of X

$\overline{Y}$ = Mean of Y

$\hat{Y}$ = Value of Y to be estimated

X_i = Any given value of X for which Y is to be estimated

This is based on the formula we have used, i.e., $\hat{Y} = a + bX_i$. The coefficient of X_i is defined as,

$$\text{Coefficient of } X_i = b = r\frac{\sigma_Y}{\sigma_X}$$

(Also known as regression coefficient of Y on X or slope of the regression line of Y on X) or b_{YX}.

$$= \frac{\sum XY - n\overline{X}\overline{Y} \times \sqrt{\sum Y^2 - n\overline{Y}^2}}{\sqrt{\sum Y^2 - n\overline{Y}^2}\sqrt{\sum X^2 - n\overline{X}^2}\sqrt{\sum X^2 - n\overline{X}^2}}$$

$$= \frac{\sum XY - n\overline{X}\overline{Y}}{\sum X^2 - n\overline{X}^2}$$

and
$$a = -r\frac{\sigma_Y}{\sigma_X}\overline{X} + \overline{Y}$$

$$= \overline{Y} - b\overline{X} \qquad \left(\text{since } b = r\frac{\sigma_Y}{\sigma_X}\right)$$

Similarly, the estimating equation of X, also known as the regression equation of X on Y, can be stated as,

$$\left(\hat{X} - \overline{X}\right) = r\frac{\sigma_X}{\sigma_Y}\left(Y - \overline{Y}\right)$$

or
$$\hat{X} = r\frac{\sigma_X}{\sigma_Y}\left(Y - \overline{Y}\right) + \overline{X}$$

and the

Regression coefficient of X on Y (or b_{XY}) $= r\dfrac{\sigma_X}{\sigma_Y} = \dfrac{\sum XY - n\overline{X}\overline{Y}}{\sum Y^2 - n\overline{Y}^2}$

If we are given the two regression equations as stated above, along with the values of 'a' and 'b' constants to solve the same for finding the value of X and Y, then the values of X and Y so obtained, are the mean values of X (i.e., $\overline{X}$) and the mean value of Y (i.e., $\overline{Y}$).

If we are given the two regression coefficients (viz., b_{XY} and b_{YX}), then we can work out the value of coefficient of correlation by just taking the square root of the product of the regression coefficients as shown,

$$r = \sqrt{b_{YX}.b_{XY}}$$

$$= \sqrt{r\frac{\sigma_Y}{\sigma_X}.r\frac{\sigma_X}{\sigma_Y}}$$

$$= \sqrt{r.r} = r$$

The (±) sign of r will be determined on the basis of the sign of the given regression coefficients. If regression coefficients have minus sign then r will be taken with minus (–) sign and if regression coefficients have plus sign then r will be taken with plus (+) sign, (Remember that both regression coefficients will necessarily have the same sign, whether it is minus or plus, for their sign is governed by the sign of coefficient of correlation.) To understand it better, see Examples 6.7 and 6.8.

Example 6.7: Given is the following information:

	$\overline{X}$	$\overline{Y}$
Mean	39.5	47.5
Standard Deviation	10.8	17.8

Simple correlation coefficient between X and Y is = + 0.42.

Find the estimating equation of Y and X.

Solution:

Estimating equation of Y can be worked out as,

□ $$\left(\hat{Y} - \overline{Y}\right) = r\frac{\sigma_Y}{\sigma_X}\left(X_i - \overline{X}\right)$$

or $$\hat{Y} = r\frac{\sigma_Y}{\sigma_X}\left(X_i - \overline{X}\right) + \overline{Y}$$

$$= 0.42\frac{17.8}{10.8}\left(X_i - 39.5\right) + 47.5$$

$$= 0.69X_i - 27.25 + 47.5$$

$$= 0.69X_i + 20.25$$

Similarly, the estimating equation of X can be worked out as

□ $$\left(\hat{X} - \overline{X}\right) = r\frac{\sigma_X}{\sigma_Y}\left(Y_i - \overline{Y}\right)$$

or $$\hat{X} = r\frac{\sigma_X}{\sigma_Y}\left(Y_i - \overline{Y}\right) + \overline{X}$$

or $$= 0.42\frac{10.8}{17.8}\left(Y_i - 47.5\right) + 39.5$$

$$= 0.26Y_i - 12.35 + 39.5$$

$$= 0.26Y_i + 27.15$$

Example 6.8: The following is the given data:

Variance of $X = 9$

Regression equations:

$$4X - 5Y + 33 = 0$$

$$20X - 9Y - 107 = 0$$

Find: (a) Mean values of X and Y.

(b) Coefficient of Correlation between X and Y.

(c) Standard deviation of Y.

Solution:

(a) For finding the mean values of X and Y, we solve the two given regression equations for the values of X and Y as follows:

NOTES

$$4X - 5Y + 33 = 0 \qquad ...(1)$$

$$20X - 9Y - 107 = 0 \qquad ...(2)$$

If we multiply Equation (1) by 5, we have the following equations:

$$20X - 25Y = -165 \qquad ...(3)$$

$$20X - 9Y = 107 \qquad ...(2)$$

$$\underline{- \quad + \quad -}$$

$$-16Y = -272$$

Subtracting Equation (2) from (3)

or $\quad Y = 17$

Putting this value of Y in Equation (1) we have,

$$4X = -33 + 5(17)$$

or $\quad X = \dfrac{-33+85}{4} = \dfrac{52}{4} = 13$

Hence, $\quad \bar{X} = 13$ and $\bar{Y} = 17$

(b) For finding the coefficient of correlation, we first presume one of the two given regression equations as the estimating equation of X. Let equation $4X - 5Y + 33 = 0$ be the estimating equation of X, then we have,

$$\hat{X} = \frac{5Y_i}{4} - \frac{33}{4}$$

and

From this we can write $b_{XY} = \dfrac{5}{4}$.

The other given equation is then taken as the estimating equation of Y and can be written as,

$$\hat{Y} = \frac{20X_i}{9} - \frac{107}{9}$$

and from this we can write $b_{YX} = \dfrac{20}{9}$.

If the above equations are correct then r must be equal to,

$$r = \sqrt{5/4 \times 20/9} = \sqrt{25/9} = 5/3 = 1.6$$

which is an impossible equation, since r can in no case be greater than 1. Hence, we change our supposition about the estimating equations and by reversing it, we re-write the estimating equations as,

$$\hat{X} = \frac{9Y_i}{20} + \frac{107}{20}$$

and $\quad \hat{Y} = \dfrac{4X_i}{5} + \dfrac{33}{5}$

Hence, $\quad r = \sqrt{9/20 \times 4/5}$

$$= \sqrt{9/25}$$

Check Your Progress

1. What is the theory of correlation?
2. What is coefficient of determination?
3. Define coefficient of correlation.
4. What are dependent and independent variables?
5. List the two methods used for estimation.

$$= 3/5$$

$$= 0.6$$

Since, regression coefficients have plus signs, we take $r = + 0.6$.

(c) Standard deviation of Y can be calculated,

□ Variance of $X = 9$ $\therefore$ Standard deviation of $X = 3$

□ $$b_{YX} = r\frac{\sigma_Y}{\sigma_X} = \frac{4}{5} = 0.6\frac{\sigma_Y}{3} = 0.2\sigma_Y$$

Hence, $\sigma_Y = 4$

Alternatively, we can work it out as,

□ $$b_{XY} = r\frac{\sigma_X}{\sigma_Y} = \frac{9}{20} = 0.6\frac{\sigma_Y}{3} = \frac{1.8}{\sigma_Y}$$

Hence, $\sigma_Y = 4$

NOTES

ACTIVITY

1. Using the various correlation methods discussed in the unit, compute the correlation for the following data:

Person	Height (x)	Self Esteem (y)
1	68	4.1
2	71	4.6
3	62	3.8
4	75	4.4
5	58	3.2
6	60	3.1

2. Two random variables have the regression with equations,

   $3X + 2Y - 26 = 0$

   $6X + Y - 31 = 0$

   Find the mean value of X as well as of Y and the correlation coefficient between X and Y. If the variance of X is 25, find $\sum$Y from the data given above.

DID YOU KNOW

The technique of **correlation** is used to test the statistical significance of the association. The *r* value will **always** lie between 1 and +1. If you have an *r* value outside of this range you have made an error in the calculations.

SUMMARY

- Correlation analysis is the statistical tool generally used to describe the degree to which one variable is related to another.
- The coefficient of determination (symbolically indicated as r^2, though some people would prefer to put it as R^2) is a measure of the degree of linear association or

correlation between two variables, say *X* and *Y*, one of which happens to be the independent variable and the other being the dependent variable.

- The coefficient of correlation, symbolically denoted by '*r*', is an important measure to describe how well one variable is explained by another. It measures the degree of relationship between two casually-related variables.
- In case of simple linear regression analysis, a single variable is used to predict another variable on the assumption of linear relationship (i.e., relationship of the type defined by $Y = a + bX$) between the given variables.
- In statistics, the Pearson Product Moment Correlation Coefficient is a measure of the correlation (linear dependence) between two variables *X* and *Y*, giving a value between +1 and –1 inclusive. It is sometimes referred the PPMCC or PCC or Pearson's *r*.
- Pearson's correlation coefficient between two variables is defined as the covariance of the two variables divided by the product of their standard deviations.
- Scatter diagram is a diagram representing two series with the known variable, i.e., independent variable plotted on the *X*-axis and the variable to be estimated, i.e., dependent variable to be plotted on the *Y*-axis on a graph paper
- The least squares method of fitting a line (the line of best fit or the regression line) through the scatter diagram is a method which minimizes the sum of the squared vertical deviations from the fitted line.
- The square of the S_e, also known as the variance of the error term, is the basic measure of reliability. The larger the variance, the more significant are the magnitudes of the *e*'s and the less reliable is the regression analysis in predicting the data.

KEY TERMS

- **Correlation analysis:** The statistical tool used to describe the degree to which one variable is related to another
- **Coefficient of determination:** A measure of the degree of linear association or correlation between two variables, say *X* and *Y*, one of which happens to be an independent variable and the other being a dependent variable
- **Coefficient of non-determination:** The ratio of unexplained variation to total variation in the *Y* variable related to the *X* variable
- **Rank correlation:** A descriptive index of agreement between ranks over individuals
- **Standard error of the estimate:** A measure developed by statisticians for measuring the reliability of the estimating equation

ANSWERS TO 'CHECK YOUR PROGRESS'

1. The theory by means of which quantitative connections between two sets of phenomena are determined is called the 'Theory of Correlation'.

2. The coefficient of determination (symbolically indicated as r^2, though some people would prefer to put it as R^2) is a measure of the degree of linear association or correlation between two variables, say X and Y, one of which happens to be independent variable and the other being dependent variable.
3. The coefficient of correlation, symbolically denoted by '*r*', is another important measure to describe how well one variable is explained by another. It measures the degree of relationship between the casually-related variables.
4. The variable to be predicted is called the dependent variable and the variable on which the prediction is based is called the independent variable.
5. The following are the two methods used for estimation:
   (a) Scatter diagram method.
   (b) Least squares method.

QUESTIONS AND EXERCISES

Short-Answer Questions

1. What is the importance of correlation analysis?
2. How will you determine the coefficient of determination?
3. When is correlation positive and when is it negative?
4. What are the assumptions in Karl Pearson's coefficient?
5. What is the relationship between coefficient of nondetermination and coefficient of alienation?
6. List the basic precautions and limitations of regression and correlation analyses.
7. Differentiate between scatter diagram and least square method.

Long-Answer Questions

1. Explain the method to calculate the coefficient of correlation using simple regression coefficient.
2. Describe Karl Pearson's method of measuring coefficient of correlation.
3. Explain Spearman's rank correlation.
4. What is regression analysis? What are the assumptions in it?
5. Explain scatter diagram and the least square method in detail. Also, mention how scatter diagram helps in studying correlation between two variables.

FURTHER READING

Best, John W. and James V. Kahn. 2005. *Research in Education,* 10th edition. New Jersey: Pearson Education.

Butcher, Harold John. 1966. *Sampling in Educational Research*, 3rd edition. United Kingdom: Manchester University Press.

Edwards, Allen Louis. 2006. *Experimental Design in Psychological Research,* 3rd edition. United States: The University of Michigan.

NOTES

Garrett, Henry Edward. 1926. *Statistics in Psychology and Education*, New Jersey: Longmans, Green and Company.

Guilford, Joy Paul. 1977. *Fundamental Statistics in Psychology and Education*, 6th edition. New York: McGraw Hill.

Kerlinger, Fred Nichols and Howard Bing Lee. 2000. *Foundations of Behavioral Research*, 4th edition. United States: Harcourt College Publishers.

UNIT 3 NORMAL PROBABILITY CURVE AND STATISTICAL SIGNIFICANCE

NOTES

Structure

Introduction
Unit Objectives
Divergence in Non-Probability Curve
Skewness
Kurtosis
Characteristics of Normal Probability Curve
Uses of Normal Probability Curve
Statistical Inference
Summary
Key Terms
Answers to 'Check Your Progress'
Questions and Exercises
Further Reading

INTRODUCTION

The *Normal Probability Curve* (NPC), simply known as normal curve, is a symmetrical bell-shaped curve. This curve is based upon the law of probability and discovered by French mathematician Abraham Demoivre (1667–1754) in the 18th century. In this curve, the mean, median and mode lie at the middle point of the distribution. The total area of the curve represents the total number of cases and the middle point represents the mean, median and mode. The base line is divided into six sigma units (**σ** units). The scores more than the mean come on the + **σ** side and the scores less than the mean come on the – **σ** side. The mean point (middle point) is marked as zero (0). All the scores are expected to lie between – 3**σ** to + 3**σ**.

UNIT OBJECTIVES

After going through this unit, you will be able to:

- Apply the normal probability curve
- Explain the characteristics of the normal probability curve
- Identify the divergences in the normal probability curve
- List the applications of normal probability curve
- Interpret the values according to the table of area under the normal probability curve

DIVERGENCE IN NON-PROBABILITY CURVE

If the distribution is not normal, the mean, median and mode do not lie on a particular point on the base line. Sometimes, the curve of NPC is either more peaked or more flat. This is known as divergence in the normality. Divergence is of two types:

(i) Skewness, and (ii) Kurtosis

Skewness

This term refers to lack of symmetry. A normal curve is a perfectly symmetrical curve. In a perfect normal curve, the mean, median and mode lie in the centre and are the same. In many distributions of data which deviate from the normal, the values of mean, median and mode are different; and also there is no symmetry between the right and left halves. This type of distribution is said to be skewed.

Skewness is of two types, positive skewness and negative skewness.

In case of positive skewness, the scores are more at the left side of the curve and less on the right side of the curve. So the mean lies to the right of the median. In this, there are more individuals in a distribution who score less than the average score.

In case of negative skewness, the scores are massed at the right end of the curve and are less on the left end of the curve, so the mean lies to the left of the median. In this there are many individuals in a group whose scores are higher than the average scores of the group.

The skewness of a given distribution is calculated by the following formulae:

(i) $$\text{SK} = \frac{3(\text{Mean} - \text{Median})}{\text{Standard Deviation}}$$

(ii) $$\text{SK} = \frac{\text{P90} + \text{P10}}{2} - \text{P50}$$

iii. $$\text{SK} = \frac{Q_3 - 2Q_2 + Q_1}{Q_3 - Q_1}$$

Figures 6.1 and 6.2 show the two types of skewed curves.

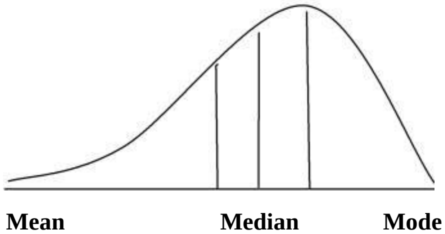


Fig. 7.1 Negative Skewed Curve

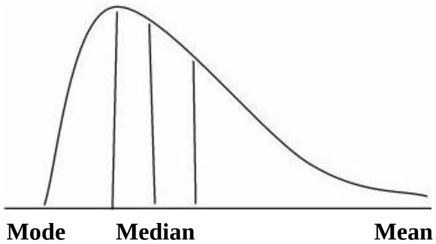


Fig. 7.2 Positive Skewed Curve

Kurtosis

When the individual score is near the average score of the group, the curve becomes flattened in the middle. On the other hand, when there are too many cases distributed in the centre, the curve becomes peaked in comparison to normal. Kurtosis is of three types:

(i) Platykurtic

(ii) Leptokurtic

(iii) Mesokurtic

A curve is said to be platykurtic when it is flatter than the normal curve as shown in Figure 7.3.

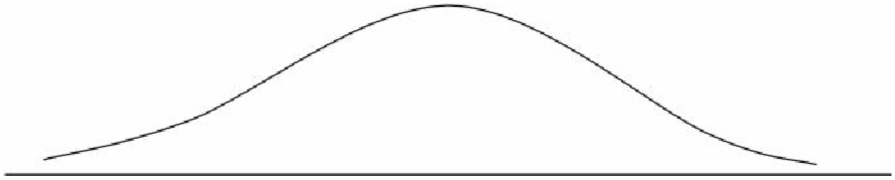


Fig. 7.3 *Platykurtic Curve*

A curve is said to be leptokurtic when it is more peaked than the normal curve as illustrated in Figure 7.4.

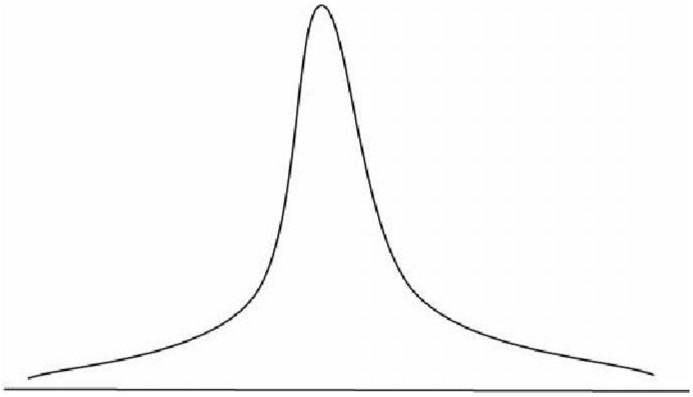


Fig. 7.4 *Leptokurtic Curve*

The normal curve is known as mesokurtic, when kurtosis is 0.263, as you can see in Figure 7.5.

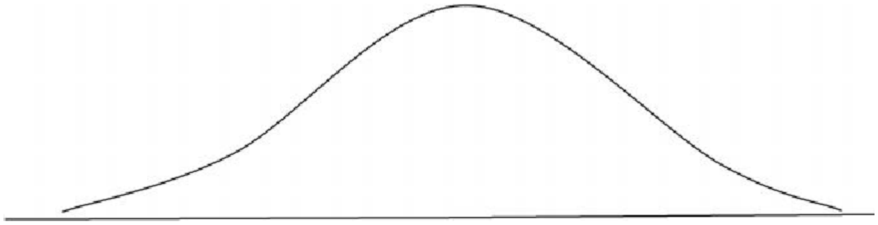


Fig. 7.5 *Mesmokurtic Curve*

Causes for Divergence from Normality

The common causes of divergence are as follows:

- The biased selection of sample.
- Development of poor tests.
- Defect in the construction and administration of the test.
- Defect in the scoring process.
- Lack of normality in the trait.

CHARACTERISTICS OF NORMAL PROBABILITY CURVE

The NPC has several features which are essential to understand for its use. The major characteristics are limited. They are as follows:

(i) It is a bell shaped curve.

(ii) The measures of central tendency are equal, i.e, mean, mode and median concentrate on one point.

(iii) The height of the curve is .3989.

(iv) It is an asymptotic curve. The ends of the curve approach but never touch the X-axis at the extremes because of the possibility of locating in the population, in cases where scores are still higher than our highest score or lower than our lowest score. Therefore theoretically, it extends from minus infinity to plus infinity as illustrated in Figure 7.6. Here, M is the mean or expectation (location of the peak) and **σ** is the standard deviation.

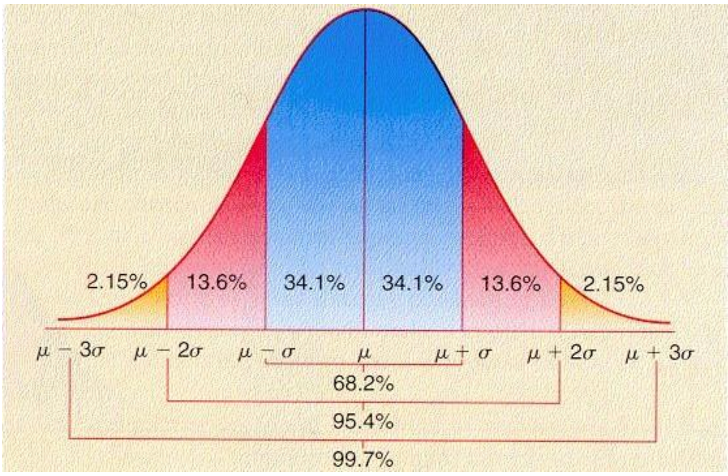


Fig. 7.6 *Normal Curve Showing Areas at Different Distances from the Mean*

(v) It has 50 per cent frequency above and 50 per cent below the mean. The mean is zero and it is always reference point.

(vi) Standard deviation of a normal curve is always 1.

(vii) The points of inflection of the curve occur at points –1 unit above and below mean.

(viii) The distribution of frequency per cent has the definite limits.

(ix) There is a definite relation between quartile deviation and standard deviation in a normal distribution curve.

(x) It is a mathematical curve and is an open-ended curve.

Some limits are:

- The middle 68 per cent frequency is between –1 and +1.
- The middle 95 per cent frequency is between –1.96 and + 1.96.
- The middle 99 per cent frequency is between –2.58 and + 2.58.

The total area under the normal curve is arbitrarily taken as 10,000. Every score should be converted into standard score (Z – score), by using the formula $Z = \frac{X - M}{\sigma}$.

The area in proportion should be converted into a percentage at the time of reading the table. From the table, we can see the areas from mean to **σ** and also we can read the value of **σ** scores from the mean for the corresponding fractional area.

USES OF NORMAL PROBABILITY CURVE

The uses of normal probability curve are discussed in this section.

(i) NPC is used to determine the percentage of cases within given limits

Example 1: Given a distribution of scores with a mean of 15 and a standard deviation of 5, what percentage of cases lie between 18 and 25? Figure 7.7 can assist in the calculation of the answer.

Solution: Both the raw scores (18 and 25) are to be converted into Z scores.

$$\text{Z score of 18} = \frac{X - M}{\sigma} = \frac{18 - 15}{5}$$

$$= \frac{3}{5}$$

$$= 0.6\sigma$$

$$\text{Z score of 25} = \frac{X - M}{\sigma} = \frac{25 - 15}{5}$$

$$= \frac{10}{5}$$

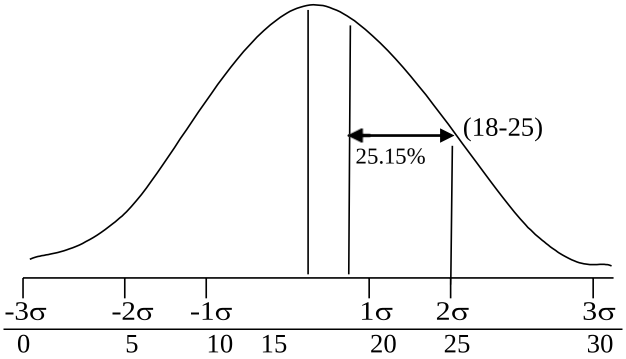
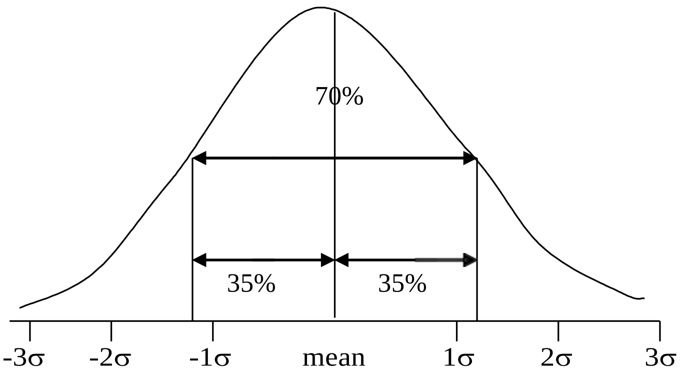


Fig. 7.7 *Determining the Percentage of Cases within Given Limits*

According to the table of area of a normal probability curve, the total percentage of cases lie between the mean and 0.6**σ** is 22.57. The percentage of cases lie between the mean and 2**σ** is 47.72. So, the total percentage of cases that fall between the scores 18 and 25 is 47.72 – 22.57 = 25.15.

(ii) NPC is used to determine the limit which includes a given percentage of cases

Example 2: Given a distribution of scores with a mean of 12 and an **σ** of 6, what limits will include the middle 70 per cent of the cases? Refer to Figure 7.8 to assist in calculating the answer.



Mean = 12, σ= 6

Fig. 7.8 *Determining the Limit of a Given Percentage of Cases*

Solution: The middle 70 per cent of the cases in a normal distribution signifies that 35 per cent cases above the mean and also 35 per cent cases below the mean. According to the table of area under NPC, 35 per cent of cases fall between the mean and 1.04 **σ**. So the middle 70 per cent of the cases will lie between –1.04**σ** to + 1.04**σ**.

The value of 1 **σ** = **6**

So 1.04 **σ** = 6 × 1.04 = 6.24

The value of mean = 12

So the lowest limit for the middle 70 per cent cases of the distribution is 12 – 6.24 = 5.76.

The highest limit for the middle 70 per cent cases of the distribution is 12 + 6.24 = 18.24.

Thus, the middle 70 per cent cases lie in between 5.76 and 18.24.

(iii)NPC is used to determine the percentile rank of a student in his class

Example 3: The score of a student in a class test is 70. The mean for the whole class is 50 and the **σ** is 10. Find the percentile rank of the student in the class. Figure 7.9 will assist in finding the answer.

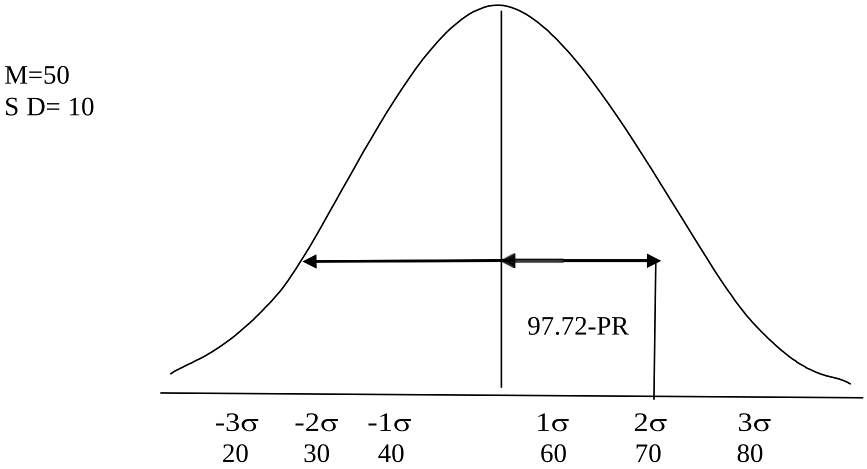


Fig. 7.9 *Determining the Percentile Rank of a Student in his Class*

Solution: The Z score for the score 70 is $\frac{70-50}{10}$ =2σ

As per the table of area under the NPC, the area of the curve that lies between mean and 2σ is 47.72 per cent. The total percentage of cases below 70 is:

50 + 47.72 = 97.72 per cent or 98 per cent.

Thus, the percentile rank of the student is 98.

(iv) NPC is used to find out the percentile value of a student whose percentile rank is known

Example 4: The percentile rank of a student in a class test is 80. The mean of the class in the test is 50 and the σ is 20. Calculate the student's score in the class test. Figure 7.10 illustrates the case.

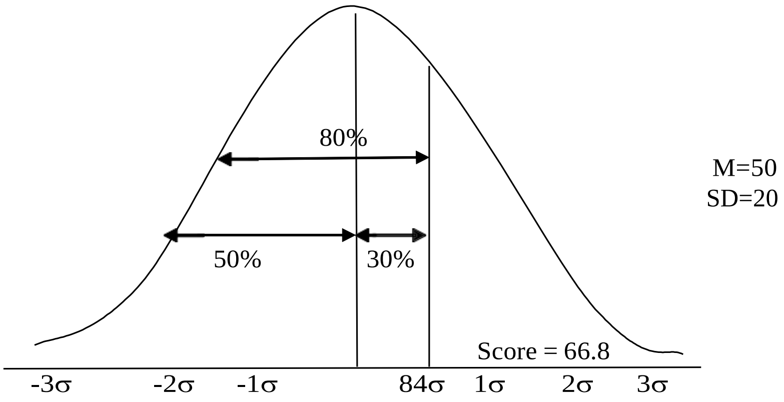


Fig. 7.10 *Determining the Percentile Value of a Student whose Percentile Rank is Known*

Solution: The student has scored 30 per cent scores above the mean. According to the table of area under NPC, 30 per cent cases from the mean is 0.84σ.

$1\sigma = 20.$

$0.84\sigma = 20 \times .84 = 16.8$

Thus, the percentile value of the student is 50 + 16.8 = 66.8.

(v) NPC is used to divide a group into sub-groups according to their capacity

Example 5: Suppose there is a group of 100 students in a Commerce class. We want to divide them into five small groups A, B, C, D and E according to their ability, the range of ability being equal in each sub-group as shown in Figure 7.11. Find out how many students should be placed in each category.

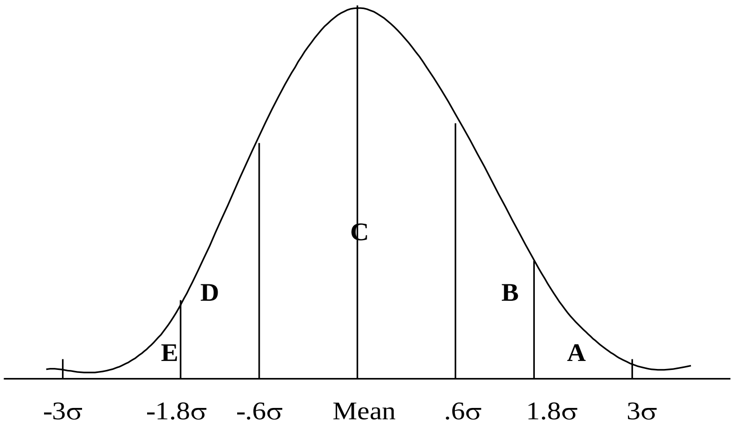


Fig. 7.11 *A Group Divided into Sub-Groups According to their Capacity*

NOTES

Solution: The total area under NPC is –3**σ** to + 3**σ**, that is 6**σ**. This 6**σ** should be divided into five parts, so 6**σ** ÷5 = 1.2**σ**.

According to the table of area under NPC:

3.5 per cent of the cases lie between 1.8 **σ** to 3**σ** (Group A, the high scorers). 23.8 per cent of the cases lie between .6σto 1.8**σ** (23.8 per cent of the cases for B and also 23.8 per cent of the cases for D), the middle 45 per cent of the cases lie – .6**σ**to + .6**σ** (Group C), and the lowest 3.5 per cent of the cases lie between –3**σ** to – 1.8**σ** (Group E)

In category 'A' the number of students = 3.5 per cent = 3 or 4 students.

In category 'B' the number of students = 23.8 per cent = 24 students.

In category 'C' the number of students = 45 per cent = 45 students.

In category 'D' the number of students = 23.8 per cent = 24 students.

In category 'E' the number of students = 3.5 per cent = 3 or 4 students.

(vi) NPC is used to compare the scores of students in two different tests

Example 6: Suppose, a student scored 60 marks in English test and 80 marks in statistics test. The mean and SD for the English test is 30 and 10 respectively, whereas for the statistics test the mean is 70 and SD is 10. Find out, in which subject the student performed better using Figure 7.12.

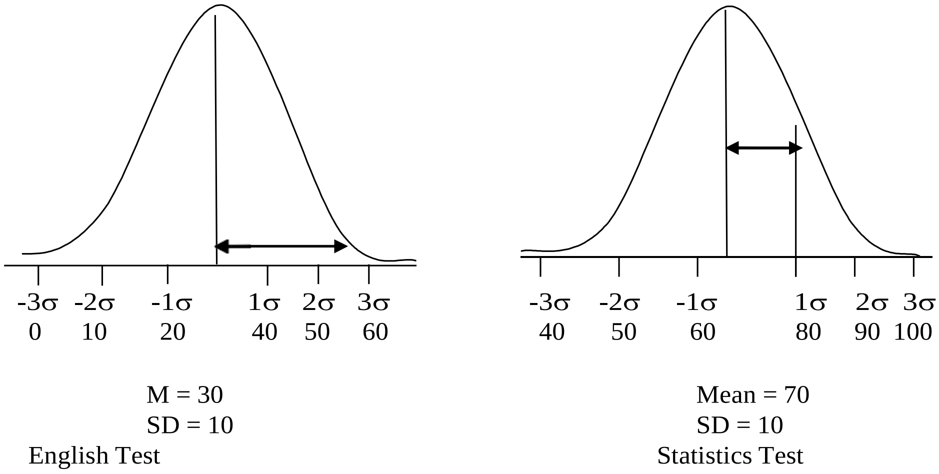


Fig. 7.12 Comparing the Scores of Students in Two Different Tests

Solution: In case of the English test:

Raw score = 60

Mean = 30

SD =10

So Z score for the English test = $\frac{X - M}{\sigma} = \frac{60 - 30}{10} = \frac{30}{10} = 3\sigma$

In case of statistics test raw score = 80

Mean = 70

SD = 10

So Z Score for the statistics test = $\frac{X-M}{\sigma} = \frac{80-70}{10} = \frac{10}{10} = 1\sigma$

So, the student has done better in the English than the statistics on.

(vii) NPC is used to determine the relative difficulty level of test items

Example 7: In a standardized test of psychology, question numbers A, B, C and D were solved by the students, 45 per cent, 38 per cent, 30 per cent and 15 per cent respectively. Assuming the normality, find out the relative difficulty level of the questions. Also explain the difficulty levels of questions. Table 7.1 displays the information in tabular form.

Solution:

Table 7.1 *Determining the Difficulty Level of Test Items*

Question Number	**Percentage of Successful students**	**Percentage of Unsuccessful Students**	**Percentage distance of Mean of Unsuccessful students**	**Difficulty level**
A	45	55	55-50=5	0.13σ
B	38	62	62-50=12	0.31σ
C	30	70	70-50=20	0.52σ
D	15	85	85-50=35	1.04σ

As we know that in an NPC, 50 – 50 cases lie both the sides of mean. The mean of NPC is that point which is shown as 0. In an NPC, the explanation of difficulty level is done on the basis of **σ** — distance. Therefore, if a question is at the positive side of the NPC and **σ** has more distance from the mean, the question of a test will be much difficult. The relative difficulty value of the test items has been shown:

The question

A to B is 0.18**σ** is more difficult (0.31**σ** –0.13**σ** = .18**σ**)

A to C is 0.39**σ** is more difficult (0.52**σ** –0.13**σ** = .39**σ**)

A to D is 0.91**σ** is more difficult (1.04**σ** –0.13**σ** = .91**σ**)

B to C is 0.21**σ** is more difficult (0.52**σ** –0.31**σ** = .21**σ**)

B to D is 0.73**σ** is more difficult (1.04**σ** –0.31**σ** = .73**σ**)

C to D is 0.52**σ** is more difficult (1.04**σ** –0.152**σ** = .52**σ**)

STATISTICAL INFERENCE

Statistical inference refers to drawing conclusions based on data that is subjected to random variation; for example, sampling variation or observational errors. The terms statistical inference, statistical induction and inferential statistics are used to describe systems of procedures that can be used to draw conclusions from datasets arising from systems affected by random variation.

There are many contexts in which inference is desirable, and there are also various approaches of performing inferences. One of the most important contexts is parametric models. For example, if you have noisy (x, y) data that you think follow the pattern $y = \beta_0 + \beta_1 x$ + error, then you can estimate β_0, β_1 and the magnitude of the error.

Check Your Progress

1. What is the shape of the normal probability curve?
2. Define ‘ordinate’.
3. From Normal table find the area between mean ordinate for which Z or sigma score has the value – 1.65.

The fundamental requirements of such set of procedures for inference are that they must be common so that it can be applied on a range of conditions and produce a logical and reasonable conclusion whenever applied to well-defined and simple situations. The result of this procedure, when used in analysis of statistical data is generally an estimate or a set of estimates of one or more parameters that describe the problem along with some indication of uncertainty with which the values are estimated. This technique is different from descriptive statistics in respect that descriptive statistics is just a straightforward presentation of facts, in which decisions are made by influence of data analyst; whereas there is no influence of analyst in statistical inference.

The method of statistical inference is generally used for point estimation, interval estimation, hypothesis testing or statistical significance testing and prediction of a random process.

Statistical inference is generally distinguished from descriptive statistics. In simple terms, descriptive statistics can be thought of as being just a straightforward presentation of facts, in which modeling decisions made by a data analyst have had minimal influence. Any statistical inference requires some assumptions. A statistical model is a set of assumptions concerning the generation of observed data and similar data. A complete statistical analysis will nearly always include both descriptive statistics and statistical inference, and will often progress in a series of steps where the emphasis moves gradually from description to inference.

ACTIVITY

1. Using the NPC determine the percentile of a student whose score is 80. The mean for the class is 60 and the **σ** is 10.

DID YOU KNOW

The terms statistical inference, statistical induction and inferential statistics are specifically used to describe systems of procedures that can be used to draw conclusions from datasets arising from systems affected by random variation, such as observational errors, random sampling or random experimentation.

SUMMARY

- The Normal Probability Curve (NPC), simply known as normal curve, is a symmetrical bell-shaped curve.
- If the distribution is not normal, the mean, median and mode do not lie on a particular point on the base line. When the curve of NPC is either more peaked or more flat, it is known as 'divergence in the normality'.
- Divergence is of two types: skewed and kurtosis. There can be different causes for divergence.
- The uses of a NPC are, to determine the percentage of cases within given limits, to determine the limit which includes a given percentage of cases, to determine

Check Your Progress

4. Enumerate two important characteristics of NPC.
5. Define the term statistical inference.

the percentile rank of a student in his class, to divide a group into sub-groups according to their capacity and to determine the relative difficulty level of test items.

- Statistical inference refers to drawing conclusions based on data that is subjected to random variation; for example, sampling variation or observational errors.
- The method of statistical inference is generally used for point estimation, interval estimation, hypothesis testing or statistical significance testing and prediction of a random process.
- Any statistical inference requires some assumptions. A statistical model is a set of assumptions concerning the generation of observed data and similar data.

KEY TERMS

- **Normal Probability Curve (NPC):** It is simply known as normal curve and is a symmetrical bell-shaped curve based upon the law of probability and discovered by French mathematician Abraham Demoivre
- **Skewness:** This term refers to lack of symmetry and is of two types, positive skewness and negative skewness
- **Leptokurtic curve:** A curve is said to be leptokurtic when it is more peaked than the normal curve

ANSWERS TO 'CHECK YOUR PROGRESS'

1. The normal probability curve is a symmetrical bell-shaped curve.
2. Ordinate is the vertical axis around which NPC is symmetrical.
3. 0.4505
4. The two important characteristics of NPC are as follows:
   (i) The curve is symmetrical around its vertical axis, i.e., ordinate.
   (ii) The highest of ordinate is maximum at mean.
5. Statistical inference means drawing conclusions based on data that is subjected to random variation; for example, sampling variation or observational errors.

QUESTIONS AND EXERCISES

Short-Answer Questions

1. When is a distribution said to skewed?
2. What are the common causes of divergence from normality in the curve?

Long-Answer Questions

1. Describe the three types of kurtosis with the help of diagrams.
2. Write a detailed note on the characteristics of the normal probability curve.
3. What are the uses of the normal probability curve?
4. What is statistical inference? Discuss with the help of an example.

FURTHER READING

NOTES

Best, John W. and James V. Kahn. 2005. *Research in Education,* 10th edition. New Jersey: Pearson Education.

Butcher, Harold John. 1966. *Sampling in Educational Research,* 3rd edition. United Kingdom: Manchester University Press.

Edwards, Allen Louis. 2006. *Experimental Design in Psychological Research,* 3rd edition. United States: The University of Michigan.

Garrett, Henry Edward. 1926. *Statistics in Psychology and Education,* New Jersey: Longmans, Green and Company.

Guilford, Joy Paul. 1977. *Fundamental Statistics in Psychology and Education,* 6th edition. New York: McGraw Hill.

Kerlinger, Fred Nichols and Howard Bing Lee. 2000. *Foundations of Behavioral Research,* 4th edition. United States: Harcourt College Publishers.

UNIT 4 HYPOTHESIS TESTING

NOTES

Structure

Introduction
- Unit Objectives
- Hypothesis
  - Null and Alternative Hypotheses
- Types of Errors
- Level of Significance
- Critical Region
- One-Tailed and Two-Tailed Tests
- Small Sample Tests
  - t-Test for Single Mean
  - Paired t-Test: Difference of Means
- Chi-Square (χ^2) Test
  - Concept of Test Statistics
- F-Test
  - Test for Equality of Two Population Variances

Large Sample Tests
- Test for Single Mean
- Test for Difference of Means

Analysis of Variance
- One-Way Classification
- Two-Way Classification

Summary

Key Terms

Answers to 'Check Your Progress'

Questions and Exercises

Further Reading

INTRODUCTION

In this unit, you will study about hypothesis testing. Hypothesis is an assumption that is tested to find its logical or empirical consequences. A hypothesis should be clear and accurate. Various concepts such as null and alternative hypotheses help to verify the testability of an assumption. You can determine whether the hypothesis is appropriate for judging the population proportion.

In this unit, you will also learn about the basic principles of experimentation and ANOVA. In business decisions, we are often involved in determining if there are significant differences among various sample means, from which conclusions can be drawn about the differences among various population means. The methodology used for such types of determinations is known as ANalysis Of VAriance or ANOVA. This technique is one of the most powerful techniques in statistical analysis and was developed by R.A. Fisher. It is also called the *F*-Test. The basic principle of ANOVA is to test for differences among the means of the populations by examining the amount of varia.

UNIT OBJECTIVES

After going through this unit, you will be able to:

- Understand the concepts of hypothesis and the types of errors
- Explain the different types of hypotheses
- Identify the critical region or region of hypothesis rejection
- Explain the tests of equality of two proportions
- Understand the concepts of standard errors of statistics
- Discuss the types of classification involved in analysis of variance
- Explain the steps involved in determining the differences within the factor of one way classification ANOVA
- Describe the two-way ANOVA techniques

HYPOTHESIS

A hypothesis is an approximate assumption that a researcher wants to test for its logical or empirical consequences. Hypothesis refers to a provisional idea whose merit needs evaluation, but having no specific meaning. Though it is often referred to as a convenient mathematical approach for simplifying cumbersome calculation. Setting up and testing hypotheses is an integral art of statistical inference. Hypotheses are often statements about population parameters like variance and expected value. During the course of hypothesis testing, some inference about population like the mean and proportion are made. Any useful hypothesis will enable predictions by reasoning including deductive reasoning. According to Karl Popper a hypothesis must be falsifiable and that a proposition or theory cannot be called scientific if it does not admit the possibility of being shown false. Hypothesis might predict outcome of an experiment in a lab setting the observation of a phenomenon in nature. Thus, hypothesis is a explanation of a phenomenon proposal suggesting a possible correlation between multiple phenomena.

The characteristics of hypothesis are:

- **Clear and accurate**: Hypothesis should be clear and accurate so as to draw a consistent conclusion.
- **Statement of relationship between variables**: If a hypothesis is relational, it should state the relationship between different variables.
- **Testability**: A hypothesis should be open to testing so that other deductions can be made from it and can be confirmed or disproved by observation. The researcher should do some prior study to make the hypothesis a testable one.
- **Specific with limited scope**: A hypothesis, which is specific, with limited scope, is easily testable than a hypothesis with limitless scope. Therefore, a researcher should pay more time to do research on such kind of hypotheses.
- **Simplicity**: A hypothesis should be stated in the most simple and clear terms to make it understandable.
- **Consistency**: Ahypothesis should be reliable and consistent with established and known facts.
- **Time-Limit**: Ahypothesis should be capable of being tested within a reasonable time. In other words, it can be said that the excellence of a hypothesis is judged by the time taken to collect the data needed for the test.

- **Empirical reference**: A hypothesis should explain or support all the sufficient facts needed to understand what the problem is all about.

A hypothesis is a statement or assumption concerning a population. For the purpose of decision-making, a hypothesis has to be verified and then accepted or rejected. This is done with the help of observations. We test a sample and make a decision on the basis of the result obtained. Decision-making plays significant role in different areas such as marketing, industry and management.

Statistical Decision-Making

Testing a statistical hypothesis on the basis of a sample enables us to decide whether the hypothesis should be accepted or rejected. The sample data enable us to accept or reject the hypothesis. Since the sample data give incomplete information about the population the result of the test need not be considered to be final or unchallengeable. The procedure, which, on the basis of sample results, enables us to decide whether a hypothesis is to be accepted or rejected, is called Hypothesis Testing or Test of Significance.

Note: Atest provides evidence, if any, against a hypothesis, usually called a mill hypothesis. The test cannot prove the hypothesis to be correct. It can give some evidence against it.

The hypothesis makes some assumption about the density function of the random variate. The sampling distribution is fundamental to this subject.

The test of a hypothesis means a procedure to decide whether to accept or reject a hypothesis.

If a sample is found to have an untenable probability (of occurrence) level (called the significance level), we reject the hypothesis. Usually the probability levels of 0.05 and 0.01 are taken. They are called 5% and 1% significance levels.

Note: The acceptance of a hypothesis implies there is no evidence from the sample that we should believe otherwise.

The rejection of a hypothesis leads us to conclude that it is false. This way of putting the problem is convenient because of the uncertainty inherent in the problem. In view of this we must always briefly state a hypothesis that we hope to reject.

A hypothesis stated in the hope of being rejected is called a null hypothesis and is denoted by H_0.

If H_0 is rejected, it may lead to the acceptance of an alternative hypothesis denoted by H_1.

For example, new fragrance soap is introduced in the market. The null hypothesis H_0, which may be rejected, is that the new soap is not better than the existing soap.

Example 1: A die is suspected to be loaded. Roll the die a number of times to test.

Solution: The null hypothesis H_0: $p = 1/6$ for showing six.

The alternative hupothesis H_1: $p \neq 1/6$

Null and Alternative Hypotheses

Hypothesis is usually considered as the principal instrument in research. The basic concepts regarding the testability of a hypothesis are as follows:

(a) **Null Hypothesis and Alternative Hypothesis**: In the context of statistical analysis, while comparing any two methods, the following concepts or assumptions are taken into consideration:

(i) **Null Hypothesis**: While comparing two different methods in terms of their superiority, wherein the assumption is that both the methods are equally good is called null hypothesis. It is also known as statistical hypothesis and is symbolised as $\mathrm{H}_{0.}$

(ii) **Alternate Hypothesis**: While comparing two different methods, regarding their superiority, wherein, stating a particular method to be good or bad as compared to the other one is called alternate hypothesis. It is symbolised as H_0.

(b) **Comparison of Null Hypothesis with Alternate Hypothesis**: Following are the points of comparison between null hypothesis and alternate hypothesis:

(i) Null hypothesis is always specific while Alternate Hypothesis gives an approximate value.

(ii) The rejection of Null hypothesis involves great risk, which is not in the case of Alternate hypothesis.

Null hypothesis is more frequently used in statistics than Alternate hypothesis because it is specific and is not based on probabilities.

The hypothesis to be tested is called the Null Hypothesis and is denoted by H_0.This is to be tested against other possible states of nature called alternative hypotheses. The alternative is usually denoted by $H_{1.}$

The null hypothesis implies that there is no difference between the statistic and the population parameter. To test whether there is no difference between the sample mean x and the population μ, we write the null hypothesis.

$$H_0\text{: } \bar{x} = \mu$$

The alternative hypothesis would be

$$H_0\text{: } \bar{x} \neq \mu$$

This means $\bar{x} > \mu$ or $\bar{x} < \mu$. This is called a two-tailed hypothesis.

The alternative H_0: $\bar{x} > \mu$ is right tailed.

The alternative H_0: $\bar{x} < \mu$ is left tailed.

These are one sided or one-tailed alternatives.

Notes:

1. The alternative hypothesis H_1 implies all such values of the parameter, which are not specified by the null hypothesis H_0.
2. Testing a statistical hypothesis is a rule, which leads to a decision to accept or reject a hypothesis.

A one tailed test requires rejection of the null hypothesis when the sample statistic is greater than the population value or less than the population value at a certain level of significance.

1. We may want to test if the sample mean *x* exceeds the population mean *p*. Then the null hypothesis is,

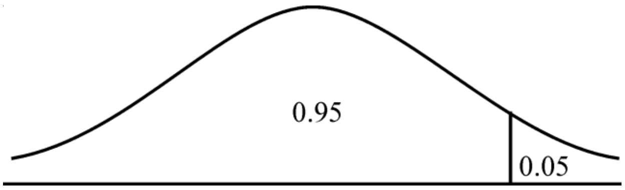
$$H_0\text{: } \bar{x} > \mu$$

2. In the other case the null hypothesis could be

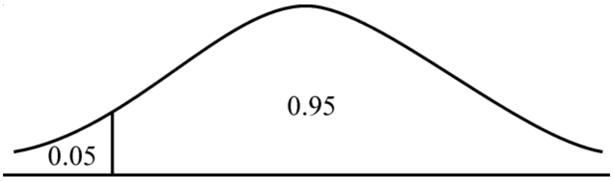
$$H_0\text{: } \bar{x} < \mu$$

Each of these two situations leads to a one tailed test and has to be dealt with in the same manner as the two tailed test. Here the critical rejection is on one side only, right for $\bar{x} > \mu$ and left for $\bar{x} < \mu$. Figure 8.1 a five per cent level of test of significance.

For example, a minister in a certain government has an average life of 11 months without being involved in a scam. A new party claims to provide ministers with an average life of more than 11 months without scam. We would like to test if; on the average the new ministers last longer than 11 months. We may write the null hypothesis H_0: $\bar{x} = 11$ and alternative hypothesis H_1: $\bar{x} > 11$ or H_1: $\bar{x} < 11$.



(a) H_0: > μ



(b) H_0: < μ

Fig. 4.1 *Five Per cent Level of Test of Significance*

TYPES OF ERRORS

There are two types of errors in statistical hypothesis, which are as follows:

(a) **Type I Error**: In this type of error, you may reject a null hypothesis when it is true. It means rejection of a hypothesis, which should have been accepted. It is denoted by α (alpha) and is also known alpha error.

(b) **Type II Error**: In this type of error, you are supposed to accept a null hypothesis when it is not true. It means accepting a hypothesis, which should have been rejected. It is denoted by β (beta) and is also known as beta error.

Type I error can be controlled by fixing it at a lower level, for example, if you fix it at 2%, then the maximum probability to commit Type I error is 0.02. But reducing Type I error, has a disadvantage when the sample size is fixed as it increases the chances of Type II error. In other words, it can be said that both types of errors cannot be reduced simultaneously. The only solution of this problem is to set an appropriate level by considering the costs and penalties attached to them or to strike a proper balance between both types of errors.

In a hypothesis test, a Type I error occurs when the null hypothesis is rejected when it is in fact true; that is, H_0 is wrongly rejected. For example, in a clinical trial of a new drug, the null hypothesis might be that the new drug is no better, on average, than the current drug; that is H_0: there is no difference between the two drugs on average. A Type I error would occur if we concluded that the two drugs produced different effects when in fact there was no difference between them.

In a hypothesis test, a Type II error occurs when the null hypothesis H_0, is not rejected when it is in fact false. For example, in a clinical trial of a new drug, the null hypothesis might be that the new drug is no better, on average, than the current drug; that is H_0: there is no difference between the two drugs on average. A Type II error would occur if it were concluded that the two drugs produced the same effect, that is,

NOTES

there is no difference between the two drugs on average, when in fact they produced different ones.

In how many ways can we commit errors?

We reject a hypothesis when it may be true. This is Type I error.

We accept a hypothesis when it may be false. This is Type II error.

The other true situations are desirable:

We accept a hypothesis when it is true. We reject a hypothesis when it is false as shown in Table 4.1.

Table 4.1 *Accept/Reject Hypothesis*

	Accept H_0	**Reject H_0**
H_0 True	Accept True H_0 Desirable	Reject True H_0 Type I Error
H_1 False	Accept False H_0 Type II Error	Reject False H_0 Desirable

The level of significance implies the probability of Type I error. A five per cent level implies that the probability of committing a Type I error is 0.05. A one per cent level implies 0.01 probability of committing Type I error.

Lowering the significance level and hence the probability of Type I error is good but unfortunately it would lead to the undesirable situation of committing Type II error.

Hence,

Type I Error - Rejecting H_0 when H_0 is true.

Type II Error - Accepting H_0 when H_0 is false.

Note:

The probability of making a Type I error is the level of significance of a statistical test. It is denoted by α.

Where,

α = Probability (Rejecting H_0 / H_0 true) 1–

α = Probability (Accepting H_0 / H_0 true)

The probability of making a Type II error is denoted by β.

Where,

β = Probability (Accepting H_0 / H_0 false)

$1–\beta$ = Probability (Rejecting H_0 / H_0 false) = Probability (The test correctly rejects H_0 when H_0 is false).

$1–\beta$ is called the power of the test. It depends on the level of significance α, sample size n and the parameter value.

LEVEL OF SIGNIFICANCE

The hypothesis is examined on a pre-determined level of significance. Generally either 5 per cent level or 1 per cent level of significance is adopted for the purpose. However, it can be stated here that the level of significance must be adequate keeping in view the purpose and nature of enquiry.

In this concept of hypothesis, you will formulate a rule provided both, null hypothesis and alternate hypothesis are given. Formulating a decision means either accepting null hypothesis and rejecting alternate hypothesis or rejecting null hypothesis and accepting alternate hypothesis. It can be easily understood with the help of an example, wherein you test 10 items and formulate a decision on the basis of the rule that states, a null hypothesis will be accepted if out of those 10 items, either none is defective or only 1 is defective otherwise alternate hypothesis will be accepted.

Suppose u is distributed normally with mean 0 and S.D. 1. Briefly we write $u \sim N$ (0, 1). If the expected value of u is written $E(u)$ the standardized normal variate is,

$$z = \frac{u - E(u)}{SE(u)}$$

The total area under the normal curve is 1 (corresponding to 100%). The area between $z = -1.96$ and $z = +1.96$ is 0.95. This is the region of acceptance with 95% confidence (refer Figure 4.2). We write

$$p(-1.96 \leq z \leq 1.96) = 0.95$$

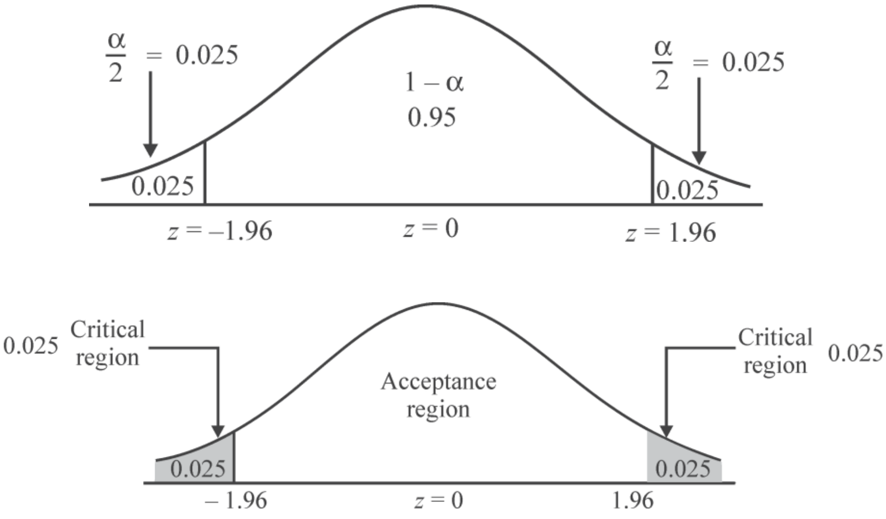


Fig. 8.2 *Region of Acceptance*

The size of the critical region is 0.05 (shaded area 0.025 on the left and 0.025 on the right). This is a two-tailed test.

If $|z|$ remains between the range ± 1.96, we are in the hypothesis acceptance region. The two values – 1.96 and 1.96 are the 5% critical values.

If $|z| > 1.96$ we are in the critical region, *i.e.*, the region of rejection of the hypothesis.

CRITICAL REGION

The Critical Region (CR), or Rejection Region (RR), is a set of values for testing statistic for which the null hypothesis is rejected in a hypothesis test. It means, the sample space for the test statistic is partitioned into two regions; one region as the critical region will lead us to reject the null hypothesis H_0, the other not. So, if the observed value of the test statistic is a member of the critical region, we conclude that "reject H_0"; if it is not a member of the critical region then we conclude that "do not reject H_0".

NOTES

We shall consider test problems arising out of Type I error.

The level of significance of a test is the maximum probability with which we are willing to take a risk of Type I error.

If we take a 5% significance level ($p = 0.05$) we are 95% confident ($p = 0.95$) that a right decision has been made.

A 1% significance level ($p = 0.01$) makes us 99% confident ($p = 0.99$) about the correctness of the decision.

The critical region is the area of the sampling distribution in which the test statistic must fall for the null hypothesis to be rejected.

We can say that the critical region corresponds to the range of values of the statistic, which according to the test requires the hypothesis to be rejected.

ONE-TAILED AND TWO-TAILED TESTS

A two-tailed test rejects the null hypothesis if the sample mean is either more or less than the hypothesised value of the mean of the population. It is considered to be apt when null hypothesis is of some specific value whereas alternate hypothesis is not equal to the value of null hypothesis. In a two-tailed curve there are two rejection regions, also called critical regions (refer Figure 8.3).

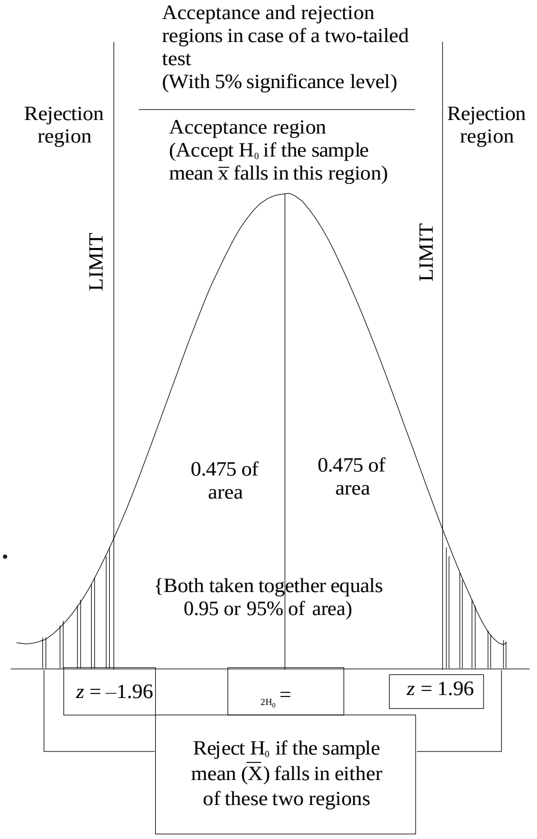

Acceptance and rejection
regions in case of a two-tailed
test
(With 5% significance level)
Rejection
region
Acceptance region
(Accept H_0 if the sample
mean $\bar{x}$ falls in this region)
Rejection
region
LIMIT
LIMIT
0.475 of
area
0.475 of
area
{Both taken together equals
0.95 or 95% of area)
$z = -1.96$
$_{2H_0} =$
$z = 1.96$
Reject H_0 if the sample
mean ($\bar{X}$) falls in either
of these two regions

Fig. 4.3 Acceptance and Rejection Regions

Conditions for the Occurrence of One-Tailed Test: When the population mean is either lower or higher than some hypothesised value, one-tailed test is considered to be appropriate where the rejection is only on the left tail of the curve. This is also known as left-tailed test (refer Figure 8.4).

Check Your Progress

1. Define the term hypothesis.
2. What do you understand by 'Type-I' and 'Type-II' errors?
3. Explain null and alternate hypothesis.

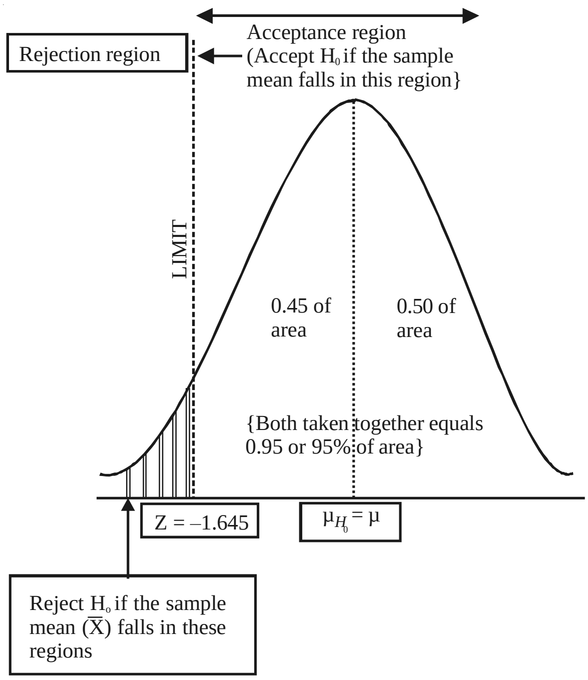

Rejection region
Acceptance region
(Accept H_0 if the sample mean falls in this region}
LIMIT
0.45 of area
0.50 of area
{Both taken together equals 0.95 or 95% of area}
Z = –1.645
$\mu_{H_0} = \mu$
Reject H_0 if the sample mean ($\overline{X}$) falls in these regions

Fig. 8.4 *Left-Tailed Test*

For example, what will happen if the acceptance region is made larger? α will decrease. It will be more easily possible to accept H_0 when H_0 is false (Type II error), i.e., it will lower the probability of making a Type I error but raise that of α, Type II error.

Note: α, β are probabilities of making an error; 1 – α, l – β are probabilities of making correct decisions (refer Figure 8.5).

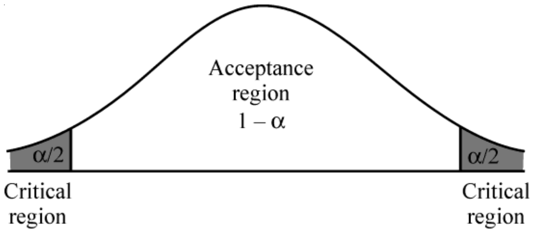


Fig. 8.5 *Acceptance and Critical Regions*

Example 2: Can we say α + β = 1?

Solution: No. Each is concerned with a different type of error. But both are not independent of each other.

SMALL SAMPLE TESTS

The sampling distribution of many statistics for large samples is approximately normal. For small samples with $n < 30$, the normal distribution, as shown above, can be used only if the sample is from a normal population with known σ.

If σ is not known we can use student's t distribution instead of the normal. We then replace σ by sample standard deviation s with some modification, given below.

Check Your Progress

4. Define critical region.
5. Explain decision rule briefly.

Let $x_1, x_2, ..., x_n$ be a random sample of size n drawn from a normal population with mean (μ) and S.D. (σ). We define,

$$t = \frac{\overline{x} - \mu}{s/\sqrt{n-1}}$$

Here t follows the student's t distribution with $n - 1$ degree of freedom.

Note: For small samples of $n < 30$ the term $\sqrt{n-1}$ in SE = $s/\sqrt{n-1}$ corrects the bias resulting from the use of sample standard deviation as an estimator of σ.

Also,

$$\frac{s^2}{S^2} = \frac{n-1}{n} \text{ or } s = S\sqrt{\frac{n-1}{n}}$$

Procedure: Small Samples

To test the null hypothesis $H_0 : \mu = \mu_0$

Against the alternative

Calculate and compare it with the table value with $n - 1$ degrees of freedom at level of significance α%

If this value > table value, reject H_0

If this value < table value, accept H_0

We can also find the 95% (or any other) confidence limits for μ.

For the two-tailed test (use the same rules as for large samples; substitute t for z) the 95% confidence limits are,

$$\overline{x} \pm t_0 s/\sqrt{n-1} \quad \alpha = 0.025$$

Rejection Region. At α% level for two-tailed test. if $|t| > t_{\alpha/2}$ reject

For one-tailed test (right) if $t > t_\alpha$ reject

(left) if $t > t_\alpha$ reject

At 5% level the three cases are

If $|t| > t_{0.025}$ reject	two-tailed
if $t > t_{0.05}$ reject	one-tailed right
if $t < -t_{0.05}$	reject one-tailed left

For proportions, the same procedure is to be followed.

Example 3: A firm produces tubes of diameter 2 cm. A sample of 10 tubes is found to have a diameter of 2.01 cm and variance 0.004. Is the difference significant? Given $t_{0.05,9} = 2.26$

$$t = \frac{\overline{x} - \mu}{s/\sqrt{n-1}}$$

$$= \frac{2.01 - 2}{\sqrt{0.004/10-1}}$$

$$= \frac{0.01}{0.021}$$

$$= 0.48$$

Since $|t| < 2.26$ the difference is not significant at 5% level.

t-Test for Single Mean

Sir William S. Gosset (pen name Student) developed a significance test and through it made significant contribution in the theory of sampling applicable in case of small samples. When population variance is not known, the test is commonly known as Student's *t*-test and is based on the *t* distribution.

Like the normal distribution, *t* distribution is also symmetrical but happens to be flatter than the normal distribution. Moreover, there is a different *t* distribution for every possible sample size. As the sample size gets larger, the shape of the *t* distribution loses its flatness and becomes approximately equal to the normal distribution. In fact, for sample sizes of more than 30, the *t* distribution is so close to the normal distribution that we will use the normal to approximate the *t* distribution. Thus, when *n* is small, the *t* distribution is far from normal, but when *n* as infinite, it is identical with normal distribution.

For applying *t*-test in context of small samples, the *t* value is calculated first of all and then calculated value is compared with the table value of *t* at certain level of significance for given degrees of freedom. If the calculated value of *t* exceeds the table value (say $t_{0.05}$), we infer that the difference is significant at 5% level, but if calculated value is *t* is less than its concerning table value, the difference is not treated as significant.

The *t*-test is used when the following two conditions are fullfilled:

(i) The sample size is less than 30, i.e., when $n \leq 30$.

(ii) The population standard deviation (σ_p) must be unknown.

In using the *t*-test we assume the following:

(i) That the population is normal or approximately normal;

(ii) That the observations are independent and the samples are randomly drawn samples;

(iii) That there is no measurement error;

(iv) That in the case of two samples, population variances are regarded as equal if equality of the two population means is to be tested.

The following formulae are commonly used to calculate the *t* value:

(i) To test the significance of the mean of a random sample

$$t = \frac{|\bar{X} - \mu|}{S \,|\, SE_{\bar{x}} \bar{X}}$$

where $\bar{X}$ = Mean of the sample

μ = Mean of the universe

$SE_{\bar{x}}$ = S.E. of mean in case of small sample and is worked out as follows:

$$SE_{\bar{x}} = \frac{\sigma_s}{\sqrt{n}} = \frac{\sqrt{\dfrac{\Sigma(x_i - \bar{x})^2}{\sqrt{n}}}}{\sqrt{n}}$$

and the degrees of freedom = $(n - 1)$

NOTES

The above stated formula for *t* can as well be stated as under:

$$t = \frac{|\bar{x} - \mu|}{SE_{\bar{x}}}$$

$$= \frac{|\bar{x} - \mu|}{\frac{\sqrt{\frac{\Sigma(x - \bar{x})^2}{n-1}}}{\sqrt{n}}}$$

$$= \frac{|\bar{x} - \mu|}{\sqrt{\frac{\Sigma(x - \bar{x})^2}{n-1}}} \times \sqrt{n}$$

If we want to work out the probable or fiducial limits of population mean (μ) in case of small samples, we can use either of the following:

(a) Probable limits with 95% confidence level,

$$\mu = \bar{X} \pm SE_{\bar{x}}\ (t_{0.05})$$

(b) Probable limits with 99% confidence level,

$$\mu = \bar{X} \pm SE_{\bar{x}}\ (t_{0.01})$$

At other confidence levels, the limits can be worked out in a similar manner, taking the concerning table value of *t* just as we have taken $t_{0.05}$ in (*i*) and $t_{0.01}$ in (*ii*) above.

(ii) To test the difference between the means of the two samples

$$t = \frac{|\bar{X}_1 - \bar{X}_2|}{\text{SE}_{\bar{x}_1 - \bar{x}_2}}$$

where $\bar{X}_1$ = Mean of the sample 1

$\bar{X}_2$ = Mean of the sample 2

$SE_{\bar{x}_1 - \bar{x}_2}$ = Standard Error of difference between two sample means and is worked out as follows,

$$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sum(X_{1i} - \bar{x}_1)^2 + \sum(X_{2i} - \bar{x}_2)^2}{n_1 + n_2 - 2}} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

and the degrees of freedom = $(n_1 + n_2 - 2)$

When the actual means are in fraction, then use of assumed means is convenient. In such a case, the standard deviation of difference, i.e.,

$$\sqrt{\frac{\Sigma(x_{1i} + x_1)^2 + \Sigma(x_{2i} - \bar{x}_2)^2}{n_1 + n_2 - 2}}$$

can be worked out by the following short-cut formula:

$$= \frac{\sqrt{\Sigma(x_{1i} - A_1)^2} + \Sigma(x_{2i} - A_1)^2 - n_1(x_{1i} - A_2)^2 - n_2(x_{2i} - A_2)^2}{n_1 + n_2 - 2}$$

NOTES

where A_1 = Assumed mean of sample 1

A_2 = Assumed mean of sample 2

X_1 = True mean of sample 1

X_2 = True mean of sample 2

(iii) To test the significance of an observed correlation coefficient

$$t = \frac{r}{\sqrt{1-r^2}} \times \sqrt{n-2}$$

Here t is based on $(n-2)$ degrees of freedom.

(iv) In context of the 'difference test'

Difference test is applied in the case of paired data and in this context t is calculated as under:

$$t = \frac{\overline{x}_{Diff} - 0}{6_{Diff}\sqrt{n}} = \frac{\overline{x}_{Diff} - 0}{6_{Diff}}\sqrt{n}$$

where $\overline{X}_{Diff}$ or $\overline{D}$ = Mean of the differences of sample items.

0 = The value zero on the hypothesis that there is no difference

$\sigma_{\text{Diff.}}$ = Standard deviation of difference and is worked out as,

$$\sqrt{\frac{\sum D - \overline{X}_{Diff})^2}{(n-1)}}$$

or

$$\sqrt{\frac{\Sigma D^2 - (\overline{D})^2 n}{(n-1)}}$$

D = Differences

n = Number of pairs in two samples and is based on $(n-1)$ degrees of freedom.

The following examples would illustrate the application of t-test using the above stated formulae.

Example 4: A sample of 10 measurements of the diameter of a sphere gave a mean X = 4.38 inches ad a standard deviation, σ = 0.06 inches. Find (a) 95% and (b) 99% confidence limits for the actual diameter.

Solution: On the basis of the given data the standard error of mean

$$= \frac{\sigma_s}{\sqrt{n-1}} = \frac{0.06}{\sqrt{10-1}} = \frac{0.06}{3} = 0.02$$

Assuming the sample mean 4.38 inches to be the population mean, the required limits are as follows:

(a) 95% confidence limits $= \bar{X} \pm SE_{\bar{x}}(t_{0.05})$ with degrees of freedom

$= 4.38 \pm .02(2.262)$

$= 4.38 \pm .04524$

i.e., 4.335 to 4.425

(b) 99% confidence limits $= \bar{X} \pm SE_{\bar{x}}(t_{0.01})$ with 9 degrees of freedom

$= 4.38 \pm .02(3.25) = 4.38 \pm .0650$

i.e., 4.3150 to 4.4450.

Example 5: The specimen of copper wires drawn from a large lot have the following breaking strength (in kg. wt.):

578, 572, 570, 568, 572, 578, 570, 572, 596, 544

Tests whether the mean breaking strength of the lot may be taken to be 578 kg. wt.

Solution: We take the hypothesis that there is no difference between the mean height of the sample and the given height of universe. In other words we can write,

$H_0 : \mu = \bar{X},\ H_0 : \mu \neq \bar{X}$. Then on the basis of the sample data the mean and standard deviation has been worked out as under:

S. No.	*X*	$(X - \bar{X})$	$(X_1 - \bar{X})_2$
1	578	6	36
2	572	0	0
3	570	–2	4
4	568	–4	16
5	572	0	0
6	578	6	36
7	570	–2	4
8	572	0	0
9	596	24	576
10	544	–28	784
$n = 10$	$\sum X_i = 5720$		$\sum(X_i - \bar{X})^2 = 1456$

$$\bar{X} = \frac{\Sigma x}{n} = \frac{5720}{10}$$

$$= 572$$

$$\sigma_s = \sqrt{\frac{\Sigma(x - \bar{x}_s)^2}{n - 1}}$$

$$= \sqrt{\frac{1456}{10 - 1}} = \sqrt{\frac{1456}{9}}$$

$$= 12.72$$

$$SE_x = \frac{\sigma_s}{\sqrt{n}} = \frac{12.72}{\sqrt{10}}$$

$$= \frac{12.72}{3.16} = 4.03$$

$$t = \frac{|\bar{x} - \mu|}{SE_x} = \frac{|572 - 578|}{4.03}$$

$$= 1.488$$

Degrees of freedom = $n - 1 = 9$

At 5% level of significance for 9 degrees of freedom the table value of $t = 2.262$ for a two-tailed test.

The calculated value of t is less than its table value and hence the difference is insignificant. The mean breaking strength of the lot may be taken to be 578 Kg. wt. with 95% confidence level.

Example 6: Sample of sales in similar shops in two towns are taken for a new product with the following results:

	Mean sales	Variance	Size of sample
Town *A*	57	5.3	5
Town *B*	61	4.8	7

Is there any evidence of difference in sales in the two towns?

Solution: We take the hypothesis that there is no difference between the two sample means concerning sales in the two towns. In other words, $H_0 : \bar{X}_1 = \bar{X}_2,\ H_0 : \bar{X}_1 \neq \bar{X}_2$. Then we work out the concerning t value as follows:

$$t = \frac{|\bar{X}_1 - \bar{X}_2|}{SE_{x_1 - x_2}}$$

where $\bar{x}_1$ = Mean of the sample concerning Town *A*

$\bar{x}_2$ = Mean of the sample concerning Town *B*

$SE_{x_1 - x_2}$ = Standard Error of the difference between two means

$$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\Sigma(x_{1i} - \bar{x}_1)^2 + \Sigma(x_{2i} - \bar{x}_2)^2}{n_1 + n_2 - 2}} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Hence,

$$t = \frac{|57 - 61|}{1.421} = \frac{4}{1.421}$$

$$= 2.82$$

Degrees of freedom = $(n_1 + n_2 - 2) = (5 + 7 - 2) = 10$

Table value of t at 5% level of significance for 10 degrees of freedom is 2.228, for a two-tailed test.

The calculated value of t is greater than its table value. Hence the hypothesis is wrong and the difference is significant.

Example 7: The sales data of an item in six shops before and after a special promotional campaign are:

Shops	*A*	*B*	*C*	*D*	*E*	*F*
Before the promotional campaign	53	28	31	48	50	42
After the campaign	58	29	30	55	56	45

Can the campaign be judged to be a success? Test at 5% level of significance.

Solution: We take the hypothesis that the campaign does not bring any improvement in sales. We can thus write:

NOTES

In order to judge this, we apply the 'difference test'. For this purpose we calculate the mean and standard deviation of differences in two sample items as follows:

Shops	*Sales before campaign* X_{Bi}	*Sales after campaign* X_{Ai}	*Difference* = *D* (*i.e., increase or decrease after the campaign*)	$(D - D)$	$(D - D)^2$
A	53	58	+5	+1.5	2.25
B	28	29	+1	–2.5	6.25
C	31	30	–1	–4.5	20.25
D	48	55	+7	+3.5	12.25
E	50	56	+6	+2.5	6.25
F	42	45	+3	–0.5	0.25
$n = 6$			$\sum D = 21$		$\sum(D - D)^2$
					$= 47.50$

$$\text{Mean of difference or } \bar{X}_{Diff} = \frac{\Sigma D}{n} = \frac{21}{6} = 3.5$$

Standard deviation of difference,

$$\sigma_{Diff} = \sqrt{\frac{\Sigma(D - \bar{D})^2}{n - 1}} = \sqrt{\frac{47.50}{6 - 1}} = 3.08$$

$$t = \frac{\bar{X}_{Diff} - 0}{\sigma_{Diff}} = \sqrt{n}$$

$$= 1.14 \times 2.45 = 2.793$$

Degrees of freedom = $(n - 1) = (6 - 1) = 5$

Table value of *t* at 5% level of significance for 5 degrees of freedom = 2.015 for one-tailed test.

Since the calculated value of *t* is greater than its table value, the difference is significant. Thus the hypothesis is wrong and the special promotional campaign can be taken as a success.

Example 8: Memory capacity of 9 students was tested before and after training. From the following scores, state whether the training was effective or not.

Student	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
Before (X_{Bi})	10	15	9	3	7	12	16	17	4
After (X_{Ai})	12	17	8	5	6	11	18	20	3

Solution: We take the hypothesis that training was not effective. We can write, $H_0 : \bar{x}_A = \bar{X}_B,\ H_0 : \bar{X} > \bar{X}_B$. We apply the difference test for which purpose first of all we calculate the mean and standard deviation of difference as follows:

Students	*Before* X_{Bi}	*After* X_{Ai}	*Difference* = *D*	D^2
1	10	12	2	4
2	15	17	2	4
3	9	8	–1	1
4	3	5	2	4
5	7	6	–1	1

6	12	11	–1	1
7	16	18	2	4
8	17	20	3	9
9	4	3	–1	1
$n = 9$			$\sum D = 7$	$\sum D^2 = 29$

$$Đ = \frac{\Sigma D}{n} = \frac{7}{9} = 0.78$$

$$\sigma_{Diff} = \sqrt{\frac{\Sigma D^2 - (\bar{D})^2 n}{n-1}} = \sqrt{\frac{29-(0.78)^2 \times 9}{9-1}} = 1.71$$

$$\therefore\ t = \frac{0.78}{1-71} = 1.369$$

Degrees of freedom = $(n - 1) = (9 - 1) = 8$

Table value of t at 5% level of significance for 8 degrees of freedom

= 1.860 for one-tailed test.

Since the calculated value of t is less than its table value, the difference is insignificant and the hypothesis is true. Hence it can be inferred that the training was not effective.

Example 9: It was found that the coefficient of correlation between two variables calculated from a sample of 25 items was 0.37. Test the significance of r at 5% level with the help of t-test.

Solution: To test the significance of r through t-test, we use the following formula for calculating t value:

$$t = \frac{r}{\sqrt{1-r^2}} \times \sqrt{n-2}$$

$$= \frac{0.37}{1-(0.37)^2} \times \sqrt{25-2}$$

$$= 1.903$$

Degrees of freedom = $(n–2) = (25–2) = 23$

The table value of at 5% level of significance for 23 degrees of freedom is = 2.069 for a two-tailed test.

The calculated value of t is less than its table value, hence r *is* insignificant.

Example 10: A group of seven week old chickens reared on high protein diet weigh 12, 15, 11, 16, 14, 14 and 16 ounces; a second group of five chickens similarly treated except that they receive a low protein diet weigh 8, 10, 14, 10 and 13 ounces. Test at 5% level whether there is significant evidence that additional protein has increased the weight of chickens. (Use assumed mean (or A_1) = 10 for the sample of 7 and assumed mean (or A_2) = 8 for the sample of 5 chickens in your calculation).

Solution: We take the hypothesis that additional protein has not increased the weight of the chickens. We can write, $H_0: X_1 > X_2\ H_0: X_1 > X_2$.

Applying t-test we work out the value of t for measuring the significance of two sample means as follows:

$$t = \frac{X_1 - X_2}{SE_{x_1 - x_2}}$$

NOTES

Calculation can be done as under:

X_1	$(X_{i1} - A_1)$	$(X_{i1} - A_1)_2$	X_2	$(X_{i2} - A_2)$	$(X_{i2} - A_2)^2$
	$A_1 = 10$			$A_2 = 8$	
12	2	4	8	0	0
15	5	25	10	2	4
11	1	1	14	6	36
16	6	36	10	2	4
14	4	16	13	5	25
14	4	16			
16	6	36			
n_1=7	$\Sigma(X_{1i} - A_1)$ =28	$\Sigma(X_{1i} - A_1)^2$ =134	n_2=5	$\Sigma(X_{2i} - A_2)$ =15	$\Sigma(X_{2i} - A_2)^2$ =69

$$X_1 = A_1 + \Sigma(x_{1i} - A_1)/n_1$$

$$\therefore \quad = 10 + \frac{28}{7} = 14$$

Similarly, $$X_2 = A_1 + \frac{\Sigma(x_{2i} - A_2)}{n_2}$$

$$= 8 + \frac{15}{5} = 11$$

Hence

$$SE_{X_1 - X_2} = \sqrt{\frac{\Sigma(X_{1i} - A_1)^2 + \Sigma(X_{2i} - A_2)^2 - n_1(\bar{X}_1 - A_1)^2 - n_2(\bar{X}_2 - A_2)^2}{n_1 + n_2 - 2}} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$= \sqrt{\frac{134 + 69 - 7\,(14 - 10)^2 - 5(11 - 8)^2}{7 + 5 - 2}} \times \sqrt{\frac{1}{7} + \frac{1}{5}}$$

$$= (2.14)\,(0.59) = 1.2626$$

We now calculate the value under *t*,

$$t = \frac{X_1 - X_2}{SE_{X_1 - X_2}} = \frac{14 - 11}{1.2626} = 2.397$$

Degree of freedom = $(n_1 + n_2 - 2) = (7 + 5 - 2) = 10$

The table value of *t* at 5% level of significance for 10 degrees of freedom = 1.812 for one-tailed test.

The calculated value of *t* is higher than its table value and hence the difference is significant which means the hypothesis is wrong. It can therefore be concluded that additional protein has increased the weight of chickens.

Paired t-Test: Difference of Means

Let (x_1, y_1), $(x_2, y_2,)$, ...,(x_n, y_n) be the pairs of values for the same subjects, e.g., Sales data before (x) and after an advertisement campaign (y).

Performance of candidates before (X) and after training (y).

We have to test the significance of the difference between x, y values.

For each pair (x_{i-}, y_i) find $d_i = x_{i-} - y_i$

H_0: $\mu_1 = \mu_2$, i.e., no difference before and after and H_0: $\mu_1 \neq \mu_2$

We find the mean $\bar{d}$ of d values and use the statistic,

$$t = \frac{\bar{d}}{S/\sqrt{n}}$$

$$S = \sqrt{\frac{\sum(d - \bar{d})^2}{n-1}}$$

CHI-SQUARE (χ^2) TEST

In the test of independence, the row and column variables are independent of each other and this is the null hypothesis. The following are properties of the test for independence

- The data are the observed frequencies.
- The data is arranged into a contingency table.
- The degrees of freedom are the degrees of freedom for the row variable times the degrees of freedom for the column variable. It is not one less than the sample size, it is the product of the two degrees of freedom.
- It is always a right tail test.
- It has a chi-square distribution.
- The expected value is computed by taking the row total times the column total and dividing by the grand total.
- The value of the test statistic doesnot change if the order of the rows or columns are switched.
- The value of the test statistic doesnot change if the rows and columns are interchanged (Transpose of the matrix).

Contingency Tables

Suppose the frequencies in the data are classified according to attribute A into r classes (rows) and according to attribute B into c classes (columns) as show in Table 8.2.

Table 8.2 *Contingency Tables*

Class	$\boldsymbol{B_1}$	$\boldsymbol{B_2}$	**…**	$\boldsymbol{B_c}$	**Total**
A_1	O_{11}	O_{12}	…	O_{1c}	$\boldsymbol{(A_1)}$
A_2	O_{21}	O_{22}	…	O_{2c}	$\boldsymbol{(A_2)}$
…	…	…	…	…	**…**
A_r	O_{r1}	O_{r2}	…	O_{rc}	$\boldsymbol{(A_r)}$
Table	$\boldsymbol{(B_1)}$	$\boldsymbol{(B_2)}$	**…**	$\boldsymbol{(B_c)}$	$\boldsymbol{N}$

The totals of row and column frequencies are (A_i), (B_j).

To test if there is any relation between A, B we set up the null hypothesis of independence between A, B.

The expected frequency in any cell is calculated by using the formula:

$$E_{ij} = \frac{(A_i)(B_j)}{N}$$

Use $\chi^2 = \dfrac{-(O_{ij} - E_{ij})^2}{E_{ij}}$ with degrees of freedom = $(r - 1)(c - 1)$

For example,

Observed Frequencies

	School	College	University	Total
Boys	10	15	25	50
Girls	25	10	15	50
Total	35	25	40	100

$$\frac{35 \times 50}{100} = 17.5$$

$$\frac{25 \times 50}{100} = 12.5$$

$$\frac{40 \times 50}{100} = 20$$

Expected Frequencies

	School	College	University	Total
Boys	17.5	12.5	20	50
Girls	17.5	12.5	20	50
Total	35	25	40	100

Degrees of freedom = (2 – 1) (3 – 1) = 2

$$\chi^2 = \sum (O - E)^2 / E = 9.9$$

This is greater than the table value. It is not true that education does not depend on sex, i.e., the two are not independent.

4.8.1 Concept of Test Statistics

In the test for given population variance, the variance is the square of the standard deviation, whatever you say about a variance can be, for all practical purposes, extended to a population standard deviation.

To test the hypothesis that a sample $x_{1,}\ x_2, \ldots, x_n$ of size n has a specified variance $\sigma^2 = \sigma_2^2$

$$H_0\text{: } \sigma^2 = \sigma^2{}_0$$

or

Null hypothesis $\sigma = \sigma_0$

$$H_1 : \sigma^2 > \sigma_0^2$$

Test statistics $\chi^2 = \dfrac{ns^2}{\sigma_0^2} = \dfrac{\sum (x - \overline{x})^2}{\sigma_0^2}$

If χ^2 is greater than the table value we reject the null hypothesis.

F-TEST

An F-test is any statistical test in which if the null hypothesis is true, the test statistic has an F-distribution. A great variety of hypotheses in applied statistics are tested by F-tests. Among these are given below:

(a) The hypothesis that the means of multiple normally distributed populations, all having the same standard deviation, are equal. This is perhaps the most well-known of hypothesis tested by means of an F-test, and the simplest problem in the ANalysis Of Variance (ANOVA).

(b) The hypothesis that the standard deviations of two normally distributed populations are equal, and thus that they are of comparable origin.

If there are two independent random samples from normal populations we have to test the hypothesis that the population variances σ_1^2, σ_2^2 are the same

$$H_0 : \sigma_1^2 = \sigma_2^2 \text{ and } H_1 : \sigma_1^2 = \sigma_2^2$$

We find

$$S_1^2 = \frac{\sum (x_1 - \overline{x}_1)^2}{n_1 - 1}, \text{ estimate of } \sigma_1^2$$

$$S_2^2 = \frac{\sum (x_2 - \overline{x}_2)^2}{n_2 - 1}, \text{ estimate of } \sigma_2^2$$

To carry out the test of significance, find

$$F = \frac{S_1^2}{S_2^2} \text{ if } S_1^2 > S_2^2$$

or $$F = \frac{S_2^2}{S_1^2} \text{ if } S_2^2 > S_1^2$$

See the F tables for $n_1 - 1$, $n_2 - l$ degree of freedom ($n_1 - l$ corresponds to the numerator of F with the greater variance, $n_2 - 1$ for the denominator) at 5% level of significance.

If the observed F is less than the table value we assume the two populations have a common variance.

At 1% level of significance, the same procedure is to be followed.

NOTES

Example 11: One sample of 10 bulbs gives a S.D. of 9 hours of life and another sample of 11 bulbs gives a S.D. of 10 hours of life. Can you say the variances are different at 1% level of significance? (Note: here $S_1 = 9$, $S_2 = 10$)

Solution:

$$S_1^2 = \frac{n_1}{n_1 - 1} S_1^2 = \frac{10}{10 - 1} \times (9)^2 = 90$$

$$S_2^2 = \frac{n_2}{n_2 - 1} S_2^2 = \frac{11}{11 - 1} \times (10)^2 = 110$$

$$F = \frac{S_2^2}{S_1^2} = \frac{110}{90} = 1.22 < \text{table value}$$

(Table value $F_{9,10,00.1} = 4.94$)

We accept the null hypothesis. The population variances may not be different.

4.9.1 Test for Equality of Two Population Variances

If p_1, p_2 are proportions of some characteristic of two samples of sizes n_1, n_2 drawn from populations with proportions P_1, P_2 then we have H_0: $P_1 = P_2$ vs H_1:$P_1 \neq P_2$

Case (a): If H_0 is true then let $P_1 = P_2 = p$

Where p can be found from the data,

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

$$q = 1 - p$$

p is the mean of the two proportions,

$$SE(p_1 - p_2) = \sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$z = \frac{P_1 - P_2}{SE(p_1 - p_2)}$$

We write z ~ $N(0, 1)$

The usual rules for rejection or acceptance are applicable here.

Case (b): If it is assumed that the proportion under question is not the same in the two populations from which the samples are drawn and that p_1, p_2 are the true proportions, we write,

$$SE(p_1 - p_2) = \sqrt{\left(\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}\right)}$$

We can also write the confidence interval *for* $p_1 - p_2$

For 2 independent samples of sizes $n_1 - n_2$ selected from two binomial populations, the 100 (1 – *a*) % confidence limits *for* $p_1 - p_2$ are,

$$(p_1 - p_2) \pm z_{\alpha/2} \sqrt{\left(\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}\right)}$$

The 90%, confidence limits would be [with α = 0.1, 100 (1 – α) = 0.90]

$$(p_1 - p_2) \pm 1.645 \sqrt{\left(\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}\right)}$$

For example, out of 5000 interviewees, 2400 are in favour of a proposal; out of another set of 2000 interviewees 1200 are in favour. Is the difference significant?

$$p_1 = \frac{2400}{5000} = 0.48$$

$$p_2 = \frac{1200}{2000} = 0.6$$

$$n_1 = 5000$$

$$n_2 = 2000$$

$$SE = \sqrt{\left(\frac{.48 \times 52}{5000} + \frac{.6 \times .4}{2000}\right)} = 0.013 \text{ Case } (b)$$

$$z = \left|\frac{P_1 - P_2}{SE}\right|$$

$$= \frac{0.12}{0.013}$$

$$= 9.2 > 1$$

The difference is highly significant at 0.27% level.

LARGE SAMPLE TESTS

Test for Single Mean

We have to test the null hypothesis that the population mean has a specified value μ, i.e., H_0: $\bar{x} = \mu$. For large *n*, if H_0 is true then,

$z = \left|\dfrac{\bar{x} - \mu}{SE(\bar{x})}\right|$ is approximately nominal. The theoretical region for *z* depending on the desired level of significance can be found out.

Example 12: A factory produces items, each weighing 5 kg with variance 4. Can a random sample of size 900 with mean weight 4.45 kg. be justified as having been taken from this factory?

Solution:

$$n = 900$$

$$\bar{x} = 4.45$$

NOTES

$\mu = 5$

$\sigma = \sqrt{4} = 2$

$$z = \left|\frac{\bar{x}-\mu}{SE(\bar{x})}\right| = \left|\frac{\bar{x}-\mu}{\sigma/\sqrt{4}}\right| = \left|\frac{4.45-5}{2/30}\right| = 8.25$$

We have $z > 3$. The null hypothesis is rejected. The sample may not be regarded as originally from the factory at 0.27% level of significance (corresponding to 99.73% acceptance region).

Test for Difference of Means

Suppose two samples of sizes n_1, n_2 are drawn from populations having means μ_1, μ_2 and standard deviations σ_1, σ_2

To test the equality of means $\bar{x}_1, \bar{x}_2$ we write,

$H_0 : \mu_1 = \mu_2$

$H_0 : \mu_1 \neq \mu_2$

If we assume H_0 is true then,

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$$

Approximately normally distributed with mean -0, S.D.-I.

We write $z \sim N(0,1)$

As usual, if $|z| > 2$, we reject H_0 at 4.55% level of significance and so on.

Example 13: Two groups of sizes 121 and 81 are subjected to tests. Their means are found to be 84 and 81 and standard directions 10 and 12.

Solution:

Test for the significance of difference between the groups,

$\bar{x}_1 = 84$

$\bar{x}_2 = 81$

$n_1 = 121$

$n_2 = 81$

$\sigma_1 = 10$

$\sigma_2 = 12$

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$$

$$z = \frac{84-81}{\sqrt{\dfrac{100}{121} + \dfrac{144}{81}}}$$

$= 1.86 < 1.96$

The difference is not significant at the 5% level of confidence.

Check Your Progress

6. Define properties of the test for independence.
7. What is F-test?
8. What is statistical decision-making?

ANALYSIS OF VARIANCE

In business decisions, we are often involved in determining if there are significant differences among various sample means, from which conclusions can be drawn about the differences among various population means. For example, we may be interested to find out if there are any significant differences in the average sales figures of 4 different salesman employed by the same company, or we may be interested to find out if the average monthly expenditures of a family of 4 in 5 different localities are similar or not, or the telephone company may be interested in checking, whether there are any significant differences in the average number of requests for information received in a given day among the 5 areas of City (Under Study), and so on. The methodology used for such types of determinations is known as ANalysis Of VAriance or ANOVA. This technique is one of the most powerful techniques in statistical analysis and was developed by R.A. Fisher. It is also called the *F*-Test.

There are two types of classifications involved in the analysis of variance. The one-way analysis of variance refers to the situations when only one fact or variable is considered. For example, in testing for differences in sales for three salesman, we are considering only one factor, which is the salesman's selling ability. In the second type of classification, the response variable of interest may be affected by more than one factor. For example, the sales may be affected not only by the salesman's selling ability, but also by the price charged or the extent of advertising in a given area.

The Basic Principle of ANOVA

The basic principle of ANOVAis to test for differences among the means of the populations by examining the amount of variation within each of these samples, relative to the amount of variation between the samples. In terms of variation within the given population it is assumed that the values of (x_{ij}) differ from the mean of this population only because of random effects i.e., there are influences on (x_{ij}) which are unexplainable, whereas in examining differences between populations we assume that the difference between the mean of the *j*th population and the grand mean is attributable to what is called a 'specific factor' or what is technically described as treatment effect. Thus, while using ANOVA, we assume that each of the samples is drawn from a normal population and that each of these populations has the same variance. We also assume that all factors other than the one or more being tested are effectively controlled. This, in other words, means that we assume the absence of many factors that might affect our conclusions concerning the factor(s) to be studied.

One-Way Classification

Under the one-way ANOVA, we consider only one factor and then observe that the reason for the said factor to be important is that several possible types of samples can occur within that factor. We then determine if there are differences within that factor. The technique involves the following steps:

(a) Obtain the mean of each sample ie, obtain $\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_K$, i.e., when there are *K* samples.

(b) Work out the mean of the sample means as follows:

$$\bar{\bar{x}} = \left(\frac{\bar{x}_1 + \bar{x}_2 + \bar{x}_3 + \cdots + \bar{x}_K}{K}\right)$$

where K = Number of samples.

(c) Take the deviations of the sample means from the mean of the sample means and calculate the square of such deviations which may be multiplied by the no. of items in the corresponding sample, and then obtain their total. This is known as the sum of squares for variance between the samples (or SS between).

Symbolically, this can be written as:

$$SS\text{ between} = \left[n_1\left(\bar{x}_1 - \bar{\bar{x}}\right)^2 + n_2\left(\bar{x}_2 - \bar{\bar{x}}\right)^2 + \cdots + n_k\left(\bar{x}_k - \bar{\bar{x}}\right)^2\right]$$

(d) Divide the result of the Step (c) by the (no. of) degrees of freedom between the samples to obtain variance or Mean Square (MS) between samples.

Symbolically, this can be written as:

$$MS\text{ between} = \left[\frac{SS\text{ between}}{(k-1)}\right]$$

where $(k - 1)$ represents degrees of freedom (d.f.) between samples.

(e) Obtain the deviations of the values of the sample items for all the samples from corresponding means of the samples and calculate the squares of such deviations and then obtain their total. This total is known as the sum of squares for variance within samples (or *SS* within). Symbolically, this can be written as:

$$SS\text{ within} = \left[\Sigma\left(x_{1i} - \bar{x}_1\right)^2 + \Sigma\left(x_{2i} - \bar{x}_2\right)^2 + \cdots + \Sigma\left(x_{ki} - \bar{x}_k\right)^2\right]$$

where $i = 1, 2, 3, ...$

(f) Divide the result of Step (e) by the degrees of freedom within samples to obtain the variance or Mean Square (*MS*) within samples. Symbolically, this can be written as:

$$MS\text{ within} = \left[\frac{SS\text{ within}}{(n-k)}\right]$$

where $(n - k)$ represents the degrees of freedom within samples, (n = total no. of items in all the samples i.e., $(n_1 + n_2 + ... n_k)$ and k = no. of samples.

(g) For a check, the sum of squares of deviations for total variance can also be worked out by adding the squares of deviations when the deviations for the individual items in all the samples have been taken from the mean of the sample means. Symbolically, this can be written as:

$$SS\text{ for total variance} = \Sigma\left(x_{ij} - \bar{\bar{x}}\right)^2$$

$$i = 1, 2, 3, \cdots$$

$$j = 1, 2, 3, \cdots$$

This should be equal to the total of the results of Step (c) and Step (e) explained before i.e., *SS* for total variance = *SS* between + *SS* within

The degrees of freedom for total variance will be equal to the no. of items in all samples minus unity i.e., $(n - 1)$. The degrees of freedom for 'between' and 'within' must add up to the degrees of freedom for total variance, i.e.,

$$(n - 1) = [(k - 1) + (n - k)]$$

This fact explains the additive property of the ANOVA technique.

(h) Finally, *F*-ratio may be worked out as under:

$$F\text{-ratio} = \frac{MS \text{ between}}{MS \text{ within}}$$

This ratio is used to judge whether the difference among several sample means is significant or is just a matter of sampling fluctuations. For this purpose we look into the table giving the values of *F* for given degrees of freedom at different levels of significance. If the worked out value of *F*, as stated above, is less than the table value of *F*, the difference is taken as insignificant, i.e., due to chance and the null hypothesis of no difference between sample means stands. In case the calculated value of *F* happens to be either equal or more than its table value, the difference is considered as significant (which means the samples could not have come from the same universe) and accordingly the conclusion may be drawn. The higher the calculated value of *F* is above the table value, the more definite and sure one can be about his conclusions/inferences.

For the sake of convenience the information obtained through various steps stated above can be put as under:

Analysis of Variance Table for One-Way ANOVA
(There are K samples having in all *n* items)

Source of Variation	*Sum of squares (SS)*	*Degrees of freedom (d.f)*	*Mean Square (MS)* $\left(MS = \frac{SS}{d.f.}\right)$	*F-ratio*
Between Samples or Categories	$\left[n_1\left(\bar{x}_1 - \bar{\bar{x}}\right)^2 + \ldots\right.$ $\left.\ldots + n_k\left(\bar{x}_k - \bar{\bar{x}}\right)^2\right]$ $(= SSC)$	$(k - 1)$	$\frac{SS \text{ between}}{(k-1)}$ i.e., $MSC = \left\{\frac{SSC}{(k\ -}\right\}$	$\left(\frac{MS \text{ between}}{MS \text{ within}}\right)$
Within Samples or Categories	$\left[\Sigma\left(x_{1i} - \bar{x}_1\right)^2 + \ldots\right.$ $\left.\ldots + \Sigma\left(x_{ki} - \bar{x}_k\right)^2\right]$ $(= SSR)$ $(i = 1, 2, 3, \cdots)$	$(n - k)$	$\frac{SS \text{ within}}{(n-k)}$ i.e., $MSR = \frac{SSR}{(n-k)}$	i.e., $\frac{MSC}{MSR}$
Total	$\Sigma\left(x_{ij} - \bar{\bar{x}}\right)^2$ $i = 1, 2, \cdots$ $(= SST)\ j = 1, 2, \ldots$	$(n - 1)$		

Short-Cut Method for One-Way ANOVA

ANOVA can be performed by following the short-cut method which is usually used in practice since the same happens to be a very convenient method, particularly when means of the samples and/or mean of the sample means happen to be non-integer values. The various steps involved in the short-cut method are as under:

(a) Take the total of the values of individual items in all the samples, i.e., work out Sx_{ij} (where $i = 1, 2, 3, \cdots$ and $j = 1, 2, 3, \cdots$) and call it as T.

(b) Work out the correction factor as under:

$$\text{Correction factor} = \frac{T^2}{n}$$

(c) Find out the squares of all the item values one by one and then take their total. Subtract the correction factor from this total and the result is the sum of squares for total variance. Symbolically, this can be written as:

$$SS \text{ total, } SST = \left[\Sigma x_{ij}^2 - \frac{T^2}{n}\right]$$

where $i = \quad 1, 2, 3, \cdots$

and $j = 1, 2, 3, \cdots$

(d) Obtain the square of each sample total $\left(T_{ij}^2\right)$ and divide such square value by each sample by the number of items in the concerning sample and take the total of the result thus obtained. Subtract the correction factor from this total and the result is the sum of squares for variance between the samples. Symbolically, we can write:

$$SS \text{ between, } SSC = \left[\Sigma \frac{Tj^2}{nj} - \frac{T^2}{n}\right]$$

$(j = 1, 2, 3, \cdots)$

where subscript j represents different samples or categories.

(e) The sum of squares within the samples can be found out by subtracting the result of Step (d) from the result of Step (c) stated above and can be written as under:

$$SS \text{ within, } SSC = \left[\left\{\Sigma x_{ij}^2 - \frac{T^2}{n}\right\} - \left\{\Sigma \frac{Tj^2}{nj} - \frac{T^2}{n}\right\}\right]$$

$$= \left[\Sigma x_{ij}^2 - \Sigma \frac{Tj^2}{nj}\right]$$

After doing all this, the ANOVA table can be set up in the same way as explained earlier.

4.3.2 Two-Way Classification

Two-way ANOVA technique is used when the data are classified on the basis of two factors. For example, the agricultural output may be classified on the basis of different varieties of seeds and also on the basis of different varieties of fertilizers used. A business firm may have its sales data classified on the basis of different salesmen and also on the basis of sales in different regions. In a factory, the various units of a product produced during a certain period may be classified on the basis of different varieties of machines used and also on the basis of different grades of labour. Such a two-way design may have repeated measurements of each factor or may not have repeated values. The ANOVA technique is little different in case of repeated measurements where we also compute the interaction variation. We shall now explain the two-way ANOVA technique in the context of both said designs with the help of examples.

NOTES

ANOVA Technique in Context of Two-Way design when Repeated Values are not There

As we do not have repeated values, we cannot directly compute the sum of squares within samples as we had done in the case of one-way ANOVA. Therefore, we have to calculate this residual or error variation by subtraction, once we have calculated (just on the same lines as we did in the case of one-way ANOVA) the sum of squares for total variance and for variance between varieties of one treatment as also for variance between varieties of the other treatment.

The various steps involved are as follows:

1. Use the coding device, if the same simplifies the task.
2. Taking the total of the values of individual items (or their coded values as the case may be) in all the samples and call it T.
3. Work out the correction factor as under:

   Correction factor = $\frac{T^2}{n}$.
4. Find out the squares of all the item values (or their coded values as the case may be) one by one and then take their total. Subtract the correction factor from this total to obtain the sum of squares of deviations for total variance. Symbolically, we can write it as:

   Sum of squares of deviations for total variance or *SS* total = $\left[\Sigma x_{ij}^2 - \frac{T^2}{n}\right]$.
5. Take the total of different columns and then obtain the square of each column total and divide such squared values of each column by the no. of items in the concerning column and take the total of the result thus obtained. Finally, subtract the correction factor from this total to obtain the sum of squares of deviations for variance between columns or (*SS* between columns)
6. Take the total of different rows and then obtain the square of each row total and divide such squared values of each row by the no. of items in the corresponding row and take the total of the result thus obtained. Finally, subtract the correction factor from this total to obtain the sum of squares of deviations for variance between rows (or *SS* between rows).
7. Sum of squares of deviations for residual or error variance can be worked out by subtracting the result of the sum of step (v) and step (vi) from the result of step (iv) stated above. In other words,

   [*SS* total – (*SS* between columns + *SS* between rows)]

   = *SS* for residual or error variance.

8. Degrees of freedom (*d·f*) can be worked out as under:

   d·f for total variance = (*cr* – 1)

   d·f for variance between columns = (*c* – 1)

   d·f for variance between rows = (*r* – 1)

   d·f for residual variance = (*c* – 1) (*r* – 1)

where c = Number of columns

and r = Number of rows.

(ix) ANOVA table can be set up in the usual fashion as shown below:

Analysis of Variance Table for Two-Way ANOVA

Source of Variation	*Sum of squares (SS)*	*Degrees of freedom (d·f)*	*Mean square (MS)*	*F-ratio of*
	$\left[\Sigma \frac{T_j^2}{nj} - \frac{T^2}{n}\right]$ (= *SSC*)	$(c-1)$	$\left\{\frac{SS \text{ between rows}}{(r-1)}\right\}$ i.e, $MS\,C = \frac{SSC}{(c-1)}$	$\left\{\frac{MS \text{ between columns}}{MS \text{ residual}}\right\}$ i.e., $\frac{MS\,C}{MS\,E}$
	$\left[\Sigma \frac{T_i^2}{n_i} - \frac{T^2}{n}\right]$ (= *SSR*)	$(r-1)$	$\left\{\frac{SS \text{ between rows}}{(r-1)}\right\}$ i.e., $MS\,R = \frac{SSR}{(r-1)}$	$\left\{\frac{MS \text{ between rows}}{MS \text{ residual}}\right\}$ i.e., $\frac{MS\,R}{MS\,E}$
	[*SS* total — *SS* between columns – SSB between rows] (= SSE)	$(c-1)$ $\times (r-1)$	$\frac{SS \text{ residual}}{(c-1)(r-1)}$ i.e., *MSE* $= \frac{SS\,E}{(c-1)(r-1)}$	
Total	$\left[\Sigma x_{ij}^2 - \frac{T^2}{n}\right]$ (= *SST*)	$(c \cdot r - 1)$		

In the table, c = Number of columns.

r = Number of rows.

SS residual = [*SS* total – (*SS* between columns + *SS* between rows)]

Thus, *MS* residual or the residual variance provides the basis for the *F*-ratios concerning variation between columns treatment and between rows treatment. *MS* residual is always due to the fluctuations of sampling, and hence serves as the basis for the significance test. Both the *F*-ratios are compared with their corresponding table values, for given degrees of freedom at a specified level of significance, as usual and if it is found that the calculated *F*-ratio concerning variation between columns is equal to or greater than its table value, then the difference among column means is considered significant. Similarly, the *F*-ratio concerning variation between rows can be interpreted.

ANOVA Technique in Context of Two-Way Design when Repeated Values are There

In case of a two-way design with repeated measurements for all the categories, we can obtain a separate independent measure of inherent or smallest variations. For this measure, we can calculate the sum of squares and degrees of freedom in the same way as we have worked out the sum of squares for variance within samples in the case of one-way ANOVA, *SS* total, *SS* between columns and *SS* between rows can also be worked out as stated above. We then find left-over sums of squares and left-over degrees of freedom which are used for what is known as '**interaction variation**'. (Interaction is the measure of inter-relationship among the two different classifications). After making all these computations, ANOVA table can be set up for drawing inferences.

ACTIVITY

1. A dice is rolled 5000 times. Of these, 2250 times it shows 3 or 4. Test the hypothesis that the die is unbiased.
2. Set up an ANOVA table for the following per acre production data for three kinds or varieties of wheat, each grown on 4 plots and state if the variety differences are significant.

Plot of land	*Per acre production data (variety of wheat)*		
	A	*B*	*C*
1	6	5	5
2	7	5	4
3	3	3	3
4	8	7	4

Did You Know

A working hypothesis is a hypothesis that is provisionally accepted as a basis for further research in the anticipation that a acceptable theory will be produced, even if the hypothesis ultimately fails. Like all hypotheses, a working hypothesis is constructed as a statement of expectations, which can be linked to the exploratory research purpose in empirical investigation and are often used as a conceptual framework in qualitative research.

Check Your Progress

9. Write the basic principle of ANOVA.
10. Define the term F-ratio.
11. When the two-way ANOVA technique is used?

SUMMARY

- A hypothesis is an approximate assumption that a researcher wants to test for its logical or empirical consequences. Hypothesis refers to a provisional idea whose merit needs evaluation, but having no specific meaning.
- Hypothesis might predict outcome of an experiment in a lab setting the observation of a phenomenon in nature. Thus, hypothesis is an explanation of a phenomenon proposal suggesting a possible correlation between multiple phenomena.

- While comparing two different methods in terms of their superiority, wherein the assumption is that both the methods are equally good is called null hypothesis. While comparing two different methods, regarding their superiority, wherein, stating a particular method to be good or bad as compared to the other one is called alternate hypothesis.
- The Type I and Type II errors cannot be reduced simultaneously. The only solution of this problem is to set an appropriate level by considering the costs and penalties attached to them or to strike a proper balance between both types of errors.
- The hypothesis is examined on a pre-determined level of significance. Generally either 5 per cent level or 1 per cent level of significance is adopted for the purpose. However, it can be stated here that the level of significance must be adequate keeping in view the purpose and nature of enquiry.
- Formulating a decision means either accepting null hypothesis and rejecting alternate hypothesis or rejecting null hypothesis and accepting alternate hypothesis.
- The Critical Region (CR), or Rejection Region (RR), is a set of values for testing statistic for which the null hypothesis is rejected in a hypothesis test. It means that the sample space for the test statistic is partitioned into two regions; one region as the critical region will lead us to reject the null hypothesis H_0, the other not.
- A two-tailed test rejects the null hypothesis if the sample mean is either more or less than the hypothesized value of the mean of the population. It is considered to be apt when null hypothesis is of some specific value whereas alternate hypothesis is not equal to the value of null hypothesis.
- The sampling distribution of many statistics for large samples is approximately normal. For small samples with $n < 30$, the normal distribution, as shown above, can be used only if the sample is from a normal population with known s.
- Sir William S. Gosset (pen name Student) developed a significance test and through it made significant contribution in the theory of sampling applicable in case of small samples. When population variance is not known, the test is commonly known as Student's *t*-test and is based on the *t* distribution.
- Like the normal distribution, *t* distribution is also symmetrical but happens to be flatter than the normal distribution. Moreover, there is a different *t* distribution for every possible sample size. As the sample size gets larger, the shape of the *t* distribution loses its flatness and becomes approximately equal to the normal distribution.
- In the test for given population variance, the variance is the square of the standard deviation, whatever you say about a variance can be, for all practical purposes, extended to a population standard deviation.
- An F-test is any statistical test in which if the null hypothesis is true, the test statistic has an F-distribution. A great variety of hypotheses in applied statistics are tested by F-tests.
- The hypothesis that the means of multiple normally distributed populations, all having the same standard deviation, are equal. This is perhaps the most well-known of hypotheses tested by means of an F-test, and the simplest problem in the ANalysis Of VAriance (ANOVA).
- Chi-square is a statistical test commonly used to compare observed data with data we would expect to obtain according to a specific hypothesis. In this, the

degrees of freedom are the degrees of freedom for the row variable times the degrees of freedom for the column variable. It is not one less than the sample size, it is the product of the two degrees of freedom.

- ANalysis Of VAriance or ANOVA technique is one of the most powerful techniques in statistical analysis and was developed by R.A. Fisher. It is also called the *F*-Test.
- The basic principle of ANOVA is to test for differences among the means of the populations by examining the amount of variation within each of these samples, relative to the amount of variation between the samples.

KEY TERMS

- **Two-tailed test:** The statistical test used in hypothesis testing and named after the 'tail' of data falling under the far left and far right of a bell-shaped normal distribution or normal bell curve
- **One-tailed test:** The statistical test used in hypothesis testing and named after the 'tail' of data falling either on the left or on the right side of a bell-shaped normal distribution or normal bell curve
- ***t*-test:** This test is commonly known as students *t*-test and is based on the *t*-distribution. This is also used for small samples when population variance is not known
- ***F*-test:** This test is generally known as variance ratio test and is mostly used in the context of analysis of variance, a technique developed by the famous statistician Prof. R. A. Fisher
- **F-ratio:** It is used to judge whether the difference among several sample means is significant or is just a matter of sampling fluctuations
- **Degrees of freedom:** It is used to estimate error for the interaction

ANSWERS TO 'CHECK YOUR PROGRESS'

1. Hypothesis is an assumption that is tested to find its logical or empirical consequence.
2. Type I Error: In this type of error, you may reject a null hypothesis when it is true. It means rejection of a hypothesis which should have been accepted. It is denoted by α (alpha) and is also known as alpha error.

   Type II Error: In this type of error, you are supposed to accept a null hypothesis when it is not true. It means accepting a hypothesis which should have been rejected. It is denoted by β (beta) and is also known as beta error.
3. Null Hypothesis: While comparing two different methods in terms of their superiority, wherein the assumption is that both the methods are equally good is called null hypothesis. It is also known as statistical hypothesis and is symbolized as H_0.

   Alternate Hypothesis: While comparing two different methods, regarding their superiority, wherein, stating a particular method to be good or bad as compared to the other one is called alternate hypothesis. It is symbolized as H_a.

4. The Critical Region (CR), or Rejection Region (RR) is a set of values for testing statistics for which the null hypothesis is rejected in a hypothesis test.
5. In this concept of hypothesis, you will formulate a rule provided both null hypothesis and alternate hypothesis are given. Formulating a decision means either accepting null hypothesis and rejecting alternate hypothesis or rejecting null hypothesis and accepting alternate hypothesis.
6. The following are the properties of the test for independence:
   - The data are the observed frequencies.
   - The data is arranged into a contingency table.
   - The degrees of freedom are the degrees of freedom for the row variable times the degrees of freedom for the column variable. It is not one less than the sample size, it is the product of the two degrees of freedom.
   - It is always a right tail test.
   - It has a chi-square distribution.
   - The expected value is computed by taking the row total times the column total and dividing by the grand total.
   - The value of the test statistic doesnot change if the order of the rows or columns are switched.
   - The value of the test statistic doesnot change if the rows and columns are interchanged (Transpose of the matrix).
7. An F-test is any statistical test in which if the null hypothesis is true, the test statistic has an F-distribution. A great variety of hypotheses in applied statistics are tested by F-tests.
8. Statistical decisions have to be made in the presence of uncertainty. In the testing of the hypothesis, the choice is between H_0 and H_1. In estimation, there are several choices available. The design of experiments requires one to choose between the nature and extent of observations.
9. The basic principle of ANOVA is to test for differences among the means of the populations by examining the amount of variation within each of these samples, relative to the amount of variation between the samples.
10. F-ratio is used to judge whether the difference among several sample means is significant or is just a matter of sampling fluctuations.
11. Two-way ANOVA technique is used when the data are classified on the basis of two factors.

QUESTIONS AND EXERCISES

Short-Answer Questions

1. What is hypothesis?
2. Explain the importance of statistical decision-making.
3. Define null and alternate hypothesis.
4. Describe the various types of errors that occur in statistical hypothesis.

5. Describe standard error.
6. What do you mean by the level of significance?
7. What is critical region?
8. Describe one-tailed test.
9. What is the importance of a two-tailed test in statistics?
10. Write the importance of a small sample test.
11. Explain t-test and Chi-square test.
12. Define F-test.
13. Differentiate between a small sample and a large sample.
14. Write the basic principle of ANOVA.
15. What are the two types of classification involved in the analysis of variance?
16. What do you mean by term F-ratio?
17. Write the steps involved in the short-cut method for one-way ANOVA.
18. Why the two-way ANOVA technique is used?

Long-Answer Questions

1. The normal rate of infection for a certain disease in cattle is known to be 50%. In an experiment with seven animals injected with a new vaccine it was found that none of the animals caught infection. Can the evidence be regarded as conclusive (at 1% level of significance) to prove the value of the new vaccine?
2. Is a particular significance level important?
3. Adice is rolled 9000 times. Of these, 3220 times it shows 3 or 4. Test the hypothesis that the die is unbiased.
4. A coin is tossed 200 times. It shows heads 116 times. Can you say the coin is biased?
5. A die was rolled 400 times. It showed a 6 coming up 80 times. Can you say the die is unbiased? p = 80/400=1/5 P=1/6
6. A company supplied 500 units of an item. The number of defectives was found to be 42 as against the company's conviction that 6 per cent items could be defective. Examine the tenability of the company claim.
7. The average score of two groups A B were found to be 25 and 22 with S.D. 4 and 5.5 respectively. Test for the equality of the two group scores. Given
$$n_1 = n_2 = 400.$$
8. 800 ore pieces from a mine were found to contain an average of 74.5 gm of gold. From a nearby mine, 1600 pieces had 75 gm gold. Test the equality of the averages from the two mines, each having an S.D. of 2.4.
9. To test the goodness of a coin, it is tossed 5 times. It is considered a bad coin if more than 4 heads show up. (a) What is the probability of Type I error? (b) If the probability of a head is 0.2, what is the probability of Type II error?
10. In a sample of 500 people, 280 are tea drinkers and the rest coffee drinkers. Are tea and coffee equally popular?

11. 10 persons randomly selected are found to have heights 63, 63, 66, 67, 68, 69, 70, 71, 71, 71 inches. Discuss the suggestion that the mean height in the population is 66 inches.
12. 360 persons out of 600 are found to suffer from pollution induced bronchitis in one city. In another, 400 out of 500 are found to suffer from bronchitis. Is there any significant difference in the incidence of bronchitis?
13. In two large populations there are 30 per cent and 25 per cent smokers. Is this difference likely to be hidden in samples of 1200 and 900 from the populations?
14. A random sample of size 20 from a normal population gives a sample mean of 42 and a sample standard deviation of 6. Test the hypothesis that the population standard deviation is 9 at 5% level of significance.
15. Explain the test to determine the differences within the factor under the one-way ANOVA.
16. Discuss the various steps involved in the short-cut method for one-way ANOVA.
17. Explain the two-way ANOVA technique in the context of repeated and non-repeated values designs.
18. Describe the significance of randomized designs with the help of examples.
19. Two random samples were drawn from two normal populations and their values are:

A	66	67	75	76	82	84	88	90	92		
B	64	66	74	78	82	85	87	92	93	95	97

Test whether the two populations have the same variance at the 5% level of significance. *F* = 3.36 at 5% level for $n_1 = 10$ and $n_2 = 8$.

20. A manufacturing company has purchased three new machines of different makes and wishes to determine whether one of them is faster than the others in producing a certain output. Five hourly production figures are observed at random from each machine and the results are given below:

Observations	A_1	A_2	A_3
1	25	31	24
2	30	39	30
3	36	38	28
4	38	42	25
5	31	35	28

Use ANOVA and determine whether the machines are significantly different in their mean speed. (Given at 5% level, $F_{2,12} = 3.89$)

FURTHER READING

Best, John W. and James V. Kahn. 2005. *Research in Education*, 10th edition. New Jersey: Pearson Education.

Butcher, Harold John. 1966. *Sampling in Educational Research*, 3rd edition. United Kingdom: Manchester University Press.

Edwards, Allen Louis. 2006. *Experimental Design in Psychological Research,* 3rd edition. United States: The University of Michigan.

Garrett, Henry Edward. 1926. *Statistics in Psychology and Education,* New Jersey: Longmans, Green and Company.

Guilford, Joy Paul. 1977. *Fundamental Statistics in Psychology and Education,* 6th edition. New York: McGraw Hill.

Kerlinger, Fred Nichols and Howard Bing Lee. 2000. *Foundations of Behavioral Research,* 4th edition. United States: Harcourt College Publishers.

NOTES

INSTITUTE OF DISTANCE EDUCATION IDE Rajiv Gandhi University

Institute of Distance Education

Rajiv Gandhi University

A Central University

Rono Hills, Arunachal Pradesh

Contact us:

☎ +91-98638 68890 Ide Rgu Ide Rgu ✉ helpdesk.ide@rgu.ac.in