بحتة 10 (إجصاء)

طلاب الفرقة الثالثة عام – شعبة انجليزي

كلية التربية

الفصل الدراسي الأول 2022-2023

3 MATH

Statistics (Pure 10)

Prepared by

Kena Math . Department

Chapter 1

Random Variable In mathematical sense, a random variable (r.v.) is a real valued function $\{f(X)\}$ defined over a specified range or over a sample space.[1]

Note: Random variable is elaborately and more specifically discussed in Chapter 6.

Continuous Random Variable A random variable which can take on a continuum of values is called a continuous r.v. In this case, the values are taken on a line within the specified range. For instance height, weight etc.

Discrete Random Variable A random variable which can take a finite or denumerable number of values e.g. the number of students in a class, the number of spots obtained in a throw of die etc.

Frequence Number of times a variate value is repeated is called frequency of the variate value e.g. suppose there are seven girl students who have secured 54 marks, 7 is the frequency of 54 marks. If there are 12 people with monthly income of Rs. 500-700, 12 is the frequency of the income group 500-700.

Now some commonly used diagrams, charts and graphs are described here with the aid of actual data.

Frequency Array If the individual items or values of a variable are given along with their corresponding frequencies, it is called a frequency array.

Example 2.3. The wages per month and the number of persons in a small scale industry are presented below.

Wages per month (Rs.)	350	490	600	780	800	1000
No. of persons	4	5	7	8	4	2

Such a presentation of data is called frequency array.

Frequency Distribution The premise of data in the form of frequency distribution describes the basic pattern which the data assumes in the mass. Frequency distribution gives a better picture of the pattern of data if the number of items is large enough.

From a frequency array, it is not possible to compare characteristics of different groups. Hence for this, the classes are established to make the series of data more compact and understandable. Class limits can sometimes arbitrarily or by Sturge's formula be delimited. The width of a class that is the difference between the upper and the lower limit of the class is termed *class interval*. Once the classes are formed, the frequencies for these classes from raw data are expedited with the help of tally marks, little slanting vertical strokes. A bunch of four tally marks is crossed by the fifth to make the counting simpler.

احزلابناء

Example 2.4. In a survey, the age of 52 women at marriage by eight parishes was reported as given below.

[1] The totality of outcomes of a random experiment is called sample space.

24,	25,	27,	26,	22,	23,	24,	25,	24,	25,	24,	23,	26,
28,	24,	25,	23,	24,	25,	25,	24,	25,	25,	22,	27,	28,
27,	26,	25,	24,	25,	28,	26,	25,	27,	25,	24,	27,	24,
25,	25,	24,	25,	24,	26,	27,	25,	27,	26,	25,	28,	26

The data can be presented in the form of frequency distribution with the help of tally marks.

Age in years (i)	Tally marks (ii)	No. of women (iii)
22	𝍷𝍷	2
23	𝍷𝍷𝍷	3
24	𝍸 𝍸 𝍷𝍷	12
25	𝍸 𝍸 𝍸 𝍷𝍷	17
26	𝍸 𝍷𝍷	7
27	𝍸 𝍷𝍷	7
28	𝍷𝍷𝍷𝍷	4
29		0

The distribution constituted by columns (i) and (iii) in the above table is known as frequency distribution. It gives the number of women according to their age at marriage i.e. two women were married at the age of 22 years, three at the age of 23 years, twelve at the age of 24 years and so on.

The frequency distribution has helped to arrange the haphazard data in a systematic manner which is easy to handle for further treatment.

Example 2.5. The birth weights (kilogram) of 30 children were recorded as follows:

2.0,	2.1,	2.3,	3.0,	3.1,	2.7,	2.8,	3.5,	3.1,	3.7,
4.0,	2.3,	3.5,	4.2,	3.7,	3.2,	2.7,	2.5,	2.7,	3.8,
3.1,	3.0,	2.6,	2.8,	2.9,	3.5,	4.1,	3.9,	2.8,	2.2

Frequency distribution can be formed in the manner described so far, using various class intervals. The width of the classes and the number of classes will be found out by Sturge's formula (2.1). The range of data is 2.0 to 4.2 i.e.

$$L = 4.2, S = 2.0.$$

The class interval

$$i = \frac{4.2 - 2.0}{1 + 3.322 \log_{10} 30} = \frac{2.2}{1 + 3.322 \times 1.4771} = \frac{2.2}{5.91} = 0.37 \doteq 0.4$$

and

$$K = 5.91 = 6.$$

Hence six classes with a width of 0.4 kg are to be taken in the frequency distribution. The distribution with the help of tally marks is,

Classes (weight in kg)	Tally marks	No. of children (Frequency)
2.0-2.4	IIII	5
2.4-2.8	IIII	5
2.8-3.2	IIII IIII	9
3.2-3.6	IIII	4
3.6-4.0	IIII	4
4.0-4.4	III	3

Notes: 1 The lower limit of a class is included in that class.

2. It is not necessary to choose the smallest value as the lower limit of the lowest class or the largest value as upper limit of the highest class. One may choose the classes as 1.0-1.4, 1.4-1.8 and so on.

Smoothening of a Grouped Distribution. In case the classes do not constitute the continuous distribution, i.e. the upper limit of the previous class is not the lower limit of the following class, it has to be made continuous. The simple way to do this is to find the difference of the upper limit of the preceding class and lower limit of the following class. Subtract half of the difference from the lower limit of each class and add the same to its upper limit. Continue this process for all the classes.

Example 2.6. The table below gives the distribution of the age of women at the time of marriage in Sri Lanka.

Age groups (years)	No. of women
15-19	11
20-24	36
25-29	28
30-34	13
35-39	7
40-44	3
44-49	2

The given distribution is not continuous as the upper limit of the preceding class is not the lower limit of the following class. Hence it is smoothened. The difference between 20 and 19 is 1. Therefore, 0.5 is to be subtracted from the lower limit of the classes and 0.5 is to be added to the upper limit of all classes. Since the difference is constant, the same quantity is subtracted and added in all classes.

Thus, the smoothened frequency distribution will be:

Smoothened age groups (years)	No. of women
14.5-19.5	11
19.5-24.5	36
24.5-29.5	28
29.5-34.5	13
34.5-39.5	7
39.5-44.5	3
44.5-49.5	2

Open End Classes. An open end class is a class lacking one limit. Generally it is the lowest class lacking the lower limit and highest class lacking the upper limit. For instance, in an age group distribution, the lowest class is taken as less than five (< 5) and highest class as more than 70 (> 70). Open end classes make it possible to accommodate values which are at large gaps without increasing the number of consecutive classes. However open end classes should be avoided as far as possible. Open ends create problem in processes like computations and graphical representations.

Cumulative Frequency (cu. fr.) It is the number of observations less than (more than) or equal to a specified value.

Cumulative Frequency Distribution It can be formed on "less than" or "more than" basis. In example 2.5, we can either base the distribution on the number of children with a birth weight less than a particular weight or the number of children with a birth weight more than a specified weight. The cumulative frequency distributions for the data given in example 2.5 are presented below.

Table 2.2: Cumulative frequency and percentage distributions

Less than type		Percentage	More than type		Percentage
Birth weight	cu.fr.		Birth weight	cu.fr.	
Less than 2.4	5	17	2.0 or more	30	100
" " 2.8	10	33	2.4 " "	25	83
" " 3.2	19	63	2.8 " "	20	67
" " 3.6	23	77	3.2 " "	11	37
" " 4.0	27	90	3.6 " "	7	23
" " 4.4	30	100	4.0 " "	3	10

Note: It must be remembered that the cumulative frequency of a less than type frequency distribution always refers to the upper limit of the class interval and for more than type it refers to the lower limit of the class interval.

DIAGRAMMATIC REPRESENTATION OF DATA

Line and Bar Diagram Such diagrams are suitable for discrete variables i.e. for data given according to some periods, places and timings. These periods, places or timings are represented on the base line(X-axis) at regular intervals and the corresponding values or frequencies are represented on the Y-axis (ordinate). The lines or bars of height proportional to these values or frequencies, as per chosen scale, are erected at the points marked on the X-axis.

In a line diagram, the vertical lines are assumed to have no width, whereas in a bar diagram, the rectangles of certain width placed centrally at the points on abscissa are erected. The width of these bars should accommodate two bars apart on the same line. Hence, the width of a bar should be less than half the distance between any two points placed on the abscissa. Bars of equal width make the comparisons simple. For any comparison, the line diagram serves the same purpose as the bar diagram. The only advantage of the bar diagram is that they are more prominent and attractive. The bars are filled with dashes, dots or colours. The shapes of the two types of diagrams are shown through an example.

Example 2.7. Aggregated figures for merchandise export (f.o.b.) in India for eight years are as follows:

Years	1971	1972	1973	1974
Export (million Rs)	1962	2174	2419	3024
Years	1975	1976	1977	1978
Export (million Rs)	3852	4688	5355	5112

(a) Data for export are depicted through line diagram *as shown* in Fig. 2.1.

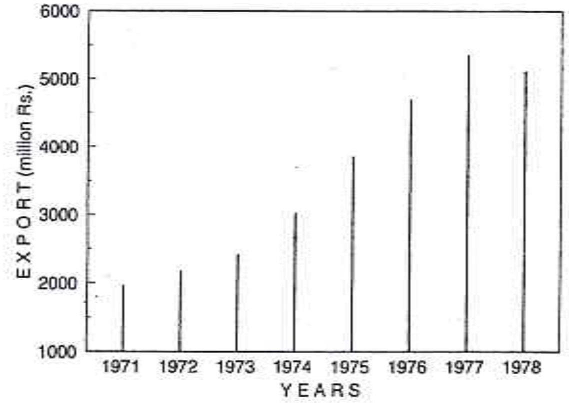


Fig. 2.1 : Line Diagram

(b) Data for export have been displayed through bar diagram in Fig. 2.2.

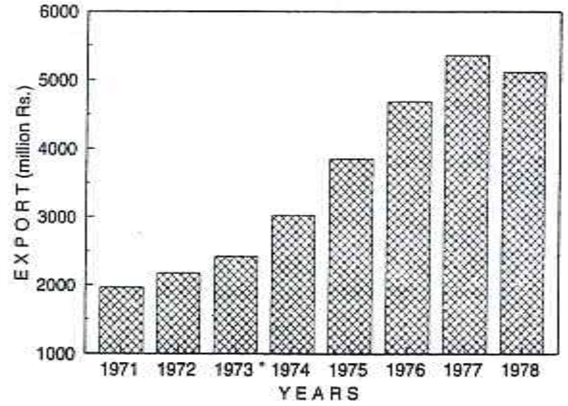


Fig. 2.2 : Bar Diagram

Histogram This type of diagrammatic representation is more suited for frequency distributions with continuous classes. In this type of distribution the upper limit of a class is the lower limit of the following class. The magnitudes of the class intervals are plotted along the abscissa and the frequencies along the ordinate according to the chosen scale. The rectangles are drawn on each class interval with height in proportion to its frequency. The number of such rectangles will be equal to the number of classes.

In case the class intervals are unequal, the area of the rectangle is considered for comparison rather than only the height of these rectangles.

A histogram for discrete frequency distribution can also be drawn by making an assumption. Here the frequency corresponding to a variate value is spread over the interval $(X - d/2)$ to $(X + d/2)$ where d is the difference from one value to the next higher (lower) value. It means we are considering an increasing (decreasing) series.

. Time series are depicted through a graph taking time on the X-axis and the variable under consideration on the Y-axis. Such a graph is called a *historigram*. It should not be confused with a histogram.

Example 2.8. Indian Cotton Mills Federation has revealed the following information about mill consumption of cotton from 1976 to 1982.

Years	Mill Consumption of Cotton ('000 bales of 170 kg. each)
1976-77	6,752
1977-78	6,616
1978-79	6,981
1979-80	7,412
1980-81	7,678
1981-82	7,035

The data about consumption of cotton can suitably be exhibited in the form of a historigram as given in Fig. 2.3.

Example 2.9. The smoothened distribution of age given in example 2.6 has been represented by a histogram. The frequency polygon has also been shown in the Fig. 2.4.

It is worth noting how the points of the first and the last rectangles are joined to the abscissa. The frequency at the mid-point of a class before the first class interval and after the last class interval is zero. Also, the area of the frequency polygon is equal to the area of the histogram in the case where class intervals are equal since the area of the histogram left out by the polygon is equal to the area encroached by the polygon outside the histogram.

Component Bar Diagram The bar diagram or chart shows an aggregate value whereas the component bar diagram gives the breakup in parts which

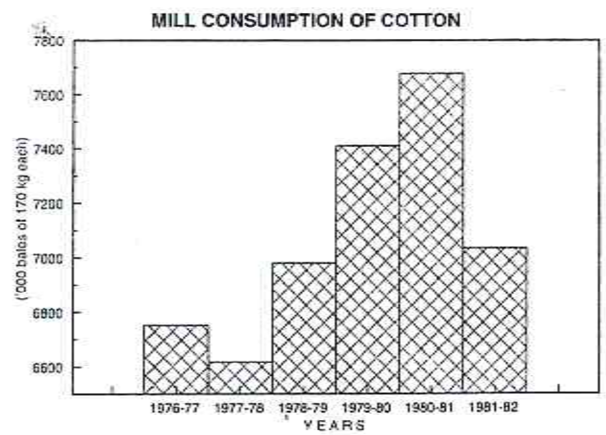


Fig. 2.3 : Historigram

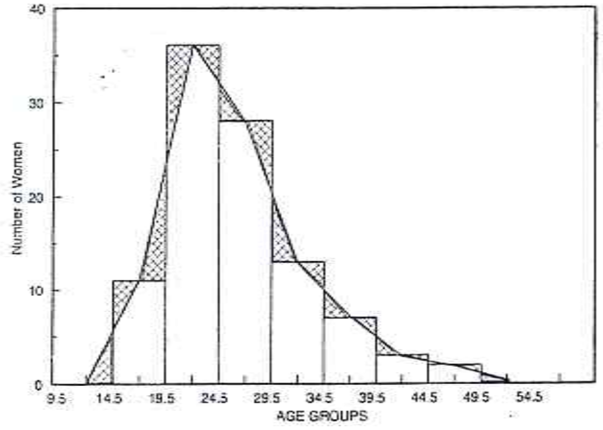


Fig. 2.4 : Histogram and Frequency Polygon

constitutes the aggregate in a year, place or sector. Such a chart makes it possible to compare the changes occurring in parts and in aggregates as well. In these types of diagrams a bar is further sub-divided into parts in proportion to the size of the sub-divisions. These sub divided rectangles are shaded differently by lines, dots and colours etc. Such charts are more informative than simple bar diagrams. Component bar diagrams are also called *subdivided bar diagrams*.

Example 2.10. The figures below give the revenue expenditure in Rajasthan on education in crore of rupees for the last four years.

Particulars	Expenditure (crore Rs.)			
	1979-80	1980-81	1981-82	1982-83
Primary education	69.7	77.9	118.2	134.7
Secondary education	36.6	44.6	73.3	87.4
Special education	2.0	2.2	3.5	4.2
University and higher education	14.1	15.7	22.9	25.1
Technical education	1.3	1.4	2.2	2.9
Sports and youth welfare	1.3	1.5	2.1	2.3
General	1.1	1.1	1.6	1.8
Total	126.1	144.4	223.8	258.4

The information about expenditure on education is displayed through sub-divided bar diagram (Fig. 2.5).

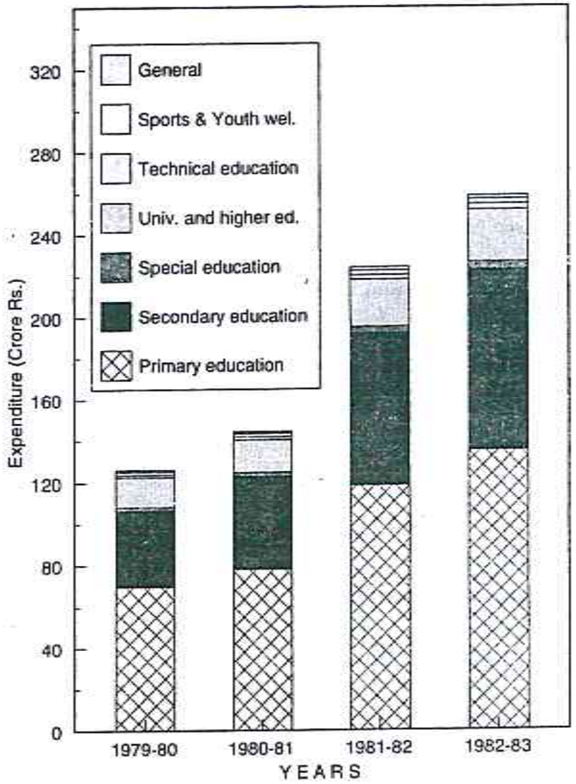


Fig. 2.5 : Subdivided Bar Diagram

The above diagram gives a clear picture of the trend of expenditure on various levels of education, and as a whole.

Pie Chart A pie chart is a circle divided into component sectors according to the break-up of components given in percentage. If each component is represented by a separate circle, large figures would need large circles. But the percentages remove this difficulty. Moreover in, a pie chart, only one circle of any size can beautifully represent all the components. We know that a circle represents an angle of 360° around the centre. So 360° angle is divided in proportion to the percentages. In the circle of a desired size, a radius, generally a horizontal line, is drawn and the calculated angles for various components are constructed one after another with the help of a protractor. Each sector is shaded differently by lines, dots or with different colours to look unique. A pie chart is good to represent the component breakup of a thing or commodity.

Example 2.11. The plan outlay of Rajasthan for the year 1983-84 is as tabulated below.

S.No.	Sector (i)	Budget estimates (crore Rs) (ii)	Percentage of total (iii)	Equivalent angles (iv)
1.	Agr. & allied services	66.3	15.4	55.5
2.	Cooperation	5.5	1.3	4.7
3.	Irrigation & Power	216.4	50.4	181.4
4.	Industries & Mining	18.8	4.4	15.8
5.	Transport & Communication	19.0	4.4	15.8
6.	Social Services	100.6	23.5	84.6
7.	Miscellaneous	2.4	0.6	2.2
	Total	429.0	100.0	360

Percentage of expenditure in different sectors is shown in column (iii) which is calculated as,

$$\text{For sector 1, Percentage} = \frac{66.3}{429} \times 100 = 15.4$$

$$\text{For sector 2, Percentage} = \frac{5.5}{429} \times 100 = 1.3$$

Similarly other percentages are calculated. Angles equivalent to percentages are shown in column (iv) of the above table and are calculated as,

$$\text{For sector 1, Angle} = \frac{15.4}{100} \times 360 = 55.5$$

$$\text{For sector 2, Angle} = \frac{1.3}{100} \times 360 = 4.7$$

The angles for other sectors have been calculated similarly. The Pie chart is drawn according to the method given in theory and displayed in Fig. 2.6.

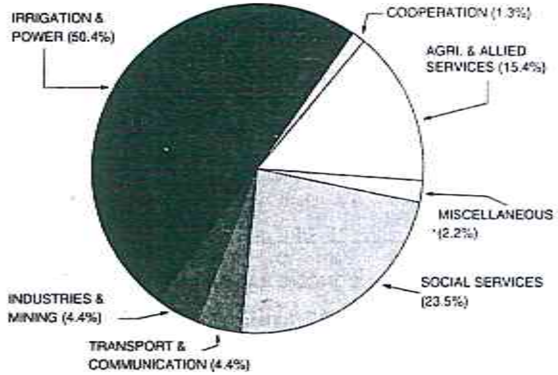


Fig. 2.6 : Pie Chart

Line Graphs In this type of graph, we have two variables under consideration. A variable is taken alone *X*-axis and the other along *Y*-axis. The variate values are suitably scaled along the axes and all distances are measured from the origin. If the smallest value in the bivariate data or frequency distribution is at a distance from zero, the origin is shifted suitably to values other than zero. The independent variable should be taken on *X*-axis and the dependent variable on *Y*-axis. The points are plotted and joined by line segments in order. These graphs depict the trend or variability occurring in the data. Sometimes two or more graphs are drawn on the same graph paper taking the same scale so that the plotted graphs are comparable.

Example 2.12. An experiment on S. cervi plants, regarding the uptake of methyl glucose at different concentrations of methyl glucose solution, was found to be as follows:

Concentration of methyl glucose (mM)	Methyl glucose uptake (μ mole/g/hr)
1.0	1.50
2.0	2.60
4.0	3.10
6.0	3.20
10.0	3.25
15.0	3.30
20.0	3.30

[*Source: Indian J. Expt. Biology*, 21(5) May, 1983]

A line graph for the given data have been drawn taking concentration on the *X*-axis and uptake on the *Y*-axis. The graph is shown in Fig. 2.7.

Chapter 2

Measures of Central Values

The collected data as such are not suitable to draw conclusions about the mass from which it has been taken. Some inferences about the population can be drawn from the frequency distribution of the observed values. This process of condensation of data reduces the bulk of data, and the frequency distribution is categorised by certain constraints known as parameters. Generally, a distribution is categorised by two parameters viz, the location parameter (central values) and the scale parameter (measures of dispersion). Hence in finding a central value, the data are condensed into a single value around which the largest number of values tend to cluster. Commonly, such a value lies in the centre of the distribution and is termed as central tendency.

R.A. Fisher has rightly said, "The inherent inability of the human mind to grasp entirely a large body of numerical data, compels us to seek relatively few constants that will adequately describe the data."

Two series of observations are not comparable because of the unsystematic variation generally present in the series, but the constants make it possible to compare the series easily. In this chapter we will continue the discussion of the measures of central values. There are three popular measures of central tendency namely, (i) mean, (ii) median, (iii) mode. Each of these will be discussed in detail here. Besides these, some other measures of location are also dealt with, such as quartiles, deciles and percentiles.

CHARACTERISTICS OF A GOOD MEASURE OF CENTRAL TENDENCY

There are various measures of central tendency. The difficulty lies in choosing the measure as no hard and fast rules have been made to select any one. However, some norms have been set which work as a guide line for choosing a particular measure of central tendency. A measure of central tendency is good or satisfactory if it possesses the following characteristics.

(1) It should be based on all the observations.

(2) It should not be affected by the extreme values.

(3) It should be as close to the maximum number of observed values as possible.

(4) It should be defined rigidly which means that it should have a definite value. The experimenter or investigator should have no discretion.

(5) It should not be subjected to complicated and tedious calculations, though the advent of electronic calculators and computers has made it possible to overlook this aspect.

(6) It should be capable of further algebraic treatment. By algebraic treatment we mean that these measures can be used further in the formulation of other formulae. For instance, a mean can be used to calculate the pooled mean of two or more series.

(7) It should be stable with regard to sampling. This means that if a number of samples of the same size are drawn from a population, the measure of central tendency having the minimum variation among the different calculated values should be preferred.

MEANS

There are three types of means which are suitable for a particular type of data. They are,

(a) Arithmetic mean or Average

(b) Geometric mean

(c) Harmonic mean.

Arithmetic Mean (A.M.) It is also popularly known as average. If mean is mentioned, it implies arithmetic mean, as the other means are identified by their full names. It is the most commonly used measure of central tendency.

Definition. Sum of the observed values of a set divided by the number of observations in the set is called a mean or an average.

If $X_1, X_2, \ldots \ldots, X_N$ are N observed values, the mean or average is given as,

$$\mu \text{ or } A = \frac{X_1 + X_2 + \ldots + X_n}{N} \qquad (3.1)$$

$$= \frac{1}{N} \Sigma_i X_i \qquad (3.1.1)$$

for $i = 1, 2, \cdots N$

Population[1] mean is usually denoted by μ or $\overline{X}$ whereas the sample[2] mean is denoted by $\bar{x}$(small letter).

When the data are arranged or given in the form of frequency distribution i.e. there are k variate values such that a value X_i has a frequency f_i ($i = 1, 2, ..., k$), the formula for the mean is,

$$\mu = \frac{f_1 X_1 + f_2 X_2 + \cdots + f_k X_k}{f_1 + f_2 + \cdots + f_k} \tag{3.2}$$

$$= \frac{\Sigma_i f_i X_i}{\Sigma_i f_i} \qquad i = 1, 2, ..., k \tag{3.2.1}$$

$$= \frac{1}{N} \Sigma_i f_i X_i \tag{3.2.2}$$

where $N = f_1 + f_2 + ... + f_k = \Sigma_i f_i$

If the data are given with k class intervals i.e. the data are in the form as follows:

Class interval	Frequency
X_1-X_2	f_1
X_2-X_3	f_2
X_3-X_4	f_3
.	.
.	.
.	.
X_k-X_{k+1}	f_k

The arithmetic mean

$$\mu = \frac{f_1 Y_1 + f_2 Y_2 + \cdots + f_k Y_k}{f_1 + f_2 + \cdots + f_k} \tag{3.3}$$

$$= \frac{1}{N} \Sigma_i f_i Y_i \tag{3.3.1}$$

where Y_i is the mid point of the class interval X_i-X_{i+1} and is given as,

$$Y_i = \frac{X_i + X_{i+1}}{2} \qquad i = 1, 2, ..., k$$

In this situation the values in the interval are considered to be centered at the mid-point of the interval.

Weighted Mean In case, k variate values $X_1, X_2, ..., X_k$ have known weights $\omega_1, \omega_2, \cdots, \omega_k$ respectively, then the weighted mean is,

$$\mu = \frac{\omega_1 X_1 + \omega_2 X_2 + \cdots + \omega_k X_k}{\omega_1 + \omega_2 + \cdots + \omega_k} \tag{3.4}$$

[1] See definition of population in Chapter 7.

[2] See definition of sample in Chapter 7.

$$= \frac{1}{\omega} \Sigma_i \omega_i X_i \tag{3.4.1}$$

where $\omega = \Sigma_i \omega_i$, $i = 1, 2, ..., k$

Weighted mean is commonly used in the construction of index numbers.

Note: If the sample values $x_1, x_2, ..., x_n$ are given, in the formulae given above, capital X will be changed to small x and N to n. The sample mean will be denoted by $\bar{x}$ e.g. the formula (3.1) will be changed to

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} \tag{3.5}$$

$$= \frac{1}{n} \Sigma_i x_i \tag{3.5.1}$$

$$i = 1, 2, ..., n$$

Similarly all other formulae will be changed to sample values. Since most of the studies are based on samples in practice, we use formulae for sample values.

Merits and Demerits

(1) Algebraic sum of the deviations of the given values from their arithmetic mean is always zero, i.e. $\Sigma_i (X_i - \overline{X}) = 0$.

(2) The sum of the squares of the deviations of the given values from their A.M. is minimum, i.e. $\Sigma_i (X_i - \overline{X})^2$ is minimum.

(3) An average possesses all the characteristics of a central value given earlier except No. 2, which is greatly affected by the extreme values.

(4) In case of grouped data if any class interval is open, arithmetic mean cannot be calculated, e.g. the classes are less than five in the beginning or more than 70 at the end of the distribution or both.

Example 3.1. Daily cash earnings of 15 workers working in different industries are as follows:

Average daily earning (Rs)

11.63,	8.22,	12.56,	12.14,	29.23,	18.23,	11.49,	11.30,
17.00,	9.16,	8.64,	27.56,	8.23,	19.77,	12.81	

Average daily earning of a worker can be calculated by the formula (3.1)

$$A = \frac{1}{15}(11.63 + 8.22 + \cdots + 12.81)$$

$$= \frac{217.97}{15}$$

$$= 14.53$$

The average daily earning of a worker is Rs. 14.53.

Example 3.2. The distribution of age at first marriage of 130 males was as given below.

Age in years (X):	18,	19,	20,	21,	22,	23,	24,	25,	26,	27,	28,	29.
No. of males (f):	2,	1,	4,	8,	10,	12,	17,	19,	18,	14,	13,	12.

The average age can be computed by the formula (3.2).

$$A = \frac{18 \times 2 + 19 \times 1 + \cdots + 29 \times 12}{2 + 1 + \cdots + 12}$$

$$= \frac{3240}{130}$$

$$= 24.92$$

The mean age of males at first marriage is 24.92 years.

Example 3.3. The distribution of the size of the holding of cultivated land, in an area, was as follows:

Size of holdings (hectares)	Mid points (y)	No. of holdings (f)
0-2	1	48
2-4	3	19
4-6	5	10
6-8	7	14
8-10	9	11
10-20	15	9
20-40	30	2
40-60	50	1

Average size of holding in the area can be calculated with the help of the formula (3.3). Mid points of the class intervals are shown in the middle column along with the data. Hence,

$$\text{A.M.} = \frac{1 \times 48 + 3 \times 19 + \cdots + 50 \times 1}{48 + 19 + \cdots + 1}$$

$$= \frac{597}{114}$$

$$= 5.237$$

The average size of holding is 5.237 hectares.

Example 3.4. The life of eighty condensers obtained in a life testing experiment has been presented below in the form of "less than" type of distribution.

Life of condensers (Years)	No. of condensers
Less than 1	3
" " 2	12
" " 3	14
" " 4	22
" " 5	33
" " 6	46
" " 7	58
" " 8	66
" " 9	75
" " 10	80

The given distribution with regular class intervals and their mid-values can be written as,

Years	Mid-values (y)	No. of condensers (f)
0-1	0.5	3
1-2	1.5	9
2-3	2.5	2
3-4	3.5	8
4-5	4.5	11
5-6	5.5	13
6-7	6.5	12
7-8	7.5	8
8-9	8.5	9
9-10	9.5	5

The average life of a condenser can be calculated by the formula (3.3) as,

$$\overline{X} = \frac{3 \times 0.5 + 9 \times 1.5 + \cdots + 5 \times 9.5}{3 + 9 + \cdots + 5}$$

$$= \frac{431}{80}$$

$$= 5.39 \text{ years}.$$

Example 3.5. The table below presents the total expenditure in the form of "more than" type frequency distribution. We known that the expenditure can not exceed Rs. 2250 crores.

Expenditure (Rs crores)	No. of Banks
More than 2000	1
" " 1750	2
" " 1500	4
" " 1250	7
" " 1000	13
" " 750	18
" " 500	28
" " 250	40

To find the average expenditure, we rewrite the given cumulative frequency distribution, with regular class intervals, as given below. Midvalues of the classes are also shown in the middle column.

Expenditure (Rs crores)	Mid-values	No. of Banks
2000-2250	2125	1
1750-2000	1875	2 – 1 = 1
1500-1750	1625	4 – 2 = 2
1250-1500	1375	7 – 4 = 3
1000-1250	1125	13 – 7 = 6
750-1000	875	18 – 13 = 5
500-750	625	28 – 18 = 10
250-500	375	40 – 28 = 12

On an average the expenditure per bank is,

$$\overline{X} = \frac{2125\times 1 + 1875 \times 1 + \cdots + 375 \times 12}{1 + 1 + \cdots + 12}$$

$$= \frac{33250}{40}$$

$$= 831.25 \text{ Rs. crores}$$

CODING OF DATA

A linear transformation of data may be regarded as coding. In coding we shift the origin and change the scale. A change can involve either a change of origin or a change of scale or a change of both, origin and scale together. The effect of coding on mean is given below.

1. If we subtract an arbitrary constant from each of the observation, the mean is also reduced by the constant value.
2. If we divide each observation of a set by an arbitrary constant, the mean is reduced as many times as the constant divisor.

Note: In case of addition or multiplication, the word 'reduced' should be replaced by 'increased' in the above statements. The above two operations cut short the calculation. But the availability of electronic calculators and computers has diminished the importance of coding of data. Anyhow, it can be used whenever needed.

Let $X_1, X_2, ..., X_N$ be N observations. An arbitrary constant a is subtracted from each of the observation and the reduced observation is divided by a constant c. Suppose the transformed observations are denoted by $X'_1, X'_2 \cdots ,X'_N$ where $X'_i = \frac{X_i - a}{c}$. The arithmetic mean of the original data with the help of coded observations is given as,

$$\overline{X} = a + \frac{\Sigma_i X'_i}{N} \times c \qquad (3.6)$$

for $i = 1, 2, ..., N$.

In case of frequency distribution with coding of data,

$$\overline{X} = a + \frac{\Sigma_i f_i X'_i}{N} \times c \qquad (3.7)$$

Where $N = \Sigma_i f_i$.

In the case of group data, generally the central mid-value of the classes is subtracted from each of the mid-value and the reduced observation is divided by the constant class interval. If the class interval is not same for all classes, any suitable value may be chosen as divisor.

Example 3.6. The production of pig iron and ferro-alloys in India from 1969 to 1975 is as given below.

Years	:	1969	1970	1971	1972	1973	1974	1975
Production (Metric Tonnes)	:	624,000	602,000	582,000	615,000	626,000	620,000	712,000

From each observation we subtract 5,80,000 and divide each subtracted value by 1000.

The coded values are,

$$X'_1 = \frac{624000 - 580000}{1000} = 44$$

$$X'_2 = \frac{602000 - 580000}{1000} = 22$$

Similarly we can calculate all other coded values. Thus, the coded observations are,

$$X' : 44, 22, 2, 35, 46, 40, 132$$

$$\Sigma_i X'_i = (44 + 22 + \cdots + 132)$$

$$= 321$$

The average production of the original observations with the help of coded values by the formula (3.6) is

$$\overline{X} = 580000 + \frac{321}{7} \times 1000$$

$$= 625857.14 \text{ metric tonnes.}$$

Example 3.7. The distribution of the marks of commerce students of a college in Business Statistics was as follows:

Class Intervals of Marks	No. of students
20-30	2
30-40	5
40-50	22

50-60	34
60-70	9
70-80	3
80-100	1

The average marks earned by a student can be calculated by the coding method. For this we can prepare the following table.

Class Intervals	Mid-values	Frequency (f)	$X' = \frac{X-55}{10}$	fX'
20-30	25	2	−3	−6
30-40	35	5	−2	−10
40-50	45	22	−1	−22
50-60	55	34	0	00
60-70	65	9	1	9
70-80	75	3	2	6
80-100	90	1	3.5	3.5
Total		76		−19.5

In the above coding process we have chosen $a = 55$ and $c = 10$. The average marks calculated by the formula (3.7) are,

$$\overline{X} = 55 + \frac{(-19.5)}{76} \times 10$$

$$= 55 - 2.56$$

$$= 52.44 \text{ marks.}$$

Pooled or Combined Mean If we have arithmetic means $\overline{X}_1$ and $\overline{X}_2$ of two groups (having the same unit of measurement of a variable), based on N_1 and N_2 observations respectively, we can compute the mean $\overline{X}_{12}$ of the variate values of the groups taken together from the individual means by the formula,

$$\overline{X}_{12} = \frac{N_1\overline{X}_1 + N_2\overline{X}_2}{N_1 + N_2} \tag{3.8}$$

The advantage of this formula is that we do not have to do the entire calculations for the mean of the combined set of observations again. Moreover the formula for two groups can be extended to any number of groups.

Geometric Mean (G.M.) In algebra geometric mean is calculated in case of geometric progression, but in statistics we need not bother about the progression. Here it is the particular type of data for which the geometric mean is of importance because it gives a good mean value. If the variate values are measured as ratios, proportions or percentages, geometric mean gives a better measure of central tendency than other means.

Definition. Geometric mean of N variate values is the Nth root of their product. Like arithmetic mean it also depends on all observations. It is affected by the extreme values but not to the extent of average. However, there is one great drawback with it, that it can not be calculated if any one or more values are zero or negative. In case an even number of observations are negative, an absurd value of geometric mean will be available from a practical point of view. Hence, if there is a zero or negative value in the set of variate values, it should not be used.

Suppose $X_1, X_2, ..., X_N$ are N variate values, then the geometric mean is given as,

$$G = \sqrt[N]{X_1 X_2 \cdots X_N} \tag{3.9}$$

In case $X_1, X_2, ..., X_k$ have the corresponding frequencies $f_1, f_2, ..., f_k$, then

$$G = \sqrt[N]{X_1^{f_1} X_2^{f_2} \cdots X_k^{f_k}} \tag{3.10}$$

where $N = \Sigma_i f_i$ for $i = 1, 2, ..., k$.

Formulae (3.9) and (3.10) are exactly similar in the sense that instead of multiplying X_i, f_i times in (3.9) X_i is raised to the power f_i in the formula (3.10). Moreover, when each f_i is unity, formula (3.10) reduces to (3.9). In case of grouped data, mid-values of the class intervals are considered as X_i and the formula (3.9) can be used as such.

Though some standard techniques are available to find out square root and cube root, yet for large value of N, N-th root is not easy to compute. To overcome this difficulty, geometric mean is computed through logarithm. Hence, some people even call it *logarithmic mean*. For logarithmic values of X's, it becomes average of log X_i values and the formula for geometric mean is

$$\log G = \frac{1}{N} \Sigma_i (\log_{10} X_i) \tag{3.11}$$

for i = 1, 2, ..., N.

In case of frequency distribution where each of X_i occurs f_i times (i = 1, 2, ..., k).

$$\log G = \frac{1}{N} \Sigma_i \; \{f_i \log_{10} X_i\} \tag{3.12}$$

where N = $\Sigma_i f_i$ for i = 1, 2, ..., k.

Taking antilog of both sides in (3.11) and (3.12), we obtain G.M. Geometric mean is usually calculated when the growth rate or increase in production etc. are given for a number of years or periods.

Example 3.8. Decadal percentage growth of urban population in India (excluding Assam and J & K) from 1921 to 1981 is given below.

Years	:	1921	1931	1941	1951	1961	1971	1981
Decadal per cent increase	:	8.25	19.08	32.09	41.49	25.85	37.91	46.02

Average per cent growth rate of urban population with the last seven decades can be obtained by calculating the geometric mean by the formula (3.11).

$$\log_{10}(G) = \frac{1}{7}\ (\log_{10} 8.25 + \log_{10} 19.08 + \log_{10} 32.09 + \log_{10} 41.49$$

$$+ \log_{10} 25.85 + \log_{10} 37.91 + \log_{10} 46.02)$$

$$= \frac{1}{7}\ (0.9165 + 1.2806 + 1.5063 + 1.6179 + 1.4125 + 1.5787 + 1.6630)$$

$$= \frac{9.9755}{7} = 1.4251$$

Taking antilog of both sides, we get the geometric mean

$$G = 26.62$$

Harmonic Mean (H.M.) In algebra, harmonic mean is found out in the case of harmonic progression only. But in statistics harmonic mean is a suitable measure of central tendency when the data pertains to speed, rates and time.

Definition. Harmonic mean is the inverse of the arithmetic mean of the reciprocals of the observations of a set.

Let $X_1, X_2, \ldots, X_N$ be N variate values in a set, then the harmonic mean,

$$H = \frac{1}{\frac{1}{N}\Sigma_i\left(\frac{1}{X_i}\right)} \qquad (3.13)$$

for $i = 1, 2, \ldots, N$.

If the data are arranged in the form of a frequency distribution in which an observation X_i has frequency f_i $(i = 1, 2, \ldots, k)$, the harmonic mean is given by,

$$H = \frac{N}{f_1/X_1 + f_2/X_2 + \cdots + f_k/X_k} \qquad (3.14)$$

$$H = \frac{1}{\frac{1}{N}\Sigma_i\ (f_i/X_i)} \qquad (3.14.1)$$

where $N = \Sigma_i f_i$ for $i = 1, 2, \ldots, k$.

It fulfills almost all properties of a good measure of central tendency, except when any observation is zero, it can not be calculated. Its main advantage is that it gives more weightage to small values and less weightage to large values.

Lemma 1. If x_1 and x_2 are two observed values, the geometric mean of their arithmetic mean and harmonic mean is equal to the geometric mean of the numbers x_1 and x_2.

We know, $A = \frac{x_1 + x_2}{2}$

$$G = \sqrt{x_1 x_2}$$

and $H = 1/\frac{1}{2}\left(\frac{1}{x_1} + \frac{1}{x_2}\right) = \frac{2x_1 x_2}{x_1 + x_2}$

$$\sqrt{\text{A.H.}} = \sqrt{\frac{x_1 + x_2}{2}\ \frac{2x_1 x_2}{x_1 + x_2}} = \sqrt{x_1 x_2} = G\,.$$

Hence proved.

Lemma 2. If A, G and H stand for A.M, G.M and H.M respectively, the relation

$$A \ge G \ge H.$$

holds.

Here this lemma is proved in the case of two observed values only. Let x_1 and x_2 be two non-negative values of a variable. We know,

$$A = \frac{x_1 + x_2}{2} \quad G = \sqrt{x_1 x_2} \quad H = \frac{2x_1 x_2}{x_1 + x_2}$$

Consider two situations (i) $x_1 = x_2$ (ii) $x_1 \neq x_2$.

(i) suppose $x_1 = x_2 = x$.

$$A = \frac{2x}{2} = x \quad G = \sqrt{x \cdot x} = x \quad H = \frac{2 \cdot x \cdot x}{x + x} = x$$

In this situation $A = G = H$. (1)

(ii) when $x_1 \neq x_2$, $x_1 - x_2$ is a real quantity and hence $\sqrt{x_1} - \sqrt{x_2}$ will also be a real quantity.

$$\therefore \qquad (\sqrt{x_1} - \sqrt{x_2})^2 \ge 0$$

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

$$\frac{x_1 + x_2}{2} \ge \sqrt{x_1 x_2}$$

i.e. $A \ge G$. (2)

Again, $x_1 + x_2 \ge 2\sqrt{x_1 x_2}$

or $(x_1 + x_2)\sqrt{x_1 x_2} \ge 2\sqrt{x_1 x_2}\ \sqrt{x_1 x_2}$

$$\sqrt{x_1 x_2} \ge \frac{2x_1 x_2}{x_1 + x_2}$$

$$G \ge H. \qquad (3)$$

Combining the results (2) and (3), we get,

$$A \ge G \ge H.$$

Example 3.9. A man travels from Jaipur to Agra by a car and takes four hours to cover the whole distance. In the first hour he maintains a speed of 50 km/h, in the second hour his speed remains 64 km/h, in the third 80 km/h and in the fourth hour he travels at the speed of 55 km/h. The average speed of the motorist can be known by calculating the harmonic mean.

$$H = 1/\tfrac{1}{4}\left(\frac{1}{50}+\frac{1}{65}+\frac{1}{80}+\frac{1}{55}\right) = \frac{4}{0.02+0.0154+0.0125+0.0182}$$

$$= \frac{4}{0.0661} = 60.5 \text{ km/hr}$$

Example 3.10. The árithmetic mean of two numbers is 13 and their geometric mean is 12. We can find (i) the numbers (ii) H.M.

Let the two numbers are x_1 and x_2.

(i) Given that,

$$\frac{x_1+x_2}{2} = 13$$

or $x_1 + x_2 = 26$

and $x_1 x_2 = 144$

Also $(x_1 - x_2)^2 = (x_1 + x_2)^2 - 4x_1 x_2$

$$= (26)^2 - 4 \times 144$$

$$= 676 - 576$$

$$= 100$$

$\therefore \quad x_1 - x_2 = \pm 10$

Taking $x_1 - x_2 = 10$ (1)

Also $x_1 + x_2 = 26$ (2)

All equations (1) and (2) we get

$$2x_1 = 36$$

or $x_1 = 18$

Putting the value of x_1 in either of the equations we get, $x_2 = 8$.

Hence the two numbers are $x_1 = 18$ and $x_2 = 8$.

Taking $x_1 - x_2 = -10$ and solving the equations we get $x_1 = 8$ and $x_2 = 18$.

(ii) we know

$$A \times H = G^2$$

Substituting the value of A and G, we get.

$$13 \times H = (12)^2$$

$$H = \frac{144}{13}$$

$$= 11.077$$

MEDIAN

It has been pointed out that mean can not be calculated whenever there is frequency distribution with open end intervals. Also the mean is to a great extent affected by the extreme values of the set of observations. Hence in such cases, there has been a search for some better measure of central tendency. For instance, there are eight persons getting salaries as Rs. 150, 225, 240, 260, 275, 290, 300 and 1500. The mean salary of the persons involved is Rs. 405. This value is not a good measure of central tendency because out of the eight people, seven get Rs. 300 or less. Hence some better measure is preferable and median is one of them.

Definitions

(i) In a distribution, median is the value of the variable which divides it into two equal halves.

(ii) In an ordered series of data, median is an observation lying exactly in the middle of the series.

(iii) In a set of observations, median is the value of a variable that have half of the number of observations below it and remaining half above it.

The median for a set of observations can easily be found out after arranging them in ascending or descending order.

Let $X_1, X_2, ..., X_N$ be N ordered observations. Now two possibilities are there: (a) N is odd, say, $N = 2p + 1$ where p is an integer. In this case $(p + 1)$-th observation will be the median value; (b) if N is even, $N = 2p$, then the average of pth and $(p + 1)$-th observations will be the median value.

Consider the case where the data are arranged in the form of frequency distribution. Suppose the ordered values $X_1, X_2, ..., X_k$ have their corresponding frequencies $f_1, f_2, ..., f_k$, the median for it can be worked out in the following manner.

(1) Find the cumulative frequencies.

(2) Find N/2 where $N = \Sigma_i f_i$ for $i = 1, 2, ..., k$.

(3) Search for the smallest cumulative frequency which contains this value N/2. The variate value corresponding to this cumulative frequency is the median.

Median for Grouped Data If the data are given with class intervals as,

Class intervals	Frequency	Cumulative frequency
Less than X_2	f_1	F_1
$X_2 - X_3$	f_2	F_2
$\vdots$	$\vdots$	$\vdots$
$X_p - X_{p+1}$	f_p	F_p
$\vdots$	$\vdots$	$\vdots$
$X_k - X_{k+1}$	f_k	F_k

where $F_k = N = \Sigma_i f_i$ for $i = 1, 2, ..., k$, we can calculate median by the procedure given here. Find $N/2$ and see in which minimum of the cumulative frequency, $N/2$ is contained. Suppose $N/2$ is contained in the minimum cumulative frequency F_p, then obviously the median class is $X_p - X_{p+1}$. To find the unique median value, we take the help of the interpolation. In this approach, it is assumed that the frequency of a class is uniformly distributed over the class interval. Let the cumulative frequency for the class just above the median class be c. Thus $(N/2 - c)$ is the frequency for the interval between the median and lower limit of the median class. The length of the interval for $(N/2 - c)$ is $\frac{1}{f}(N/2 - c) \times I$ where f—frequency of the median class, I—class interval of the median class and say L_0—lower limit of the median class.

Hence the median

$$M_d = L_0 + \frac{N/2 - c}{f} \times I \qquad (3.15)$$

Properties

(1) Median is a positional average and hence it is not influenced by the extreme values.

(2) Median can be calculated even in the case of open end intervals.

(3) Median can be located even if the data are incomplete.

(4) It is not a good representative of data if the number of items is small.

(5) It is not amenable to further algebraic treatment.

(6) It is susceptible to sampling fluctuations.

Example 3.11. Actual waiting time for the first job on the selected sample of nine people having different field of specialisations was as given below.

Waiting time (in months):11.6, 11.3, 10.7, 18.0, 3.3, 9.2, 8.3, 3.8, 6.8

The median waiting time can be calculated by arranging the data first in ascending order and then taking the mid-value.

3.3, 3.8, 6.8, 8.3, 9.2, 10.7, 11.3, 11.6, 18.0

Hence $N = 9$, $\therefore p = 4$.

Hence 5-th value is the median value that is 9.2 months.

Example 3.12. The export of agricultural products in million dollars from a country during eight quarters in 1974 and 1975 was,

29.7, 16.6, 2.3, 14.1, 36.6, 18.7, 3.5, 21.3.

To find the median of the given set of values, we arrange the data in descending order

36.6, 29.7, 21.3, 18.7, 16.6, 14.1, 3.5, 2.3

Here $N = 8$, $\therefore p = 4$

The mean of 4th and 5th values will be median value.

4th value = 18.7 and 5th value = 16.6

$$\text{Median} = \frac{18.7 + 16.6}{2}$$

$$= 17.65$$

Example 3.13. Given the distribution of income of different occupational groups for the families in a region as:

Professional groups	Income per year ('000 Rs)	Number of families	Cu.fr.
Manager	169.1	82	82
Professional	136.3	62	144
Middle management	79.1	235	379
Manual work	35.7	179	558
Shopkeeper	34.0	96	654
Self-employed	24.9	195	849
Small farmer	19.2	714	1563
Farm labour	14.4	147	1710

The data are written in descending order and the cumulative frequencies are shown in the last column.

$$N = 1710 \text{ and } N/2 = 1710/2 = 855$$

The number 855 is contained in the smallest cumulative frequency 1563. Hence the corresponding value, 19.2 is the median value. It means that the median income is Rs. 19.2 thousand per year.

MODE

It is another measure of central tendency. Mode is a value of a particular type of items which occur most frequently. For instance if shoe size No. 7 has maximum demand, size No. 7 is the modal value of shoe sizes.

Definition. Mode is a variate value which occurs most frequently in a set of values.

In case of discrete distribution, one can find mode by inspection. The variate value having the maximum frequency is the modal value. For instance, consider the discrete distribution.

Variate value (X)	:	3	4	7	8	9	11	12
Frequency (f)	:	2	6	5	14	10	6	3

Clearly $x = 8$ has maximum frequency 14. Hence 8 is the modal value.

Remark: If in a set of observed values, all values occur once or equal number of times, there is no mode.

In cases where maximum frequency is repeated for more than one variate value or occurs for extreme values, the variate value should not be taken as modal value.

In case of frequency distributions in which the maximum frequency differs minutely from its adjoining class frequencies, the modal value or class can not be correctly adjudged and hence the variate value or class corresponding to largest frequency should not be accepted as modal value or modal class merely by inspection.

If there is an irregular distribution, that is, if the trend of frequency changes all of a sudden for certain value(s), the variate value or class corresponding to the maximum frequency should not be accepted as mode.

Example 3.14. The distribution of marks of 174 students out of 25 marks is,

Marks (*X*):	3	4	6	7	9	10	13	15	18	20
Frequency (*f*):	4	8	15	20	32	16	14	35	10	6

The given distribution is an irregular distribution because the frequencies are increasing up to the value of $X = 9$ and then gradually decreasing except for $X = 15$ corresponding to which $f = 35$. This frequency is not consistent with the trend of data. Hence to consider $X = 15$ as mode is not proper.

In all the above three situations, mode obtained by mere inspection is not the correct value due to certain vagaries in sampling. Moreover, a measure of central tendency is considered good provided most of the variate values cluster around it. Therefore, a better modal value can be worked out by the *method of grouping*. Various steps involved in the method of grouping are:

(i) Write the variate values, in order, in column (1) and the frequencies corresponding to them in column (2).

(ii) Add frequencies in pairs starting from the first and place them in a position between the two frequencies in column (3).

(iii) Omit the first frequency and repeat the step (ii) and place the added frequencies in column (4).

(iv) Again group the frequencies in threes starting from the first and place the sum of each group against the mid frequency in column (5).

(v) Leave first frequency and repeat step (iv) creating column (6).

(vi) Again leave first two frequencies and repeat step (iv) placing the added values in column (7). Draw brackets in each column against added frequencies for two's or three's.

(vii) Parenthesise the maximum frequency of each column. The end frequencies which are not used in grouping are left out. For any distribution, the above mentioned seven columns are to be created.

Once the frequency table is prepared, another table known as *analysis table* has to be prepared. In this table, the variate values are written in the caption (column heads) and column numbers along the stub head. In the body of the table, give value 1 to each of the variate value which has been summed up in constituting the maximum frequency. Total 1's of each column of the analysis table. The variate value having the maximum column total is the mode.

Note: In case of tie, choose the variate value or class having maximum frequency.

Marks	Freq.	Added Frequencies				
(1)	(2)	(3)	(4)	(5)	(6)	(7)
3	4	12				
4	8		23	27		
6	15	35			43	
7	20		(52)			(67)
9	32	48		(68)		
10	16		30		(62)	65
13	14	(49)		59		
15	(35)		45			
18	10	16			51	
20	6					

Analysis table

Columns	Marks (*X*)									
	3	4	6	7	9	10	13	15	18	20
(2)								1		
(3)							1	1		
(4)				1	1					

(5)				1	1	1				
;(6)					1	1	1			
(7)			1	1	1					
Total			1	3	4	2	2	2		

In the above table for X = 9, the maximum sum of 1's is 4, hence the modal value is 9.

Remarks

(1) It is worth pointing out that by inspection one would have concluded that the mode is 15 as it has maximum frequency 35, however this is not correct, as revealed by the analysis table.

(2) The given distribution is *unimodal* as it has only one modal value. There may be two or more columns having equal maximum frequency in the analysis table, in such case, each corresponding variate value would have been taken as mode. The distribution having two modes is known as *biomodal* and with more than two modes is known as *multimodal*.

Mode of a Continuous Distribution If the distribution is with continuous class intervals, mode can be easily calculated in the manner described here. One must take care that the distribution is continuous and in order (ascending or descending). The class intervals for all the classes are equal. If they are unequal, they should be made equal presuming that the frequencies are uniformly distributed throughout the class interval.

Let the grouped frequency distribution be as follows:

Classes	Frequency
X_1-X_2	f_1
X_2-X_3	f_2
X_3-X_4	f_3
⋮	⋮
X_{p-1}–X_p	f_{p-1}
X_p-X_{p+1}	f_p
X_{p+1}–X_{p+2}	f_{p+1}
⋮	⋮
X_k–X_{k+1}	f_k

Assume that the distribution is in order and the maximum frequency is f_p. Then, the modal class is X_p-X_{p+1}. The exact value of mode can be found out by the interpolation formula,

$$M_0 = X_p + \frac{f_p - f_{p-1}}{(f_p - f_{p-1}) + (f_p - f_{P+1})} \quad (X_{p+1} - X_p) \qquad (3.16)$$

If we denote the lower limit of the modal class by L_0, the maximum frequency by f, the frequency preceding f and by f+ following f by f_{-1} and the class interval by I, the formula (3.16) can be written as,

$$M_0 = L_0 + \frac{f - f_{-1}}{(f - f_{-1}) + (f - f_{+1})} \times I \qquad (3.16.1)$$

Putting $f - f_{-1} = \Delta_1$, $f - f_{+1} = \Delta_2$, the formula for mode is

$$M_0 = L_0 + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times I \qquad (3.16.2)$$

Example 3.15. We find the mode for the frequency distribution given in example 2.5. The distribution is,

Classes (wt in kg)	Number of children
2-0-2.4	5
2.4-2.8	5
2.8-3.2	9
3.2-3.6	4
3.6-4.0	4
4.0-4.4	3

Obviously by inspection the modal class is (2.8–3.2). We calculate mode by the formula (3.16.2). In this case,

$L_0 = 2.8$, $\Delta_1 = 9 - 5 = 4$, $\Delta_2 = 9 - 4 = 5$,

$I = 3.2 - 2.8 = 0.4$

Hence,

$$M_0 = 2.8 + \frac{4}{4+5} \times 0.4$$

$$= 2.8 + 0.178$$

$$= 2.978 \text{ kg}$$

In case the frequency distribution is such that the modal class can not be ascertained merely by inspection, the method of grouping should be adapted.

Many frequency distributions have more than one mode. But, we are interested in a single central value and hence for such distributions mode is considered as an ill-defined measure of central tendency. For a moderately skewed[3] or asymmetrical frequency distribution, mode can be calculated by Karl Pearson's empirical formula,

$$\text{Mean} - \text{Mode} = 3\,(\text{Mean} - \text{Median}) \qquad (3.17)$$

$$\text{Mode} = 3\,\text{Median} - 2\,\text{Mean} \qquad (3.17.1)$$

In case where mode is ill defined, formula (3.17) can be used to determine the modal value.

Graphical Method of Finding Mode If we draw a histogram for the given distribution, naturally the highest bar will possess the modal value. To find the

[3] Skew distribution is discussed in Chapter 4.

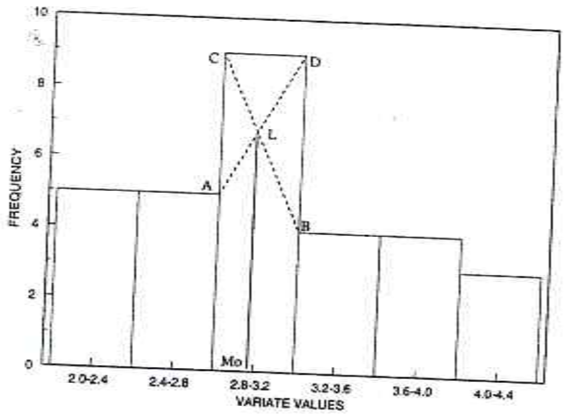


Fig. 3.1 : Mode by Graphical Method

exact modal value consider only three bars namely, the highest bar and the bars adjacent to it on both the sides. In the middle bar draw two diagonal lines joining the point *A* of the preceding bar to *D* and *B* of the following bar to *C* as shown in Fig. 3.1. Suppose these diagonals *AD* and *BC* intersect each other at the point *L*. Draw a perpendicular line from *L* on the axis of *X* which meets it at the point M_0. The distance of M_0 from origin on the aforesaid scale is the modal value.

Merits and Demerits

(1) It is not affected by extreme values of a set of observations.

(2) It can be calculated for distributions with open end classes.

(3) The main drawback of mode is that often it does not exist.

(4) Often its value is not unique.

(5) It does not fulfil most of the requirements of a good measure of central tendency.

Remark. After going through the details of the three measures of central tendency namely, the mean, median and mode, it is apparent that no measure is absolutely good. All these measures have some good and bad points. The choice of the measure depends more upon the purpose of information and the situation in which that average value is to be used. Therefore, a measure should be used judiciously.

Example 3.16. Given the number of families in a locality according to monthly per capita expenditure classes in rupees as,

Monthly per capita expenditure classes (Rs)	Number of families
140-150	17
150-160	29
160-170	42
170-180	72
180-190	84
190-200	107
200-210	49
210-220	34
220-230	31
230-240	16
240-250	12

we calculate the modal per capita expenditure by the graphical method.

Clearly the modal class is 190-200 as it has maximum frequency 107. We draw the following diagram and determine the modal value.

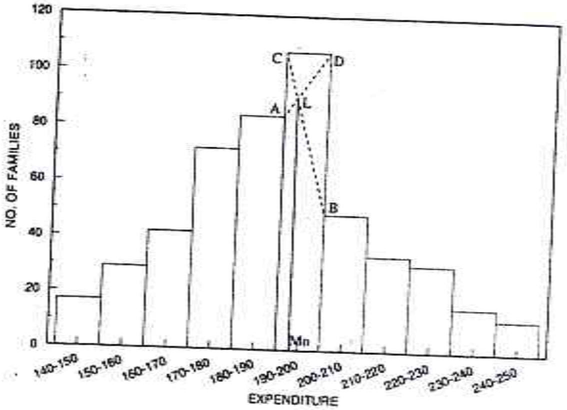


Fig. 3.2 : Modal Expenditure by Graph

The value of M_0 on *X*-axis is 193. Hence, mode = 193.

Note: It can be verified that the value of mode by the formula (3.16.2) is 192.84 which is almost equal to the value obtained by graph.

FRACTILES

α-fractile of a continuous distribution of a random variable *X* is a point, X_α, such that for the distribution of *X*, the random variable has probability α of being less than or equal to X_α. For a discrete distribution, fractile may be defined as that variate value which has α-proportion of items up to this value for an increasing ordered set of values. Some people use the term *quantile* in place of fractile. For instance, 1/4-fractile is called first quartile, 1/2-fractile is known as second quartile and 3/4-fractile is called third quartile and are denoted by Q_1, Q_2 and Q_3, respectively. Similarly 1/10-fractile is known as first decile, 2/10-fractile is the second decile and so on. In all we have nine deciles which are denoted by

D_i (i = 1, 2, ..., 9). In the same manner, the multiples of 1/100-fractile are called percentiles and are denoted by P_i (i = 1, 2, ..., 99).

Quartiles From the definition of fractile, it is apparent that three variate values of the variable X which divide the series into four equal parts are called quartiles for the corresponding distribution of X. Hence Q_1 is a value which has 25% items which are less than or equal to Q_1. Similarly Q_2 has 50% items with values less than or equal to Q_2 and Q_3 has 75% items whose values are less than or equal to Q_3.

For discrete data it is simple to locate the fractiles. Arrange the data in order if they are not and work out the cumulative frequencies. To find out Q_1 calculate $(N + 1)/4$ where N is the total number of observations. Search for the minimum cumulative frequency in which $(N + 1)/4$ is contained. The variate value against this cumulative frequency is the value of Q_1. For Q_2, find $(N + 1)/2$ and search for the minimum cumulative frequency in which $(N + 1)/2$ is contained. The variate value corresponding to this cumulative frequency is the second quartile Q_2. Calculate $3(N + 1)/4$ and locate Q_3 in the same manner as Q_1 and Q_2.

To find the decile D_i(i = 1, 2, ..., 9) we calculate the value $i(N + 1)/10$ and search for the minimum cumulative frequency which contains the value i $(N + 1)/10$. The variate value corresponding to this cumulative frequency is the i-th decile. In a similar manner, calculate $i(N + 1)/100$ (i = 1, 2, ..., 99) for percentiles and proceeding on as for quartiles and deciles, the percentiles are located.

Continuous Distribution Case Let the observations, arranged in order, in the form of a continuous distribution be as given below.

Classes	Frequency	Cu.frequency
X_1-X_2	f_1	$F_1 = f_1$
X_2-X_3	f_2	F_2
X_3-X_4	f_3	F_3
⋮	⋮	⋮
X_p-X_{p+1}	f_p	F_p
⋮	⋮	⋮
X_k-X_{k+1}	f_k	$F_k = N$

To locate the quartile class, calculate $iN/4$ instead of $i(N+1)/4$ and proceed as we do for discrete distribution, i.e., search that minimum cumulative frequency in which $iN/4$ is contained. The class corresponding to this cumulative frequency is called *quartile class*. The unique value of the i-th quartile is calculated by the formula

$$Q_i = l_0 + \frac{iN/4 - c}{f} \times I \qquad (3.18)$$

where i = 1, 2, 3

Q_i — i-th quartile which is to be worked out

l_n — lower limit of the i-th quartile class

N — total of all the frequencies

c — cumulative frequency for the class just above the quartile class

f — frequency of the quartile class

I — class interval

Deciles The procedure for locating the i-th decile class is to calculate $iN/10$ and search that minimum cumulative frequency in which this value is contained. The class corresponding to this cumulative frequency is i-th decile class. The unique value of i-th decile can be calculated by the formula,

$$D_i = l_n + \frac{iN/10 - c}{f} \times I \qquad (3.19)$$

where i = 1, 2, ..., 9.

All the terms in (3.19) can be decoded as explained in (3.18) by changing the word quartile class by decile class.

Percentiles To locate i-th percentile class, calculate $iN/100$ (i = 1, 2, ..., 99) and find that minimum cumulative frequency which contains this value. The class corresponding to this minimum cumulative frequency is the percentile class. Unique value of i-th percentile can be calculated by the formula,

$$P_i = l_n + \frac{iN/100 - c}{f} \times I \qquad (3.20)$$

where (i = 1, 2, ..., 99).

All the terms in (3.20) are decoded as in (3.18) simply by replacing the word quartile class by percentile class.

Note that Q_2, D_5 and P_{50} are equivalent to the median of the given distribution.

Example 3.17. Consider again the continuous distribution given in example 3.16 and calculate (i) all the quartiles (ii) 70th decile and (iii) 90-th percentile.

The frequency distribution and the cumulative frequencies in the last column are presented below.

Monthly per capita expenditure classes (Rs.)	Number of families	Cumulative frequency
140-150	17	17
150-160	29	46
160-170	42	88
170-180	72	160
180-190	84	244
190-200	107	351
200-210	49	400
210-220	34	434

$\frac{n}{4} = 123.75$

$\frac{n}{2} = 246.5$

$\frac{3n}{4} = 369.75$

220-230	31	465
230-240	16	481
240-250	12	493

(i) For Q_1, $\frac{N}{4} = \frac{493}{4} = 123.25$

The number 123.25 is contained in the minimum cu. freq. 160. Hence the class 170-180 is the first quartile class. By the formula (3.18) we have,

$$Q_1 = 170 + \frac{123.25 - 88}{72} \times 10$$

$$= 170 + 4.90$$

$$= 174.90\text{Rs.}$$

Similarly for $Q_2, \frac{2N}{4} = \frac{2 \times 493}{4} = 246.50$

The number 246.50 is contained in the minimum cu.freq. 351. Hence the class 190-200 is the second quartile class. Thus, by the formula (3.18),

$$Q_2 = 190 + \frac{246.50 - 244}{107} \times 10$$

$$= 190 + 0.23$$

$$= 190.23\text{Rs.}$$

Again for $Q_3, \frac{3N}{4} = \frac{3 \times 493}{4} = 369.75$

The number 369.75 is contained in the minimum cu.fr. 400.

Hence the class 200-210 is the third quartile class. By the formula (3.18).

$$Q_3 = 200 + \frac{369.75 - 351}{49} \times 10$$

$$= 200 + 3.83$$

$$= 203.83\text{Rs.}$$

(ii) For D_7, $\frac{7N}{10} = \frac{7 \times 493}{10} = 345.1$

The number 345.1 is contained in the minimum cu.fr. 351. Hence the class 190-200 is the 7-th decile class. By formula (3.19) we have,

$$D_7 = 190 + \frac{345.1 - 244}{107} \times 10$$

$$= 190 + 9.45$$

$$= 199.45\ \text{Rs.}$$

(iii) For P_{90}, $\frac{90N}{100} = \frac{90 \times 493}{100} = 443.70$

The number 443.70 is contained in the minimum cu.fr. 465. Hence, the 90-th percentile class is 220-230. By formula (3.20) we have,

$$P_{90} = 220 + \frac{443.70 - 434}{31} \times 10$$

$$= 220 + 3.13$$

$$= 223.13\text{Rs.}$$

Concluding Remarks Different measures of central tendency and fractiles (quantiles) have been discussed in this chapter. Out of mean, median and mode, the mean (average) is the most commonly used measure of central tendency. But the other two namely, the median and mode are not any less important. Median is a largely used central measure in psychology, education and other social sciences. It is a suitable average for qualitative information like the attitude towards disabled people, beauty or intelligence of certain individuals, etc. Mode is a useful measure for manufacturers.

QUESTIONS AND EXERCISES

1. What do you understand by a measure of central tendency? Explain with examples.
2. What are the desirable properties which an average should possess? Which of the average to your mind possesses most of these properties and why?
3. Under what circumstances, would you use the following instead of any other measure of central tendency.
   (a) Mode.
   (b) Geometric mean.
   (c) Median.
4. In what respect is the weighted mean superior to the simple average?
5. Which type of average is most suitable for the following problems and why?
   (a) Average income per month in a year of an advocate.
   (b) Normal size of shirts for a readymade garment's manufacturer.
   (c) Consumption per head in a family consisting of 7 men, and 4 women and 9 children.
   (d) The average marks of a mediocre student in a class.
   (e) Average speed of a plane in flight from Delhi to New York.
6. Name different kinds of averages and discuss their merits and demerits.
7. Explain the meaning of a fractile and give its uses. What information do we obtain by quartiles, deciles and percentiles?
8. What considerations will you weigh in choosing a suitable average for studying a phenomenon? Give a few typical cases in which your choice will fall on any average other than the arithmetic mean. *(C.A., 1965)*
9. How will you find (a) the average marks of a class of students to show the level of intelligence, (b) the average cost of goods purchased in different lots to determine the selling prices, (c) the average size of groups of items for the purpose of

classification and (d) the average rate of increase in prices when the prices increase as different rates during successive periods. Explain why you adopt a particular method in each case? *(B.Com., Agra and Raj., 1948)*

10. What is the effect of reducing each observation of a decreasing series by 10 on the following:
(a) the average
(b) the median
(c) the mode
(d) the quartiles.

11. Prove that
(a) H.M. ≤ G.M. ≤ A.M.
(b) $\sqrt{A.M. \times H.M.} = G.M.$
where A.M., G.M. and H.M. are the usual abbreviations.

12. Define the following and give one appropriate example of your own for the use of each.
(a) Mode.
(b) Third quartile.
(c) Geometric mean.
(d) Median.
(e) Average.

13. Additional irrigation utilisation from major and medium schemes at the end of various plans in India is,

Additional irrigation (Lakh hectares)

97,	110,	130,	152,	168,	187,	22,	266,	226

Find the average additional irrigation utilisation during this period.

14. Following are the percentages of literates in six villages situated at six different distances from the district headquarters.
Percentage of literates: 52.22, 46.59, 21.36, 30.17, 22.87, 17.77
Find the mean percentage of literates.

15. The distribution of Labrum length of workers of Apis cerena measured in millimetre is as given below.

Labrum length (mm)	No. of workers
0.30	8
0.31	4
0.32	2
0.33	2
0.34	3
0.35	2
0.36	3
0.37	3
0.38	1
0.40	5

Find the average Labrum length of workers and the mode.

16. The distribution of the age of patients visiting an outdoor patients department in a dispensary on Sunday is as follows:

Age (years)	No. of patients
More than 10	152
" "20	128
" "30	113
" "40	77
" "50	36
" "60	22
" "70	5
and up to 80	

Calculate (i) mean age of persons visiting the dispensary, (ii) median age and (iii) modal age.

17. The distribution of age of males at the time of marriage was as follows:

Age (years)	No. of males
18-20	5
20-22	18
22-24	28
24-26	37
26-28	24
28-30	22

Find at the time of marriage (i) the average age, (ii) the modal age, (iii) the median age, (iv) third quartile, (v) sixth decile and (vi) ninetieth percentile.

18. The earnings of five nationalised banks in crore rupees are as given below:

217.40,	330.50	682.55,	1263.59,	2249.63

Find the geometric mean of the earnings.

19. The prices of wheat, dal, rice, vanaspati ghee, milk, sugar and their weights are as given below:

Prices (Rs. per kg)	:	1.57,	5.77,	3.94,	17.00,	3.50,	4.50
Weights	:	34	6	4	9	24	23

Find the weighted average of prices of the commodities.

20. In a factory a mechanic takes 15 days to fabricate a machine, the second mechanic takes 18 days, the third mechanic takes 30 days and the fourth mechanic takes 90 days. Find the average number of days taken by the workers to fabricate the machine.
[Hint: Find harmonic mean]

Chapter 3

Measures of Dispersion

In the third chapter, we concentrated upon a central value, which gives an idea of the whole mass that is a complete set of variate values. However, the information so obtained is neither exhaustive nor comprehensive, as the mean does not lead us to know whether the observations are close to each other or far apart. Median is a positional average and has nothing to do with the variability of the observations in the series. Mode is the largest occurring value independent of other values of the set. This leads us to conclude that a measure of central tendency alone is not enough to have a clear idea about the data unless all observations are almost the same. Moreover, two or more sets may have the same mean and/or median but they may be quite different. To clear this point consider the three sets as follows:

Set A	30	30	30	30	30
Set B	28	29	30	31	32
Set C	3	5	30	37	75

All the three sets *A, B* and *C* have mean 30 and median is also 30. But by inspection it is apparent that the three sets differ remarkably from one another. Thus to have a clear picture of data, one needs to have a measure of dispersion or variability (scatteredness) amongst observations in the set. Commonly used measures of dispersion are:

(1) Range

(2) Interquartile range and Quartile deviation

(3) Mean deviation

(4) Variance

(5) Standard deviation

(6) Coefficient of variation.

Before giving the details of various measures of dispersion, it is worthwhile to point out that a measure is to be adjudged on the basis of all those properties which are discussed for the measures of central tendency. Hence, their repetition

is superfluous. Each of the measures of dispersion is adequately elucidated in the subsequent discussion.

RANGE

Definition. It is the difference between the largest and the smallest observation in a set.

If we denote the largest observation by L and the smallest observation by S, the formula is,

$$\text{Range } R = L - S \tag{4.1}$$

A relative measure known as *coefficient of range* is given as,

$$\text{Coeff. of range} = \frac{L - S}{L + S} \tag{4.2}$$

Lesser the range or coefficient of range, better the result.

Properties

(1) It is the simplest measure and can easily be understood.

(2) Besides the above merit, it hardly satisfies any property of a good measure of dispersion e.g. it is based on two extreme values only, ignoring the others. It is not liable to further algebraic treatment.

Example 4.1. The population in eighteen panchayat samities of a district is as given below:

Population ('000 number)

77,	76,	83,	68,	57,	107,	80,	75,	95
100,	113,	119,	121,	121,	83,	87,	46,	74.

We can find the range by the formula (4.1)

$$L = 121 \text{ and } S = 46$$

The range,

$$R = 121 - 46 = 75$$

Some people also write the range as 46-121.

Also,

$$\text{Coeff. of range} = \frac{121 - 46}{121 + 46} = \frac{75}{167} = 0.449$$

INTERQUARTILE RANGE (I.R.)

Definition. The difference between the third quartile and first quartile is called interquartile range. Symbolically,

$$\text{I.R.} = Q_3 - Q_1 \tag{4.3}$$

QUARTILE DEVIATION (Q.D.)

This is half of the interquartile range, i.e.

$$\text{Q.D.} = \frac{Q_3 - Q_1}{2} \tag{4.4}$$

Also the coefficient of quartile deviation is given by the formula,

$$\text{Coeff. of Q.D.} = \frac{Q_3 - Q_1}{Q_3 - Q_1} \tag{4.5}$$

Coefficient of quartile deviation is an absolute quantity (unitless) and is useful to compare the variability among the middle 50% observations.

Properties

(1) It is a better measure of dispersion than range in the sense that it involves 50% of the mid values of a series of data rather than only two extreme values of a series.

(2) Since it excludes the lowest and highest 25% values, it is not affected by the extreme values.

(3) It can be calculated for the grouped data with open end intervals.

(4) It is not capable of further algebraic treatment.

(5) It is susceptible to sampling fluctuations.

(6) This measure does not take into account the individual values occurring between Q_1 and Q_3. It means that no idea about the variation of even 50% mid values is available from this measure. Anyhow, it provides some idea if the values are uniformly distributed between Q_1 and Q_3.

(7) It is not considered a good measure of dispersion as it does not show the scattering of the central value. In fact it is a measure of partitioning of distribution. Hence, it is not commonly used.

Example 4.2. The value of Q_1, Q_2 and Q_3 as worked out in example 3.17 are,

$Q_1 = 174.90$ $Q_2 = 190.23$ $Q_3 = 203.83$

Interquartile range, I.R. $= 203.83 - 174.90 = 28.93$

$$\text{Quartile deviation,Q.D.} = \frac{203.83 - 174.90}{2} = \frac{28.93}{2} = 14.465$$

$$\text{Coeff. of Q.D.} = \frac{203.83 - 174.90}{203.83 + 174.90} = \frac{28.93}{378.73} = 0.076$$

MEAN DEVIATION (M.D.)

The measures of dispersion discussed so far are not satisfactory in the sense that they lack most of the requirements of a good measure. Mean deviation is a better measure than range and Q.D.

Definition. It is the average of the absolute deviations taken from a central value, generally the mean or median.

Consider a set of N observations $X_1, X_2, ..., X_N$. Then the mean deviation.,

$$\text{M.D.} = \frac{1}{N} \Sigma_i |X_i - A| \tag{4.6}$$

for i = 1, 2, ..., N where A is a central value.

Let $|X_i - A| = d_i$,

$$\text{Then,} \qquad \text{M.D.} = \frac{1}{N} \Sigma_i d_i \tag{4.6.1}$$

In case of data given in the form of a frequency distribution where the variate values $X_1, X_2, ..., X_k$ occur $f_1, f_2, ..., f_k$ times respectively, the formula for mean deviation is,

$$\text{M.D.} = \frac{1}{N} \Sigma_i f_i |X_i - A| \tag{4.7}$$

where $\Sigma_i f_i = N$ for $i = 1, 2, ..., k$.

In case of grouped data, the mid-point of each class interval is treated as X_i and we can use formula (4.7).

Properties

(1) Mean deviation removes one main objection of the earlier measures, that it involves each value of the set.

(2) It is not affected much by extreme values.

(3) It has no relationship with any of the other measures of dispersion.

(4) Its main drawback is that algebraic negative signs of the deviations are ignored which is mathematically unsound.

(5) Mean deviation is minimum when the deviations are taken from median.

Note: If the deviations are taken from mean and the signs of the deviations are taken into consideration, the sum of the deviations is zero i.e. $\Sigma_i (X_i - \bar{X}) = 0$.

Example 4.3. The production of all crops in India from 1971 to 1978 is given below:

Production (million tonnes)							
111.5,	111.2,	102.3,	112.4,	108.8,	125.3,	116.5,	132.7

Mean deviation taking deviations from the mean and also from the median has been calculated.

(i) $\text{Mean} = \frac{1}{8}(111.5 + 111.2 + ... + 132.7) = \frac{920.7}{8} = 115.09$

Mean deviation taking deviations from the mean using the formula (4.6) is,

$$\text{M.D.} = \frac{1}{8}\,(|111.5 - 115.09| + |111.2 - 115.09| + \ldots + |132.7 - 115.09|)$$

$$= \frac{1}{8}\,(3.59 + 3.89 + 12.79 + 2.69 + 6.29 + 10.21 + 1.41 + 17.61)$$

$$= \frac{58.48}{8} = 7.31 \text{ million tonnes.}$$

(ii) Now calculate the mean deviation taking the deviations from the median. For finding out the median, arrange the data in ascending order.

102.3, 108.8, 111.2, 111.5, 112.4, 116.5, 125.3, 132.7

$$\text{Median} = \frac{111.5 + 112.4}{2} = \frac{223.9}{2} = 111.95$$

Mean deviation taking deviations from the median by the formula (4.6) is,

$$\text{M.D.} = \frac{1}{8}\,(|102.3 - 111.95| + |108.8 - 111.95| + \ldots + |132.7 - 111.95|)$$

$$= \frac{1}{8}\,(9.65 + 3.15 + 0.75 + 0.45 + 0.45 + 4.55 + 13.35 + 20.75)$$

$$= \frac{53.1}{8} = 6.64 \text{ million tonnes.}$$

The mean deviation about median is less than the mean deviation about mean. This further substantiates the statement that mean deviation about median is minimum.

Example 4.4. The distribution of age at the marriage of grooms with brides of age group 15-39 is displayed here.

Age groups: (years)	15-19	19-23	23-27	27-31	31-35	35-39
No. of grooms:	8	59	47	23	6	4

Mean deviation taking deviations from the mean has been calculated. The calculations are shown in the table given below.

Class intervals	Midpoints (X)	Frequency (f)	fX	$\|X - \bar{X}\|$	$f\,\|X - \bar{X}\|$
15-19	17	8	136	7.24	57.92
19-23	21	59	1239	3.24	191.16
23-27	25	47	1175	0.76	35.72
27-31	29	23	667	4.76	109.48
31-35	33	6	198	8.76	52.56
35-39	37	4	148	12.76	51.04
Total		147	3563		497.88

where $\bar{X} = \dfrac{3567}{147} = 24.24$

Mean deviation about mean by the formula (4.7) is,

$$\text{M.D.} = \frac{497.88}{147} = 3.39 \text{ years}$$

VARIANCE

The main objection of mean deviation, that the negative signs are ignored, is removed by taking the square of the deviations from the mean.

Definition. The variance is the average of the squares of the deviations taken from mean.

Let $X_1, X_2, \ldots, X_N$ be the measurements on N population units, the population variance,

$$\sigma^2 = \frac{1}{N}\,\Sigma_i\,(X_i - \bar{X})^2 \qquad (4.8)$$

for $i = 1, 2, 3, \ldots, N$.

$$\sigma^2 = \frac{1}{N}\,\{\Sigma_i\,X_i^2 - (\Sigma_i\,X_i)^2/N\} \qquad (4.8.1)$$

where $\bar{X}$ is the population mean.

If the data are given in the form of frequency distribution in which the variate value X_i has its corresponding frequency f_i $(i = 1, 2, \ldots, k)$, the variance,

$$\sigma^2 = \frac{1}{N}\,\Sigma_i f_i\,(X_i - \bar{X})^2 \qquad (4.9)$$

for $i = 1, 2, \cdots k$.

where $N = \Sigma_i\,f_i$ and $\bar{X} = \dfrac{1}{N}\,\Sigma_i\,f_i\,X_i$.

$$\sigma^2 = \frac{1}{N}\,\{\Sigma_i\,f_i\,X_i^2 - (\Sigma_i\,f_i\,X_i)^2/N\} \qquad (4.9.1)$$

In case of grouped data, mid-values of the classes are considered as X_i and consequently we can make use of the formula (4.9).

The sample[1] variance of the set $x_1, x_2, ..., x_n$ of n observations is given by the formula,

$$s^2 = \frac{1}{n-1} \Sigma_i (x_i - \bar{x})^2 \tag{4.10}$$

$$= \frac{1}{n-1} \{\Sigma_i x_i^2 - (\Sigma_i x_i)^2/n\} \tag{4.10.1}$$

for $i = 1, 2, \cdots, n$.

where $\bar{x} = \frac{1}{n} \Sigma_i x_i$.

If the observation x_i occurs f_i times for $i = 1, 2, ..., k$, then the sample variance,

$$s^2 = \frac{1}{n-1} \Sigma_i f_i (x_i - \bar{x})^2 \tag{4.11}$$

$$= \frac{1}{n-1} \{\Sigma_i f_i x_i^2 - (\Sigma_i f_i x_i)^2/n\} \tag{4.11.1}$$

where $n = \Sigma_i f_i$.

Example 4.5. The following nine measurements are the heights in inches in a sample of nine soldiers.

Height (X):	69,	66,	67,	69,	64,	63,	65,	68,	72

The sample variance of height of soldiers can be computed by formula (4.10).

$$\sum_{i=1}^{9} x_i = 69 + 66 + \ldots + 72 = 603$$

$$\bar{x} = \frac{603}{9} = 67 \text{ inches}$$

$(x-\bar{x}):$, 2, −1, 0, 2, −3, − 4, −2, 1, 5

$(x-\bar{x})^2:$, 4, 1, 0, 4, 9, 16, 4, 1, 25

$$\sum_{i=1}^{9} (x_i-\bar{x})^2 = 4+1+0+4+9+16+4+1+25$$

$$= 64$$

$$s^2 = \frac{64}{8}$$

$$= 8 \text{ inches}^2$$

CODING OF DATA

If the values in a series or mid-values of the classes are large enough, coding of values is a good device to simplify the calculations. In coding subtract a constant A (generally the middle X_i value of the series) from each value or mid-value in case of grouped data and then divide the reduced values by a suitable constant 'C' $(C \neq 0)$, generally the class interval in case of grouped data.

Thus, the coded value,

$$X' = \frac{X-A}{C}.$$

We present elaborately the method of calculating the variance for the grouped data because it will automatically cover the case of ungrouped frequency distribution.

Classes	Frequency	Mid-values	$X' = \frac{X-A}{C}$	fX'	fX'^2
$Y_1 - Y_2$	f_1	X_1	$X'_1 = \frac{X_1-A}{C}$	$f_1X'^1$	$f_1 X'^2_1$
$Y_2 - Y_3$	f_2	X_2	$X'_2 = \frac{X_2-A}{C}$	$f_2X'^2$	$f_2 X'^2_2$
⋮	⋮	⋮	⋮	⋮	⋮
$Y_p - Y_{p+1}$	f_p	X_p	$X'_p = \frac{X_p-A}{C}$	$f_p X'_p$	$f_p X'^2_p$
⋮	⋮	⋮	⋮	⋮	⋮
$Y_k - Y_{k+1}$	f_k	X_k	$X'_k = \frac{X_k-A}{C}$	$f_k X'_k$	$f_k X'^2 k$
Total	N			$\Sigma_i f_i X'_i$	$\Sigma_i f_i X_i'^2$

for $i = 1, 2, ..., k$.

Here the mean of uncoded data in terms of coded data can be obtained as,

$$\bar{X}' = \frac{1}{N} \Sigma_i f_i X'_i = \frac{1}{N} \Sigma_i f_i \left(\frac{X_i - A}{C}\right)$$

$$= \frac{1}{CN} \Sigma_i f_i X_i - \frac{1}{CN} \Sigma_i f_i A = \frac{1}{C} \bar{X} - \frac{A}{C}$$

Since $\frac{1}{N} \Sigma_i f_i X_i = \bar{X}$ and $\Sigma_i f_i = N$

or $$\bar{X} = A + C\bar{X}' \tag{4.12}$$

Variance of uncoded data 'σ^2' in terms of variance of coded data 'σ'^2' will be obtained as,

$$\sigma'^2 = \frac{1}{N} \Sigma_i f_i (X'_i - \bar{X}')^2 \tag{4.13}$$

$$= \frac{1}{N} \Sigma_i f_i \left(\frac{X_i - A}{C} - \frac{\bar{X} - A}{C}\right)^2 = \frac{1}{N} \Sigma_i f_i \left(\frac{X_i}{C} - \frac{\bar{X}}{C}\right)^2$$

or $$C^2 \sigma'^2 = \frac{1}{N} \Sigma_i f_i (X_i - \bar{X})^2 = \sigma^2 \tag{4.13.1}$$

[1] Sample : see Chapter 7.

Also we can write

$$\sigma'^2 = C^2 \frac{1}{N} \{\Sigma_i f_i X_i'^2 - (\Sigma_i f_i X'_i)^2/N\} \qquad (4.13.2)$$

when each $f_i = 1$, the variance

$$\sigma^2 = C^2 \frac{1}{N} \{\Sigma_i X_i'^2 - (\Sigma_i X'_i)^2/N\} = C^2 \sigma'^2 \qquad (4.13.3)$$

From (4.12) it is easily inferred that the mean is affected by the shift of origin, and also by the change of scale. Hence on subtracting A from each observation and then dividing by C, the mean of the coded data is to be multiplied by C, and then added to the constant A to get the mean of the uncoded data.

The relation (4.13.1) does not involve A. It means that the variance is not affected by the shift of origin, but the change of scale does affect it, as the relation involves C. Moreover, if the variance of coded variable is multiplied by the square of the scale constant C, the variance of the uncoded variable is obtained. In most of the situations coding makes the calculations simple if A and C are properly chosen.

Properties

(1) The variance has mostly removed the lacunae which are present in the measures of dispersion given before it.

(2) The main demerit of variance is, that its unit is the square of the unit of measurement of variate values. For clarity, say, the variable X is measured in cms, the unit of variance is cm^2. Generally this value is large and makes it difficult to decide about the magnitude of variation.

(3) The variance gives more weightage to the extreme values as compared to those which are near to mean value, because the difference is squared in variance.

STANDARD DEVIATION (S.D.)

The drawbacks of variance are overcome in this measure of dispersion.

Definition. The positive square root of the variance is called standard deviation.

$$\text{S.D.} = \sqrt{\sigma^2} = \sigma \qquad (4.14)$$

For a sample,

$$\text{S.D.} = \sqrt{s^2} = s \qquad (4.14.1)$$

In simple words, we can say that standard deviation explains the average amount of variation on either side of the mean.

Properties

(1) Standard deviation is considered to be the best measure of dispersion and is used widely.

(2) There is however one difficulty with it. If the unit of measurement of variables of two series is not the same, then their variability can not be compared by comparing the values of standard deviation.

(3) An empirical relation between Q.D., M.D. and S.D. is,

$$6\,\text{Q.D.} = 5\,\text{M.D.} = 4\,\text{S.D.} \qquad (4.15)$$

COEFFICIENT OF VARIATION (C.V.)

All the measures of dispersion discussed so far have units. If two series differ in their units of measurement, their variability cannot be compared by any measure given so far. Also, the size of measures of dispersion depends upon the size of values. Hence in situations where either the two series have different units of measurements, or their means differ sufficiently in size, the coefficient of variation should be used as a measure of dispersion. It is a unitless measure of dispersion and also takes into account the size of the means of the two series. It is the best measure to compare the variability of two series or sets of observations. A series with less coefficient of variation is considered more consistent or stable.

Definition. Coefficient of variation of a series of variate values is the ratio of the standard deviation to the mean multiplied by 100.

If σ is the standard deviation and $\overline{X}$ is the mean of the set of values, the coefficient of variation is,

$$\text{C.V.} = \frac{\sigma}{\overline{X}} \times 100 \qquad (4.16)$$

This measure was given by Professor Karl Pearson.

Note: In case of sample studies, we use s, the sample S.D. and $\bar{x}$, the sample mean instead of σ and $\overline{X}$ respectively.

Properties

(1) It is one of the most widely used measure of dispersion because of its virtues.

(2) Smaller the value of C.V., more consistent are the data and vice versa. Hence a series with smaller C.V. than the C.V. of other series is more consistent, i.e., it possesses less variability.

(3) For field experiments, C.V. is generally reported. If C.V. is low, it indicates more reliability of experimental findings.

Example 4.6. The following figures give the crude birth rate per 1000 people in Switzerland from 1968 to 1980.

Crude birth rate (X):	17.1,	16.5,	15.8,	15.2,	14.3,	13.6,	12.9,	12.3
:*,	11.7,	11.5,	11.3,	11.3,	11.6			

The variance, standard deviation and coefficient of variation for birth rate are computed below.

We calculate the quantities,

$$\sum_{i=1}^{13} X_i = (17.1 + 16.5 + \ldots + 11.6) = 175.1$$

$$\bar{X} = \frac{175.1}{13} = 13.47$$

$$\sum_{i=1}^{13} X_i^2 = (17.1^2 + 16.5^2 + \ldots + 11.6^2) = 2411.57$$

Variance by the formula (4.8.1) is,

$$\sigma^2 = \frac{1}{13}\left[2411.57 - (175.1)^2/13\right] = \frac{53.1077}{13} = 4.085$$

Standard deviation by the formula (4.14) is,

$$\sigma = \sqrt{4.085} = 2.021$$

Coefficient of variation by the formula (4.16) is,

$$\text{C.V.} = \frac{2.021}{13.47} \times 100 = 15.004 \text{ per cent}$$

Example 4.7. The prices of wheat at different centres were found to be as follows:

Prices of wheat (Rs/kg)	No. of centres
1.75	3
1.72	2
1.73	4
1.76	5
1.71	6
1.80	2
1.87	7
2.34	1

we can measure the variation in prices of wheat by calculating the standard deviation. The computations are shown in the table below:

Prices of wheat (X)	No of centers (f)	fX	$X - \bar{X}$	$f(X - \bar{X})$	$f(X - \bar{X})^2$
1.75	3	5.25	–.04	–.12	.0048
1.72	2	3.44	–.07	–.14	.0098
1.73	4	6.92	–.06	–.24	.0144
1.76	5	8.80	–.03	–.15	.0045
1.71	6	10.26	–.08	–.48	.0384
1.80	2	3.60	.01	.02	.0002
1.87	7	13.09	.08	.56	.0448
2.34	1	2.34	.55	.55	.3025
Total	30	53.70		00	.4194

$$\bar{X} = \frac{53.70}{30} = 1.79$$

The variance of prices by formula (4.9) is,

$$\sigma^2 = \frac{0.4194}{30} = 0.01398$$

$$\sigma = 0.1182$$

Example 4.8. Following table gives the expenditure per month on food per family of six persons in ten regions:

Regions:	I	II	III	IV	V	VI	VII	VIII	IX	X
Expenditure (X): (Rs)	1090	1270	1260	1200	1170	1080	1000	1310	1210	1130

The variance of above set of observations will be calculated with the help of coding. Subtract 1000 from each value and divide by 10 i.e. take $A = 1000$ and $C = 10$.

We show the computation in the following table.

X	$X - 1000$	$X' = \frac{X - 1000}{10}$	X'^2
1090	90	9	81
1270	270	27	729
1260	260	26	676
1200	200	20	400
1170	170	17	289
1080	80	8	64
1000	00	0	00
1310	310	31	961
1210	210	21	441
1130	130	13	169
Total		172	3810

The variance of the (coded) variable X' by formula (4.8.1) is

$$\sigma'^2 = \frac{1}{10}\{3810 - (172)^2/10\} = \frac{851.6}{10} = 85.16$$

The variance of the uncoded variable by the relation (4.13.3) is

$$\sigma^2 = (10)^2 \times 85.16 = 8516$$

Example 4.9. Monthly wages of employees in a factory are distributed as given below.

Wages (Rs.)	No. of employees
300-400	15
400-500	22
500-600	18
600-700	14
700-800	9
800-900	7
900-1000	5
1000-1100	4

We show the calculation of variance of the given distribution with and without coding. The method of computation is shown in the following table. For coding we have chosen $A = 650$ and $C = 100$.

Wage (Rs.)	Mid-points (X)	Freq. (f)	fX	fX^2	$X' = \frac{X - 650}{100}$	fX'	fX'^2
300-400	350	15	5250	1837500	−3	−45	135
400-500	450	22	9900	4455000	−2	−44	88
500-600	550	18	9900	5445000	−1	−18	18
600-700	650	14	9100	5915000	0	00	00
700-800	750	9	6750	5062500	1	9	9
800-900	850	7	5950	5057500	2	14	28
900-1000	950	5	4750	4512500	3	15	45
1000-1100	1050	4	4200	4410000	4	16	64
Total		94	55800	36695000		−53	387

The variance of the uncoded variable X by formula (4.9.1) is,

$$\sigma^2 = \frac{1}{94}\{36695000 - (55800)^2/94\}$$

$$= \frac{1}{94}\{36695000 - 33123829\}$$

$$= \frac{3571171}{94}$$

$$= 37991.18$$

Now the variance of the coded variable by formula (4.9.1) is,

$$\sigma'^2 = \frac{1}{94}\{387 - (-53)^2/94\}$$

$$= \frac{357.11703}{94}$$

$$= 3.7991173$$

The variance of the uncoded variable by relation (4.13.1) is,

$$\sigma^2 = (100)^2 \times 3.7991173$$

$$= 37991.17$$

This example throws light on two points.

(1) The result obtained by the method of coding is the same as we get without coding the variable.

(2) This shows how much labour is saved through the procedure of coding.

Example 4.10. Find the number of items lying within the interval, mean ± S.D. of the following distribution.

Class intervals:	10-12	13-15	16-18	19-21	22-24	25-27	28-30	31-33	34-36
Frequency:	24	44	65	27	19	14	8	5	4

To find out the number of items within the interval, mean ± S.D., we calculate the mean and S.D. using the method of coding. Take $A = 23$ and $C = 3$.

Class intervals	Continuous intervals	Mid-values	$X' = \frac{X - 23}{3}$	f	fX^2	fX'^2
10-12	9.5-12.5	11	−4	24	−96	384
13-15	12.5-15.5	14	−3	44	−132	396
16-18	15.5-18.5	17	−2	65	−130	260
19-21	18.5-21.5	20	−1	27	−27	27
22-24	21.5-24.5	23	0	19	00	00
25-27	24.5-27.5	26	1	14	14	14
28-30	27.5-30.5	29	2	8	16	32
31-33	30.5-33.5	32	3	5	15	45
34-36	33.5-36.5	35	4	4	16	64
Total			0	210	−324	1222

The mean by formula (3.7) is

$$\bar{X} = 23 + \frac{(-324)}{210} \times 3$$

$$= 23 - 4.628 = 18.372$$

The variance by formula (4.13.2) is,

$$\sigma^2 = 3^2 \times \frac{1}{210}\left\{1222 - \frac{(-324)^2}{210}\right\}$$

$$= \frac{9}{210}\{1222 - 499.886\}$$

$$= \frac{9 \times 722.11}{210} = \frac{6499.03}{210} = 30.95$$

$$\sigma = 5.56$$

$$\overline{X} - \sigma = 18.372 - 5.56 = 12.81$$

$$\overline{X} + \sigma = 18.372 + 5.56 = 23.93$$

The number of items which spread between 12.5 and 15.4 is 44.

Number of items per unit length of interval = 44/3.

The number of items lying between 12.81 and 15.5.

$$= \frac{15.5 - 12.81}{3} \times 44 = \frac{118.36}{3} = 39.45$$

The number of items lying between 21.5 and 24.5 is 19.

Number of items lying between 21.5 and 23.93.

$$= \frac{23.93 - 21.5}{3} \times 19 = \frac{46.17}{3} = 15.39$$

Hence, the number of items lying between 12.81 and 23.93 is equal to sum of items lying between 12.81 to 15.5, 15.5 to 21.5 and 21.5 to 23.93 i.e.

$$= 39.45 + (65 + 27) + 15.39 = 146.84$$

$$= 147$$

POOLED OR COMBINED VARIANCE

By the combined variance of two groups, we mean the variance of the observations of the two groups taken together. Let us consider two groups consisting of N_1 and N_2 observations respectively. Suppose the means of the groups are $\overline{X}_1$ and $\overline{X}_2$ and the variances are σ_1^2 and σ_2^2 respectively. We know by formula (3.8) that the pooled mean of both the groups is,

$$\overline{X}_{12} = \frac{N_1\overline{X}_1 + N_2\overline{X}_2}{N_1 + N_2}$$

The combined variance of the two groups is given by the formula,

$$\sigma_{12}^2 = \frac{N_1\{\sigma_1^2 + (\overline{X}_1 - \overline{X}_{12})^2\} + N_2\{\sigma_2^2 + (\overline{X}_2 - \overline{X}_{12})^2\}}{N_1 + N_2} \qquad (4.17)$$

$$= [N_1(\sigma_1^2 + d_1^2) + N(\sigma_2^2 + d_2^2)]/(N_1 + N_2) \qquad (4.17.1)$$

where $d_1 = \overline{X}_1 - \overline{X}_{12}$ and $d_2 = (\overline{X}_2 - \overline{X}_{12})$.

The advantage of the formula of combined variance is that once we know the individual mean and variance of each group, we can calculate the variance of the combined groups without redoing the entire calculation.

Obviously the combined standard deviation can be found by taking the square root of the combined variance.

Formula (4.17) can be extended for more than two groups easily.

Example 4.11. The mean and variance of scores earned by two groups, one of the boys and the other of the girls, on computation yielded the following results:

$$N_1 = 62 \qquad \overline{X}_1 = 108.2 \qquad \sigma_1^2 = 524.41$$

$$N_2 = 45 \qquad \overline{X}_2 = 105.4 \qquad \sigma_2^2 = 355.32$$

The variance of scores earned by boys and girls taken as one group of students can be calculated by the formula (4.17).

Pooled mean by formula (3.8) is

$$\overline{X}_{12} = \frac{62 \times 108.2 + 45 \times 105.4}{62 + 45} = \frac{11451.4}{107} = 107.02$$

Pooled variance,

$$\sigma_{12}^2 = \frac{62 \times \{524.41 + (108.2 - 107.02)^2\} + 45 \times \{355.32 + (105.4 - 107.02)^2\}}{62 + 45}$$

$$= \frac{62 \times \{524.41 + 1.3924\} + \{355.32 + 2.6244\} \times 45}{107}$$

$$= \frac{48707.24}{107} = 455.21$$

Combined standard deviation,

$$\sigma_{12} = 21.34$$

CONCLUDING REMARKS

A measure of dispersion, specially the variance, is the backbone of statistics. As a matter of fact, statistics involves variance almost in every study in one way or the other. Most of the surveys or experiments are considered as a study of sample units. Hence the formulae for sampling are mostly used. In cases, where no inference has to be drawn for a larger group other than the observations under study, we should use the formulae given for the population. Moreover, all the formulae except variance are not affected whether we consider a population or a sample. Of course, the interpretation of values has to be made accordingly. Readers will come across the use of variance vis-a-vis the standard deviation in the chapters ahead.

QUESTIONS AND EXERCISES

1. What does dispersion indicate about the data? Why is this of great importance?
2. Which measure of dispersion do you consider the best and why?
3. Range gives very little information about data, still used widely, why?
4. In what respects is the coefficient of variation superior to other measures of dispersion?

Excersises

5. Define and discuss the following terms.
   (a) Quartile deviation.
   (b) Mean deviation.
   (c) Variance.
   (d) Coefficient of variation.
6. What are the requirements of a good measure of dispession?
7. Explain why standard deviation is considered superior to other measures of dispersion?
8. What is the effect of subtracting 25 from each observation and dividing the result of each observation by 5 on the following.
   (a) Mean deviation
   (b) Range
   (c) Standard deviation
   (d) Coefficient of variation
9. Distinguish between absolute and relative measures of dispersion.
10. Fill in the blanks:
    (a) Mean deviation is minimum about ________________.
    (b) Variance is zero when ________________.
    (c) Range is zero when ________________.
    (d) Coefficient of variation is infinity when ________________.
    (e) Coefficient of variation is zero when ________________.
11. Discuss the relative merits of range, standard deviation and mean deviation. *(M. Com., Saugar, 1963)*
12. What is meant by dispersion? What are the methods of computing measures of dispersion? Illustrate the practical utility of these methods. *(M. Com., Alld., 1956; B. Com., Agra, 1958; M.A., Vikram, 1961)*
13. Define 'mean deviation'. How does it differ from standard deviation?*(C.A., 1968)*
14. Following figures give the production of non-fatty dry milk during the twelve months of 1975.

Months:	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
Production (million Lbs.):	83.5	81.6	95.8	111.5	131.4	126.5	98.7	76.2	53.2	50.3	49.3	67.1

Calculate (i) range, (ii) mean deviation about mean, (iii) variance and (iv) coefficient of variation.

15. The following table gives the number of branches and number of plants in a patch of field.

No. of branches (X)	No. of plants (f)
2	16
3	13
4	19
5	12
6	5
7	4

Calculate the mean deviation about median and variance of the number of branches

16. The following table gives the distribution of monthly expenditure per head of the residents of a village.

Monthly expenditure (Rs.)	No. of persons
75-71	20
70-66	17
65-61	15
60-56	14
55-51	13
50-46	12
45-41	14
40-36	15
35-31	10

For the given distribution, calculate (i) quartile deviation, (ii) mean deviation about mean and (iii) standard deviation.

17. Two persons participated in five shooting competitions and were able to hit the target correctly out of fifteen shots as given below:

Competitor A	Competitor B
6	12
12	15
12	7
10	7
7	4

Find which of the competitors is more consistent in shooting performance.

18. The mean and variance of marks of a group of 120 students are 38.7 and 510.76, respectively. The mean and variance of marks of another group of 100 students are 54.4 and 412.09, respectively. Calculate the standard deviation of both the groups taken together.
19. Following table gives the distribution of the age of lady teachers of a school as revealed by records.

Age groups (years)	No. of lady teachers
15-19	3
20-24	13
25-29	21
30-34	15
35-39	5
40-44	4
45-49	2

Compute (i) quartile deviation, (ii) mean deviation about mean, (iii) coefficient of variation and (iv) number of teachers between the age of 26 and 33 years.

20. From the data given below, giving arithmetic average and standard deviation of four sub-groups, calculate the average and standard deviation of the whole group.

not cause any change in the other variable, according to any rule. Scatter diagram pertaining to independent variables is shown in Fig. 13.2.

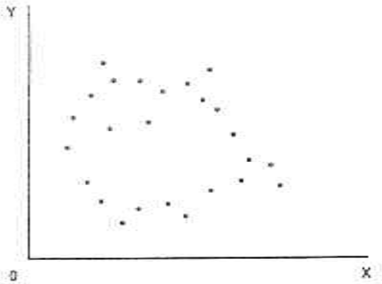


Fig. 13.2 Scatter Diagram

From the above figure, it is easily verifiable that if any line is drawn through the plotted points, not more than two points will be lying on the line and most of the other points will be at a considerable distance from this line.

Once it is decided on the basis of prior information or scatter diagram that the two variables are linearly related, the problem arises on deciding which of the many possible lines is the best fitted line. To cope with this problem, mathematical basis leads to the most logical and accurate solution. The least square method is the most widely accepted method of fitting a straight line and is discussed here adequately.

LEAST SQUARE METHOD OF FITTING A REGRESSION LINE

The equation for a regression line of Y on X for the population is given as

$$Y = \alpha + \beta X + e \qquad (13.1)$$

Equation (13.1) is also known as the *mathematical model* for linear regression. The main difference between the cartesian equation of a line and a regression line is that a regression line is a probabilistic model which enables one to develop procedures for making inferences about the parameters α and β of the model. In this model, the expected value of Y is a linear function of X, but for fixed X, the variable Y differs from its expected value by a random amount. As a special case, the form $y = \alpha + \beta x$ is called the *deterministic model.* In this model, the actual observed value of y is a linear function of x. In this equation α is the intercept which the line cuts on the axis of Y and β is the slope of the line. β is also called the *regression coefficient* and is defined as, "*β is the measure of change in the dependent variable (Y) corresponding to a unit change in the independent variable (X)*". β is often written as β_{YX} to indicate that it is the regression coefficient of Y on X. In case no suffix is attached to β, it is considered by itself. β can take any real value within the range $-\infty$ to ∞.

Suppose the regression line given by (13.1) is to be fitted on the basis of n pairs of sample observations, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Each pair (x_i, y_i) for $i = 1, 2, \ldots, n$ will satisfy the regression line (13.1).

Thus, $$y_i = \alpha + \beta x_i + e_i \qquad (13.1.1)$$

or $$e_i = (y_i - \alpha - \beta x_i) \qquad (13.1.2)$$

e_i may be positive or negative in case y_i is greater than or less than $(\alpha + \beta x_i)$ respectively. Whether the error is positive or negative, it does not matter as an error is after all an error. So to avoid the sign of error and confining to its magnitude only, square both sides of (13.1.2) and take the sum over n pairs of observations. This gives,

$$\Sigma_i\, e_i^2 = \Sigma_i\, (y_i - \alpha - \beta x_i)^2 \qquad (13.1.3)$$

In Legendre's principle of least squares, the quantity which is minimized is the *residuals* or *error sum of squares.* Here the assumption is that each e_i is normally distributed with mean zero and variance σ_e^2. Thus, the quantity which is to be minimized here $\Sigma_i\, e_i^2$. Let us denote this quantity by Q. Hence,

$$Q = \Sigma_i\, (y_i - \alpha - \beta x_i)^2 \qquad (13.1.4)$$

To get the least square estimates of α and β, so that Q is minimum, differentiate Q partially with respect to α and β respectively and equate to zero. Also replace α and β by their estimated values, say a and b respectively. Thus, we get two equations as given below. These equations are called normal equations.

$$\frac{\partial Q}{\partial \alpha} = -\,2\ \Sigma_i\, (y_i - a - bx_i) = 0 \qquad (13.2)$$

$$\frac{\partial Q}{\partial \beta} = -\,2\ \Sigma_i\ (y_i - a - bx_i)\, x_i = 0 \qquad (13.3)$$

On rearranging we get,

$$na + b\,\Sigma_i\ x_i = \Sigma_i\, y_i \qquad (13.4)$$

$$a\,\Sigma_i\ x_i + b\,\Sigma_i\ x_i^2 = \Sigma_i\ y_i x_i \qquad (13.5)$$

From (13.4)

$$a + b\,\frac{1}{n}\,\Sigma_i\ x_i = \frac{1}{n}\,\Sigma_i\ y_i \qquad (13.6)$$

$$a + b\bar{x} = \bar{y}$$

or $$a = (\bar{y} - b\bar{x}) \qquad (13.7)$$

Substituting the value of a from (13.7) in (13.5), we get,

$$\left(\frac{1}{n}\,\Sigma_i\ y_i\ -\ b\,\frac{1}{n}\,\Sigma_i\ x_i\right)\Sigma_i\ x_i\ +\ b\ \Sigma_i\ x_i^2 = \Sigma_i\ y_i\ x_i$$

or $$b\,\left\{\Sigma_i\, x_i^2\ -\ \frac{1}{n}\ (\Sigma_i\ x_i)^2\right\} = \Sigma_i\ x_i\ y_i\ -\ \frac{1}{n}\ (\Sigma_i\ x_i)\ (\Sigma_i\ y_i)$$

or $$b = \frac{\Sigma_i\, x_i\, y_i - \frac{1}{n}\,(\Sigma_i\, x_i)\,(\Sigma_i\, y_i)}{\Sigma_i\, x_i^2 - \frac{1}{n}\,(\Sigma_i\, x_i)^2} \qquad (13.8)$$

Expression (13.8) can easily be written as,

$$b = \frac{\Sigma_i\,(x_i - \bar{x})(y_i - \bar{y})}{\Sigma_i\,(x_i - \bar{x})^2} \qquad (13.8.1)$$

Suppose $x_i - \bar{x} = u_i$ and $y_i - \bar{y} = v_i$, under this transformation,

$$b = \frac{\Sigma_i\, u_i\, v_i}{\Sigma_i\, u_i^2} \qquad (13.8.2)$$

If we divide the numerator and denominator by n in (13.8.1), we get

$$b = \frac{\text{cov}\,(X, Y)}{\text{var}\,(X)} \qquad (13.8.3)$$

or $$b = \frac{s_{XY}}{s_X^2} \qquad (13.8.4)$$

b is the estimated regression coefficient of Y on X and is also symbolised as b_{YX}.

With the help of the formula (13.36.1) given ahead we can write,

$$s_{XY} = r\, s_X\, s_Y$$

Substituting the value of s_{XY} in terms of r, s_X and s_Y in (13.8.4) we obtain,

$$b_{YX} = r\,\frac{s_Y}{s_X} \qquad (13.8.5.)$$

where r is the sample correlation coefficient between X and Y.

Properties of Regression Coefficient (1) It can take any value between $-\infty$ and ∞. Its sign is same as that of s_{XY} i.e. cov (X, Y).

Further, if the whole population has been studied, i will vary from 1 to N for all the N units of the population. In this situation we get population regression coefficient β directly, for which the formula is

$$\beta = \frac{\sigma_{XY}}{\sigma_X^2} \qquad (13.9)$$

In case of population regression coefficient β of Y on X, which can elaborately be specified as β_{YX} we can express it as,

$$\beta_{YX} = \rho\,\frac{\sigma_Y}{\sigma_X} \qquad (13.9.1)$$

where ρ is the population correlation coefficient between X and Y.

As a and b are the estimated values of α and β respectively, the equation of the estimated regression line is

$$\hat{Y} = a + bX \qquad (13.10)$$

where, the hat (^) over Y indicates that Y is an estimated value. Substituting the value of a from (13.7), the line of best fit is,

$$\hat{Y} = (\bar{y} - b\bar{x}) + bx \qquad (13.10.1)$$

or $$(\hat{Y} - \bar{y}) = b\,(X - \bar{x}) \qquad (13.10.2)$$

Prediction Equation The regression line is also known as *prediction equation*. Once the constants a and b are calculated, there remains two unknown variables in the regression equation viz., Y and X. Moreover, we know Y depends on X in the case of regression equation of Y on X. Under the presumption that the trend of change in Y corresponding to X remains the same, the value of Y can be estimated for any value of X. But such a presumption rarely holds good for a very wide range of X values. Hence, the fitted regression line gives a better estimate of Y for a given value of X, which is within the range of X values, taken into consideration at the time of calculation of a and b, or not much beyond the values of X involved in the calculations. For example, we know that the crop yield increases with the increases in the quantity of fertilizers applied in the field. But beyond certain fertilizer-dose, the increase in yield is negligible. Hence the estimation of the yield of a crop for a fertilizer dose should be restricted only for doses within certain limits. In industry, the production of a product depends on the consumption of electricity. But it does not mean that if we go on increasing the consumption of electricity, the production of a product of a product will keep on increasing in the same proportion. It holds true only up to a certain limit because many other factors besides consumption of electricity affect the production. The fitting of a regression line through actual data is given below.

افتراضي

اسئلة

Example 13.1. The table below, gives the data regarding industrial consumption index of electricity and industrial production index (taking 1960 = 100) from 1951 to 1970.

Year	Index of industrial consumption of electricity (X)	Index of industrial production (Y)
1951	36.6	54.8
1952	39.5	57.2
1953	43.4	58.1
1954	47.6	63.4
1955	53.4	72.5
1956	58.5	78.4
1957	66.1	82.7
1958	74.9	84.4
1959	87.1	90.3

1960	100.0	100.0
1961	115.1	109.2
1962	131.7	119.8
1963	150.0	129.7
1964	162.6	140.8
1965	176.3	153.8
1966	190.4	153.2
1967	209.4	152.6
1968	233.6	163.0
1969	255.7	175.3
1970	271.4	184.3

It is known that the production (Y) depends on the consumption of electricity (X) and the relation between the two variables is linear. Hence, a regression line can be fitted to the bivariate data by calculating the values of a and b. First make the following calculations.

$$\Sigma_i x_i = (36.6 + 39.5 + \cdots + 271.4) = 2503.3$$

$$\Sigma_i y_i = (54.8 + 57.2 + \cdots + 184.3) = 2223.5$$

$$\Sigma_i x_i^2 = (36.6^2 + 39.5^2 + \cdots + 271.4^2) = 425211.85$$

$$\therefore \quad \bar{x} = \frac{2503.3}{20} = 125.17$$

$$\bar{y} = \frac{2223.5}{20} = 111.18$$

Also $\quad n = 20$

$$\Sigma_i x_i y_i = (36.6 \times 54.8 + 39.5 \times 57.2 + \cdots + 271.4 \times 184.3)$$

$$= 339769.87$$

Using the formula (13.8) we get

$$b = \frac{339769.87 - \frac{1}{20}(2503.3)(2223.5)}{425211.85 - \frac{1}{20}(2503.3)^2}$$

$$= \frac{61465.49}{111886.31}$$

$$= 0.55$$

and from (13.7)

$$a = 111.18 - 0.55 \times 125.17$$

$$= 42.34$$

Hence the estimated equation of the regression line is,

$$\hat{Y} = 42.34 + 0.55\,X$$

Given the value of $X = 150$, we can estimate the value of Y from the estimated equation, that is

$$\hat{Y} = 42.34 + 0.55 \times 150 = 124.84$$

If we see the data, we find that the actual value of Y for $X = 150$ is 129.7. The difference between the actual and estimated value is not much. Hence, the line seems to be a good fit. Again, if we want the projection for the increased consumption of electricity, i.e. $X = 350$, the estimated value of Y is,

$$\hat{Y} = 42.34 + 0.55 \times 350$$

$$= 234.84$$

From this we infer, that, if the industrial consumption index of electricity rises to the level of 350, production will rise to the level of 234.84.

Example 13.2. A departmental store gives in-service training to salesmen followed by a test. It is experienced that the performance regarding sales of any salesman is linearly related to the scores secured by him. The following data give test scores and sales made by nine salesmen during fixed period.

أداء
يحرز

Test scores (X) :	16	22	28	24	29	25	16	23	24
Sales ('00 Rs) (Y):	35	42	57	40	54	51	34	47	45

The sales Y of any salesman are considered to depend on his ability as judged by his test scores X. The regression line of Y on X can be fitted to the data in the following manner.

$$\Sigma_i x_i = 207 \text{ and } \Sigma_i y_i = 405 \text{ for } i = 1, 2, \ldots, 9.$$

$$\bar{x} = \frac{207}{9} = 23 \text{ and } \bar{y} = \frac{405}{9} = 45$$

Since the means of x and y observations are whole numbers, it is preferable to use the formula (13.8.1). To show the calculations clearly, it is better to prepare the following table.

Observation number	x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(x - \bar{x})(y - \bar{y})$
1	16	35	−7	−10	49	70
2	22	42	−1	−3	1	3
3	28	57	5	12	25	60
4	24	40	1	−5	1	−5
5	29	54	6	9	36	54
6	25	51	2	6	4	12
7	16	34	−7	−11	49	77
8	23	47	0	2	00	00
9	24	45	1	0	1	00
Total	207	405	00	00	166	271

From (13.8.1)

$$b = \frac{271}{166}$$

$$= 1.63$$

Hence, the estimated regression line as given by (13.10.2) is,

$$(\hat{Y} - 45) = 1.63\,(X - 23)$$

$$\hat{Y} = 7.51 + 1.63\,X$$

The predicted sales due to a salesman having the scores = 30 is,

$$\hat{Y} = 7.51 + 1.63 \times 30$$

$$= 56.41$$

Thus, the estimated sale is Rs. 56.41.

Regression Line of X on Y Often we come across situations in which two variables (Y and X) are such that not only Y depends on X but X also depends on Y. For example, the heights and weights of people are two variables where heights of people depend on weights and weights depend on heights. In such a case we can find not only the regression line of Y on X but also of X on Y. Suppose the regression line of X on Y is,

$$X = \alpha_1 + \beta_1\,Y + e_1 \tag{13.11}$$

The parameters α_1 and β_1 can be estimated in the same way as α and β in (13.1). Instead of repeating the derivation, it will be worthwhile to write directly the estimated values of α_1 and β_1, say, a_1 and b_1 (by interchanging the variable Y by X and X by Y in the formulae for a and b respectively). Thus, the estimates are,

$\Sigma X = n\alpha_1 + \beta_1 \Sigma y$

$\Sigma xy = \alpha_1 \Sigma y + \beta \Sigma y^2$

$$a_1 = (\bar{x} - b_1\,\bar{y}) \tag{13.12}$$

$$b_1 = \frac{\Sigma_i\,(x_i - \bar{x})\,(y_i - \bar{y})}{\Sigma_i\,(y_i - \bar{y})^2} \tag{13.13}$$

$$= \frac{\Sigma_i x_i y_i - (\Sigma_i x_i)\,(\Sigma_i y_i)\,/\,n}{\Sigma_i y_i^2 - (\Sigma_i y_i)^2/n} \tag{13.13.1}$$

$$= \frac{s_{XY}}{s_Y^2} \tag{13.13.2}$$

We can express b_1, which is the estimated regression coefficient of X on Y and symbolically denoted as b_{XY}. In terms of r, s_X, s_Y, as in case of (13.8.5), we can write,

$$b_{XY} = r\,\frac{s_X}{s_Y} \tag{13.13.3}$$

where r is the sample correlation coefficient between X and Y.

The equation of the estimated regression line of X on Y is,

$$(\hat{X} - \bar{x}) = b_1\,(Y - \bar{y}) \tag{13.14}$$

$\hat{X} = (\bar{x} - b\bar{y}) + bY$

$(\hat{x} - \bar{x}) = b(y - \bar{y})$

Also, the population regression coefficient of X on Y may be given as follows,

$$\beta_1 = \frac{\sigma_{XY}}{\sigma_Y^2} \tag{13.15}$$

The population regression coefficient β_1 of X on Y, which is often symbolised as β_{XY}, can be expressed as,

$$\beta_{XY} = \rho\,\frac{\sigma_X}{\sigma_Y} \tag{13.15.1}$$

where ρ is the population correlation coefficient between X and Y.

It is trivial to prove that the two regression lines given by (13.10.2) and (13.14) intersect each other at a point having coordinates $(\bar{x}, \bar{y})$, i.e. at the mean of two variables.

Example 13.3. Using the data and partial calculations of example (13.1), we fit in the regression line.

$$(\hat{X} - \bar{x}) = b\;(Y - \bar{y})$$

First, we calculate,

$$\Sigma_i y_i^2 = (54.8^2 + 57.2^2 + \cdots + 184.3^2)$$

$$= 281790.03$$

Now from (13.13.1),

$$b_1 = \frac{61465.49}{281790.03 - \frac{1}{20}\,(2223.5)^2}$$

$$= \frac{61465.49}{34592.42}$$

$$= 1.78$$

Hence, the required equation of regression line is,

$$(\hat{X} - 125.17) = 1.78\,(Y - 111.18)$$

or

$$\hat{X} = 1.78\,Y - 72.73$$

The value of X can be estimated for any given value of Y, in the manner followed in example (13.1).

Regression Coefficient from Coded Data Coding of data has already been discussed in Chapter 4. We have to calculate the variance and covariance for calculating β or b and the methodology given in Chapter 4 is applicable. Coding is useful to reduce the labour of calculations. This not only saves time but also reduces the chances of errors in calculations. Since ages, coding has been a very popular device as a short cut method of calculations, but now it is loosing its importance with the increasing use of modern electronic calculators. Still in many examinations calculators are not provided and hence it is worthwhile to discuss it here.

Let a constant c_1 be subtracted from X observations and c_2 from Y observations. Then the reduced values of X and Y are divided by d_1 and d_2, respectively. Thus, the coded variate values are

$$dx_i = \frac{x_i - c_1}{d_1} \text{ and } dy_i = \frac{y_i - c_2}{d_2}$$

for $i = 1, 2, ..., n$.

For calculating the regression coefficient from coded data, prepare the Table 13.1 given on next page.

Under the transformation,

$$\bar{dx} = \frac{1}{n} \Sigma_i dx_i = \frac{1}{n} \Sigma_i \frac{(x_i - c_1)}{d_1}$$

$$= \frac{1}{d_1} \left\{ \frac{1}{n} \Sigma_i x_i - \frac{1}{n} \Sigma_i c_1 \right\}$$

$$\bar{dx} = \frac{\bar{x} - c_1}{d_1} \qquad (13.16)$$

or

$$\bar{x} = d_1 \bar{dx} + c_1 \qquad (13.16.1)$$

Similarly,

$$\bar{dy} = \frac{\bar{y} - c_2}{d_2} \qquad (13.17)$$

or

$$\bar{y} = d_2 \bar{dy} + c_2 \qquad (13.17.1)$$

Now the regression coefficient from coded observations is,

$$b_c = \frac{\Sigma_i (dx_1 - \bar{dx}) \ (\bar{dy}_i - dy)}{\Sigma_i (dx_i - \bar{dx})^2} \qquad (13.18)$$

$$= \frac{\Sigma_i dx_i dy_i - (\Sigma_i dx_i) \ (\Sigma_i dy_i)/n}{\Sigma_i dx_i^2 - (\Sigma_i dx_i)^2/n} \qquad (13.18.1)$$

Substituting the transforms for dx_i and dy_i etc. we get,

$$b_c = \frac{\Sigma_i \left(\frac{x_i - c_1}{d_1} - \frac{\bar{x} - c_1}{d_1} \right) \left(\frac{y_i - c_2}{d_2} - \frac{\bar{y} - c_2}{d_2} \right)}{\Sigma_i \left(\frac{x_i - c_1}{d_1} - \frac{\bar{x} - c_1}{d_1} \right)^2}$$

$$= \frac{\frac{1}{d_1 d_2} \Sigma_i (x_i - \bar{x}) \ (y_i - \bar{y})}{\frac{1}{d_1^2} \Sigma_i (x_i - \bar{x})^2} \qquad (13.18.2)$$

$$= \frac{d_1}{d_2} \ b_{YX} \qquad (13.18.3)$$

Table 13.1

x	y	$x - c_1$	$y - c_2$	$dx = \frac{x - c_1}{d_1}$	$dy = \frac{y - c_2}{d_2}$	$dx\ dy$	d^2x	d^2y
x_1	y_1	$x_1 - c_1$	$y_1 - c_2$	$dx_1 = \frac{x_1 - c_1}{d_1}$	$dy_1 = \frac{y_1 - c_2}{d_2}$	$dx_1\ dy_1$	d^2x_1	d^2y_1
x_2	y_2	$x_2 - c_1$	$y_2 - c_2$	$dx_2 = \frac{x_2 - c_1}{d_1}$	$dy_2 = \frac{y_2 - c_2}{d_2}$	$dx_2 - dy_2$	d^2x_2	d^2y_2
. . . . x_i	. . . . y_i	. . . . $x_i - c_1$	. . . . $y_i - c_2$	. . . . $dx_i = \frac{x_i - c_1}{d_1}$	. . . . $dy_i = \frac{y_i - c_2}{d_2}$	. . . . $dx_i\ dy_i$	. . . . d^2x_i	. . . . d^2y_i
. . . . x_n	. . . . y_n	. . . . $x_n - c_1$	. . . . $y_n - c_2$	. . . . $dx_n = \frac{x_n - c_1}{d_1}$	. . . . $dy_n = \frac{y_n - c_2}{d_2}$	. . . . $dx_n\ dy_n$	. . . . d^2x_n	. . . . d^2y_n
Total $\Sigma_i x_i$	$\Sigma_i y_i$	$\Sigma_i x_i - nc_1$	$\Sigma_i y_i - nc_2$	$\Sigma_i dx_i$	$\Sigma_i dy_i$	$\Sigma_i dx_i dy_i$	$\Sigma_i d^2 x_i$	$\Sigma_i d^2 y_i$

where i varies from 1 to n.

or $$b_{YX} = \frac{d_2}{d_1}\, b_c \qquad (13.18.4)$$

From (13.18.4) it is clear that the regression coefficient is independent of the change of origin but is affected by the change of scale. To obtain the value of regression coefficient for original observations, from regression coefficient based on coded data, it should be multiplied by d_2/d_1.

Note : (i) c_1, c_2, d_1 and d_2 need not necessarily be different. Any of them may be equal if found appropriate. d_1 and d_2 should never be taken as zero because the coded value will become infinity and d_2/d_1 will become an indeterminate quantity.

(ii) Often we like to do only one operation, i.e. we subtract a constant and no divisor is taken. In this situation $d_1 = d_2 = 1$. Hence the value of b_{YX} will directly be equal to the value obtained from coded data. When the divisors are taken but no constants are subtracted from the observed values, then $c_1 = c_2 = 0$.

Example 13.4. The table below gives the total grain production and cereal production of cereals in Lakh tonnes (rounded figures) for nine years.

Year:	1	2	3	4	5	6	7	8	9
Total grain production (y):	400	440	480	550	620	650	660	740	760
Cereal production (x):	50	60	70	85	95	100	105	115	120

We know that total production is directly proportional to the cereal production. Hence a regression line of total production (y) on cereal production (x) can be fitted. Since the figures are large enough, we will use coding of data. Therefore, subtract 620 from each value of y. Also divide each subtracted figure by 10. Also subtract 95 from each value of x and divide the reduced figure by 5. In other words, $c_2 = 620$, $c_1 = 95$, $d_2 = 10$ and $d_1 = 5$. The calculations for fitting of regression line, using the coding technique, are presented in the following table.

Year	Total production y	Cereal production x	$\frac{y-620}{10}$ $= dy$	$\frac{x-95}{5}$ $= dx$	$dxdy$	dx^2
1.	400	50	−22	−9	198	81
2.	440	60	−18	−7	126	49
3.	480	70	−14	−5	70	25
4.	550	85	−7	−2	14	4
5.	620	95	0	0	00	00
6.	650	100	3	1	3	1
7.	660	105	4	2	8	4
8	740	115	12	4	48	16
9.	760	120	14	5	70	25
Total			−28 $= \Sigma_i dy_i$	−11 $= \Sigma_i dx_i$	537 $= \Sigma_i dx_i dy_i$	205 $= \Sigma_i d^2 x_i$

From (13.6.1),

$$\bar{x} = 95 - \frac{11}{9} \times 5$$
$$= 95 - 6.1$$
$$= 88.9$$
$$\bar{y} = 620 - \frac{28}{9} \times 10$$
$$= 620 - 31.1$$
$$= 588.9$$

Now the regression coefficient from (13.18.1) is,

$$b_c = \frac{537 - \frac{(-28)(-11)}{9}}{205 - \frac{(-11)^2}{9}}$$
$$= \frac{537 - 34.22}{205 - 13.44}$$
$$= \frac{502.78}{191.56}$$
$$= 2.6247$$

From (13.18.4),

$$b_{YX} = \frac{10}{5} \times 2.6247$$
$$= 5.2494$$
$$= 5.25$$

Equation of the estimated regression line is,

$$(\hat{Y} - 588.9) = 5.25\,(X - 88.9)$$
$$\hat{Y} = 588.9 + 5.25\,X - 466.72$$
$$= 122.18 + 5.25X$$

From the fitted regression line, we can estimate the total production for the cereal production of 90 Lakh tonnes.

Thus,

$$\hat{Y} = 122.18 + 5.25 \times 90$$
$$= 122.18 + 472.50$$
$$= 594.68$$
$$\doteq 595 \text{ Lakh tonnes.}$$

Test of Significance of Regression Parameters The estimators of α and β have already been given by (13.7) and (13.8) respectively based on n paired observations. The experimenter is interested to test whether the parameters α and β, involved in the regression line, are of practical relevance or not. To do so, we have to test the hypotheses $\beta = 0$ and $\alpha = 0$. These hypotheses can be tested provided we know the distribution of b and a. The estimators a and b are random variables such that,

$$b \sim N(\beta, \sigma_b^2) \qquad (13.19)$$

and
$$a \sim N(\alpha, \sigma_a^2) \qquad (13.20)$$

where
$$\sigma_b^2 = \sigma_e^2 / \Sigma_i (x_i - \bar{x})^2 \qquad (13.21)$$

and
$$\sigma_a^2 = \sigma_e^2 \left\{ \frac{1}{n} + \frac{\bar{x}^2}{\Sigma_i (x_i - \bar{x})^2} \right\} \qquad (13.22)$$

The variances σ_b^2 and σ_a^2 are not known and are thus estimated from sample observations. In (13.21) and (13.22) if σ_e^2 is estimated, say, by s_e^2, then the estimators s_b^2 and s_a^2 and σ_b^2 and σ_a^2 respectively, are obtained. There are only n paired observations (x_i, y_i) where $i = 1, 2, ..., n$ and s_e^2 is to be estimated from $\Sigma_i e_i^2$. The expression (13.1.3) for $\Sigma_i e_i^2$ involves two parameters α and β which are estimated by a and b. The estimate of $\Sigma_i e_i^2$ is,

$$\Sigma_i e_i^2 = \Sigma_i (y_i - a - bx_i)^2$$

The appropriate divisor to obtain s_e^2 from $\Sigma_i e_i^2$ is $(n-2)$. Thus,

$$s_e^2 = \frac{1}{(n-2)} \; \Sigma_i (y_i - a - bx_i)^2 \qquad (13.23)$$

The divisor $(n-2)$ is used because the two parameters α and β are estimated resulting into a loss of two d.f. Moreover, s_e^2 so obtained is an unbiased estimate of σ_e^2. s_e^2 is also called the mean square error (MSE).

With the help of simple algebra, it is easy to shown that

$$s_e^2 = \frac{1}{(n-2)} \; \{\Sigma_i (y_i - \bar{y})^2 - b\, \Sigma_i (x_i - \bar{x})(y_i - \bar{y})\} \qquad (13.23.1)$$

Putting $x_i - \bar{x} = u_i$ and $y_i - \bar{y} = v_i$, we get

$$s_e^2 = \frac{1}{(n-2)} \; \{\Sigma_i v_i^2 - b \; \Sigma_i u_i v_i\} \qquad (13.23.2)$$

In (13.23.2), the quantity $b\, \Sigma_i u_i v_i$ is called the sum of squares due to regression and the quantity within the braces is called the residual sum of squares.

Substituting the value of b as $\Sigma_i u_i v_i / \Sigma_i u_i^2$, we obtain,

$$s_e^2 = \frac{1}{(n-2)} \left\{ \Sigma_i v_i^2 = \frac{(\Sigma_i u_i v_i)^2}{\Sigma_i u_i^2} \right\} \qquad (13.23.3)$$

Now using the estimate s_e^2 for σ_e^2, we get the estimates s_b^2 and s_a^2 for σ_b^2 and σ_a^2 respectively, viz.,

$$s_b^2 = \frac{s_e^2}{\Sigma_i (x_i - \bar{x})^2} \qquad (13.24)$$

$$= \frac{s_e^2}{\Sigma_i u_i^2} \qquad (13.24.1)$$

and
$$s_a^2 = s_e^2 \left\{ \frac{1}{n} + \frac{\bar{x}^2}{\Sigma_i (x_i - \bar{x})^2} \right\} \qquad (13.25)$$

$$= s_e^2 \left\{ \frac{1}{n} + \frac{\bar{x}^2}{\Sigma_i u_i^2} \right\} \qquad (13.25.1)$$

The test for testing the hypothesis,

$$H_0 : \beta_{YX} = 0 \text{ vs. } H_1 : \beta_{YX} \neq 0$$

will be as follows:

H_0 against H_1 can be tested by t-statistic which is given as:

$$t_{n-2} = \frac{b}{s_b} \qquad (13.26)$$

where the suffix $(n-2)$ indicates the degrees of freedom for t and s_b is the standard error of b which is the square root of s_b^2 given by (13.24). If $t_{n-2} > t_\alpha$, reject H_0 where t_α is the table value of t for $(n-2)$ d.f. at prefixed level of significance α. If $t_{n-2} \leq t_\alpha$, accept H_0. Accepting H_0 means that the regression co-efficient of Y on X has no practical significance i.e. the change in Y corresponding to a unit change in X is practically meaningless.

Similarly we can perform the test for testing the hypothesis.

$$H_0 : \alpha = 0 \text{ vs. } H_1 : \alpha \neq 0$$

by using the statistic,

$$t_{n-2} = \frac{a}{s_a} \qquad (13.27)$$

where s_a is the standard deviation of a, which is the square root of s_a^2 given by (13.25). If the value of $\alpha = 0$ is tenable, then it would be desirable to use the regression equation $Y = \beta X + e$, i.e. the line is passing through the origin.

Alternative Test. The hypothesis

$$H_0 : \beta_{YX} = 0 \text{ vs. } H_1 : \beta_{YX} \neq 0$$

can also be tested by the F-test using the analysis of variance technique. For this the ANOVA table is as given below:

Table 13.2: ANOVA table

Source	d.f.	S.S.	M.S.	F-value
Due to regression	1	$b\,\Sigma_i u_i v_i$	$\frac{b\,\Sigma_i u_i v_i}{1}$	$\frac{b\,\Sigma_i u_i v_i}{s_e^2} = F$
Deviation from regression	$(n-2)$	$(\Sigma_i v_i^2 - b\,\Sigma u_i v_i)$	$\frac{(\Sigma_i v_i^2 - b\,\Sigma u_i v_i)}{(n-2)} = s_e^2$	
Total	$(n-1)$	$\Sigma_i v_i^2$		

If $F > F_{\alpha,(1,\,n-2)}$, reject H_0 otherwise accept H_0. The physical interpretation for rejection or acceptance of H_0 remains the same as given with t-test of H_0.

Confidence Limits for Regression Parameters Following the same principle as given in Chapter 9, an expression analogous to (9.10) is given for $(1-\alpha)$ per cent confidence limits.

Confidence limits for β_{YX} are,

$$b \pm s_b\, t_{\alpha,(n-2)} \qquad (13.28)$$

where $t_{\alpha,(n-2)}$ is the table value for two-tailed t-test at α level of significance and for $(n-2)$ degree of freedom. b and s_b are as given earlier.

Again $(1-\alpha)$ per cent confidence limits for α, the intercept, are,

$$a \pm s_a\, t_{\alpha,(n-2)} \qquad (13.29)$$

where a and s_a are as given earlier. $t_{\alpha,(n-2)}$ has been explained just before.

Example 13.5. We give the tests of significance for the data given in example (13.1).

Test of significance of regression coefficient amounts to testing of hypothesis,

$$H_0 : \beta_{YX} = 0 \text{ vs. } H_1 : \beta_{YX} \neq 0$$

To test H_0, we make use of the test statistic,

$$t_{n-2} = \frac{b}{s_b}$$

We will make use of all the calculations made in example (13.1).

To perform the test, we need b and s_b. We have $b = 0.55$ and now calculate s_b. From (13.23.2),

$$s_e^2 = \frac{1}{18}\left[\left\{281790.03 - \frac{(2223.5)^2}{20}\right\} - 0.55 \times 61465.49\right]$$

Since $\Sigma_i y_i^2 = 54.8^2 + 57.2^2 \cdots + 184.3^2 = 281790.03$

$$s_e^2 = \frac{1}{18}[34592.42 - 33806.02]$$

$$= \frac{786.40}{18}$$

$$= 43.69$$

From (13.24),

$$s_b^2 = \frac{43.69}{111886.31}$$

$$= 0.00039$$

or $s_b = 0.0197$

The statistic,

$$t_{18} = \frac{0.55}{0.0197}$$

$$= 27.92$$

The table value of t for 18 d.f. and 5 per cent level of significance is 2.101. Since the table value of t is less than the calculated t-value, we reject H_0 which means that the regression coefficient plays a significant role in determining Y through X.

95 per cent confidence interval for β_{YX} from (13.28) is,

$$0.55 \pm 0.0197 \times 2.101$$

i.e. 0.55 ± 0.0414

Upper limit for β_{YX} is 0.5914 and lower limit is 0.5086.

Again the hypothesis, whether or not the line passes through the origin, is equivalent to testing,

$$H_0 : \alpha = 0 \text{ vs. } H_1 : \alpha \neq 0$$

H_0 can be tested by the statistic,

$$t_{n-2} = \frac{a}{s_a}$$

we know, $a = 42.34$

Now we calculate s_a by (13.25)

$$s_a^2 = 43.69\left\{\frac{1}{20} + \frac{125.17^2}{111886.31}\right\}$$

$$= 43.69\ (0.05 + 0.14)$$

$$= 8.30$$

or $s_a = 2.88$

Now, $t_{18} = \frac{42.34}{2.88}$

$$= 14.70$$

The calculated value of t is greater than the table value of $t_{0.05, 18} = 2.101$. Hence, we reject H_0. This means that the regression line does not pass through the origin.

95 per cent confidence limits for α from (13.29) are,

$$42.34 \pm 2.88 \times 2.101$$

i.e. $$42.34 \pm 6.05$$

The upper limit for α is 48.39 and lower limit is 36.29.

CURVILINEAR REGRESSION

It has already been stated in the beginning of this chapter that the relationship between the dependent variable Y and the independent variable X can be curvilinear in many cases. The shape of the curve depends on the rate of change in Y corresponding to the change in the value of X. Some of the commonly used curves are given here along with their mathematical equations. These curves may be fitted to the data and used.

Second Degree Curve Mathematical equation of the curve is,

$$Y = \alpha + \beta X + \gamma X^2 \qquad (13.30)$$

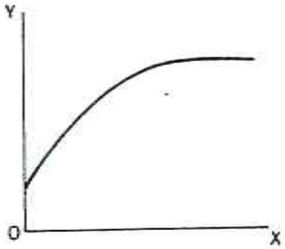


Fig. 13.3 Parabola

The shape of the curve is the upper half of the parabola. It is generally used for a relationship between the production of a crop and the quantity of fertilizer applied per unit area.

Exponential Growth Curve Mathematical equation of the curve is,

$$Y = \alpha \beta^X \qquad (13.31)$$

If we put $\beta = 1 + i$, where i is the rate of interest and X the number of years, then, Y gives the amount to which the initial amount α will rise. By taking the logarithm of both sides, the model becomes linear in log terms.

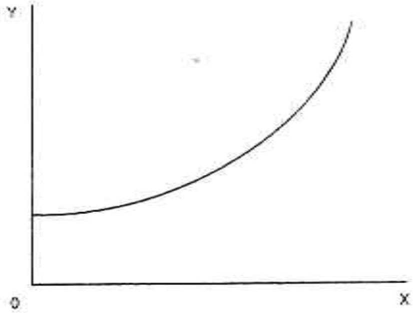


Fig. 13.4 Growth Curve

Mathematical equation of exponential growth curve is also given as

$$Y = \alpha e^{\delta X} \qquad (13.32)$$

Exponential Decay Curve Mathematical model is,

$$Y = \alpha \beta^{-X} \qquad (13.33)$$

If $\beta < 1$, then it represents the decay curve. The decay of the emission of particles of a radioactive element follows this law.

Another mathematical form of the above curve is,

$$Y = \alpha e^{-\delta X} \qquad (13.34)$$

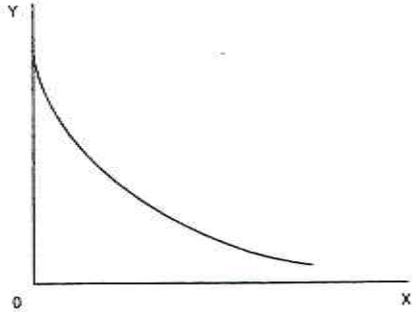


Fig. 13.5 Decay Curve

Logistic Growth Curve The mathematical model is,

$$\frac{1}{Y} = \alpha \beta^X + \gamma \qquad (13.35)$$

This curve is often suitable to the growth of human populations of some countries.

It is worth pointing out, that it is not possible to enumerate all situations in which a particular linear or curvilinear regression model is suitable. Moreover, neither one can enunciate all possible equations. It is through experie:nce and certain

CORRELATION COEFFICIENT

The correlation between two variables is termed as simple correlation and its general measure is *Karl Pearson coefficient of correlation*. The estimated value of population correlation coefficient ρ between two variables X and Y is denoted by

Here we shall give all the formulae for r. If in some study all the units of the population are measured, we should use the formulae given for r over all the observations of the population and, instead of r, use the symbol ρ. But in rare cases we study the whole population. Therefore, we have given the formulae for sample observations. Sometimes the suffix XY is added to r, i.e. r_{XY}, to connote that it is the correlation coefficient between the variables X and Y. Theoretically, the sample correlation coefficient is given as,

$$r_{XY} = \frac{\text{cov }(X, Y)}{\sqrt{v(X)\,v(Y)}} \qquad (13.36)$$

$$= \frac{s_{XY}}{s_X\, s_Y} \qquad (13.36.1)$$

where s_X and s_Y are the sample standard deviations of variables X and Y respectively, and s_{XY} is the estimated covariance between X and Y. If we have n pairs of sample observations, $(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots (x_n, y_n)$, then the correlation coefficient,

$$r = \frac{\frac{1}{n-1}\Sigma_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left\{\frac{1}{n-1}\Sigma_i (x_i - \bar{x})^2\right\}\left\{\frac{1}{n-1}\Sigma_i (y_i - \bar{y})^2\right\}}} \qquad (13.37)$$

For $i = 1, 2, \ldots, n$

$$= \frac{\Sigma_i (x_i - \bar{x})\ (y_i - \bar{y})}{\sqrt{\Sigma_i (x_i - \bar{x})^2\ \Sigma_i\ (y_i - \bar{y})^2}} \qquad (13.37.1)$$

$$= \frac{\Sigma_i x_i y_i - \frac{1}{n}(\Sigma_i x_i)(\Sigma_i y_i)}{\sqrt{\left\{\Sigma_i x_i^2 - \frac{(\Sigma_i x_i)^2}{n}\right\}\left\{\Sigma_i y_i^2 - \frac{(\Sigma_i y_i)^2}{n}\right\}}} \qquad (13.37.2)$$

for $i = 1, 2, \ldots, n$.

If we take the deviations from mean and denote

$$x_i - \bar{x} = u_i \text{ and } y_i - \bar{y} = v_i,$$

then the coefficient of correlation,

$$r = \frac{\Sigma_i u_i v_i}{\sqrt{\Sigma_i u_i^2\, \Sigma_i v_i^2}} \qquad (13.37.3)$$

The correlation coefficient r is a pure number i.e. r is independent of the units in which X and Y are measured.

Note: All said about ρ holds true for r except that r is an estimated value of the population parameter ρ

Assumptions about Correlation Coefficient There are three assumptions made in giving the correlation coefficient by the above formulae. They are:

1. The random variables X and Y are distributed normally.
2. The variables X and Y are linearly related.
3. There is a cause and effect relationship between factors affecting the values of X and Y in the series of data.

Limits of Correlation Coefficient Formula (13.37.3) gives,

$$r^2 = \frac{(\Sigma_i u_i v_i)^2}{\Sigma_i u_i^2 \Sigma_i v_i^2} \qquad (13.38)$$

We know from the Schwartz inequality that if u_i and v_i are real quantities for all $i = 1, 2, \ldots, n$, then

$$(\Sigma_i u_i v_i)^2 \le (\Sigma_i u_i^2)\ (\Sigma_i v_i^2) \qquad (13.39)$$

The sign of equality holds if and only if

$$\frac{u_1}{v_1} = \frac{u_2}{v_2} = \cdots = \frac{u_n}{v_n} \qquad (13.40)$$

Thus using (13.39), we can write that

$$r^2 \le 1 \qquad (13.41)$$

which implies that $-1 \le r \le 1$. (13.42)

Hence it is clear that the correlation coefficient can never be greater than 1 and less than − 1. When ρ (or r) = 1, it means that there exists a perfect positive correlation between two variables and when ρ (or r) = − 1, it is called the perfect negative correlation. When two variables are independent, the correlation between them is zero. But its converse is not true i.e. if the correlation between two variables is zero, they are not necessarily independent. Zero correlation coefficient shows the absence of linear relationship between the two variables. The values between 1 and − 1 are interpreted accordingly.

Example 13.6. The age in years of fourteen young couples is given below:

Husband (X):	21	25	26	24	22	30	19	24	28	32	31	29	21	18
Wife (Y):	19	20	24	21	21	24	18	22	19	30	27	26	19	18

To know the extent of relationship between the age of husbands and wives, we calculate the coefficient of correlation r from the given data. For the given variate values,

$$n = 14,$$

$$\Sigma_i x_i = 350,$$

$$\bar{x} = 25$$

$$\Sigma_i y_i = 308,$$

$$\bar{y} = 22$$

Since means are whole numbers, it is convenient to make use of the formula (13.37.1). To do the calculations systematically, prepare the following table.

	x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
	21	19	−1	−3	16	9	12
	25	20	0	−2	00	4	00
	26	24	1	2	1	4	2
	24	21	−1	−1	1	1	1
	22	21	−3	−1	9	1	3
	30	24	5	2	25	4	10
	19	18	−6	−4	36	16	24
	24	22	−1	0	1	00	00
	28	19	3	−3	9	9	−9
	32	30	7	8	49	64	56
	31	27	6	5	36	25	30
	29	26	4	4	16	16	16
	21	19	−4	−3	16	9	12
	18	18	−7	−4	49	16	28
Total	350	308	0	0	264	178	185

From the above table we have,

$$\Sigma(x_i-\bar{x})^2 = 264, \Sigma_i(y_i-\bar{y})^2 = 178, \Sigma_i(x_i-\bar{x})(y_i-\bar{y}) = 185$$

Putting these values in the formula (13.37.1) we get,

$$r = \frac{185}{\sqrt{264\times 178}} = \frac{185}{216.78} = 0.85$$

The coefficient of correlation between the age of husband and that of the wife is 0.85 which is close to 1. Hence it can be said that there is a high degree of relationship between the age of husbands and wives.

Example 13.7. The birth rate and death rate per thousand persons in Switzerland from 1968 to 1980 were as follows:

Birth rate (X):

17.1 16.5 15.8 15.2 14.3 13.6 12.9 12.3 11.7 11.5 11.3 11.3 11.6

Death rate (Y):

9.3 9.3 9.1 8.2 8.9 8.9 8.5 9.7 9.0 8.7 9.1 9.0 9.2

The correlation between the birth rate and the death rate can be calculated in the following manner.

First, we make the following table and compute various values involved in the formula.

Obs. no.	x	y	x^2	y^2	xy
1	17.1	9.3	292.41	86.49	159.03
2	16.5	9.3	272.25	86.49	153.45
3	15.8	9.1	249.64	82.81	143.78
4	15.2	8.2	231.04	67.24	124.64
5	14.3	8.9	204.49	79.21	127.27
6	13.6	8.9	184.96	79.21	121.04
7	12.9	8.5	166.41	72.25	109.65
8	12.3	9.7	151.29	94.09	119.31
9	11.7	9.0	136.89	81.00	105.30
10	11.5	8.7	132.25	75.69	100.05
11	11.3	9.1	127.69	82.81	102.83
12	11.3	9.0	127.69	81.00	101.70
13	11.6	9.2	134.56	84.64	106.72
Total	175.1 $=\Sigma x_i$	116.9 $=\Sigma y_i$	2411.57 $=\Sigma_i x_i^2$	1052.93 $=\Sigma_i y_i^2$	1574.77 $=\Sigma_i x_i y_i$

There are 13 pairs of observations, hence $n = 13$. Now we make use of formula (13.37.2) because the means of x and y values are not exact.

$$r = \frac{1574.77-(175.1)(116.9)/13}{\sqrt{\left(2411.57-\frac{175.1^2}{13}\right)\left(1052.93-\frac{116.9^2}{13}\right)}}$$

$$= \frac{0.22}{\sqrt{53.11\times 1.73}} = \frac{0.22}{9.58} = 0.02$$

The coefficient of correlation between the birth rate and the death rate is 0.02, which is very close to zero. Hence it can be inferred that they are not linearly related to each other.

Correlation Coefficient from Coded Data Using the same coding as in case of regression coefficient, i.e. $dx_i = (x_i - c_1)/d_1$ and $dy_i = (y_i - c_2)/d_2$ and also making use of table (13.1) as such and relations from (13.16) to (13.17.1), the correlation coefficient from coded data as per formula (13.37.1) is,

$$r_c = \frac{\Sigma_i(dx_i-\overline{dx})(dy_i-\overline{dy})}{\sqrt{\Sigma_i(dx_i-\overline{dx})^2\ \Sigma_i(dy_i-\overline{dy})^2}} \qquad (13.43)$$

$$= \frac{\Sigma_i\left(\frac{x_i-c_1}{d_1}-\frac{\bar{x}-c_1}{d_1}\right)\left(\frac{y_i-c_2}{d_2}-\frac{\bar{y}-c_2}{d_2}\right)}{\sqrt{\Sigma_i\left(\frac{x_i-c_1}{d_1}-\frac{\bar{x}-c_1}{d_1}\right)^2\ \Sigma_i\left(\frac{y_i-c_2}{d_2}-\frac{\bar{y}-c_2}{d_2}\right)^2}}$$

$$= \frac{\Sigma_i \left(\frac{x_i - \overline{x}}{d_1}\right)\left(\frac{y_i - \overline{y}}{d_2}\right)}{\sqrt{\Sigma_i \left(\frac{x_i - \overline{x}}{d_1}\right)^2 \Sigma_i \left(\frac{y_i - \overline{y}}{d_2}\right)^2}} \qquad (13.43.1)$$

$$= \frac{\Sigma_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\Sigma_i (x_i - \overline{x})^2 \Sigma_i (y - \overline{y})^2}} \qquad (13.43.2)$$

$$= r_{YX} \qquad (13.44)$$

Relation (13.44) reveals that the correlation from coded data is independent of the constants involved in coding of the data. This result helps in reducing the labour of calculation of r to a great extent by choosing the suitable constants c_1, c_2, d_1 and d_2. Some or none or all of them may be equal. Moreover, c_i and/or c_2 should be taken as zero if the origin is not shifted. Also d_1 and d_2 should be taken as one if the scale of measurement is not to be changed.

Example 13.8 First we give an example taking assumed set of paired observations showing that the correlation coefficient calculated with and without coding remains the same.

X:	8	6	12	14	16	10
Y:	15	10	20	25	30	20

We have done the coding by taking $c_1 = 10$, $d_1 = 2$, $c_2 = 20$ and $d_2 = 5$. The calculations are shown in the following table:

First we will calculate r with the help of formula (13.37.2) using the columns from (i) to (v).

$$r = \frac{1450 - \frac{66 \times 120}{6}}{\sqrt{\left(796 - \frac{66^2}{6}\right)\left(2650 - \frac{120^2}{6}\right)}}$$

$$= \frac{130}{\sqrt{70 \times 250}}$$

$$= \frac{13}{13.23}$$

$$= 0.98$$

Now we calculate r with the help of coded data from columns (vi) to (x) of the above table. The formula for r is,

$$r = \frac{\Sigma_i dx_i\, dy_i - \frac{(\Sigma_i dx_i)\ (\Sigma_i dy_i)}{n}}{\sqrt{\left\{\Sigma_i d^2 x_i - \frac{(\Sigma_i dx_i)^2}{n}\right\}\left\{\Sigma_i d^2 y_i - \frac{(\Sigma_i dy_i)^2}{n}\right\}}}$$

	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)	(viii)	(ix)	(x)
	x	y	x^2	y^2	xy	$(x-10)/2$ $= dx$	$(y-20)/5$ $= dy$	d^2x	d^2y	$dxdy$
	8	15	64	225	120	−1	−1	1	1	1
	6	10	36	100	60	−2	−2	4	4	4
	12	20	144	400	240	1	0	1	0	0
	14	25	196	625	350	2	1	4	1	2
	16	30	256	900	480	3	2	9	4	6
	10	20	100	400	200	0	0	0	0	0
Total	66	120	796	2650	1450	3	0	19	10	13
	$\equiv \Sigma_i x_i$	$\equiv \Sigma_i y_i$	$\equiv \Sigma_i x_i^2$	$\equiv \Sigma_i y_i^2$	$\equiv \Sigma_i x_i y_i$	$\equiv \Sigma_i dx_i$	$\equiv \Sigma_i dy_i$	$\equiv \Sigma_i d^2 x_i$	$\equiv \Sigma_i d^2 y_i$	$\equiv \Sigma_i dx_i dy_i$

$$= \frac{13 - \frac{3 \times 0}{6}}{\sqrt{\left(19 - \frac{3^2}{6}\right)\left(10 - \frac{0}{6}\right)}}$$

$$= \frac{13}{\sqrt{17.5 \times 10}}$$

$$= 0.98$$

We get the same value of r in both ways. Hence, it justifies numerically that $r_c = r_{XY}$.

Example 13.9. We calculate the correlation coefficient between the variables, Total grain production and cereal production for the data given in example (13.4).

We also make use of coded values, since the data is too heavy. The partial calculations made in the example (13.4) have been used as such. Formula for r from coded values is used in the form as given in the example (13.8).

$$r = \frac{537 - \frac{(-28) \times (-11)}{9}}{\sqrt{\left(205 - \frac{(-11)^2}{9}\right)\left(1418 - (-\frac{28)^2}{9}\right)}}$$

Since, $\Sigma_i d^2 y_i = (-22)^2 + (-18)^2 + (-14)^2 + (-7)^2 + 0^2 + 3^2 + 4^2 + 12^2 + 14^2 =$ 1418

$$r = \frac{502.78}{\sqrt{191.56 \times 1330.89}}$$

$$= \frac{502.78}{504.92}$$

$$= 0.9958$$

The correlation between the area sown and the production is of high order.

Test of Significance of Correlation Coefficient As already stated in Chapter 9, the random sample (s) is/are drawn from the population or universe under consideration. Whatever conclusions are derived or deduced from the sample values, are meant do draw inferences about the parent population. The estimates are not unique and hence a sort of confirmation is sought by way of test of significance, for the validity of inferences drawn from the sample about the population. Of course, these results are subjected to certain probability of wrong decision which is covered by the level of significance.

The test of significance of correlation coefficient means to test the hypothesis, whether or not the correlation coefficient is zero in the population i.e. we test,

$$H_0 : \rho = 0 \text{ vs. } H_1 : \rho \neq 0$$

The test statistic for testing H_0 is,

$$t_{n-2} = \frac{r}{s_r} \tag{13.45}$$

where r is the estimated value of ρ based on the n paired observations and s_r is the standard error of r. Suffix $(n-2)$ denotes the degrees of freedom of t. Also,

$$s_r = \sqrt{\frac{1-r^2}{n-2}}$$

Thus,

$$t_{n-2} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \tag{13.45.1}$$

If the calculated value of t is greater than the table value of t for α level of significance and $(n-2)$ d.f., reject H_0, otherwise accept H_0. Rejection of H_0 leads to the conclusion that the two variables are not independent. This means that the correlation between them is worth considering. If H_0 is accepted, it means that the value of r is due to sampling whereas in reality two variables are uncorrelated in the population.

Example 13.10. In the example (13.9),

$$r = 0.9958 \text{ and } n = 9.$$

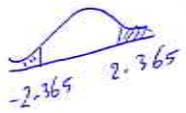
To test the significance of the population correlation coefficient, we test

$$H_0 : \rho = 0 \text{ vs. } H_1 : \rho \neq 0$$

by the statistic (13.45.1). Hence,

$\alpha = 0.05$

$$t = \frac{0.9958\ \sqrt{9-2}}{\sqrt{1-(0.9958)^2}}$$



$$= \frac{0.9958 \times \sqrt{7}}{\sqrt{.0084}}$$

$$= \frac{2.63}{0.092}$$

$$= 28.59$$

The table value of t at $\alpha = 0.05$ and 7 d.f. is 2.365. Since the calculated value of t is greater than the table value of t, we reject H_0. It means that there is a signifciant correlation between the area sown and the production.

Example 13.11. From example (13.6) we have,

$$r = 0.85 \text{ and } n = 14.$$

$t_{cal} = 5.587$ 12 d.f.

We test the hypothesis,

$$H_0 : \rho = 0 \text{ vs. } H_1 : \rho \neq 0$$

$t_{tab.} = 3.499$

⇒ reject H_0

$$= \frac{23 - 39.09}{\sqrt{(139 - 43.51)(111 - 35.11)}}$$

$$= \frac{-16.09}{85.12}$$

$$= -\ 0.19.$$

The correlation between the age of workers and the number of double duty days is negative. It means that as the age increases, the workers avoid double duty. Of course, the correlation is of a low degree.

RELATIONSHIP BETWEEN CORRELATION COEFFICIENT AND REGRESSION COEFFICIENTS

It has been said that simple correlation is expressed only when there exists a linear relation between two variables. Hence, it should be possible to establish the mathematical relationship between the correlation coefficient and the two regression coefficients namely, b_{YX} and b_{XY}. We know,

$$r = \frac{\Sigma_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma_i (x_i - \bar{x})^2 \, \Sigma_i (y_i - \bar{y})^2}}$$

for $i = 1, 2, \ldots, n$

or

$$r^2 = \frac{\{\Sigma_i (x_i - \bar{x})(y_i - \bar{y})\}^2}{\Sigma_i (x_i - \bar{x})^2 \, \Sigma_i (y_i - \bar{y})^2}$$

Also

$$b_{YX} = \frac{\Sigma_i (x_i - \bar{x})(y_i - \bar{y})}{\Sigma_i (x_i - \bar{x})^2}$$

and

$$b_{XY} = \frac{\Sigma_i (x_i - \bar{x})(y_i - \bar{y})}{\Sigma_i (y_i - \bar{y})^2}$$

$\therefore$

$$B_{YX} \ b_{XY} = \frac{\{\Sigma_i (x_i - \bar{x})(y_i - \bar{y})\}^2}{\Sigma_i (x_i - \bar{x})^2 \, \Sigma_i (y_i - \bar{y})^2}$$

$$= r^2$$

or

$$r = \sqrt{b_{YX} \cdot b_{XY}} \qquad (13.53)$$

Relation (13.53) shows that the correlation coefficient is the geometric mean of the two regression coefficients. The sign of r will be the same as that of either b_{YX} or b_{XY}.

Again,

$$b_{YX} = \frac{s_{XY}}{s_X^2}$$

and

$$r = \frac{s_{XY}}{s_X \, s_Y}$$

$$s_{XY} = r s_X \, s_Y$$

Substituting for s_{XY} we get,

$$b_{YX} = r \, \frac{s_Y}{s_X} \qquad (13.54)$$

Similarly,

$$b_{XY} = r \, \frac{s_X}{s_Y} \qquad (13.55)$$

RANK CORRELATION

It is not always possible to take measurements on units or objects. Many characters are expressed in comparative terms such as beauty, smartness, temperament, etc. In such cases the units are ranked pertaining to that particular character instead of taking measurements on them. Sometimes, the units are also ranked according to their quantitative measure. In these types of studies, two situations arise, (i) the same set of units is ranked according to two characters A and B, (ii) two judges give ranks to the same set of units independently, pertaining to one character only. In both these situations, we get paired ranks for a set of units. For example, two judges rank the girls independently in a beauty competition. The students are ranked according to their marks in Mathematics and Statistics. In all these situations, the usual Pearsonian correlation coefficient can not be obtained. Hence, the psychologist, Charles Edward Spearman (1906) developed a formula for correlation coefficient, which is known as *rank correlation* or *Spearman's correlation*. It is denoted by r_s. Suffix S to r is a connotation for Spearman, the name of the inventor.

The formula for r_s is derived as the ratio of covariance to the product of standard deviation of two series of ranks. Here the derivation is omitted and thus the formula for rank correlation is,

$$r_s = 1 - \frac{6\,\Sigma_i d_i^2}{n\,(n^2 - 1)} \qquad (13.56)$$

where $i = 1, 2, \ldots, n$,
d_i is the difference between the ranks of the i-th units and n is the total number of units.

The value of r_s also lies between −1 and 1.

Example 13.16. Two judges gave the following ranks (from the highest to the lowest) to eleven girls who contested in a beauty competition. Whether or not, there is an agreement between the independent rankings of the two judges, can be ascertained only by finding out the rank correlation between the ranks awarded by two judges.

Girl No.	1	2	3	4	5	6	7	8	9	10	11
	Ranks										
Judge A	3	4	1	2	5	10	11	7	9	8	6
Judge B	2	4	3	1	7.	9	6	11	10	5	8

Following the usual procedure, the rank correlation is calculated.

d	1	0	−2	1	−2	1	5	−4	−1	3	−2	Total
d^2	1	0	4	1	4	1	25	16	1	9	4	66

Here $\Sigma_i d_i^2 = 66$ and $n = 11$

From (13.56),

$$r_s = 1 - \frac{6 \times 66}{11 \times 120} = 1 - 0.30 = 0.70$$

The value of rank correlation $r_s = 0.70$, which is quite high. Hence it can be concluded that there is an agreement between judges with regard to the beauty of the girls.

Example 13.17. The ranks of 12 students according to their marks in Mathematics and Statistics were as follows:

Student No.	1	2	3	4	5	6	7	8	9	10	11	12
Mathematics	5	2	1	6	8	11	12	4	3	9	7	10
Statistics	4	3	2	7	6	9	10	5	1	11	8	12

The interest lies to know whether or not students who are good in Mathematics also excel in Statistics and vice versa. This objective can be met out by finding out the rank correlation coefficient. For this the calculations are,

													Total
d	1	−1	−1	−1	2	2	2	−1	2	−2	−1	−2	0
d^2	1	1	1	1	4	4	4	1	4	4	1	4	30

Here $\Sigma_i d_i^2 = 30$ and $n = 12$

From (13.56),

$$r_s = 1 - \frac{6 \times 30}{12 \times 143}$$

$$= 1 - 0.105$$

$$= 0.895$$

The correlation between the ranks of marks in the two subjects is very high. From this, it is inferred that the students who are good in Mathematics are also good in Statistics.

Problem of Tied observations

Formula (13.56) is based on the assumption that n units received n sets of paired ranks from 1 to n independently assigned in each set. As a matter of fact no ties should occur in case of samples from continuous populations. But ties seldom occur due to rounding of measurements. Hence, the problem of ties comes in while calculating spearman's rank correlation.

If ties occur between measurements across sets x and y, it creates no problem. But if ties occur within a set of sample values, the problem of ties can not be ignored. In this situation, of course, the ranks are assigned by mid rank method. Under this process, the sum of ranks remain the same as $n\,(n + 1)/2$, but there is a decrease in the sum of squares of ranks. Hence, adjustment has to be made in the formula of spearman's rank correlation. It is given here without proof.

Suppose m_h is the number of tied mesurements in the ith group of tied values $(h = 1, 2, ..., k)$ for the set x. It means there are k groups of tied values in the set x. Then the correction factor u_x for ties for the set x is given as,

$$u_x = \sum_{h=1}^{K} m_h\,(m_h^2 - 1)$$

A similar expression v_y i.e. the correction factor for groups of tied measurements within the set y can be given as,

$$v_y = \sum_{j=1}^{l} g_j\,(g_j^2 - 1)$$

where g_j is the number of tied measurements in the jth group of tied values for $j = 1, 2, ..., l$. In case of ties in both the sets, the amended formula for Spearman's rank correlation is,

$$r = \frac{n\,(n^2 - 1) - 6\sum_{i=1}^{n} d_i^2 - (u_x + v_y)/2}{\sqrt{(n^2\,(n^2 - 1)^2 + (u_x + v_y)\,(n^3 - n) + u_x\,v_y}} \qquad (13.56.1)$$

$$= \frac{(n^3 - n) - 6\sum_{i=1}^{n} d_i^2 - (u_x + v_y)/2}{\sqrt{(n^3 - n)^2 - (n^3 - n)\,(u_x + v_y) + u_x\,v_y}} \qquad (13.56.2)$$

In the above formula, if no ties occur in the set x, $u_x = 0$ and if no ties occur in the set y, $v_y = 0$. If no ties occur in both the sets, $u_x = 0$, $v_y = 0$ and thus formula (13.52.2) reduces to formula (13.56).

Example 13.18. The percentage marks secured by eleven students in I.C.W.A. and C.A. examinations are as follows:

Students	1	2	3	4	5	6	7	8	9	10	11
I.C.W.A. Marks (x)%	40	55	40	60	62	78	58	96	85	60	60
C.A. Marks (y)%	65	45	68	65	70	75	69	88	75	72	82

Excersises

1- Mean morphometic values of workers of Apis Cerana measured in mm were as follows at thirteen locations.

Labrum length:	0.42	0.43	0.40	0.33	0.36	0.40	0.32	0.40	0.38	0.36	0.40	0.33	0.35
Labrum breadth:	1.10	1.02	1.00	1.00	1.02	0.99	1.00	1.00	1.00	1.00	1.01	1.00	0.99

(*a*) Find the regression of labrum breadth on labrum length,

(*b*) Estimate labrum breadth for labrum length = 0.5,

(*c*) Test the significance of the regression coefficient.

2- Aggregate figures for merchandise exports and imports in India for eight years in millions of rupees are as follows:

Export:	1960	2170	2420	3020	3850	4690	5360	5210
Import:	2330	2110	2580	3540	4070	4000	4550	5920

Calculate correlation coefficient between the import and the export using the method of coding of data.

3- Following are the data pertaining to the production and export of sugar in lakh tonnes in India from 1971 to 1982.

Production (*X*):	37.4	31.1	38.7	39.5	47.9	42.6	48.4	64.6	58.4	38.6	51.4	84.0
Export (*Y*):	3.90	1.33	1.10	4.39	9.41	9.67	3.41	2.51	8.62	9.90	6.64	6.50

(*a*) Find the regression of *Y* on *X*.

(*b*) Test the significance of the regression coeffcient.

(*c*) Test whether the regression line in the population passes through the origin

(*d*) What export of sugar can be expected when the production is 50 lakh tonnes?

4. The age groups and monthly income of 80 workers of a manufacturing unit were as given in the following bivariate frequency table.

	Age groups				
Monthly pay	20 – 30	30 – 40	40 – 50	50 – 60	60 – 70
300 – 350	5	2	1	—	—
350 – 400	4	4	2	3	—
400 – 450	2	6	4	1	—
450 – 500	2	3	5	2	—
500 – 550	1	4	7	5	—
550 – 600	—	2	3	2	—
600 – 650	—	—	1	—	1
650 – 700	1	1	2	2	—
700 – 750	—	—	1	1	—

Calculate the coefficient of correlation between age and monthly income.

5.–Fourteen singers in a music competition were ranked by two judges as follows:

Singers:	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*	*K*	*L*	*M*	*N*
Judge I:	13	4	5	11	2	6	8	9	12	1	3	7	10	14
Judge II:	10	9	7	8	1	3	6	14	11	2	4	5	12	13

Using the method of rank correlation, find whether the ranks given by the two judges have concordance.

6.- The I.Q.'s of a group of 6 persons were measured, and then they were made to appear in a certain examination. Their I.Q's and examination marks were as follows:

Person:	*A*	*B*	*C*	*D*	*E*	*F*
I.Q.:	110	100	140	120	80	90
Exam. Marks:	70	90	80	60	10	20

Compute the coefficient of correlation and rank correlation. Why are the correlation figures obtained different? (*B.Sc., Chennai, 1975*)

7. The following calculations have been made for closing prices of twelve stocks (*X*), on the Mumbai Stock Exchange on a certain day, along with the volume of sales in thousands of Shares (*Y*). From these calculations, find the regression equation:

ΣX = Rs 580, ΣY = Rs 370, ΣXY = Rs 11494, ΣX^2 = Rs 41658 and ΣY^2 = Rs. 17206

(*B.A. Hon. Econ., Delhi, 1971*)

8.–From the two regression equations, find the value of *r*, $\overline{X}$ and $\overline{Y}$:

$4Y = 9X + 15$ and $25X = 6Y + 7$ (*M.A., Agra, 1973*)

9.- A department store gives in-service training to its salesmen followed by a test to consider whether it should terminate the services of any of the salesman who do not qualify in the test. The following data give the test scores and sales made by nine salesmen during a certain period.

Test score:	14	19	24	21	28	22	15	[illegible]	19
Sales ('00 Rs.):	31	36	48	37	50	45	33	[illegible]	39

Calculate the coefficient of correlation between the test scores and the sales. Does it indicate that the termination of services of the low test scorer is justified? If the firm wants a minimum sales volume of Rs. 3000, what is the minimum test score that will ensure continuation of the service? (*C.A., Nov. 1974*)

10. (*a*) Given,

	X series	*Y series*
Mean	24	140
S.D.	16	48
	r = 0.6	

(*i*) Find out the most probable value of *Y* if *X* is 50 and the most probable value of *X* if *Y* is 180.

(*ii*) What would be the coefficient of correlation if the two regression coefficient are 0.6 and 0.4.

Chapter(5)

Sampling

As per statistical dictionary by Kendall and Buckland (1975), the population is defined as, "*In statistical usage, the term population is applied to any finite or infinite collection of individuals.*"

For example, the population of students in a University. It means anybody who is enrolled in a University belongs to this population. The number of plants in a field, persons suffering from cancer, workers in textile industry, persons in army services in India, are some other examples of population in statistical sense.

From the definition it is evident that population can be finite as well as infinite. Obviously, the finite population consists of individuals or items which are finite in number. An infinite population is one which either possesses the infinite property through some limiting process or is non-enumerable. The population of all real numbers between 0 and 1, the population of all integers are examples of infinite population. In case of random sampling with replacement (discussed later), any population is always infinite. Without going too deep in the matter, we may say that a population consisting of a large number of units or items may be deemed to possess the properties of an infinite population.

From the above discussion it is evident that every population consists of individuals or items which are known as *sampling units*. The formal definition of sampling unit is given below.

Sampling Unit The population may be regarded as consisting of units which are to be used for the purpose of sampling. Each unit is regarded as individual and indivisible when the selection is made. Such a unit is known as a sampling unit. A sampling unit may be specified on some natural basis or any other criterion fixed for it. The criteria for it depend on the purpose of the survey and the sampling scheme to be followed. A person, an animal, a household, an orchard, a factory or a village are few examples of sampling units.

Sample A finite part of a population or a subset of a set of sampling units, selected by some process, usually by deliberate selection with the object of investigating the properties of the parent population or set, is called a sample. For example, if we select five students from a class of forty students, five selected students constitute a sample. We select fifty bulbs to test the life of bulbs, from a lot of bulbs manufactured by a factory in a week. Fifty selected bulbs constitute a sample.

Sampling is a device which makes one able to draw inferences about the whole population simply by observing or measuring a few of the sampling units. However, this creates many doubts like (a) whether the conclusions drawn on the basis of sample observations really hold good for the whole population or the whole mass? (b) would the results not be unreliable? Such questions always surface in the mind of applied scientists. But the fact remains that the sampling has served

the purpose of all the scientists to a great extent. Many arguments can be put forth in its support. A few of them are as given below.

(1) It is difficult to handle a population which usually consists of a large number of units.

(2) Too much time is required to study the whole population and often the study becomes outdated by the time it is complete.

(3) Finances required to cover the whole population can hardly be made available.

(4) In a study where individuals are killed or perished under observation, studying the population serves no purpose. To clarify this point further we give an example. If all the battery cells of a manufacturing concern are put to life testing, nothing will be left for use.

(5) In case, the population is infinite or consists of uncountable number of units, its study is impossible.

Some people also think that complete enumeration yields better results than the sampling studies. However, this is not correct because complete enumeration (census studies) adds many errors which are reduced or eliminated by sampling. Hence, in many cases sample studies yield better results than population studies. The reliability of results depends more on the quality of the sample. If the sample is a true representative of the population, the results obtained from it are very near to the true value. In the view of mathematicians, a 'random sample' can always be deemed as a true representative of the population. To a great extent it is correct. But the moment one starts identifying various sampling units belonging to a population, one resorts to thinking of various sampling schemes. The use of a particular type of sampling scheme depends on: (i) type of population, (ii) information available about sampling units, (iii) object of the study, (iv) availability of resources like sampling frame, time, money and trained personnel, and (v) last but not the least, the knowledge and experience of the person selecting the sample.

Errors in Surveys In any survey two types of errors are likely to occur (i) sampling errors (ii) non-sampling errors.

Sampling Errors The errors which are introduced due to errors in selection of a sample or the discrepancies between population parameters and estimates which are derived from a random sample. This discrepancy generally decreases as the sample size is quite rapid but gradually the decrease in error becomes negligible with increasing sample size. Hence a sample of optimum size must be obtained for a study. In this way we can minimize the error or keep the error as small as we please and at the same time minimizing the cost of the survey. To determine the optimum sample size, a tolerable amount of error is prefixed and the smallest sample size is determined to help keep the error within tolerable limit.

Non-Sampling Error An error in sample estimates cannot be attributed to sampling fluctuations. It is experienced that the studies based on complete enumeration do not yield similar results in repeated enumerations. Such a discrepancy occurs due to many errors which are termed as non-sampling errors. Various sources of such errors which may be visualized are,

(i) Observational error or response error: if the observations are taken repeatedly on the same unit, the observed values generally differ or otherwise even the same respondent is asked the same question repeatedly, his response may differ.

(ii) Lack of preciseness of definition also adds to the non-sampling errors. For example, in judging the loss of crop due to a disease like wilt or rust, will be subject to error due to definitions of what we call severely diseased, moderately diseased and a low intensity of disease. Moreover this measure of intensity will vary from person to person depending on the maturity, qualifications and training, the person has.

(iii) Errors are also introduced in editing and tabulation of data. Since the population data are large, the chances of errors in complete enumeration are more compared to a sample data.

Some better ways of minimising the sampling errors are the choice of an appropriate sampling scheme, selecting a sample of optimum size and the use of standard techniques of estimation. Whereas the non-sampling errors can be minimized through superior management of survey or investigation, employing befitting personnel and by using modern computational aids.

Here we define a few more terms which will be used often in this chapter and elsewhere.

Parameter Any population constant is called a parameter. For example, population mean (μ) and population variance (σ^2) etc. are parameters. To obtain a parameter value, the observations are taken on each and every unit of the population and a value of a constant pertaining to a characteristic is calculated from those observations. This constant value is termed as parameter. Such constant measure(s) of a population characterise a population.

Estimator An estimator is a rule or method of estimating a population parameter. It is generally expressed as a function of sample variates. An estimator is itself a random variable.

Estimate A particular value of an estimator obtained from a set of values of a random sample is known as estimate. As an explanation, generally few units are selected from a population and then observations are taken on these selected units. The constant for a characteristic is calculated from these sample observations. The constant, so obtained, is known as an estimate and stands for a population

parameter. For example the sample mean $\bar{x}$ is an estimate of population mean μ and sample variance s^2 is an estimate of population variance σ^2.

Statistics A statistic is a function of observable random variables and does not involve any unknown parameter. All the more, the function itself is a random variable. But a statistic is not necessarily an estimator of some population parameter. For example, $\frac{1}{n} \Sigma_i X_i$ $(i = 1, 2, \cdots n)$ is a statistic, student-*t*, i.e. $\sqrt{n}\ (\bar{X} - \mu)/s$ is a statistic etc.

Selection with Replacement (swr) In this case, a unit is selected from a population with a known probability and the unit is returned to the population before the next selection is made (after recording its characteristic(s)). Thus, in this method at each selection, the population size remains constant and the probability at each selection or draw remains the same. Under this sampling plan, a unit has chances of being selected more than once. For example, a card is randomly drawn from a pack of cards and placed back in the pack, after noting its face value, before the next card is drawn or from an urn containing balls of different colours a ball is drawn, its colour is noted and kept back in the urn before another ball is drawn. Such a sampling method is known as sampling with replacement. There are N^n possible samples of size *n* from a population of *N* units in case of sampling with replacement.

Sampling without Replacement (swor) In this selection procedure, if a unit from a population of size *N* is selected, it is not returned to the population. Thus, for any subsequent selection, the population size is reduced by one. Obviously, at the time of the first selection, the population size is *N* and the probability of a unit being selected randomly is 1/*N*; for the second unit to be randomly selected, the population size is (*N* – 1) and the probability of selection of any one of the remaining sampling unit is 1/(*N* – 1), similarly at the third draw, the probability of selection is 1/(*N* – 2) and so on.

If we consider the limiting case that *N* is very large $N \to \infty$, the probabilities $\frac{1}{N}, \frac{1}{N-1}, \frac{1}{N-2}, \cdots,$ become constant and in the limiting situation, the selection procedures with replacement and without replacement become equivalent. Whereas in the case of small populations, the values $\frac{1}{N}, \frac{1}{N-1}, \frac{1}{N-2}, \cdots, \frac{1}{N-n+1}$ differ considerably. The sampling from small and large populations is generally expressed as sampling from finite and infinite populations respectively. There are $\binom{N}{n}$ possible samples, in case of sampling without replacement.

Size of a Sample The size of a sample is the number of sampling units which are selected from a population by a random method. The problem that arises is how to decide what should actually be the number of sampling units to be selected from a population. As a matter of fact, the sample size depends on a number of consideration which are as follows:

(1) The purpose for which the sample is drawn.

(2) The type of population from which the sample is to be drawn. That is, if the sampling units constituting the population are highly variable, then a large sample is required and conversely, if the population comprises of less variable units, then a small sample is good enough. For a perfectly homogeneous population, a single unit is sufficient to get the correct results for the whole population. For example, the blood of a person is perfectly homogeneous and hence a drop of blood taken for investigation gives the true picture of blood constitution in the body.

(3) Availability of technical people or equipment needed.

(4) Resources allotted for the study in terms of time and money.

(5) Precision required: If we want to detect very minute differences, most probably a large sample will be required and vice-versa. The mathematical formulation for determining sample size is as given below.

$$d = u_R \frac{\hat{s}_x}{\sqrt{n}} \quad (7.1)$$

or

$$n = \frac{(u_R \hat{s}_x)^2}{d^2} \quad (7.1.1)$$

where d is the precision required to detect the differences to the extent of d; u_R — the number of standard deviations necessary for the required level of reliability R, which is generally in terms of probability. The value of u_R is obtained from the table of the probability distribution which the data follow. For example, if the data are selected from a normal distribution, then for a 95 per cent level of confidence $u_R = 1.96$ and $\hat{s}_x$ — the standard deviation of x. Its value is substituted from experience or is based on the information obtained from some earlier study.

From (7.1.1) it is evident that if d is small i.e. greater precision is required, we will have to select a large sample. Also, if $\hat{s}_x$ is large, n is to be large. Moreover, as the level of reliability increases, n also increases.

SAMPLING METHODS

In the selection of a sample, always the effort is to make the sample a true representative of the population. A large number of schemes have been worked out to achieve this objective. In general, we do probability sampling which is free from human bias. But in some situations, judgement sampling or purposive sampling is preferred to probability sampling. For example, if we want a sample of persons who are suffering from cancer, we have to select cancer patients who happen to come to the hospital(s). But, in general, probability sampling is in use, which enables the investigator to control sampling errors and avoid human bias.

Probability sampling scheme leads to two types of samples (i) Unrestricted random samples (ii) Restricted random samples.

UNRESTRICTED RANDOM SAMPLING

If from a population, selection of sampling units is done in such a manner that each and every unit in the population has the same probability (chance) of being selected, such a probability sampling plan is called random sampling. Suppose there are N sampling units in a population, the probability of selection of any unit is $1/N$.

Simple Random Sampling (srs) In this type of sampling, n units from a population of N units are selected, without replacement, in such a way that the probability of selection of any one sample, out of $\binom{N}{n}$ possible samples is same i.e. $1/\binom{N}{n}$. In practice, the units are drawn one by one from the population numbered from 0 to (N – 1). This process gives an equal chance of being selected to all units not previously selected. A random number table is always helpful in selecting a simple random sample.

Method of Selection A random sample may be selected either by drawing the chits or by the use of random numbers. The *chit method* is a random method but is subject to many human biases as people can identify chits in many ways. The Roulette wheel is another device for drawing numbers randomly. It is more prevalent in lotteries. But the best of all is the use of random numbers given by Fisher and Yates — known as Fisher and Yates random numbers. The random number table is also given by Tippett. This is known as Tippett's random number table. But Fisher and Yates' random number tables are more popular. Which random number tables one uses, does not matter in any way. For selecting a random sample from a population of size N, each sampling unit is numbered from 0 to (N – 1). If N is a two-digit figure, the units can be numbered as 00, 01, 02, ... up to the maximum of 98. In case N is a three-digit figure, the units can be numbered as 000, 001, 002, ... up to the maximum of 998 and so on. In this way, the list of serially numbered sampling units is known as the *sampling frame*. Sometimes, a sampling frame is a map showing the position of sampling units with identification marks. Once the sampling frame is ready, a sample of size n can be selected in the following manner. Let N be a d-digit number. Make use of a d-digit random number table. Now read numbers one by one from the random number table. If this observed number, say K, is less than or equal to (N – 1), then select k-th unit. If the observed number is equal to N, select the unit at serial number 00. In case, K is greater than N, ($K > N$), divide K by N and get the remainder 'R'. Now the R-th unit is selected. This process continues till n sampling units are selected.

The sample may be selected with replacement or without replacement. In the case of sampling with replacement, a number occurring more than once is accepted. A unit is repeated as many times as a random number occurs. But in the case of sampling without replacement, if a random number directly or by way of a remainder occurs more than once is omitted at any subsequent stage. In the above selection procedure numbering of units from 00 onwards and making use of remainders have an advantage, as no random number is being wasted during the selection procedure. This saves time and labour.

To explain the selection procedure further consider an example where a population consists of 18 units and a sample of size 5 is to be selected from this population.

Since 18 is a two-digit figure, units are numbered as 00, 01, ..., 17. Five random numbers are obtained from a two digit random number table.

They are as given below:

65, 43, 62, 54, 46.

On dividing 65 by 18, the remainder is 11, hence select the unit on serial No. 11. Similarly dividing 43, 62, 54 and 46 by 18, the respective remainders are 7, 8, 0 and 10. Hence, select units of serial numbers, 11, 07, 08, 00 and 10 and these selected units constitute the sample.

ESTIMATION OF POPULATION PARAMETERS

Let $X_1, X_2, ..., X_n$ be a random sample of size n from a population of N units and $x_1, x_2, ..., x_n$ be the corresponding observed values. The values of the known parameter(s) should be estimated from the set of these sample observations. These estimates are often single valued which are known as point estimates. Also, many times an interval is to be estimated in which the parameter is expected to lie with a certain level of confidence. Here we will not pursue the estimation theory and hence the formulae for sample mean, sample variance, etc. are given directly.

Some properties of estimates are also discussed in this chapter. Formulae are presented in terms of $x_i(i = 1, 2, ..., n)$, the actual observed value corresponding to X_i.

Formulae for Mean and Variance Let $U_1, U_2, ..., U_N$ be N population units. A random sample of n units is selected. Suppose the observations on the sampled units are $x_1, x_2, ..., x_n$. The sample mean,

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} \qquad (7.2)$$

$$= \frac{1}{n} \sum_i \cdot x_i \qquad (7.2.1)$$

for $i = 1, 2, ..., n$.

The sample variance,

$$s^2 = \frac{1}{(n-1)} \sum_i (x_i - \bar{x})^2 \qquad (7.3)$$

$$= \frac{1}{(n-1)} \{\Sigma\, x_i^2 - n\,\bar{x}^2\} \qquad (7.3.1)$$

$$= \frac{1}{(n-1)} \left\{\sum_i x_i^2 - (\sum_i x_i)^2/n\right\} \qquad (7.3.2)$$

Let the population mean be μ and variance be σ^2. $\bar{x}$ is an estimated value of μ and s^2 is an estimated value σ^2. Formulae for μ and σ^2 are given by (3.1) and (4.8). $\sigma^2 = \frac{1}{N}\sum(X_i - \bar{X})^2$

Now, we define a special term S^2, which is slightly different from σ^2 and is given as,

$$\mu = \frac{X_1 + X_2 + \cdots + X_N}{N}$$

$$S^2 = \frac{1}{(N-1)} \sum_i (X_i - \mu)^2 \qquad (7.4)$$

$$= \frac{1}{(N-1)} \left\{\sum_i X_i^2 - N\mu^2\right\} \qquad (7.4.1)$$

where $i = 1, 2, ..., N$.

PROPERTIES OF ESTIMATES

Unbiasedness An estimate is said to be unbiased if its expected value is equal to its parameter value. For example, if $\bar{x}$ is an estimate of μ, $\bar{x}$ will be an unbiased estimate if and only if,

$$E(\bar{x}) = \mu \qquad (7.5)$$

The expectation may well be understood with this example. If from a population, we take all possible samples of size n and take the mean of all samples, the mean of the estimates is known as the expected value. It can theoretically be proved that $\bar{x}$ is an unbiased estimate of μ. In formula (7.3) for s^2, the divisor is $(n - 1)$ instead of n. The reason for this is that using the divisor $(n - 1)$ makes s^2 an unbiased estimate of S^2,, i.e.,

$$E(s^2) = S^2$$

Mathematical proofs are avoided in this book. Hence, the above conjecture of unbiasedness will be shown through an example.

Example 7.1. A population of five units has the observations.

7, 6, 8, 4, 10

Random samples of three units are drawn such that no unit is repeated in a sample and the order of selection does not matter. Out of the 5 units, there can be 10 samples each consisting of 3 units.

Possible samples, their means and variances are given in the following table.

Sample Nos.	1	2	3	4	5
Observations	7, 6, 8	7, 6, 4	7, 6, 10	7, 8, 4	7, 8, 10
$\bar{x}$	7	17/3	23/3	19/3	25/3
s^2	1	7/3	13/3	13/3	7/3
	6	7	8	9	10
	7, 4, 10	6, 8, 4	6, 8, 10	6, 4, 10	8, 4, 10
	7	6	8	20/3	22/3
	9	4	4	28/3	28/3

$$\text{Mean of } \bar{x}\text{'s} = \frac{1}{10}\left(7 + \frac{17}{3} + \frac{23}{3} + \frac{19}{3} + \frac{25}{3} + 7 + 6 + 8 + \frac{20}{3} + \frac{22}{3}\right)$$

$$= \frac{210}{30} = 7$$

Similarly, mean of s^2's $= \frac{50}{10} = 5.0$

Population mean $\mu = \frac{35}{5} = 7$

and variance

$$S^2 = \frac{1}{4}\{(7-7)^2 + (6-7)^2 + (8-7)^2 + (4-7)^2 + (10-7)^2\}$$

$$= \frac{20}{4} = 5.0$$

The above calculations verify the statement that the sample mean is an unbiased estimate of population mean whereas, s^2 is an unbiased estimate of S^2.

Simple Consistency The notion of consistency is mainly concerned with infinite population. If T_n is an estimator of a parameter θ where T_n is based on a random sample $X_1, X_2 ..., X_n$ such that $T_n = t_n (X_1, X_2,.... X_n)$ then the sequence $\{T_n\}$ is said to be a consistent estimator of θ if for every ε > 0, the following condition holds:

$$\lim_{n\to\infty} P\{\theta - \varepsilon < T_n < \theta + \varepsilon\} = 1 \qquad (7.7)$$

or

$$\lim_{n\to\infty} P\{|T_n - \theta| > \varepsilon\} = 0 \qquad (7.7.1)$$

Mean Squared Error If $T_n = t_n(X_1, X_2,......, X_n)$ be an estimator of a parameter θ, then $E\{(T_n - \theta)^2\}$ is said to be the mean-squared error of the estimator T_n. The sequence of estimators $\{T_n\}$ is said to be a mean squared error consistent if the following condition holds

$$\lim_{n\to\infty} E\{(T_n - \theta)^2\} = 0 \qquad (7.8)$$

Note : When the observed values $x_1, x_2,...., x_n$ for $X_1, X_2,...., X_n$ are substituted in the estimator, we get the estimated value.

Standard Error (S.E.) Standard deviation has been discussed in Chapter 4 as a measure of variability. Another measure is standard error, which is the standard deviation of the sampling distribution of an estimator. The idea is that if we draw a number of repeated samples of fixed size n from a population having a mean μ and variance σ^2, each sample mean, say $\bar{x}$, will have a different value. Here $\bar{x}$ itself is a random variable and hence it has a distribution. The standard deviation of $\bar{x}$ is called *standard error*. It has been proved that the standard error '$\sigma_{\bar{x}}$' of the mean $\bar{x}$ based on a sample of size n is,

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \qquad (7.9)$$

From formula (7.9), it is obvious that the larger the sample size, the smaller the standard error and vice-versa. The advantage of considering standard error instead of a standard deviation is that this measure is not influenced by the extreme values present in a population under consideration. Moreover, it upholds all the virtues of standard deviation.

In practice we avoid studying or surveying the whole population. The process of drawing repeated samples is still more cumbersome. Hence, in reality neither we use σ to calculate the standard error of $\bar{x}$ nor we take more than one sample. As a matter of fact, what we do is, that we select only one sample, find its standard deviation s and use the following formula to find out the standard error of $\bar{x}$ i.e.

$$\text{S.E.}(\bar{x}) = \frac{s}{\sqrt{n}} \qquad (7.10)$$

Standard error is commonly used in testing of hypothesis and interval estimation. Many distributions, which are originally not normally distributed, have been taken as normal by considering the distribution of mean $\bar{x}$ for a large n. This result follows from a most celebrated theorem known as *central limit theorem*. This theorem has been stated below.

CENTRAL LIMIT THEOREM

If $X_1, X_2,...,X_n$ are n identically and independently distributed random variables with mean $\bar{X}_n$ and variance σ_n^2, the standardized variable $\{\bar{X}_n - E(\bar{X}_n)\} / \frac{\sigma_x}{\sqrt{n}}$ approaches a standard normal distribution as n approaches infinity.

Remark. The standard error may be considered as the key to the sampling theory.

Standard Error of Mean Suppose a simple random sample of size n is drawn from a normal population of N units. The variance of the sample mean is ,

$$V(\bar{x}) = \left(\frac{1}{n} - \frac{1}{N}\right) S^2 \qquad (7.11)$$

$$= \frac{N-n}{N} \cdot \frac{S^2}{n} \qquad (7.11.1)$$

An unbiased estimate of $V(\bar{x})$ is

$$s_{\bar{x}}^2 = \left(\frac{1}{n} - \frac{1}{N}\right) s^2 \qquad (7.12)$$

where s^2 is given by (7.3). If $1/N$ is negligible, we obtain

$$s_{\bar{x}}^2 = \frac{s^2}{n} \qquad (7.13)$$

or

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} \qquad (7.13.1)$$

Chapter (6)

Some Sampling Distributions

In Chapter 7 we have studied the samples and estimates, which represent the corresponding population specified by the respective parameters. It means that the sample and its parent population are related to each other. Hence our study can always be two-fold, i.e. (i) what we can derive from the population regarding the sample(s) to be taken from it, and (ii) what kind of information can be gathered from a sample or a series of samples about the population from which they have been drawn. The first point has been covered to a great extent in Chapter 7. therefore, we shall concentrate on the second point in this chapter. Here the objective is to know the manner in which a statistic or a series of statistics vary from one sample to another.

Statistic A statistic is a function of one or more random variables not involving any unknown parameter, e.g. $\Sigma_i X_i / n$, i.e. $\overline{X}$, $\Sigma_i (X_i - \overline{X})^2 / (n-1)$, i.e. s^2, s and $\overline{X}/s$ etc. Moreover, a statistic itself is a random variable and follows some distribution.

Generally, we draw a random sample from a population and calculate its mean, range, median, variance and standard deviation. We know that there are $\binom{N}{n}$ possible samples of size n from a population of N units. One way is to draw all possible samples and then calculate for each sample, the mean $\overline{X}$, S.D. s, etc. And find their frequency distribution. But this is possible only for small populations. Another approach is to draw a large number of samples and collect information on the sample distribution of various statistics. The studies from this approach have become easier with the help of electronic computers. But this technique is not always feasible because of the cost and ambiguity of results involved with it. Moreover, it is not an exact approach as we are concerned with the theoretical sampling distribution of various statistics.

Definition *Sampling distribution* describes the way in which a statistic or a function of statistics, which is/are the function(s) of the random variables $X_1, X_2, \ldots, X_n$, will vary from one sample to another sample of the same size. Such sampling distributions have given a fillip to the number of test statistics for hypotheses testing.[1] Hence, some important sampling distributions are discussed in this chapter.

STUDENT'S t-DISTRIBUTION

This has been discussed in Chapter 7, under the heading 'Standard Error', that if a random sample $X_1, X_2, \ldots, X_n$ of size n, with observed values $x_1, x_2, \ldots, x_n$ is drawn from a normal population having mean μ and S.D. σ, the mean $\overline{x}$ is distributed normally with mean μ and S.D. $\sigma/\sqrt{n}$ *i. e.* $\overline{x} - N(\mu, \sigma/\sqrt{n})$.

Also the variable Z, where,

$$Z = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}} \tag{8.1}$$

is a normal variate with mean 0 and standard deviation 1, i.e. $Z \sim N(0, 1)$.

In practice, the standard deviation σ is not known and in such a situation the only alternative left is to, use s, the sample estimate of standard deviation σ. Thus the variate $\sqrt{n}\ (\overline{x} - \mu)/s$ is approximately normal provided n is sufficiently large. If n is not sufficiently large, the variate $\sqrt{n}(\overline{x} - \mu)/s$ is distributed as t and hence

$$t = \frac{\overline{x} - \mu}{s/\sqrt{n}} \tag{8.2}$$

where $$s^2 = \frac{1}{n-1}\ \Sigma_i (x_i - \overline{x})^2$$

is a widely used variable and the distribution of t is called student's t-distribution. This distribution was discovered by W.S. Gosset in 1908. The statistician Gosset is better known by the pseudonym 'student' and hence t-distribution is called student's t-distribution. He drived the distribution of $\overline{x}$ $s_{\overline{x}}$ (where $s_{\overline{x}} = s/\sqrt{n}$) to find an exact test of a mean by making use of estimated standard deviation s, based on a random sample of size n. R.A. Fisher in 1925 published that t-distribution can also be applied to the test of regression coefficient and other practical problems. Here, we give the density function and properties of t-distribution without derivation. The readers interested in the derivation should consult a book on mathematical statistics. The density function of variable t with $k = n - 1$ degrees of freedom[2] is,

$$f_k(t) = \frac{1}{\sqrt{k}\, B\left(\frac{1}{2}, \frac{k}{2}\right)} \frac{1}{\left(1 + \frac{t^2}{k}\right)^{\frac{k+1}{2}}} \qquad -\infty < t < \infty \tag{8.3}$$

where $$B\left(\frac{1}{2}, \frac{k}{2}\right) = \frac{\Gamma 1/2 \Gamma k/2}{\Gamma\left(\frac{k+1}{2}\right)} = \frac{\sqrt{\pi}\, \Gamma k/2}{\Gamma\left(\frac{k+1}{2}\right)}$$

[1] It has been discussed in Chapters 9 and 10.

[2] Degree of freedom are the number of independent observations in a set of observations.

Therefore,

$$f_k(t) = \frac{\Gamma(k+1)/2}{\sqrt{k\pi}\,\Gamma k/2}\left(1+\frac{t^2}{k}\right)^{\frac{-(k+1)}{2}} \qquad (8.3.1)$$

Properties of *t*-Distribution

(1) *t*-distribution is a unimodal distribution.

(2) The probability distribution curve is symmetrical about the line $t = 0$.

(3) It is a bell-shaped curve just like a normal curve with its tails a little higher above the abscissa than the normal curve. Its spread increases as degrees of freedom 'K' decreases. This means that for the same value of *t*-variate and *x*, the normal variate, the area beyond *t* is larger than the area beyond *x*.

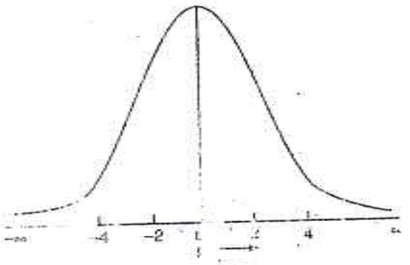


Fig. 8.1: *t*-distribution curve.

(4) *t*-distribution has only one parameter *k*. the degrees of freedom equal to $(n-1)$.

(5) The constants of *t*-distribution are as follows:

(Mean) $\mu = 0$ for $k \geq 2$.

(Variance) $\sigma^2 = \frac{k}{k-2}$ for $k \geq 3$.

(Skewness) $\alpha_3 = 0$ for $k \geq 4$.

(Kurtosis) $\alpha_4 = \frac{3(k-2)}{(k-4)}$ for $k \geq 5$.

(6) The area under *t*-distribution curve for $t < t'$ is determined by the equation,

$$F_k(t) = p(t < t') = \int_{-\infty}^{t'} f(t)\ dt \qquad (8.4)$$

Students and other readers need not integrate actually for the area as the tables of area under the curve for different values of *t* are available and vice-versa.

Equation (8.4) is given to acquaint the reader with the know-how of getting the area if *t* is given and the value of *t* if area is given.

(7) *t*-distribution tends to normal distribution as *k* increases. For practical purposes, *t* is taken as equivalent to the normal distribution provided $k \geq 30$.

(8) Moment generating function for *t*-distribution does not exist.

t-distribution has tremendous utility in testing of hypothesis about one population mean or about equality of two population means when standard deviation is not known. Other uses of *t*-distribution will be known to the readers in the chapters ahead. *t*-distribution table has been provided in Appendix B.

CHI-SQUARE DISTRIBUTION

So far, we have been discussing the distribution of mean obtained from all possible samples, or a large number of samples drawn from a normal population, distributed with mean μ and variance σ^2/n. Now we are interested in knowing the distribution, of sample variances s^2 of these samples. Consider a random sample $X_1, X_2, \ldots, X_n$ of size n. Let the observations of this sample be denoted by $x_1, x_2, \ldots, x_n$. We know that the variance,

$$s^2 = \frac{1}{n-1}\Sigma_i (x_i - \bar{x})^2 \text{ for } i = 1, 2, \cdots n.$$

or

$$\Sigma_i (x_i - \bar{x})^2 = (n-1)s^2 = ks^2 \qquad (8.5)$$

where $k = (n-1)$.

A quantity ks^2/σ^2, which is a pure number, is defined as χ^2_k. Now we will give the distribution of the random variable χ^2_k, which was first discovered by Helmert in 1876 and later independently given by Karl Pearson in 1900. Another way to understand chi-square is: if $X_1, X_2, \ldots X_n$ are n independent normal variates with mean zero and variance unity, the sum of squares of these variates is distributed as chi-square with n d.f. The chi-square distribution was discovered mainly as chi-square with n d.f. The chi-square distribution was discovered mainly as a measure of goodness of fit in case of frequency distribution, i.e., whether the observed frequencies follow a postulated distribution or not. The probability density function (p.d.f.) of χ^2-variate is,

$$f_k(\chi^2) = \frac{1}{2^{k/2}\,\Gamma(k/2)}\,(\chi^2)^{\frac{1}{2}k-1}\,e^{-\frac{1}{2}\chi^2} \qquad (8.6)$$

Properties of Chi-square Distribution

(1) The whole chi-square distribution curve lies in the first quadrant since the range of χ^2 is from 0 to ∞.

(2) From the density function given by (8.6), it is evident that χ^2-distribution has only one parameter k, the degrees of freedom for χ^2.
Thus, the shape of the probability density curve mainly depends on the parameter k. The shape of the curves for four different degrees of freedom say, $k = 2$, $k = 7$, $k = 12$, $k = 20$ are given below.

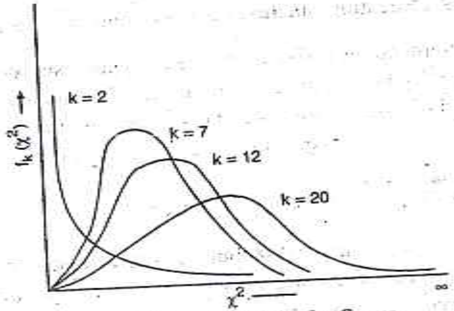


Fig. 8.2: Chi-square Distribution Curves.

(3) Chi-square distribution curve is highly positive skewed.

(4) It is an unimodal curve and its mode is at the point, $\chi^2 = (k - 1)$.

(5) The shape of the curve varies immensely specially when k is small. For $k = 1$ and 2, it is just like an hyperbola.

(6) Chi-square distribution is completely defined by one parameter k, which is known as the *degrees of freedom* of chi-square distribution.

(7) The constants for chi-square distributions are as follows:

(Mean) $\mu = k$

(Variance) $\sigma^2 = 2k$

(Skewness) $\alpha_1 = 2\left(\frac{2}{k}\right)^{\frac{1}{2}}$

(8) The moment generating function for chi-square distribution is,

$$\phi_{\chi^2}(t) = (1 - 2t)^{-k/2}$$

(9) r-th raw moment of chi-square distribution is,

$$\mu'_r = \frac{2^r\,\Gamma\left(\frac{k}{2}+r\right)}{\Gamma k/2}$$

Putting $r = 1$, $\mu'_1 = \dfrac{2\,\Gamma\left(\frac{k}{2}+1\right)}{\Gamma k/2} = k$

$$r = 2, \quad \mu'_2 = \frac{2^2\,\Gamma\left(\frac{k}{2}+2\right)}{\Gamma k/2}$$

$$= 4\left(\frac{k}{2}+1\right)\frac{k}{2} = k\,(k+2)$$

Again,

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

$$= k(k+2) - k^2 = 2k$$

and so on.

(10) It can be shown that for large degrees of freedom say, $k \geq 100$, the variable $(\sqrt{2\chi^2} - \sqrt{2k-1})$ is distributed normally with mean 0 and variance 1.

(11) If χ_1^2 and χ_2^2 are two independent chi-squares with degrees of freedom k_1 and k_2 respectively, their sum $(\chi_1^2 + \chi_2^2)$ will be distributed as chi-square with $(k_1 + k_2)$ d.f. This additive property of independent chi-squares, holds good for any number of chi-squares, i.e. if there are m independent chi-squares with $k_1, k_2, \ldots, k_m$ d.f. respectively, the sum $\Sigma_i \chi_i^2$ $(i = 1, 2, \cdots, m)$ will be distributed as chi-square with $\Sigma_i k_i$ d.f.

FISHER'S z-DISTRIBUTION

Mr Ronald A. Fisher defined a statistic z which is based upon the ratio of two-sample variances. Suppose s_1^2 and s_2^2 are two variances of random samples of sizes n_1 and n_2 respectively, drawn from two normal populations.

The statistic z is defined as,

$$z = \frac{1}{2}\log_e\left(\frac{s_1^2}{s_2^2}\right) \tag{8.7}$$

or

$$z = \log_e s_1 - \log_e s_2 \tag{8.7.1}$$

or

$$e^{2z} = \frac{s_1^2}{s_2^2} \tag{8.7.2}$$

Putting

$$\frac{s_1^2}{s_2^2} = F, \text{ we get}$$

$$e^{2z} = F \tag{8.7.3}$$

The letter F is the first letter of Fisher's name as a mark of respect. z-distribution is of the form,

$$y = \frac{ce^{k_1 z}}{(k_2 + k_1 e^{2z})^{(k_1+k_2)/2}} \tag{8.8}$$

where c is a constant, $k_1 = (n_1 - 1)$ and $k_2 = (n_2 - 1)$.

Properties of z-Distribution

(1) z-distribution is symmetrical about the point $z = 0$.

(2) z-distribution is a family of distributions in which each one exists for a pair of degrees of freedom k_1 and k_2.

Here the readers are acquainted with Fisher's z-distribution as it is closely associated with F-distribution given below.

F-DISTRIBUTION

From (8.7.3) it is evident that the ratio of two independent sample variances is denoted as F. Now we will consider the distribution of the ratio of two-sample variances in another manner and based on this new approach, the distribution of F is given. The distribution of F was worked out by G.W. Snedecor.

Consider two normal populations $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$. An independent sample of size n_1 is drawn from a population $N(\mu_1, \sigma_1^2)$. and of size n_2 from a population $N(\mu_2, \sigma_2^2)$. Let the sample variances be s_1^2 and s_2^2 respectively. From the chi-square distribution theory we know $k_1 s_1^2/\sigma_1^2$ is distributed as chi-square with k_1 d.f., i.e.

$$\frac{k_1 s_1^2}{\sigma_1^2} \sim \chi_1^2 \qquad (8.9)$$

where χ_1^2 has d.f. $k_1 = (n_1 - 1)$.

Similarly,

$$\frac{k_2 s_2^2}{\sigma_2^2} \sim \chi_2^2 \qquad (8.10)$$

where χ_2^2 has d.f. $k_2 = (n_2 - 1)$.

From (8.9), and (8.10) we can write that,

$$\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} = \frac{\chi_1^2/k_1}{\chi_2^2/k_2} \qquad (8.11)$$

$$= F_{k_1, k_2} \qquad (8.11.1)$$

In (8.11.1), k_1 and k_2 are called the degrees of freedom of F. Dividing s_u^2 ($u = 1, 2$) by its corresponding population variance standardizes the sample variance, in the sense that on the average both the numerator and denominator approach 1. Now we may be interested in testing the hypothesis that both the normal populations have the same variance, i.e. $\sigma_1^2 = \sigma_2^2$. Under this hypothesis,

$$\frac{s_1^2}{s_2^2} = F_{k_1, k_2} \qquad (8.12)$$

As a norm, the greater sample variance is taken as the numerator.

From (8.11) it is apparent that the ratio of two independent chi-squares is distributed as F and (8.12) reveals that under the hypothesis $\sigma_1^2 = \sigma_2^2$, the ratio of two independent sample variances is distributed as F. The probability density function of F-distribution is,

$$f_{k_1, k_2}(F) = \frac{(k_1/k_2)^{k_1/2}}{B\left(\frac{k_1}{2}, \frac{k_2}{2}\right)} \frac{F^{k_1/2-1}}{\left(1 + \frac{k_1}{k_2} F\right)^{(k_1+k_2)/2}} \qquad (8.13)$$

$$0 \le F < \infty$$

Obviously, F is always a positive number and the F-distribution curve wholly lies in the first quadrant. There are two parameters of F-distribution namely, k_1 and k_2. Hence the shape of F-distribution curve depends on k_1 and k_2. For two pairs of degrees of freedom (k_1, k_2), the shape of the F-distribution are shown in Fig. 8.3.

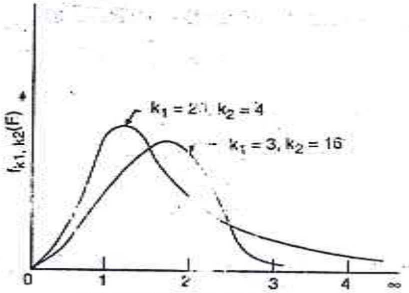


Fig. 8.3: F-distribution Curves.

Properties of F-Distribution

1. F-distribution curve extends on abscissa from 0 to ∞.
2. It is an unimodal curve and its mode lies on the point

$$F = \frac{k_2(k_1 - 2)}{k_1(k_2 + 2)}$$

which is always less than unity.

3. F-distribution curve is a positive skew curve. Generally the F-distribution curve is highly positive skewed when k_2 is small.
4. The constants of F-distribution are,

(mean) $\mu = \dfrac{k_2}{k_2 - 2}$ for $k_2 \ge 3$.

(variance) $\sigma^2 = \dfrac{2k_2^2(k_1 + k_2 - 2)}{k_1(k_2 - 2)(k_2 - 4)}$ for $k_2 \ge 5$.

5- Useful relation for F

$$F_{1-\alpha,(k_1,k_2)} = \frac{1}{F_{\alpha}(k_2, k_1)}$$

Chapter (7)

Tests of Significance

A research worker or an experimenter has always some fixed ideas about certain population(s) vis-a-vis population parameter(s) based on prior experiments, surveys or experience. Sometimes these ideas might have been fixed in the mind vicariously. There is a need to ascertain whether these ideas or claims are correct or not by collecting information in the form of data. In this way, we come across two types of problems, first is to draw inferences about the population on the basis of sample data and the other is to decide whether our sample observations have come from a postulated population or not. The first type of problem has already been covered in Chapter 7. In this chapter, we would be dealing with the second type of problem.

Generally, a hypothesis is established beforehand. By hypothesis we mean to give postulated or stipulated value(s) of a parameter. Also, instead of giving values, some relationship between parameters is postulated in the case of two or more populations. On the basis of observational data, a test is performed to decide whether the postulated hypothesis be accepted or not. This involves certain amount of risk. This amount of risk is termed as a level of significance. When the hypothesis is accepted, we consider it a non-significant result and if the reverse situation occurs, it is called a significant result. The tests, which are dealt within this chapter pertain to parametric tests. A test is defined as, "*A statistical test is a procedure governed by certain rules, which leads to take a decision about the hypothesis for its acceptance or rejection on the basis of sample values.*"

Statistical tests of hypotheses play an important role in industry, biological sciences, social sciences and economics, etc. The use of tests has been made clear through a number of practical problems.

1. A feed manufacturer announces that his feed contains forty per cent protein. Now to make sure whether his claim is correct or not, one has to take a random sample of the product and by chemical analysis, find the protein percentages in the samples. From these observed values he would decide about the manufacturer's claim for his product. This is done by performing a test of significance.

2. There is a process *A* which produces certain items. It is considered that a new process *B* is better than process *A*. Both the processes are put under operation and then the items produced by them are sampled and observations are taken on them. A statistical test is performed based on these observations which enables us to decide whether process *B* is better than *A* or not.

3. Often we are interested to know what is the best dose of a chemical treatment? Two or more doses of the chemical are applied or administered on a number of subjects and response is observed. Now it is tested statistically whether the doses differ significantly or not.

4. Psychologists are often interested in knowing whether the level of IQ of a group of school boys is up to a certain standard or not. In this case, some boys are selected and an intelligence test is conducted. The scores obtained by them pass through a statistical test and a decision is made whether their IQ is up to the standard or not.

There is no end to such types of practical problems where statistical tests can be applied. These are only a few examples. Here, one very important point is to be noted. Whatever conclusions are drawn about the population(s), they are always subjected to some error. Hence there is always some risk involved in these decisions. Thus, a level of significance is always associated with these decisions. Now we will discuss various items involved in the testing of hypothesis in an exact way before describing the statistical tests.

TYPES OF HYPOTHESIS

"A hypothesis is an assertion or conjecture about the parameter(s) of population distribution(s)".

In case we are considering more than one population, it may be about the relationship between the similar parameters of the distributions. For example, the mean μ of a distribution is fifty i.e., H: $\mu = 50$. The variance σ^2 of a distribution is 36. i.e. H: $\sigma^2 = 36$. For two populations, the hypothesis may be that the means μ_1 and μ_2 or variances σ_1^2 and σ_2^2 are equal i.e. H: $\mu_1 = \mu_2$ or H: $\sigma_1^2 = \sigma_2^2$. Many times, the statement in the notational form can be given in the following manner as well.

$$H\colon \mu > 0,\ H\colon \mu < 0,\ H\colon \mu = 0,\ H\colon \mu = c,$$

$$H\colon \mu_1 > \mu_2,\ H\colon \mu_1 \le \mu_2, \text{ etc.}$$

where *c* is a known constant value. Similarly for variance(s), the hypotheses may be given as,

$$H : \sigma^2 = \sigma_0^2,\ H : \sigma^2 > \sigma_0^2,\ H : \sigma_1^2 = \sigma_2^2,\ H : \sigma_1^2 < \sigma_2^2, \text{ etc}.$$

where σ_0^2 is a known fixed value. A hypothesis is further classified according to its nature and usage.

Null Hypothesis A hypothesis which is to be actually tested for acceptance or rejection is termed as null hypothesis. It is denoted by H_0.

Alternative Hypothesis It is a statement about the population parameter or parameters, which gives an alternative to the null hypothesis (H_0), within the range of pertinent values of the parameter, i.e., If H_0 is accepted, what hypothesis is to be rejected and vice versa. An alternative hypothesis is denoted by H_1 or H_A. The idea of alternative hypothesis was originated by Neyman. For instance,

if $H_0 : \mu = 0$, the alternatives are, $H_1 : \mu \neq 0$, $H_1 : \mu > 0$
or $H_1 : \mu < 0$,

if $H_0 : \mu_1 = \mu_2$, the alternatives are, $H_1 : \mu_1 \neq \mu_2$, $H_1 : \mu_1 > \mu_2$
or $H_1 : \mu_1 < \mu_2$.

if $H_0 : \sigma^2 = \sigma_0^2$, the alternatives are, $H_1 : \sigma^2 \neq \sigma_0^2$, $H_1 : \sigma^2 < \sigma_0^2$
or $H_1 : \sigma^2 > \sigma_0^2$

if $H_1 : \sigma_1^2 = \sigma_2^2$, the alternatives are, $H_1 : \sigma_1^2 \neq \sigma_2^2$, $H_1 : \sigma_1^2 > \sigma_2^2$
or $H_1 : \sigma_1^2 < \sigma_2^2$.

Simple and Composite Hypothesis If the statistical hypothesis completely specifies the distribution, it is called a *simple hypothesis*, otherwise it is called a *composite hypothesis*. For instance, we consider a normal population $N(\mu, \sigma)$ where σ^2 is known and we want to test the hypothesis, $H_0 : \mu = 25$ against $H_1 : \mu =$ 30. From these hypotheses we know that μ can take either of the two values, 25 or 30. In this case, H_0 and H_1 are both simple. But generally H_1 is composite, i.e. of the form, $H_1 : \mu \neq 25$, $H_1 : \mu < 25$ or $H_1 : \mu > 25$. Likewise, simple and composite hypothesis for any other parameter(s) can be stated.

TWO TYPES OF ERRORS

After applying a test, a decision is taken about the acceptance or rejection of null hypothesis vis-a-vis the alternative hypothesis. There is always some possibility of committing an error in taking a decision about the hypotheses. These errors can be of two types:

Type I error. Reject null hypothesis (H_0) when it is true.

Type II error. Accept null hypothesis (H_0) when it is false.

These two types of errors can be better understood with an example where a patient is given a medicine to cure some disease and his condition is scrutinised for some time. It is just possible that the medicine has a positive effect but it is considered that it has no effect or adverse effect. Thus, it is the *first kind of error* or *type I error*. On the contrary, if the medicine has an adverse effect but is considered to have had a positive effect, it is called the *second kind of error* or *type II error*. Now let us consider the implications of these two types of error. If type I error is committed, the patient will be given another medicine, which may or may not be effective. But if type II error is committed, i.e. the medicine is continued in spite of an adverse effect, the patient is likely to develop some other complications or may even die. This means that the type II error is much more severe than the type I error. Hence in drawing inference about the null hypothesis, practice is followed that type II error be minimized even at certain risk of type I error.

Level of Significance It is the quantity of risk of the type I error which we are ready to tolerate in making a decision about H_0. In other words, it is the probability of type I error which is tolerable. The level of significance is denoted by α and is conventionally chosen as 0.05 or 0.01. $\alpha = 0.01$ is used for high precision and $\alpha = 0.05$ for moderate precision.

P-Value Concept Another approach is to find out the *P*-value at which H_0 is significant, i.e., to find the smallest level α at which H_0 is rejected. In this situation, it is not inferred whether H_0 is accepted or rejected at level 0.05 or 0.01 or any other level. But the statistician only gives the smallest level α at which H_0 is rejected. This facilitates an individual to decide for himself as to how much significant the data are. This approach avoids the imposition of a fixed level of significance. About the acceptance or rejection of H_0, the experimenter can himself decide the level α by comparing it with the *P*-value. The criterion for this is that if the *P*-value is less than or equal to α, reject H_0 otherwise accept H_0.

CRITICAL REGION (C.R.)

A statistic is used to test the hypothesis H_0. The test statistic follows some known distribution. In a test, the area under the probability density curve is divided into two regions, viz., the *region of acceptance* and the *region of rejection*. The region of rejection is the region in which H_0 is rejected. It means that if the value of test statistics lies in this region, H_0 will be rejected. The region of rejection is called a *critical region*. Moreover, the area of the critical region is equal to the level of significance α. The critical region is always on the tail of the distribution curve. It may be on both the tails or on one tail, depending upon the alternative hypothesis.

One- and Two-tailed Tests If the alternative hypothesis, H_1 is of the type $\mu > \mu_0$; $\mu_1 <$ or $> \mu_2$; $\sigma^2 > \sigma_0^2$ or $\sigma_1^2 < \sigma_2^2$; etc., the critical region lies on only one tail of the probability density curve. In this situation the test is called *one-tailed test*. If H_1 is of the type $\mu > \mu_0$; $\mu_1 < \mu_2$; $\sigma^2 > \sigma_0^2$; $\sigma_1^2 > \sigma_2^2$; etc., the critical region is towards the right tail as shown in Fig. 9.1.

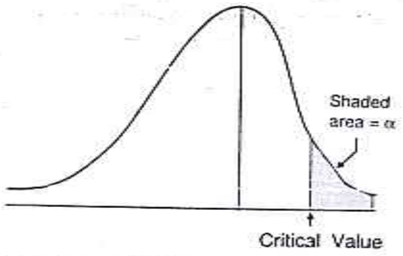


Fig. 9.1: One-sided Right Tailed Critical Region.

On the contrary, if H_1 is of the type $\mu < \mu_0$; $\sigma^2 < \sigma_0^2$, $\mu_1 > \mu_2$; $\sigma_1^2 < \sigma_2^2$;etc., the critical region lies on the left tail as depicted in Fig. 9.2.

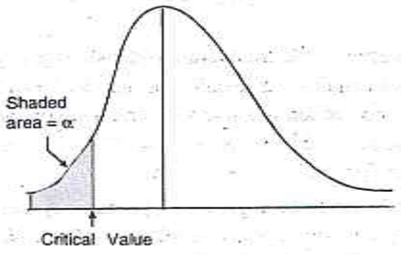


Fig. 9.2: One-sided Left Tailed Critical Region.

However, if H_1 is of the type $\mu \neq 0$; $\mu_1 \neq \mu_2$; $\sigma^2 \neq \sigma_0^2$ or $\sigma_1^2 \neq \sigma_2^2$ etc., the critical region lies on both the tails. In this situation a test is called a *two-tailed test*. In a two-tailed test, an area equal to α/2 lies on both the tails, for a test of significance level α. The critical regions are shown in Fig. 9.3.

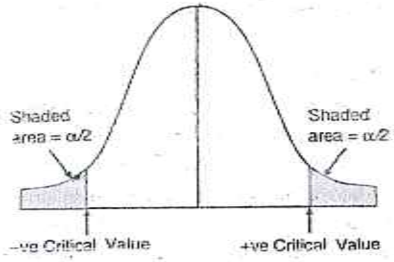


Fig. 9.3: Two-tailed Critical Regions.

SIZE AND POWER OF A TEST

The size of a test is the probability of rejecting the null hypothesis when it is true, and is usually denoted by α. The level of significance and size are synonymous in a practical sense. Therefore,

$$P\,(\text{reject } H_0 / H_0) = \alpha \tag{9.1}$$

The power of a test is defined as the probability of rejecting the null hypothesis when it is actually false, i.e. when H_1 is true. In short,

$$\begin{aligned} \text{Power} &= P\,(\text{reject } H_0 / H_1) \\ &= P\,(\text{accept } H_0 / H_1) \\ &= 1 - (\text{Prob. of type II error}) \\ &= 1 - \beta \end{aligned} \tag{9.2}$$

where β is the probability of type II error Among a class of tests, the best test is the one which has the maximum power for the same size.

RANDOMIZED TEST

A randomized test (*T*) is one in which no test statistic is used. The decision about the rejection of H_0 is taken, if it satisfies some predecided criterion. For instance, if it is decided that H_0 will be rejected if on tossing a coin it falls with the head on the upper side and will be accepted if it falls with the tail on the upper side. Since randomized test is rarely used, it is not explained further.

NON-RANDOMIZED TEST

A test T of a hypothesis H is said to be non-randomized if the hypothesis H is rejected on the basis that a test statistic belongs to the critical region c_T, i.e. $\psi_T(X_1, X_2, ..., X_n) \in c_T$. Some of the commonly used non-randomized tests are given in this chapter as well as in the next chapter.

DEGREES OF FREEDOM (d.f.)

It is apparent from the discussion made so far that in a test of hypothesis, a sample is drawn from the population of which the parameter is under test. The size of the sample varies since its depends either on the experimenter or on the resources available. Moreover, the test statistic involves the estimated value of the parameter which depends on the number of observations. Hence, the sample size plays an important role in testing of hypothesis and is taken care of by degrees of freedom.

Definition- Degrees of freedom is the number of independent observations in a set.

Note: The table values for the distribution of test statistics are provided in the appendix-B for various levels of significance and degrees of freedom. These table values make us decide about the rejection of H_0.

STUDENTS *t*-TEST

t-distribution has already been discussed in Chapter 8 where sampling distribution of $\bar{x}/s_{\bar{x}}$ was discussed. Here we shall make use of *t*-distribution in testing of hypothesis about the population mean or means.

Suppose, a small random sample $(X_1, X_2, ..., X_n)$ of size n has been drawn from a normal population having mean μ and variance σ^2 which are unknown. We want to test the hypothesis,

$$H_0 : \mu = \mu_0 \text{ against } H_1 : \mu \neq \mu_0$$

where μ_0 is some assumed value considered fit for μ. Let the observed values on random sample $(X_1, X_2, ..., X_n)$ be $(x_1, x_2, ..., x_n)$. Statistic t is given as,

$$t_{n-1} = \frac{\bar{x} - \mu_0}{s_{\bar{x}}} \tag{9.3}$$

For the sample values, expression for t is,

$$t_{n-1} = \frac{\sqrt{n}\,(\bar{x} - \mu_0)}{s} \tag{9.3.1}$$

where $\overline{x}$ is the sample mean, s is the standard deviation of the sample, $(n-1)$ in suffix indicates the d.f. of t.

On substituting formula (7.3) for s, (9.3.1) can be written as

$$t_{n-1} = (\overline{x} - \mu_0)\sqrt{\frac{n(n-1)}{\Sigma_i (x_i - \overline{x})^2}} \qquad (9.3.2)$$

where $i = 1, 2, ..., n$.

Definition Student-t is the deviation of estimated mean from its population mean expressed in terms of standard deviation.

To decide about the acceptance or rejection of H_0 vis-a-vis H_1, the calculated value of t is compared with the table value of t for $(n-1)$ d.f. and level of significance α. t-distribution table is provided in Appendix B (Table V) for different d.f. and various levels of significance. The tabulated t-value gives the critical value of t. More clearly, if $t_{cal} \geq t_{\alpha/2}$ for $(n-1)$ d.f., reject H_0, otherwise accept it.

Note: (1) In the above situation, a two-tailed test has to be applied since H_1 is $\mu \neq \mu_0$.

(2) In case of a one tailed test, i.e., for the alternative hypothesis H_1:$\mu > \mu_0$ reject H_0, if $t_{cal} \geq t_\alpha$ for $(n-1)$ d.f. C.R. lies on the right tail. If the alternative hypothesis is, H_1:$\mu < \mu_0$, reject H_0 if $t_{cal} \leq -t_\alpha$ for $(n-1)$ d.f. The C.R. lies on left tail.

(3) If t-table is provided for a two tailed critic region, it should be consulted as such. If a one tailed t-table has been provided, we sh..d consult the table for level α/2 where α is a prefixed level of significance.

(4) If H_0 is rejected at $\alpha = 0.01$, calculated t-va..e is called a highly significant value.

Assumptions about t-Test t-test is based on the following five assumptions:

1. The random variable X follows normal distribution. In other words the random sample has been drawn from a normal population.
2. All observations in the sample are independent.
3. The sample size is not large. There is no hard-and-fast rule which can be given to call a sample large. But, as a practice, a sample of size 30 or more is considered a large sample. At the same time one should note that at least five observations are desirable for applying a t-test.
4. The assumed value μ_0 of the population mean is the correct value.
5. The sample values are correctly taken and recorded.

In case the above assumptions do not hold good, the reliability of the test decreases. Further, some tests are very sensitive and some are not. The test which gives quite a satisfactory result in spite of some departure from the basic assumptions is called a *robust test*. It is interesting to point out that the student's t-test is a robust test.

Example 9.1. A breeder claims that his variety of cotton contains, at the most, 40 per cent lint in seed cotton. Eighteen samples of 100 grams each were taken, and after ginning the following quantity of lint was found in each sample.

Sample No:	1	2	3	4	5	6	7	8	9
Quantity of Lint in 100 g sample:	36.3	37.0	36.6	37.5	37.5	37.9	37.8	36.9	36.7
	10	11	12	13	14	15	16	17	18
	38.5	37.9	38.8	37.5	37.1	37.0	36.3	36.7	35.7

To check the breeder's claim, a t-test is performed as under. Here we have to test,

$$H_0 : \mu = 40 \text{ against } H_1 : \mu < 40$$

To test H_0, the test statistic is

$$t_{n-1} = \frac{\sqrt{n}(\overline{x} - \mu_0)}{s}$$

Given $\mu_0 = 40$

Now we compute $\overline{x}$ and s.

$$\overline{x} = \frac{669.7}{18} = 37.206$$

$$s^2 = \frac{1}{n-1}\{\Sigma_i x_i^2 - (\Sigma_i x_i)^2/n\}$$

$$= \frac{1}{17}\left\{24927.33 - \frac{(669.7)^2}{18}\right\}$$

$$= \frac{10.77}{17}$$

$$= 0.633$$

$$s = 0.796$$

$$t = \frac{\sqrt{n}\ (\overline{x} - \mu_0)}{s}$$

$$= \frac{\sqrt{18}\ (37.206 - 40)}{0.796}$$

$$= \frac{-11.854}{0.796} = -14.89$$

The table value of t at the prefixed $\alpha = 0.01$ and 17 d.f. is 2.567. Here, $t_{cal} < -2.567$. Hence H_0 is rejected. It means that the average percentage of lint in this cotton variety is less than 40 per cent.

Example 9.2. The life expectancy of people in the year 1970 in Brazil is expected to be 50 years. A survey was conducted in eleven regions of Brazil and the data obtained are given below. Do the data confirm the expected view?

Life expectancy: (years)	54.2,	50.4,	44.2,	49.7,	55.4,	57.0,
	58.2,	56.6,	61.9,	57.5,	53.4.	

Here we have to test,

$$H_0 = \mu = 50 \text{ against } \mu \neq 50$$

To test H_0, the statistic t is

$$t_{n-1} = \frac{\sqrt{n}(\bar{x}-\mu_0)}{s}$$

Now, we first complete $\bar{x}$ and s.

$$\bar{x} = \frac{598.5}{11} = 54.41$$

$$s^2 = \frac{1}{10}\{32799.91 - (598.5)^2/11\}$$

$$= \frac{1}{10}(236.07)$$

$$= 23.607$$

$$s = 4.859$$

$$t = \frac{\sqrt{11}\,(54.41-50)}{4.859}$$

$$t = \frac{14.626}{4.859}$$

$$= 3.01$$

The table value of t at $\alpha = 0.05$ and 10 d.f. is 2.228. Since $t_{cal} > 2.228$, reject H_0. It means that the life expectancy is more than 50 years. The chance of this result holding good for the whole population is 95 per cent.

TEST OF EQUALITY OF TWO POPULATION MEANS

Often, we have samples from two normal populations and want to test the hypothesis,

$$H_0 : \mu_1 = \mu_2 \text{ against } H_1 : \mu_1 \neq \mu_2$$

i.e., $$H_0 : \mu_1 - \mu_2 = 0 \text{ against } H_1 : \mu_1 - \mu_2 \neq 0$$

To test H_0 we come across either of the two situations, (i) $\sigma_1^2 = \sigma_2^2$, i.e., both the populations are distributed with the same variance, (ii) $\sigma_1^2 \neq \sigma_2^2$, i.e., two populations have unequal variances and are unknown for both the populations.

To test H_0 against H_1 in situation (i), the test statistic under H_0 based on two independent samples $X_{11}, X_{12}, \ldots, X_{1n_1}$ and $X_{21}, X_{22}, \ldots, X_{2n_2}$ of size n_1 and n_2, respectively from the two population is,

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\boxed{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}} \quad \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} \qquad (9.4)$$

Let the sample values be denoted by

$$(x_{11}, x_{12}, \ldots, x_{1n_1}) \text{ and } (x_{21}, x_{22}, \ldots, x_{2n_2}).$$

Then for sample observations, the expression for t is,

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \qquad (9.4.1)$$

Since under H_0, $\mu_1 = \mu_2$.

Statistic t has $(n_1 + n_2 - 2)$ d.f.,

where $\bar{x}_1$ and $\bar{x}_2$ are the means of the samples I and II respectively, s_p is the pooled standard deviation which is equal to $\sqrt{s_p^2}$, whereas

$$s_p = \frac{\sum_{=1}^{n_1}(x_{1i}-\bar{x}_1)^2 + \sum_{j=1}^{n_2}(x_{2j}-\bar{x}_2)^2}{(n_1+n_2-2)} \qquad (9.5)$$

s_p^2 can also be cal lated without taking the deviations from the means by the formula

$$s_p^2 = \frac{\left\{\sum_{i=1}^{n_1} x_{1i}^2 - (\sum_{i=1}^{n_1} x_{1i})^2/n_1\right\} + \left\{\sum_{j=1}^{n_2} x_{2j}^2 - (\sum_{j=1}^{n_2} x_{2j})^2/n_2\right\}}{(n_1+n_2-2)} \qquad (9.5.1)$$

In case the sample variances s_1^2 and s_2^2 are known or calculated earlier, then s_p^2 may be obtained by the formula.

$$s_p^2 \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1+n_2-2)} \qquad (9.5.2)$$

Formulae (9.5.1) and (9.5.2) can easily be obtained from (9.5) using the formulae for the sample variance as given below.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

$$= \frac{1}{n-1}\left\{\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2/n\right\}$$

It should be noted that the pooled variance has been calculated because the two populations possess the same variability. The pooled variance s_p^2 will be a good estimate of the pooled variance of $(\bar{x}_1 - \bar{x}_2)$ and the expression $s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$ is the standard deviation of $(\bar{x}_1 - \bar{x}_2)$. This is also called the standard error[1] of $(\bar{x}_1 - \bar{x}_2)$. Decision about the acceptance on rejection is taken according to the following rule:

(i) Given H_1: $\mu_1 \neq \mu_2$ and α level of significance, reject H_0

$$\left[\begin{array}{l} \text{if } t_{cal} \geq t_{\alpha/2},\ (n_1+n_2-2) \\ \text{or } t_{cal} \leq -t_{\alpha/2},\ (n_1+n_2-2) \end{array}\right.$$

otherwise H_0 is not rejected.

(ii) Given H_1: $\mu_1 > \mu_2$, and α level of significance, reject H_0

$$\text{if } t_{cal} \geq t_{\alpha},\ (n_1+n_2-2)$$

otherwise, H_0 is not rejected.

(iii) Given H_1: $\mu_1 < \mu_2$, and α level of significance, reject H_0

$$\text{if } t_{cal} \leq t_{\alpha},\ (n_1+n_2-2).$$

otherwise, H_0 is not rejected.

Example 9 3. The following table gives the monthly average of total solar radiation on a horizontal and an inclined surface at a particular place.

Month	Mean daily total radiation on horizontal surface (cal/cm²/day)	Mean daily total radiation on inclined surface (cal/cm²/day)
Jan.	363	536
Feb.	404	474
Mar.	518	556
Apr.	521	549
May.	613	479
Jun.	587	422
Jul.	365	315
Aug.	412	414
Sept.	469	505
Oct.	468	552
Nov.	371	492
Dec.	330	507

To test whether the average daily total radiations in a year on a horizontal and an inclined surface are equal, amounts to testing H_0: $\mu_1 = \mu_2$ against H_1: $\mu_1 \neq \mu_2$. To test H_0 under the assumption that the population variances of average monthly total radiation on horizontal and inclined surfaces are equal i.e. $\sigma_1^2 = \sigma_2^2$, the test statistic is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

$$S_1^2 = \frac{~~2543558~~\ 94646.3}{11} = 8604.21$$

$$S_2^2 = \frac{55197}{11} = 5017.91$$

Now make the following computations:

$$\sum_{i=1}^{12} x_{1i} = 5.421;\quad \bar{x}_1 = \frac{5421}{12} = 451.75;\quad \sum_{i=1}^{12} x_{1i}^2 = 2543583$$

$$\sum_{j=1}^{12} x_{2j} = 5801;\quad \bar{x}_2 = \frac{5801}{12} = 483.42;\quad \sum_{j=1}^{12} x_{2j}^2 = 2859497$$

$$s_p^2 = \frac{\left\{\sum_{i=1}^{n_1} x_{1i}^2 - (\Sigma_i x_{1i})^2/n_1\right\} + \left\{\sum_{j=1}^{n_2} x_{2j}^2 - (\Sigma_j x_{2j})^2/n_2\right\}}{(n_1+n_2-2)}$$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

$$= \frac{\{2543583 - (5421)^2/12\} + \{2859497 - (5801)^2/12\}}{22}$$

$$= \frac{94646.3 + 55197.0}{22}$$

$$= 6811.06$$

$$s_p = 82.53$$

$$t = \frac{451.75 - 483.42}{\sqrt{\frac{94646.3/11}{12} + \frac{55197/11}{12}}}$$

The statistic

$$t = \frac{451.75 - 483.42}{82.53\sqrt{\frac{1}{12}+\frac{1}{12}}} \quad \frac{-31.67}{\sqrt{717.017 + 418.16}} = -0.94$$

$$= \frac{-31.67}{33.69} ✓$$

$$= -0.94$$

The table value of t at $\alpha = 0.05$ and 22 d.f. is 2.074. Since $t_{cal} > -2.074$; H_0 is not rejected. It means that on the whole, the average daily total radiations on the horizontal surface and the inclined surface are equal.

Situation (ii) when $\sigma_1^2 \neq \sigma_2^2$. In this situation the problem that arises is that the variances cannot be pooled and hence the degrees of freedom for t cannot be obtained. The test of difference between two means in this case is known as *Berhans-Fisher* problem. Berhans and Fisher showed that the test statistic is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} \qquad (9.6)$$

[1] By definition, standard deviation of an estimate is called its standard error.

where s_1^2 are s_s^2 are estimates of σ_1^2 and σ_2^2, respectively, obtained from two independent random samples of sizes n_1 and n_2, selected randomly from two populations. The statistic (9.6) does not follow the student's t-distribution under H_0. The theoretical considerations of Berhans-Fisher problem are omitted in this book. Here, we give the Cochran's approximation in which case an ordinary t-table can be used and this is sufficiently accurate for testing H_0.

For taking a decision about H_0, the calculated value of t is compared with t^* which is obtained by the formula

$$t^* = \frac{t_1 s_1^2/n_1 + t_2 s_2^2/n_2}{s_1^2/n_1 + s_2^2/n_2} \qquad (9.7)$$

where t_1 and t_2 are the table values of t at a prefixed α level of significance with $(n_1 - 1)$ and $(n_2 - 1)$ d.f. respectively.

In case when $n_1 = n_2 = n$ (say), we obtain

$$t^* = t_\alpha \text{ with } (n-1) \text{ d.f.}$$

If $|t_{cal}| > t^*$, reject H_0, otherwise H_0 is not rejected.

Note. The test of significance by a t-test of the hypotheses about correlation coefficient, regression coefficient, etc., have been given in the respective chapters.

Example 9.4: The following data give the gain in body weight (kilograms) per heifer of two different breeds under a grazing treatment:

	Gain in body weight (kg)						
Breed 1:	57.3,	26.9,	53.2,	16.8,	44.8,	54.2,	71.4
Breed 2:	64.2,	52.2,	48.6,	26.6,	44.5,	71.8	

To test the equality of mean gain in the body weight of the two breeds, amounts to the testing of hypothesis,

$$H_0: \mu_1 = \mu_2 \text{ against } H_1: \mu_1 \neq \mu_2 .$$

Here we assume that the population variances of gain in weight in the two breeds are different, i.e., $\sigma_1^2 \neq \sigma_2^2$.

H_0 can be tested by the statistic,

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Now we calculate different constants in the formula:

Given $n_1 = 7$, $n_2 = 6$.

$$\sum_{i=1}^{7} x_{1i} = 324.6 ,\ \bar{x}_1 = 46.37 ,\ \sum_{i=1}^{7} x_{1i}^2 = 17162.02$$

$$\sum_{j=1}^{6} x_{2j} = 307.9 ,\ \bar{x}_2 = 51.32 ,\ \sum_{j=1}^{6} x_{2j}^2 = 17051.49$$

$$s_1^2 = \frac{1}{6}\{17162.02 - (324.6)^2/7\} = 351.64$$

$$s_2^2 = \frac{1}{5}\{17051.49 - (307.9)^2/6\} = 250.22$$

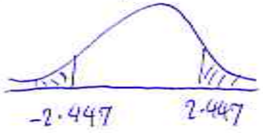
Substituting the values in the formula for t, we get

$$t = \frac{46.37 - 51.32}{\sqrt{\frac{351.64}{7} + \frac{250.22}{6}}}$$

$$= \frac{-4.95}{\sqrt{50.23 + 41.70}}$$

$$= \frac{-4.95}{9.59}$$

$$= -0.516$$



To take the decision about H_0, at 5 per cent level of significance, we calculate From Table V in appendix B, we find

$t_{0.025,5} = 2.447$ ~~and $t_{.025,5} = 2.571$~~

By formula (9.7),

$$t^* = \frac{2.447 \times 50.23 + 2.571 \times 41.70}{50.23 + 41.70}$$

$$= \frac{230.12}{91.93}$$

$$= 2.50$$

The calculated value of $t > -2.50$. Hence, hypothesis H_0 is not rejected. It means that the average gain in weight in the two breeds is equal.

PAIRED t-TEST

This test is applicable only when two samples are not independent and the observations are taken in pairs. In this case, for each observation in one sample there is a corresponding observation in the other sample pertaining to the same character. It means that the paired observations are on the same unit or matching units. For example, a manufacturer wants to test the circularity of ball-bearing. For this he selects some ball-bearings randomly and measures the diameter along two mutually perpendicular diameters. Another example may be, say, there are sixteen students divided into eight pairs such that in each pair, both the students have the same IQ. Two groups are formed so that one student of each pair is taken in a group. The two groups are exposed to two teaching methods and a test is performed after one month. The experimenter is interested to test whether there is any difference in the effectiveness of teaching methods on the basis of test scores. From the discussion, it is amply clear that the paired t-test is applicable when the paired observations are taken on the same or matching units or items.

Let the mean difference among the paired observations in the population be denoted by $\overline{D}$. The null hypothesis

$H_0: \overline{D} = 0$ against $H_1: \overline{D} \neq 0$ or $\overline{D} > 0$ or $\overline{D} < 0$ is tested as follows.

Suppose n observations in pairs and their differences, $(X - X') = d$, in the sample are as given below:

Pair No.	Sample values (X)	Sample values (X')	Difference $(X - X') = d$
1	x_1	x'_1	d_1
2	x_2	x'_2	d_2
3	x_3	x'_3	d_3
$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	x_i	x'_i	d_i
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	x_n	x'_n	d_n

The test statistic for testing H_0 is

$$t_{n-1} = \frac{\overline{d}}{s_{\overline{d}}} \tag{9.8}$$

since $\overline{D} = 0$,

$$t_{n-1} = \frac{\sqrt{n}\,\overline{d}}{s_{\overline{d}}} \tag{9.8.1}$$

Suffix $(n-1)$ denotes the d.f. of t.

where $\quad \overline{d} = \sum_{i=1}^{n} d_i / n$

and

$$s_{\overline{d}}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (d_i - \overline{d})^2 \tag{9.9}$$

$$= \frac{1}{n-1} \left\{ \sum_{i=1}^{n} d_i^2 - (\Sigma_i d_i)^2 / n \right\} \tag{9.9.1}$$

while calculating $\overline{d}$ or $s_{\overline{d}}$ etc., the sign of the differences is taken into consideration.

The criterion for decision about H_0 is

Given $H_1: \overline{D} \neq 0$, reject H_0 $\left[\begin{array}{l} \text{if } t_{cal} \geq t_{\alpha/2,\,n-1} \\ \text{if } t_{cal} \leq -t_{\alpha/2,\,n-1} \end{array}\right.$

otherwise H_0 is not rejected

Given $H_1: \overline{D} > 0$, reject H_0 if $t_{cal} \geq t_{\alpha,\,n-1}$

otherwise H_0 is not rejected

Given $H_1: \overline{D} < 0$, reject H_0 if $t_{cal} \leq -t_{\alpha,\,n-1}$

otherwise H_0 is not rejected.

Rejecting H_0 means that the difference is meaningful and can not be ignored by considering it as a casual difference or the difference due to random error.

Note: We can take $(X - X')$ or $(X' - X)$ as difference d. But whatever we take once, it should be maintained uniformly.

Example 9.5. The following table gives the pulsality index (Pi) of 11 patients

Patient No.	PI value		Difference
	During Seizure (X)	After Seizure (X')	$(X' - X) = d$
1	0.45	0.60	0.15
2	0.54	0.65	0.11
3	0.48	0.63	0.15
4	0.62	0.78	0.16
5	0.48	0.63	0.15
6	0.60	0.80	0.20
7	0.45	0.69	0.24
8	0.46	0.62	0.16
9	0.35	0.68	0.33
10	0.40	0.50	0.10
11	0.44	0.57	0.13

[*Source of data*: Gastroenterology, Vol. 84, 1983.]

We want to test whether there is a significant increase on the average in PI values, after seizure as compared to during seizure. This amounts to testing the hypothesis,

$H_0: \overline{D} = 0$ against $H_1: \overline{D} > 0$.

H_0 will be tested by the statistic,

$$t = \frac{\sqrt{n}\,\overline{d}}{s_{\overline{d}}}$$

Now we calculate $\overline{d}$ and $s_{\overline{d}}$. The differences d's are entered in the last column along with the data.

$$\sum_{i=1}^{11} d_i = 1.88\ ,\ \overline{d} = \frac{1.88}{11} = 0.171\ ,\ \sum_{i=1}^{11} d_i^2 = 0.3642$$

By formula (9.9.1).

$$s_d^2 = \frac{1}{10}\left\{0.3642 - \frac{(1.88)^2}{11}\right\}$$

$$= 0.004289$$

$$s_d = 0.065$$

Thus. $$t = \frac{\sqrt{11} \times 0.171}{0.065}$$

$$= 8.72$$

The table value of t at $\alpha = 0.05$ and 10 d.f. is 1.812. Since $t_{cal} > 1.812$, H_0 is rejected. It means that the PI value after seizure increases significantly as compared to during seizure

Example 9.6. The data below show the infiltration rate (cm/h) at an upstream and a downstream of the experimental plots.

Plot No.	Infiltration Rate		Difference
	Upstream (X)	Downstream (X')	$(X - X') = d$
1	15.66	13.56	2.10
2	15.54	13.62	1.92
3	13.00	12.30	0.70
4	13.62	13.20	0.42
5	20.46	18.78	1.68
6	19.80	11.94	7.86
7	13.02	14.76	–1.74
8	22.08	15.18	6.90
9	17.40	10.86	6.54
10	13.80	11.52	2.28
11	18.18	15.30	2.88
12	11.16	12.00	–0.84
13	16.80	6.30	10.50
14	20.10	9.12	10.98
15	13.20	9.84	3.36
16	11.16	10.74	0.42

[*Source of data: Annals of Arid Zone*, Vol. 22, 1983]

To know whether the average difference in infiltration rate upstream and downstream is significant or not, amounts to testing the hypothesis,

$$H_0: \overline{D} = 0 \text{ against } H_1 : \overline{D} \neq 0$$

H_0 can be tested by the statistic

$$t = \frac{\sqrt{n}\ \overline{d}}{s_{\overline{d}}}$$

The difference d has been entered in the last column of the above table. Now compute $\overline{d}$ *and* $S_{\overline{d}}$.

$$\sum_{i=1}^{16} d_i = 55.96\ ,\ \overline{d} = 3.50\ ,\ \sum_{i=1}^{16} d_i^2 = 423.25$$

By formula (9.9.1),

$$s_d^2 = \frac{1}{15}\left\{423.25 - \frac{(55.96)^2}{16}\right\}$$

$$= \frac{227.53}{15}$$

$$= 15.17$$

$$s_d = 3.89$$

Thus $$t = \frac{\sqrt{16} \times 3.50}{3.89}$$

$$= 3.60$$

The table value of t at $\alpha = 0.01$ and 15 d.f is 2.602. Since $t_{cal} > 2.602$, H_0 is rejected. It shows that there is a highly significant difference between the infiltration rate along upstream and downstream.

INTERVAL ESTIMATION

The theory of interval estimation was developed by Jerzy Neyman. It was a breakthrough in the field of statistics. In point estimation, only one value is found through sample observations, as this represents the parameter. This value may or may not be a good representative of the parameter. But in an interval estimation, two limits (lower and upper) are found out through sample values within which any point may be taken as an acceptable value of the parameter, with certain confidence probability. The two limits are called the *confidence limits* (C.L.) and the difference between the upper and the lower confidence limits is called the *confidence interval*.

In case, different samples are drawn from the same population, then each sample will generally provide different estimates and these estimates themselves behave as a random variable. The standard deviation obtained by these estimated values is called the standard error and helps in finding the confidence interval.

Suppose, we are interested in finding out $(1-\alpha)$ per cent confidence interval where α is the tolerable probability of the event that an estimated value may lie outside. Consider a parameter θ of a distribution. Let x' be the value of the deviate for the distribution. $f(x, \theta)$ corresponding to the confidence probability $(1-\alpha)$. Let the sample estimate of θ be $\hat{\theta}$ and σ_θ the standard error of $\hat{\theta}$. The confidence interval is given by $\hat{\theta} \pm x' \sigma_{\hat{\theta}}$. Generally σ_θ is not known and hence it is replaced by $s_{\hat{\theta}}$, the estimated value of $\sigma_{\hat{\theta}}$. Thus, the confidence limits are given by $(\hat{\theta} \pm x' s_{\hat{\theta}})$. The lower limit is $(\hat{\theta} - x' s_{\hat{\theta}})$ and upper limit is $(\hat{\theta} + x' x_{\hat{\theta}})$. The confidence interval is $\{(\hat{\theta} + x' s_{\hat{\theta}}) - (\hat{\theta} - x' s_{\hat{\theta}})\}$ i.e. $2x' s_{\hat{\theta}}$.

The above general ideas will be implemented to the specific problems which will further elucidate the concepts about interval estimation. Let us draw a small sample of size $n (n < 30)$ from a normal population $N(\mu, \sigma^2)$. To find out the confidence limits for the mean μ when σ is unknown, we have to consider the statistic t. A value of sample mean is an acceptable value of population mean if and only if the statistic t lies between $-t_\alpha$ and t_α, i.e.,

$$-t_\alpha \leq \frac{\bar{x}-\mu}{s/\sqrt{n}} \leq t_\alpha \qquad (9.10)$$

where t_α is the table value of t at α level of significance and $(n-1)$ d.f.

Taking the left side of inequality (9.10) we get,

$$-t_\alpha \frac{s}{\sqrt{n}} \leq (\bar{x}-\mu)$$

or

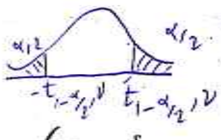
$$\mu \leq \bar{x} + t_\alpha \cdot \frac{s}{\sqrt{n}} \qquad (9.11)$$

Similarly, taking the right side of inequality (9.10) we get

$$\frac{\bar{x}-\mu}{x/\sqrt{n}} \leq t_\alpha$$

or

$$\bar{x} - t_\alpha \cdot \frac{s}{\sqrt{n}} \leq \mu \qquad (9.12)$$



combining (9.11) and (9.12) we obtain, $\bar{x} - t_\alpha \cdot \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_\alpha \cdot \frac{s}{\sqrt{n}}$ (9.13)

Thus, $\left(\bar{x} - t_\alpha \cdot \frac{s}{\sqrt{n}}\right)$ is the lower limit for μ and $\left(\bar{x} + t_\alpha \cdot \frac{s}{\sqrt{n}}\right)$ is the upper limit for μ. $(1-\alpha)$ per cent confidence interval is $2t_\alpha \cdot \frac{s}{\sqrt{n}}$. Similarly $(1-\alpha)$ per cent confidence limits for the difference between two normal population means i.e. $(\mu_1 - \mu_2)$ are given as follows. Let small samples $(x_{11}, x_{12}, \cdots x_{1n_1})$ and $(x_{21}, x_{22}, \cdots x_{2n_2})$ of size n_1 and n_2 be independently drawn from populations $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ respectively.

Consider the situation $\sigma_1^2 = \sigma_2^2$. In this situation the confidence limits for $(\mu_1 - \mu_2)$ are

$$(\bar{x}_1 - \bar{x}_2) \pm t_\alpha \cdot \cancel{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \ \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \qquad (9.14)$$

where $\bar{x}_1$ and $\bar{x}_2$ are sample means, t_α is the table value of t at α probability and $(n_1 + n_2 - 2)$ d.f. s_p is given by the formula (9.5) through (9.5.2). In a similar manner, the confidence limits for $(\mu_1 - \mu_2)$ in the situation $\sigma_1^2 \neq \sigma_2^2$ are,

$\nu = n_1 + n_2 - 2$

$$(\bar{x}_1 - \bar{x}_2) \pm t^*_{1-\alpha/2, \nu} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \qquad (9.15)$$

where s_1^2 and s_2^2 are two sample variances and t^* is as given by (9.7). Similarly $(1-\alpha)$ per cent confidence limits for $\bar{D}$ are,

$$\bar{d} \pm t_\alpha \cdot \frac{s_d}{\sqrt{n}} \ \checkmark \qquad (9.16)$$

All the notations in (9.16) are as given with paired t-test.

Example 9.7. For the data given in example 9.2, we find the 95 per cent confidence limits for μ.

All the calculations, made there, are used here directly. Thus the confidence limits for μ by the formula (9.13) are,

$$\text{C.L.} = 54.41 \pm 2.228 \times \frac{4.859}{\sqrt{11}}$$

$$= 54.41 \pm 3.26$$

Upper limit = 57.67 and lower limit = 51.15.

Example 9.8. Confidence interval for $(\mu_1 - \mu_2)$ for the data given in example (9.3) is computed here. All the calculations made in the solution of example (9.3) are used here directly.

Thus the 95 per cent confidence limits for $(\mu_1 - \mu_2)$ by the formula (9.14) are,

$$\text{C.L.} = (451.75 - 483.42) \pm 82.53 \sqrt{\frac{1}{12} + \frac{1}{12}} \times 2.074$$

$s_1^2 = \frac{94646.3}{11} = 8604.2$

$s_2^2 = \frac{55197}{11}$

$$= -31.67 \pm 33.69 \times 2.074$$

$\sqrt{717.017 + 418.16}$

$$= -31.67 \pm \boxed{69.87}$$

69.8774

33.69237

Upper limit = 38.20 and lower limit = −101.54

lengths

The confidence interval = 38.20 − (−101.54)

= 139.74

LARGE SAMPLE TESTS

The test of hypothesis about a population mean or two population means, by the t-test, is applicable under the circumstances that population variance(s) is/are not known and the sample(s) is/are of small size. In cases where the population variance(s) is/are known, we use Z-test (normal test). Moreover, when the sample size is large, sample variance approaches population variance and is deemed to be almost equal to population variance. In this way, the population variance is known even if we have sample data and hence the normal test is applicable. The distribution of Z is always normal with a mean zero and a variance 1. The value of Z can be read from the table for the area under the normal curve, e.g., given $\alpha = 0.05$, $Z = 1.96$ and given $\alpha = 0.01$, $Z = 2.58$, when we are applying two-tailed test. Any other value of Z, for any given value of α, can be read from Table IV. For testing H_0: $\mu = \mu_0$ against $H_1 : \mu \neq \mu_0$, the test statistic is

$$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} \qquad (9.17)$$

whereas in (9.17), $\bar{x}$ is the sample mean and σ is the standard deviation based on large sample of size n.

Also for testing $H_0 : \mu_1 = \mu_2 + \Delta$ vs. $H_1 : \mu_1 \neq \mu_2 + \Delta$, the expression is,

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \qquad (9.18)$$

where Δ is a known quantity.

In (9.18), σ_1^2 and σ_2^2 are the two population variances and all other notations are as in (9.6). Here it should be remembered that variances calculated from large samples are treated as population variances.

Note: In case of a one-tailed test, the value of Z should be obtained from the table for the area under the normal curve, corresponding to the prescribed value of α. For traditional value of $\alpha = 0.05$, Z for one-tailed test is 1.645 and for $\alpha = 0.01$, Z is 2.33.

The test criterion for the three commonly encountered alternative hypotheses at α level of significance is,

Given H_1: $\mu \neq \mu_0$ or $\mu_1 \neq \mu_2 + \Delta$.

$$\text{reject } H_0 \left[\begin{array}{l} \text{if } Z \geq Z_{\alpha/2} \\ \text{or if } Z \leq -Z_{\alpha/2} \end{array} \right.$$

Given H_1: $\mu > \mu_0$, reject H_0 if $Z \geq Z_\alpha$.

Given H_1: $\mu < \mu_0$, reject H_0 if $Z \leq Z_\alpha$.

Example 9.9 The table below gives the total income in thousand rupees per year of 36 randomly selected persons from a particular class of people.

Income (thousand Rs.)					
6.5	10.5	12.7	13.8	13.2	11.4
5.5	8.0	9.6	9.1	9.0	8.5
4.8	7.3	8.4	8.7	7.3	7.4
5.6	6.8	6.9	6.8	6.1	6.5
4.0	6.4	6.4	8.0	6.6	6.2
4.7	7.4	8.0	8.3	7.6	6.7

On the basis of the sample data, can it be concluded that the mean income of a person in this class of people is Rs. 10,000 per year?

We have to test the hypothesis,

$$H_0: \mu = 10 \text{ against } H_1: \mu \neq 10$$

Since the sample size is 36, we will use a normal test for which the statistic is

$$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

Now we compute $\bar{x}$ and σ.

$$\sum_{i=1}^{36} x_i = 280.7\,, \; \bar{x} = 7.80\,, \; \sum_{i=1}^{36} x_i^2 = 2368.75$$

$$\sigma^2 = \frac{1}{35}\left\{2368.75 - \frac{(280.7)^2}{36}\right\}$$

$$= \frac{180.07}{35}$$

$$= 5.14$$

$$\sigma = 2.27$$

Thus, $$Z = \frac{\sqrt{36}\,(7.80 - 10)}{2.27}$$

$$= \frac{-13.2}{2.27}$$

$$= -5.81$$

The table value of Z from Table IV at $\alpha = 0.05$ for a two-tailed test is 1.96. Since $Z < -1.96$, reject H_0. It means that the average annual income is less than rupees ten thousand.

Example 9.10. Two samples were drawn from two normal populations $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$. The following information was available on these samples regarding the expenditure in rupees per month per family.

Sample 1:	$n_1 = 42\,, \; \bar{x}_1 = 744.85\,, \; \hat{\sigma}_1^2 = 158165.43$
Sample 2:	$n_2 = 32\,, \; \bar{x}_2 = 516.78\,, \; \hat{\sigma}_2^2 = 26413.61$

On the basis of the available information it is required to test whether the average expenditure per month per family is equal. For the said problem we have to test the hypothesis,

$H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$

Since the sample sizes are large, H_0 is tested by the statistic,

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Substituting the values of various terms in the formula we get

$$Z = \frac{744.85 - 516.78}{\sqrt{\frac{158165.43}{42} + \frac{26413.61}{32}}}$$

$$= \frac{228.07}{67.76}$$

$$= 3.36$$

The table value of Z at $\alpha = 0.05$ is 1.96. Since $Z > 1.96$, we reject H_0. It means that the average expenditure per month per family in the two populations is not equal.

Test of Hypothesis for Proportions If the observations on various items or objects are categorised into two classes c_1 and c_2 (binomial population), we often want to test the hypothesis, whether the proportion of items in a particular class, say c_1, is P_0 or not. For example, the management of a manufacturing concern introduces the new bonus scheme. Then the management wants to test whether 60 per cent employees will favour the new scheme. Thus for binomial population, the hypothesis

$$H_0 : P = P_0 \text{ against } H_1 : P \neq p_0 \text{ or } H_1 : P > p_0 \text{ or } H_1 : P < p_0$$

can be tested by z-test where P is the actual proportion of items in the population belonging to class c_1. Proportions are mostly based on large samples and hence z-test is given here. The test for a small sample case which involves a hyper-geometric distribution is omitted. Let a large sample of n items be selected randomly. Out of n items, n_1 belong to class c_1 and n_2 belong to c_2 where $n_2 = (n - n_1)$. Also let the estimated proportion in class c_1 be denoted by $\hat{p}$. In this way we have the following configuration.

Class	c_1	c_2	Total
No. of items:	n_1	n_2	n
Proportion:	$\hat{p}$	$\hat{q}$	1

where $\hat{p} = \frac{n_1}{n}$, $\hat{q} = \frac{n_2}{n}$, $\hat{p} = 1 - \hat{q}$

H_0 can be tested by the statistic,

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0\, q_0 / n}}, \qquad (9.19)$$

where $q_0 = 1 - p_0$

Decision about H_0 can be taken according to the following criterion.

Given $H_1 : P \neq p_0$, reject H_0 if $z \geq z_{\alpha/2}$ or $Z \leq Z_{\alpha/2}$

Given $H_1 : P > p_0$, reject H_0 if $z \geq z_\alpha$

Given $H_1 : P < p_0$, reject H_0 if $z \leq z_\alpha$

For prefixed $\alpha = 0.05$, $z_{\alpha/2} = 1.96$ and $z_\alpha = 1.645$ and for $\alpha = 0.01$, $z_{\alpha/2} = 2.58$ and $Z_\alpha = 2.33$.

For any other level of significance α, the readers should obtain the value of z_α or $z_{\alpha/2}$ from Table IV for area under the normal curve given in the appendix.

Example 9.11. To test the conjecture of the management that 60 per cent employees favour a new bonus scheme, a sample of 150 employees was drawn and their opinion was taken whether they favoured it or not. Only 55 employees out of 150 favoured the new bonus scheme.

Thus, we test the hypothesis,

$$H_0: P = 0.60 \text{ against } H_1: P \neq 0.60$$

The test for H_0 through (9.19) is,

$$z = \frac{0.367 - 0.60}{\sqrt{\frac{0.60 \times .40}{150}}} \qquad \text{Since } \hat{P} = \frac{55}{150} = 0.367$$

$$= -5.825 \quad \boxed{-5.92} \quad 25.4164 \quad q = .633$$

At $\alpha = 0.01$, $z < -2.58$. Hence H_0 is rejected. It means that 60 per cent employees do not favour the new bonus scheme.

Test of Equality of Proportions If we have two populations and each item of a population belonged to either of the two classes c_1 and c_2. A person is often interested to know whether the proportion of items in class c_1 in both the populations is the same or not i.e., we want to test the hypothesis,

$$H_0: P_1 = P_2 \text{ against } H_1: P_1 \neq P_2 \text{ or } H_1: P_1 > P_2$$

$$\text{or } H_1: P_1 < P_2$$

where P_1 and P_2 are the proportions of items in the two populations belonging to class c_1.

Two independent random samples of large sizes n_1 and n_2 are drawn from populations A and B respectively. Let the number of items belonging to classes c_1 and c_2 be as follows:

	Classes		
	c_1	c_2	Total
Sample from A	O_1	O_2	n_1
Sample from B	O'_1	O'_2	n_2
Total	$O_1 + O'_1$	$O_2 + O'_2$	$n_1 + n_2$

The estimated proportion in c_1 for the population A is $p_1 = \frac{O_1}{n_1}$, $q_1 = \frac{O_2}{n_1}$ and for population B is $p_2 = \frac{O'_1}{n_2}$, $q_2 = \frac{O'_2}{n_2}$.

H_0 against H_1 can be tested by the statistic,

$$z = \frac{|p_1 - p_2|}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \qquad (9.20)$$

under H_0 i.e. $P_1 = P_2$,

where
$$\hat{p} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} \text{ and } \hat{q} = 1 - \hat{p} \qquad (9.21)$$

$$= \frac{O_1 + O'_1}{n_1 + n_2} \qquad (9.21.1)$$

and $\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$ is the standard error of $(p_1 - p_2)$ under the assumption $P_1 = P_2 = p$. In case H_0 is doubted to be true, standard error of $(p_1 - p_2)$ should be restandardised using,

$$\sigma_{p_1 - p_2} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}} \quad ✓$$

The decision about H_0 against any of the three commonly encountered alternative hypotheses can be taken according to the rules discussed with one sample case.

Example 9.12. A sample of 400 families in an old city is selected randomly, and a sample of 500 families is randomly selected from several new colonies of the same city. A survey is conducted for the number of houses possessing television (TV) sets. The number of TV holders in the old city is 48 out of 400 selected families and 120 in new colonies out of 500 families. Then the hypothesis whether the proportion of TV holders in old city and in the new colonies is the same, i.e.

$$H_0: P_1 = P_2 \text{ vs. } H_1: P_1 \neq P_2$$

can be tested as under:

	Television		
	Holders	Non-holders	Total
Old city	48	352	400
New colonies	120	380	500

From the data,

$$p_1 = \frac{48}{400} = \frac{3}{.25}, \; p_2 = \frac{120}{500} = \frac{6}{25}$$

From (9.21.1),

$$\hat{p} = \frac{168}{900} = \frac{14}{75}$$

$$\hat{q} = 1 - \frac{14}{75} = \frac{61}{75}$$

$$\sqrt{\frac{.12(.88)}{400} + \frac{.24(.76)}{500}}$$

Thus, Z is computed by formula (9.20) as given below,

$$Z = \frac{|3/25 - 6/25|}{\sqrt{\frac{14}{75} \times \frac{61}{75}\left(\frac{1}{400} + \frac{1}{500}\right)}}$$

$$\frac{|.12 - .24|}{\sqrt{.000365 + .000264}}$$

$$= \frac{3 \times 75}{25\sqrt{14 \times 61 \times 0.004}}$$

$$= 4.8$$

$$Z = 4.785$$

The table value of Z at α = 0.05 for two-tailed test is 1.96. Since $Z > 1.96$, H_0 is rejected, which means that the proportion of TV holders in old city area and in new colonies is not the same at 5 per cent level of significance.

This chapter has covered the general theory of testing of hypothesis, t and Z tests. Besides these tests, there are a great number of test procedures suitable for test of significance in a variety of cases. It is impossible to cover all of them in one or two chapters. Hence, only two more tests namely, χ^2 and F test, which are widely applied and used, are discussed in the next chapter.

QUESTIONS AND EXERCISES

1. Throw light on the need of the testing of hypothesis.
2. Discuss a hypothesis. What types of hypotheses do you know? Discuss each of them.
3. Discuss two types of errors in the testing of hypothesis. What is their role in testing?
4. What is a critical region and on what basis, are we able to know about the position of critical region(s)?
5. Why are the degrees of freedom so important in taking a decision about the rejection or acceptance of a hypothesis?

Test the hypothesis whether he is doing the job correctly: Also estimate 95% confidence interval for average of the rice bag.

[$t_{.005,5} = 4.023$, $t_{.025,5} = 2.57$]

[$t_{.005,6} = 3.707$, $t_{.025,6} = 2.44$] *(M.Com, Calcutta, 1985)*

41. Is it likely that a sample of size 300 whose mean is 12, is a random sample from a large population with mean 12.5 and S.D. 5.2. *[I.C.W.A. (Inter); Dec, 1992]*

SUGGESTED READING

Goon, A.M., Gupta, M.K. and Dasgupta, B., *An Outline of Statistical Theory, Vol. II*, The World Press, Calcutta, 1980.

Harshbarger, T.R., *Introductory Statistics*, Macmillan Publishing Company, New York, 1977.

Huntsberger, D.V. and Billingsley, P., *Elements of Statistical Inference*, Allyn and Bacon, London, 1977.

Kendall, M.G., and Stuart., A, *The Advanced Theory of Statistics, Vol. II*, Charles Griffin London, 1961.

Lehman, E.L., *Testing Statistical Hypothesis*, Wiley Eastern, New Delhi, 1976.

Lindgren, B.W., *Statistical Theory*, Collier Macmillan Publishers, London, 3rd ed., 1976.

Meyer, P.L., *Introductory Probability and Statistical Applications*, Addison-Wesley Publishing Company, Philippines, 1965.

Rahman, N.A., *Practical Exercises in Probability and Statistics*, Charles Griffin, London, 1972.

Rao, C.R., *Linear Statistical Inference and its Applications*, Wiley Eastern, New Delhi, 2nd. ed., 1973.

Sadowski, W., *Statistics for Economist*, Pergamon Press, Oxford, 1967.

Walker, H.M., and Lev, J., *Statistical Inference*, Henry Holt Company, New York, 1953.

Wilks, S.S., *Mathematical Statistics*, John Wiley, New York. 1962.

Chapter (8)

Chi-Square and F-Test

Chi-square test is one of the most commonly used tests of significance. All basic ideas concerning the test of significance remain the same as discussed in Chapter 9. The chi-square distribution as discussed in Chapter 9 has its importance in getting the critical values of χ^2-variate. For convenience, the table for critical values of χ^2at various levels of significance, and for different degrees of freedom, is provided in appendix B.

The chi-square test is applicable to test the hypotheses of the variance of a normal population, goodness of fit of the theoretical distribution to observed frequency distribution, in a one way classification having k-categories. It is also applied for the test of independence of attributes, when the frequencies are presented in a two-way classification called the contingency table. The chi-square test dates back to 1900, when Karl Pearson used it for frequency data classified into K-mutually exclusive categories. It is also a frequently used test in genetics, where one tests whether the observed frequencies in different crosses agree with the expected frequencies or not. Now we give chi-square test of various hypotheses in sufficient details one by one.

TEST OF HYPOTHESIS FOR POPULATION VARIANCE

Suppose, on the basis of previous knowledge, we have a preconceived value, σ_0^2, of variance of a normal population. Draw a random sample of size $n(n < 30)$ from this population. On the basis of n sample observations $(x_1, x_2, ..., x_n)$, the postulated value σ_0^2 of the population variance σ^2 is to either be substantiated or refuted with the help of a statistical test. For this the hypothesis

$$H_0 : \sigma^2 = \sigma_0^2 \text{ vs } H_1 : \sigma^2 \neq \sigma_0^2$$

is tested by the statistic

$$\chi^2 = \frac{\Sigma_i (x_i - \bar{x})^2}{\sigma_0^2} \qquad (10.1)$$

$i = 1, 2, ..., n$

$$= \frac{(n-1)\,s^2}{\sigma_0^2} \qquad (10.1.1)$$

where s^2 is the variance of the sample, χ^2-statistic has $(n-1)$ d.f. Reject H_0 at pre-decided level of significance α if $X^2_{cal} \ge \chi^2_{\alpha/2,n-1}$ or if $\chi^2_{cal} \le \chi^2_{(1-\alpha/2),n-1}$.

In case of one-tailed test, i.e., testing H_0 against $H_1 : \sigma^2 > \sigma_0^2$, the test criterion is that reject H_0 if $\chi^2_{cal} \ge x^2_{\alpha,n-1}$.

Again for testing H_0 against $H_1 : \sigma^2 < \sigma_0^2$, the test criterion is that reject H_0 if $\chi^2_{cal} \le \chi^2_{(1-\alpha),n-1}$.

صارم يشتري

Example 10.1. An owner of a big firm agrees to purchase the product of a factory, if the produced items do not have variance of more than 0.5 mm^2 in their length. To make sure of the specifications, the buyer selects a sample of 18 items from his lot. The length of each item was measured to be as follows:

Length (mm):

18.57,	18.10,	18.61,	18.32,	18.33,	18.46,
18.12,	18.34,	18.57,	18.22,	18.63,	18.43,
18.37,	18.64,	18.58,	18.34,	18.43,	18.63

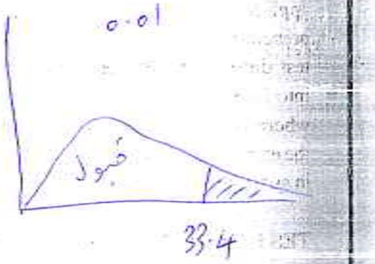
On the basis of the sample data, the hypothesis

$$H_0 : \sigma^2 = 0.5 \text{ vs. } H_1 : \sigma^2 > 0.5$$

can be tested by the statistic

$$\chi^2 = \frac{\Sigma_i (x_i - \bar{x})^2}{\sigma_0^2}$$

$i = 1, 2, ..., 18$



We calculate, $\Sigma_i (x_i - \bar{x})^2 = \Sigma_i x_i^2 - \frac{(\Sigma_i x_i)^2}{n}$

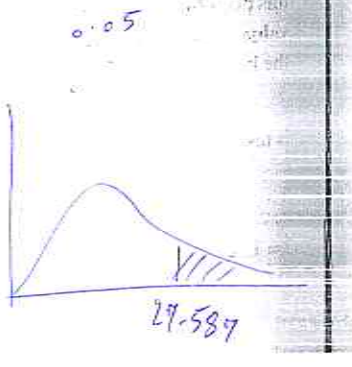
For the given data,

$$\Sigma_i x_i^2 = 6112.64 \; ; \; \Sigma_i x_i = 331.69$$

$$\therefore \quad \Sigma_i (x_i - \bar{x})^2 = 6112.64 - \frac{(331.69)^2}{18}$$

$$= 6112.640 - 6112.125$$

$$= 0.515 .$$



Thus,

$$\chi^2 = \frac{0.515}{0.5}$$

$$= 1.03$$

For $\alpha = 0.05, \chi^2_{0.05,17} = 27.587$. Since the calculated value of χ^2 is 1.03 which is not greater than 27.587, we accept the null hypothesis, $\alpha^2 = 0.5$ at $\alpha = .05$. It means that the buyer should purchase the lot.

TEST OF GOODNESS OF EIT

Generally the population under study has been taken to follow a known distribution such as normal, binomial or Poisson distribution. There is neither enough evidence nor enough logic, in the number of cases, to assume a particular distribution for the data. In such a situation one has every right to question the validity of such an assumption. To test the assertion of how closely the actual distribution approximates to a particular theoretical distribution, chi-square test is appropriate. To assume that the population is distributed normally is a common practice and hence we explain the test of goodness of fit of normal population first.

Let there be k class intervals and the corresponding frequencies be $f_1, f_2, ..., f_k$. The area of normal curve within each interval is found from the table of area under the normal curve (see appendix). On multiplying this area by the total of frequencies, we get the expected (theoretical) frequencies. In this way, the expected frequency for each interval is obtained. Since the tables are provided for standard normal curve, we first change each limit of class interval into a standard normal deviate by using the formula,

$$Z = \frac{x - \bar{x}}{s}$$

where $\bar{x}$ is the sample mean and s is the sample standard deviation.

Let the expected frequencies, as worked out in k classes, be $f_1', f_2', \cdots f_k'$ respectively.

Chi-square statistic is

$$\chi^2 = \Sigma_i \frac{(f_i - f_i')^2}{f_i'} \qquad (10.2)$$

$i = 1, 2, ..., k.$

Degrees of freedom for χ^2 in (10.2) are $(k - p - 1)$ where k is the number of class intervals and p is the number of parameters of the distribution estimated. One d.f. is reduced since $\Sigma_i f_i$ is a constant. For normal distribution two parameters μ and σ are estimated by $\bar{x}$ and s. Hence, in this case, chi-square has $(k-3)$ degrees of freedom.

If the calculated value of χ^2 is greater than the table value of χ^2 for $(k - p - 1)$ d.f. and level of significance α, reject H_0. Rejection of H_0 means, that the postulated theoretical distribution is not fit to the observed data, or in other words the data do not support the assertion about the theoretical distribution.

Example 10.2. The data regarding supplemental security income (SSI) programme, to escape from poverty, the poor people over 65 years as enrolled up to 1975 in an area are as follows:

(5 × 5), the maximum value of C is 0.894. it is meaningless to calculate the value of C if the hypothesis of independence is not rejected.

F-TEST

A large number of surveys or experiments are conducted to draw conclusions about the effect of certain factors or treatments. Observations are taken pertaining to the character under study. *F*-test is used either for testing the hypothesis about the equality of two population variances or the equality of two or more population means. The equality of the two population means has been dealt with *t*-test. Besides a *t*-test, we can also apply a *F*-test for testing equality of two population means. *F*-distribution has already been discussed in Chapter 8. The expression (8.12) clearly indicates that the ratio of two sample variances is distributed as *F*. The same has been used in this chapter. The *F*-test is given below in adequate details for testing various hypothesis.

TEST OF EQUALITY OF TWO POPULATION VARIANCES

Let there be two normal populations $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$. The hypothesis,

$$H_0 : \sigma_1^2 = \sigma_2^2 \text{ vs. } H_1 : \sigma_1^2 \neq \sigma_2^2$$

can be tested by *F*-test.

Let an independent sample of size n_1 be selected from population $N(\mu_1, \sigma_1^2)$ and of size n_2 from population $N(\mu_2, \sigma_2^2)$. Let the observations for these two samples be $(x_{11}, x_{12}, \ldots, x_{1n_1})$ and $(x_{21}, x_{22}, \ldots, x_{2n_2})$. Then the sample variances are,

$$s_1^2 = \sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1)^2/(n_1 - 1) \text{ and } s_2^2 = \sum_{j=1}^{n_2} (x_{2j} - \bar{x}_2)^2/(n_2 - 1)$$

s_1^2 and s_2^2 are the unbiased estimates of σ_1^2 and σ_2^2 respectively.

The statistic to test H_0 is,

$$F_{k_1, k_2} = \frac{s_1^2}{s_2^2} \tag{10.11}$$

where $\qquad k_1 = (n_1 - 1)$ and $k_2 = (n_2 - 1)$

As a norm, larger variance is taken in the numerator of (10.11) and the d.f. corresponding to it is denoted as k_1. If the calculated value of F is greater than the table value of F for (k_1, k_2) d.f. and α level of significance, reject H_0, i.e. if $F_{cal} > F_{\alpha/2}, (k_1, k_2)$, reject H_0 or if $F_{cal} < F_{1-\alpha/2}, (k_1, k_2)$, reject H_0.

For $H_1 : \sigma_1^2 > \sigma_2^2$, $\qquad$ reject H_0 if $F_{cal} > F_\alpha, (k_1, k_2)$.

For $H_1 : \sigma_1^2 < \sigma_2^2$, $\qquad$ reject H_0 if $F_{cal} < F_{1-\alpha}, (k_1, k_2)$

and in the reverse situation H_0 is not rejected.

Alternative Method $\quad H_0$ can also be tested by normal deviate test. We know from (8.7) that

$$z = \frac{1}{2} \log_e (s_1^2/s_2^2) \tag{10.12}$$

and conspicuously

$$\sigma_z^2 = \frac{1}{2}\left(\frac{1}{k_1} + \frac{1}{k_2}\right) \tag{10.13}$$

The statistic (z/σ_z) is approximately a standard normal deviate for large or moderately large d.f. k_1 and k_2.

If the value of (z/σ_z) is greater or equal to the normal deviate value for α level of significance, reject H_0. For $\alpha = 0.05$, the normal deviate value is 1.96.

In case the experimenter knows whether $\sigma_1^2 < \sigma_2^2$ or $\sigma_1^2 > \sigma_2^2$, he should use one-tailed test. Table value of F or z be obtained accordingly and decision about H_0 be taken in the usual manner.

Example 10.12. Life expectancy in 9 regions of Brazil in 1900 and in 11 regions of Brazil in 1970 was as given in the table below:

Regions	Life expectancy (years)	
	1900	1970
1	42.7	54.2
2	43.7	50.4
3	34.0	44.2
4	39.2	49.7
5	46.1	55.4
6	48.7	57.0
7	49.4	58.2
8	45.9	56.6
9	55.3	61.9
10		57.5
11		53.4

[*Source: The Review of Income and Wealth*, Series 29, No. 2, June 1983].

It is desired to confirm, whether the variation in life expectancy in various regions in 1900 and in 1970 is same or not.

Let the population in 1900 and 1970 be considered as $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, respectively.

The hypothesis,

$$H_0 : \sigma_1^2 = \sigma_2^2 \text{ vs. } H_1 : \sigma_1^2 \neq \sigma_2^2$$

can be tested by *F*-test.

First we calculate s_1^2 and s_2^2.

$$s_1^2 = \frac{1}{8}\left\{\sum_{i=1}^{9} x_{1i}^2 - \frac{(\Sigma_i x_{1i})^2}{9}\right\}$$

$$\Sigma_i x_{1i} = 405, \Sigma_i x_{1i}^2 = 18527.78$$

$$s_1^2 = \frac{1}{8}\left\{18527.78 - \frac{(405)^2}{9}\right\}$$

$$= \frac{302.78}{8} = 37.848$$

$$s_2^2 = \frac{1}{10}\left\{\sum_{j=1}^{11} x_{2j}^2 - \frac{(\Sigma_j x_{2j})^2}{11}\right\}$$

$$\Sigma_j x_{2j} = 598.5, \Sigma_j x_{2j}^2 = 32799.91$$

$$s_2^2 = \frac{1}{10}\left\{32799.91 - \frac{(598.5)^2}{11}\right\}$$

$$= \frac{236.07}{10} = 23.607$$

The test statistic,

$$F = \frac{s_1^2}{s_2^2}$$

$$= \frac{37.848}{23.607}$$

$$= 1.603$$

The table values of F at $\alpha = 0.05$ and (8, 10) d.f. for two-tailed test are $F_{0.025,(8,10)} = 3.85$ *and* $F_{0.975,(8,10)} = 0.233$. Calculated value of F is less than 3.85 and greater than 0.233. Hence, H_0 is not rejected. This confirms the equality of variances in 1900 and 1970 in regions of Brazil.

TEST OF EQUALITY OF SEVERAL POPULATION MEANS

Frequently we come across situations where we have to test the validity of the hypothesis of equality of k normal population means i.e. we want to test,

$H_0 : \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$

vs. H_1 : at least two of the means ($\mu' s$) are not equal,

where $k > 2$.

Suppose the observations in k random samples of size $n_1, n_2, \ldots, n_k$ from k normal populations $N(\mu_i, \sigma_i^2)$ for $i = 1, 2, \ldots, k$ are as given below.

	Samples					
	1	2	3	...	k	
	x_{11}	x_{21}	x_{31}	...	x_{k1}	
	x_{12}	x_{22}	x_{32}	...	x_{k2}	
	$\vdots$	$\vdots$	$\vdots$		$\vdots$	
	x_{1n_1}	x_{2n_2}	x_{3n_3}	...	x_{kn_k}	
Total	$x_{1.}$	$x_{2.}$	$x_{3.}$	...	$x_{k.}$	$x_{..} = G$

where x_{ij} denotes the j-th observation in the i-th sample for $i = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, n_i$. If we assume that the population variances $\sigma_1^2, \sigma_2^2 \ldots, \sigma_k^2$ are homogeneous, F statistic is,

$$F = \frac{\sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x})^2/(k-1)}{\sum_{i=1}^{k}\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 / \sum_{i=1}^{k} (n_i - 1)} \qquad (10.14)$$

where mean of i-th sample $\bar{x}_i = \sum_{j=1}^{n_i} x_{ij}/n_i$

Over mean, $\bar{x} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}/n = \frac{G}{n}$

where, $n_1 + n_2 + \ldots + n_k = n$

Statistic F has $\{(k-1), (n-k)\}$ d.f.

The expression in the numerator of (10.14) denotes the sample variance between samples and the expression in the denominator denotes the variance within samples. The test criterion is that reject H_0 if $F_{cal} > F_{\alpha,\, k-1,\, n-k}$ where $F_{\alpha,\, k-1,\, n-k}$ is the table value of F at α level of significance and $(k-1, n-k)$ d.f.

Calculations involved in F-test can conveniently be carried out through a table known as *analysis of variance table*. Analysis of variance is extensively used in analysis of data pertaining to agricultural and biological experiments, the details of which are kept out of scope of this book. Here we give details of analysis of variance as applicable to survey designs, simple experiments and regression analysis, etc., in this chapter and the chapters ahead.

ANALYSIS OF VARIANCE (ANOVA)

When a number of populations are under study and from each population a random sample or a group of units is selected, analysis of variance is a powerful tool to analyse the data. The purpose of analysis of variance is two-fold

(i) The total variance with respect to a variable (factor) is splitted into number of independent component variances, which are responsible for the total variance. Still, the sum of variances due to component factors never equals the total variance. Such a difference is attributed to error variance. This is the variance that occurs due to certain extraneous factors which can not be held responsible for any known component.

(ii) The next step is to test the null hypothesis about each of the component factors individually. This hypothesis is tested by finding out the ratio of variance of a component factor to the error variance. In ANOVA table, estimated variances are termed as *mean sum of square* (M.S.). The skeleton of analysis of variance table is given below.

Table 10.3: ANOVA

Source of variation	Degrees of freedom	Sum of squares	Mean sum of square	*F*-value
Due to	d.f.	S.S.	M.S.	*F*-value
A				
B				
Error				
Total				

Table (10.3) in practice is given with abbreviated notations. Degrees of freedom for various components are written in the usual way. Error degrees of freedom are obtained by subtracting the components d.f. from total d.f. Similarly the error S.S. is obtained by subtracting the components S.S. from total S.S. whereas the total sum of square is calculated by taking the total of the square of each individual value and subtracting from its value a factor which is known as *correction for mean* or *correction factor* (C.F.). Hence, the sum of squares for testing equality of *k*-population means, using the same notations as given in the preceding section are,

$$\text{(Correction for mean) C.F.} = \left(\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}\right)^2 \Big/ n$$

$$= G^2/n.$$

$$\text{Total S.S.} \quad = \sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}^2 - \frac{G^2}{n} = T_{xx}$$

$$\text{Between samples S.S.} \quad = \sum_{i=1}^{k} \frac{x_{i.}^2}{n_i} - \frac{G^2}{n} = S_{xx}$$

$$\text{Error S.S.} \quad = \sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}^2 - \sum_{i=1}^{k} \frac{x_{i.}^2}{n_i} = T_{xx} - S_{xx}$$

$$= E_{xx}$$

Mean sum of squares are obtained by dividing the sum of squares by its corresponding d.f. *F*-value for a component is obtained by taking the ratio of a component M.S. to error M.S. Analysis of variance table with full details is as presented below.

Table 10.4: ANOVA

Due to	d.f.	S.S.	M.S.	*F*-value
Bet. samples	$(k-1)$	$\sum_{i=1}^{k} \frac{x_{i.}^2}{n_i} - \frac{G^2}{N} = S_{xx}$	$\frac{S_{xx}}{k-1} = S_x$	$S_x/E_x = F$
Within samples (Error)	$(n-k)$	$\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}^2 - \sum_{i=1}^{k} \frac{x_{i.}^2}{n_i} = E_{xx}$	$\frac{E_{xx}}{n-k} = E_x$	
Total	$n-1$	$\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}^2 - \frac{G^2}{n} = T_{xx}$		

The calculated value of *F* is compared with the table value of *F* for α level of significance and $(k-1, n-k)$ d.f. Traditionally α is chosen to be 0.05, and for more precision, α is chosen to be 0.01, thought there is no hard-and-fast rule about it. We may choose any other value of α if it sounds more logical.

The above ANOVA is meant for one way classification. The ANOVA table may be extended for two or more way classification. In that situation, the component factors will increase in the ANOVA table accordingly. The methodology of analysis of variance is further explicated through a numerical example.

Example 10.13. The following table gives the gain in body weight (kg) per heifer during four grazing treatments.

Heifer No.	Gain in body weight (kg) Treatments				Total
	T_1	T_2	T_3	T_4	
			(x_{ij})		
1	67.3	74.2	63.1	48.7	
2	36.9	42.2	32.9	49.0	
3	63.2	58.6	59.2	62.0	
4	26.8	36.6	42.4	38.8	
5	54.8	54.6	34.0	48.2	
6	64.2	81.8	65.6		
7	81.4				
Total	394.6	348.0	297.2	246.7	1286.5

The hypothesis that the mean gain in weight of heifers under four treatments is equal or not, can be tested by *F*-test, i.e. the hypothesis,

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

against H_1 : at least two means are different,

can be tested through analysis of variance technique.

For the given data, $n_1 = 7, n_2 = 6, n_3 = 6, n_4 = 5$ and $n = 24$,

$$G = 1286.5,\ x_{1.} = 394.6,\ x_{2.} = 348.0,\ x_{3.} = 297.2,\ x_{4.} = 246.7$$

$$\sum_{i=1}^{4} \sum_{j=1}^{n_i} x_{ij}^2 = (67.3^2 + 36.9^2 + \ldots + 38.8^2 + 48.2^2)$$

$$= 74357.57$$

$$G^2/n = \frac{(1286.5)^2}{24} = 68961.76$$

Total S.S. $= 74357.57 - 68961.76$

$$= 5395.81$$

$$\sum_{i=1}^{4} \frac{x_{i.}^2}{n_i} = \frac{(394.6)^2}{7} + \frac{(348.0)^2}{6} + \frac{(297.2)^2}{6} + \frac{(246.7)^2}{5}$$

$$= 22244.16 + 20184.00 + 14721.31 + 12172.18$$

$$= 69321.65$$

Between treatment S.S.,

$$\sum_{i=1}^{4} \frac{x_{i.}^2}{n_i} - \frac{G^2}{n} = 69321.65 - 68961.76$$

$$= 359.89$$

Error S.S. $= 5395.81 - 359.89$

$$= 5035.92$$

Deu to	d.f.	S.S.	M.S.	F-value
Treatments	3	359.89	119.96	$\frac{199.96}{251.79} = 0.48$
Error	20	5035.92	251.79	
Total	23	5395.81		

From Table VII (ii), value of $F_{0.05,(3,20)} = 3.10$.

Since the calculated value of F is less than the table value, H_0 is not rejected. It leads to the result that the mean increase in weight of heifers, under four grazing treatments, is not significantly different at 5 per cent level of significance.

RELATION BETWEEN t, χ^2, F AND z

It is worth pointing out that for some hypothesis, more than one test can be used. For instance, the equality of two population means can be tested by t-test as well as F-test. The reason is that some relationship exists between t, χ^2 and F-distribution in particular situations and hence these tests become equivalent, permitting the use of either of these. Thus, the relationship for $k_1 = 1, k_2 = n$ is,

$$t_n^2 = F_{1,n} \qquad (10.15)$$

$$t_n = e^z = \sqrt{F_{1,n}} \qquad (10.15.1)$$

where z is given by (10.12).

If $k_1 = n, k_2 = \infty$, z is related to chi-square and the relation is

$$z = \frac{1}{2} \log_e \left(\frac{\chi^2}{n}\right). \qquad (10.16)$$

$$e^{2z} = \chi^2/n \qquad (10.16.1)$$

or

$$F = \chi^2/n \qquad (10.16.2)$$

Since we know, $F = e^{2z}$

If $k_1 = 1, k_2 = \infty$, then

$$F_{1,\infty} = \chi_1^2 \qquad (10.17)$$

where suffix 1 denotes the d.f. for χ^2.

Also from (10.15),

$$t_\infty^2 = \chi_1^2 \qquad (10.18)$$

If F-table is available only for the upper percentage points, the following identity enables us to obtain the F-values on the left-tail distribution. Let α be the level of the test and F be distributed with (k_1, k_2) d.f., the identity is,

$$F_{\alpha,(k_1,k_2)} = F_{1-\alpha,(k_1,k_2)} \qquad (10.19)$$

The identity (10.19) is very useful and easy to prove.

If the degrees of freedom for a chi-square are large i.e. more then 100, then the chi-square can be approximated to a standard normal variate using the relation,

$$Z = \sqrt{2\chi^2} - \sqrt{2k - 1} \qquad (10.20)$$

where k is the d.f. for a chi-square and $Z \sim N(0, 1)$. In this situation, the significance of null hypothesis can be tested by using the normal table for one tail.

These tests will be used in subsequent chapters also as and when the need arises to test various hypotheses.

QUESTIONS AND EXERCISES

1. What are the kinds of hypotheses that can be tested by the chi-square test?
2. What are the types of observational data suitable for the chi-square test, in a contingency table?
3. What do you understand by the test of goodness of fit?
4. Discuss a contingency table.
5. What is Yates' correction and its need?
6. Answer the following in not more than three lines.
   (a) Expected frequencies are obtained under which hypothesis?
   (b) Why can the chi-square not be negative?