UNIT 3 STATISTICAL TOOLS

Structure

3.1 INTRODUCTION

Why a learner of urban planning and development needs to know about statistical tests is simply because statistical tests will help him/her in analyzing data and drawing up inferences about the data. Those who are middle as well as at the decision making level need some understanding of statistical analysis to understand the strengths and weaknesses of published data to take decisions on whether to apply it in decision making. With the availability of several user friendly software, use of statistical tests has now become a reality, even for non-statisticians, provided they are computer literate and understand the basic principles of statistical analysis. This unit will help you to acquire knowledge about some basic statistical tools which you can use in data analysis.

After reading this unit you will be able to:

- define data, types of data and variables
- explain measures of central tendency
- calculate measures of dispersion, correlation and regression
- describe various inferential statistical tools.

3.2 DATA: MEANING AND TYPES

You know that some basic statistical tools need to be applied for the analysis of data while writing a report of urban development studies. Before describing the meaning of data, let us, know what we mean by statistics. According to Netter and Wasserman "statistics refers to the body of technique or methodology which has been developed for the collection, presentation and analysis of quantitative data and for the use of such data in decision-making". Statistical tools and techniques are used by the researchers to analyse and interpret data. Thus 'data' is a fundamental requirement for any decision making. Data is generally defined

as the evidence of fact which describes a group or a situation and from which conclusion is drawn. The data is the plural from the word 'datum', which means fact? Data is usually classified into two types:

i) Primary data and Secondary data

ii) Discrete data and continuous data

i) **Primary data and Secondary data:** Primary data is the first hand information gathered by an investigator or observer regarding a situation. Researcher collects primary data keeping problems in mind. According to P.V. Young, there are two types of sources of primary data i.e. direct primary sources and indirect primary sources. In direct primary sources, researchers have direct interaction of first hand filed work observation through interview schedule and questionnaire. While in indirect primary sources, he uses the medium of radio broadcasting, television appeal and other valuable documents for gathering information. Some of the advantages of primary data are: (i) Flexibility in collecting data; (ii) more appropriate for large area.

   The secondary data are gathered from personal or public documents. The various sources of secondary data are books, journals, reports, letters and diaries etc.

ii) **Discrete Data and Continuous Data:** Discrete data can take only a discrete value, that can be divided into categories or group such as male and female, white and black; boys and girls, etc.

   On the other hand, the continuous data can take any value including decimal. This is a type of data usually associated with some sort of physical measurement. The height of trees in a nursery is an example of continuous data.

3.3 VARIABLES AND TESTS

While dealing with the statistical tools and data you have to acquire knowledge about two important concepts i.e. variables and tests. Let us discuss them one by one.

i) **Variables:** Variables represents persons or objects which can be manipulated, controlled or measured for the sake of research. There are two types of variable in research such as independent and dependent variables.

   The independent variable is the variable that is varied or manipulated by the researcher. On the other hand, dependent variable is the response that is measured. In other words, an independent variable is the presumed cause; where as the dependent variable is the presumed effect. For example diseases among children are the independent variable, while infant mortality is the dependent variable.

ii) **Statistical Test:** Generally these are two types of tests applicable for statistical interpretation of data for testing hypothesis and drawing inferences, i.e. parametric test and non-parametric test. A parametric test is a test whose model specifies certain conditions about the parameters of the parent

population from which the sample was drawn. On the other hand, non-parametric test is a test whose model does not specify conditions about the parameters of the parent population from which sample was drawn.

In this session you read about data, variables and statistical tests, now answer the questions given in Check Your Progress-1

Check Your Progress 1

Note: a) Write your answer in about 50 words.

b) check your progress with possible answers given at the end of the unit.

1) What are the important types of data?

.....

.....

.....

.....

.....

2) What do you understand by non-parametric test?

.....

.....

.....

.....

.....

3.4 MEASURES OF CENTRAL TENDENCY

Measures of central tendency help the researcher to provide quantitative description of objects and events. Here numbers are assigned as per the rules and after the assignment of numbers, the individual score are compared with the average score to know the position of the individual in the group. Here average is called as the "Central Value". The score which represents the average performance of a group is known as central tendency. The main benefits to study the measures of central tendency are: (i) to get a single value that describe the characteristics of the entire group; (ii) to get a clear idea about the entire data; and (iii) lastly, it facilitates comparison.

There are three common measures of central tendency:

i) Mean or Arithmetic mean

ii) Median and

iii) Mode

3.4.1 Mean

Generally mean of a distribution is called as arithmetic mean. It is the average value of the group. Mean is the sum of the scores divided by the number of

scores. It is defined as the point on the scale of measurement obtained by dividing sum of all scores by the number of scores.

Mean is calculated from two types of data. (i) Ungrouped Data and (ii) Grouped Data

i) **Calculation of Mean from Ungrouped Data:** the formula for calculation of mean for ungrouped data is:

$$\overline{X} = \frac{\Sigma X}{N}$$

$\overline{X}$ = Mean

X = Individual score

N = Total number of scores

Ó = Indicates "sum of"

Example: The following marks 70, 30, 20, 90, 40 are secured by the 5 candidates in a term end examination conducted by a Municipality School. Calculate Mean.

Calculation of Mean

Candidates	Marks
A	70
B	30
C	20
D	90
E	40
N=5	Óx=250

$$\text{Mean} = \frac{\Sigma x}{N}$$

=

Mean = $\overline{X}$ = 50

ii) **Calculation of Mean from Grouped Data:** The mean from grouped data is calculated by applying following formula:

$$\text{Mean} = \overline{X} = \frac{\Sigma fx}{N}$$

Ó = stands for "sum of"

f = Stands for frequency

X = Stand for the mid point of class intervals

N = Total number of cases

Calculate mean value of the following group data:

Calculation Mean Value

Class Interval	Frequency
30-34	2
25-29	3
20-24	6
15-19	4
10-14	5
	N=20

At first you have to calculate the mid point of the class interval. The method of calculating mid point is

$$\text{Mid Point} = \text{LL} + \frac{\text{UL} - \text{LL}}{2}$$

LL = Lower Limit

UL = Upper Limit

$$\text{The Mid Point} = 30 + \frac{34 - 30}{2} = 32$$

For the first class interval 30-34

Class Interval	Frequency	X	fx
30-34	2	32	64
25-29	3	27	81
20-24	6	22	132
15-19	4	17	68
10-14	5	12	60
	N=20		Ófx= 405

$\overline{X} = \frac{\Sigma fx}{N}$

Now by using the formula you can calculate the mean of data given you.

Mean =

$$= \frac{405}{2} = 20.25$$

Let us know some of the important properties of mean.

Following are some of the important properties of mean:

i) The mean is used when a reliable and accurate measure of central tendency is needed.

ii) The mean is used when scores are distributed symmetrically around the central point.

Merits

i) It is easy to compute

ii) It is the best representative of the group.

iii) It is reliable.

Demerits

i) The value of mean depends on value of each item in the series.

ii) When scores are widely discrepant this measurement cannot be used.

iii) When scores are skewed mean can not be used.

3.4.2 Median

The median is a value that divides a distribution into two equal halves. The median is useful when the data is in ordinal scale, i.e., some measurements are much bigger or much smaller than the other measurement value. The mean of such data will be biased toward these extreme values. Thus, the mean is not a good measure of distribution, in this case. The median is not influenced by extreme values. The median value, also called the central or halfway value, (50th percentile, i.e., 50% value below median value, and 50% above it) is obtained in the following way:

- List the observations in order of magnitude (from the lowest to the highest value, or vice versa).
- Count the number of observations = n.
- The median value is the middle value, if n is odd {i.e., (n+1)/2} and the mean of two middle values, if n is even {i.e., (n/2) and the next value}

i) Calculation of Median from ungrouped data

Below we have given a few examples of how to calculate Median.

Example :

Case 1: The weights of 7 women are given in Table below, then calculate median value.

S.No.	Weight of women (kg)
1	40
2	41
3	42
4	43
5	44
6	47
7	72

The median value is the value belonging to observation number (7 + 1)/2, which is the fourth one value: 43 kg.

Case: If there are 8 observations as given in Table below then what will be median:

S.No.	Weight of women (kg)
1	40
2	41
3	42
4	43
5	44
6	47
7	49
8	72

The median would be 43.5 kg {the average of '(n/2=8/2) 4th value i.e. 43' and 'next value, i.e., 44'}; the median in this case would be (43+44)/2 = 43.5 kg}.

ii) Calculation of Median from Grouped Data

Let us calculate median for a grouped data given in below table.

Number of patients	Number of clinics	Cumulative frequency
0 - 19	5	5
20 - 39	8	13
40 - 59	10	23
60 - 79	11	34
80 - 99	19	53
100 - 119	10	63
120 - 139	9	72
140 - 159	8	80
Total	80	

$$L+\left[\left(\frac{N}{2}-F\right)\times d\right]/f$$

The steps for calculation of median from grouped are as follows:

Step1: The total of frequency is first divided by 2, i.e., 80/2 (=40). The cumulative frequency 40 will correspond to the class interval (80-99). This is called the median interval.

Step2: The formula is Median =

Step3: Record all values of symbol variables from the table as given below:

L (=80) is the lower limit of the median interval,

F (=34) is the cumulative frequency of the class, preceding to median class,

d (=20) is the width of class interval,

f (=19) is the frequency of median class.

Step4: Replace the symbol values with numeric values as noted in step3 in the formula,

Therefore, Median = 80 + [(40-34) × 20] / 19 = 80 + 6.32 = 86.32 patients.

Merits and Demerits

Merits

i) It is rigidly defined

ii) It is easily understood and easy to calculate. In some cases it can be located merely by inspection.

iii) It is not at all affected by extreme values.

iv) It can be calculated for distribution with open end classes.

Demerits

i) In case of even member of observation, median cannot be determined exactly. We merely estimate it by taking the mean of two middle terms.

ii) It is not based on all the observation for example the median of 10,25,50,60 and 65 is 50. We can replace the observations 10 and 25 by any two values which are smaller than 50 and the observation 60 and 65 by any two values greater than 50, without affecting the value of median. This property is sometimes described by saying that median is insensitive.

iii) It is not amenable to algebraic treatment.

iv) As compared with mean, it is affected much by fluctuations of sampling.

Uses

i) Median is the only average to be used while dealing with qualitative data which cannot be measured quantitatively but still can be arranged in ascending on descending order of magnitude, e.g. to find the average intelligence and average honesty among a group of people.

ii) It is to be used for determining the typical value in problems concerning wages, distribution of wealth, etc.

3.4.3 MODE

Let us consider the following statements.

i) The average height of an Indian (male) is 5’6”.

ii) The average size of the shoes sold in a shop 7.

iii) An average student in a hostel spends Rs. 150 p.m.

In all above cases, the average referred to its mode.

Mode is the value which occurs most frequently in a set of observations and around which the other items of the set, cluster densely. In other word, mode is the value of the variable which is predominant in the series. According to AM Tuttle “mode is the value which has the greatest frequency density in this immediate neighborhood”. Thus in case of dissent frequency distribution mode is the value of X corresponding to maximum frequency. Let us calculate mode from the data given below.

X:	12	3	4	5	6	7	8
F:	49	16	25	22	15	7	3

The value corresponding to the maximum frequency, viz 25 is 4. Hence mode is 4

Let us calculate mode of a grouped data given in the table below:

Class Interval	Frequency
30-34	2
25-29	3
20-24	6
15-19	4
10-14	5
	N=20

Following step will be used in the calculation of mode:

Step-1 Formula for calculating mode

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

L = (20) Lower limit of the model class

F_1 = (6) frequency of the modal class

F_0 = (3) frequency of the class proceeding modal class

F_2 = (4) frequency of the class succeeding modal class

h = (4) magnitude of class interval

Step-2 Replace the symbol values with the numeric values as noted in the step-1 in the formula

The calculated value of mode is:

$$\text{Mode} = 20 + \left[\frac{4-3}{2\times 20-4-3}\right]\times 4$$

$$= 20 + 1.08 = 21.08$$

If the distribution is moderately asymmetrical, the mean, median and mode obey the following empirical relationship:

Mode=3 Median – 2 Mean

Merits and Demerits of Mode

Merits:-

i) Mode is relatively comprehensible and easy to calculate.

ii) Mode is not at all affected by extreme values.

iii) Mode can be conveniently located even if the frequency distribution has class intervals of unequal magnitude classes proceeding and succeeding it are of the same magnitude.

Demerits:

i) Mode is ill defined. It is not always possible to find an early defined mode. In some cases, we may come across distribution with two modes, such distribution are called bimodal. If a distribution has more than two modes it is said to be multimodal.

ii) It is not based upon all the observation.

iii) It is not capable of further mathematical treatment.

iv) As compared with mean, mode is affected to a greater extent by fluctuation of sampling.

3.5 MEASURES OF DISPERSION

The mean, median, and mode are measures of the central tendency of a variable, but they do not provide any information of how much the measurements vary or are spread. This module will describe some common measures of variation (or variability), which in statistical text books are often referred to as measures of dispersion. Measures of dispersion or variability of a data give an idea up to which extent the values are clustered or spread out. In other words, it gives an idea of the homogeneity and heterogeneity of data. Two sets of data can have similar measures of central tendency but different measures of dispersion. Therefore, measures of central tendency should be reported along with measures of dispersion. There are various measures of dispersion. Those are discussed below:

3.5.1 Range

It is the simplest measure of dispersion. This can be represented as the difference between maximum and minimum values, or simply, as the maximum and minimum values for all observations.

Example : If the weights of 7 women are as given in Table below, then what is the range?

S.No.	Weight of women (kg)
1	40
2	41
3	42
4	43
5	44
6	47
7	72

The range would be 72 – 40 = 32 kg.

Although simple to calculate, the range does not tell us anything about the distribution of the values between the two extreme ones.

3.5.2 Percentiles

A second way of describing the variation or dispersion of a set of measurements is to divide the distribution into percentiles (100 parts). As a matter of fact, the concept of percentiles is just an extension of the concept of the median, which may also be called the 50th percentile. Percentiles are points that divide all the measurements into 100 equal parts. The 30th percentile (P30) is the value below which 30% of the measurements lie. The 50th percentile (P50), or the median, is the value below which 50% of the measurements lie. To determine percentiles, the observations should be first listed from the lowest to the highest just like when finding the median. However, in case of grouped data, percentile can be calculated on similar lines of calculating the median.

3.5.3 Mean Deviation

It is the average of deviation from arithmetic mean , where | | denotes

Mod, considering all differences 'as positive' or 'in absolute value'.

a) Calculation of Mean Deviation (A.D.)

i) Ungrouped Data- The formula used for calculation y mean deviation is:

$$\text{Average Deviation} = \frac{\text{Sum of all deviations}}{N}$$

$$\text{A.D.} = \frac{Ó\,|x|}{N} \qquad \sum \frac{|X_i - \bar{X}|}{n}$$

X = deviation of the raw score Mean

|x| = absolute deviation (disregarding the positive and negative sign)

N = Number of scores

Ó = sum total

Example: Calculate Mean Deviation (A.D.) from the following scores

10, 20, 30, 40, 50

Table: Calculation of Mean Deviation form ungrouped data

Score	Deviation (Raw Score –Mean)	x	|x|
10	10-30	-20	20
20	20-30	-10	10
30	30-30	0	0
40	40-30	10	10
50	50-30	20	20
Ó x=150			Ó|x|=60

$$\text{Mean} = \frac{Ó|x|}{N} = \frac{150}{5} = 30$$

$$\text{A.D.} = \frac{Ó|x|}{N} = \frac{60}{5} = 12$$

Thus A.D. = 12

3.5.4 Standard Deviation (S.D.)

Standard deviation is the only measure of dispersion which has algebraic treatment. It is the most stable measure of variability. The concept of S.D. was first suggested by Karl Pearson is 1893. Here all the deviations of the scores from mean are taken into account. In short it is considered as 'Root-Mean-Squarer-Deviation from Mean'. When the deviation are squared positive and negative signs become positive. When we take positive square root of the deviations, it is known as S.D. It is usually known as ó (sigma).

The formula used to calculate standard deviation is

$$\text{S.D.} = \sqrt{\frac{\Sigma(X - M)^2}{N}}$$

$$= \sqrt{\frac{\Sigma d^2}{N}}$$

ó =

Where,

Ó = sum total

d = deviation (score-mean)

N = total number of cases

a) Calculation of S.D. from Ungrouped Data

The formula to calculate S.D. from ungrouped data is

ó =

Example. Find out the S.D. of the following scores:

8, 9, 10, 11, 12, 13, 14, 15

Procedure for the calculation of S.D are as follows.

- Calculate Mean
- Calculate deviation against each score
- square the deviations
- find the total or sum of squared deviations
- divide sum of squared deviation by N
- Find the square root of the division.

Table: Calculation of S.D from Ungrouped Data

Score (X)	Deviation d(X-M)	d	d^2
8	8-12	-4	16
9	9-12	-3	9
10	10-12	-2	4
11	11-12	-1	1
12	12-12	0	0
13	13-12	1	1
14	14-12	2	4
19	19-12	7	49
Ó x=96			Ód²l=84

Mean = = 8

$$\text{S.D.} = \sqrt{\frac{\Sigma d^2}{N}} = \sqrt{\frac{84}{12}} = \sqrt{7} = 2.64$$

b) Calculation of S.D. from Grouped Data

$\frac{\Sigma x}{N} = \frac{96}{12}$

In grouped data, deviations are taken from the mid points of the class intervals. The deviations are squared and multiplied by frequency of the said class interval. Then the root, mean of square deviations is to be calculated.

The formula to calculate S.D. is

$$ó = \sqrt{\frac{\Sigma d^2}{N}}$$

Where,

Ó = sum total

f = frequency

d^2 = Square of deviation

N = total number of frequencies

Table: Calculation of Standard Deviation from Grouped Data

Class Interval (C.I)	Frequency (f)
10-14	2
15-19	3
20-24	4
25-29	5
30-34	6
	N=20

Computation of S.D. is given below.

C.I.	f	X	fx	X-M	d	d^2	fd^2
10-14	2	12	24	12-20.5	-8.5	72.25	144.50
15-19	8	17	136	17-20.5	-3.5	12.25	98.00
20-24	6	22	132	22-20.5	1.5	2.25	13.50
25-29	2	27	54	27-20.5	6.5	42.25	84.50
30-34	2	32	64	32-20.5	11.5	132.25	264.50
	N=20		Óf x=410				Ófd²=605.00

Mean =

$$= \frac{410}{20} = 20.5$$

S.D. =

$$= \sqrt{\frac{605.00}{20}} = \sqrt{30.25}$$

$$= 5.5$$

The standard deviation is 5.5

3.5.5 Coefficient of Variation

100 times the coefficient of dispersion based upon standard deviation is called coefficient variation (c.v), i.e.,

$$C.V. = 100 \times \frac{\sigma}{X}$$

According to profession Karl Pearson who suggested this measure, C.V.is the percentage variation in the mean, standard deviation being considered as then total variation in the mean.

Fun comparing the variability of two series, we calculate the co-efficient of variations for each series. The series having greater c.v. is said to be more variable than the other and the series having lesser c.v. is said to be more consistent than the other.

In this session you read about measures of central tendency and measures of dispersion, now answer the questions given in Check Your Progress-2

Check Your Progress 2

Note: a) Write your answer in about 50 words.

b) check your progress with possible answers given at the end of the unit.

1) What are the different measures of central tendency?

.....

.....

.....

.....

.....

2) What are the different measures of dispersion

.....

.....

.....

.....

.....

3.6 CORRELATION AND REGRESSION

3.6.1 Correlation: Concept and Meaning

Correlation is relationship between the two sets of continuous data; for example the relationship between height and body weight. Correlation statistics are used to determine the extent to which two independent variables are related and can be expressed by a measure called 'coefficient of correlation'. The correlation coefficient may be positive or negative and therefore it may vary from '-1' to '+1'. Positive correlation means that values of two different variables increase and decrease together. For example, height and weight correlate positively. Negative correlation means that if the value of one variable decreases then the value of the other variable increases (inverse relationship). For example, literacy and number of children in family may correlate negatively.

The strength of a correlation is determined by the absolute value of the correlation coefficient; the closer the value to 1, the stronger the correlation. For example, a correlation of -0.9 indicates an inverse relationship between two variables and shows a stronger relationship than that associated with a correlation of +0.2 or -0.5. Correlation between two variables is shown by scatter plot (Figure 1) below.

Correlation analysis is important because it can be used to predict values of one variable on the basis of value of other variable. A correlation does not mean causation but it also does not mean absence of causation, that is, if two variables exhibit strong correlation, then, one of the variables may cause the other. Correlation data is, therefore, not sufficient evidence for causation.

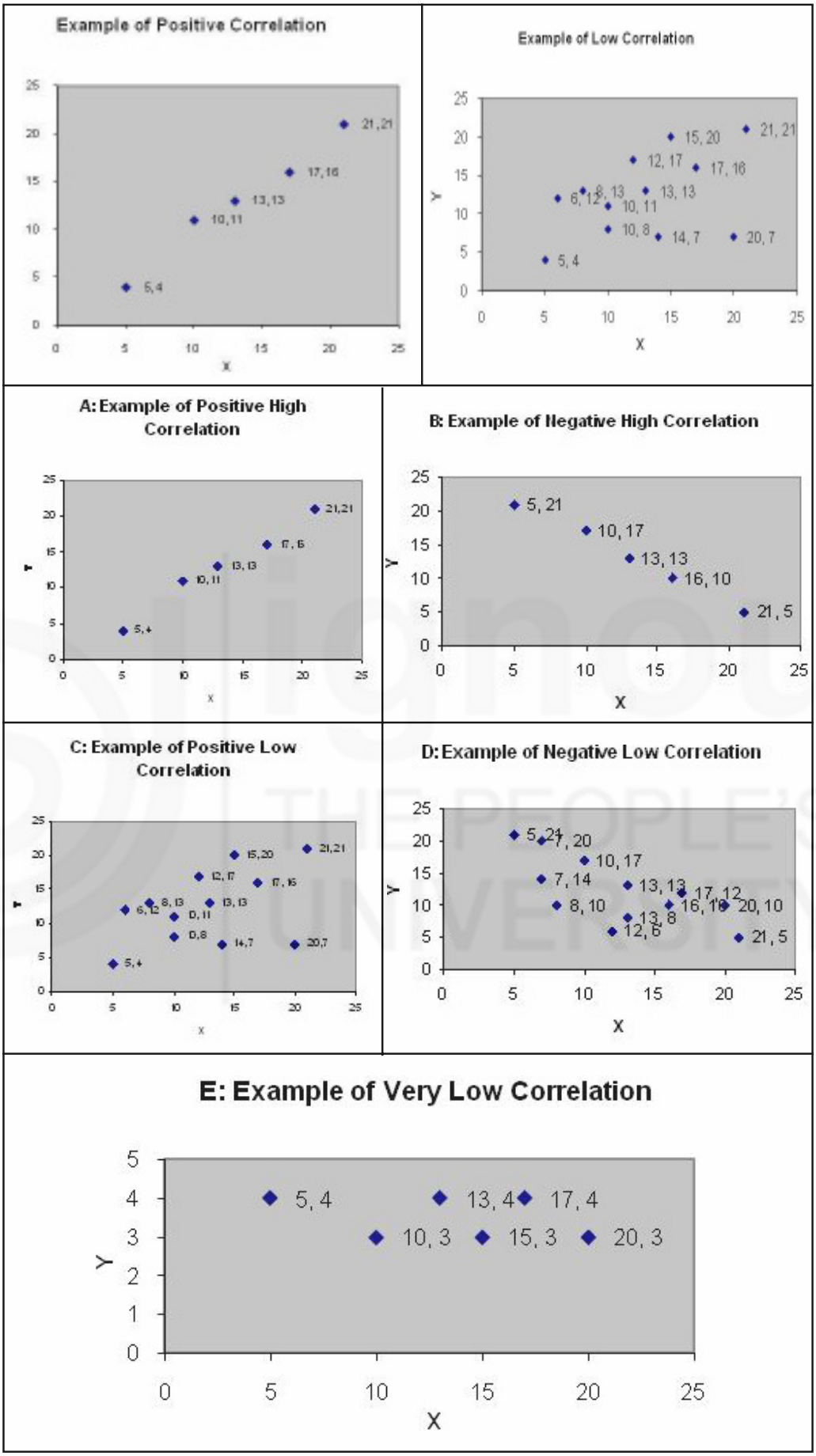

Fig. 3.1 : Scatter Diagram showing relation between two variables

The slopes of both the lines are identical in these two examples, but the scatter around the line is much greater in the second. Clearly the relationship between variables y and x is much closer in the first diagram.

If we are interested only in **measuring** the association between the two variables, then **Pearson's Correlation Coefficient (r)** gives us an estimate of the strength of the linear association between two numerical variables. Pearson's Correlation Coefficient can either be calculated by hand or the value of r can be obtained using either a calculator with built in capability to do the calculation or a variety of computer software programs. Note that in case there is curvilinear relationship, the value of r will be shown to be zero. The correlation coefficient has the following **properties**:

1) For any data set, r lies between '-1' and '+1'.

2) If r = +1, or -1, the linear relationship is perfect, that is, all the points lie exactly on a straight line. If most of the points lie on the line, then it is very strong relationship and r is near to 1. If r = +1, variable y increases as x increases (i.e., the line slopes upwards). (See Diagram A.) If r = -1, variable y decreases as x increases (i.e., the line slopes downward). (See Diagram B.)

3) If r lies between 0 and +1, the regression line slopes upwards, but the points are scattered about the line. (See Diagram C.) The same is true of negative values of r, between 0 and -1, but in this case the regression line slopes downward. (See Diagram D.)

4) If r = 0, there is very low linear relationship between y and x. This may mean that there is no relationship at all between the two variables (i.e., knowing x tells us nothing about the value of y). (See Diagram E.).

$\frac{\Sigma xy}{N\sqrt{\frac{\Sigma x^2}{N}}\sqrt{\frac{\Sigma y^2}{N}}}$

Calculation of the Pearson's Correlation Coefficient

Formula for calculation of Karl Pearson's correlation co-efficient is:

$$r = \quad =$$

$$= \frac{\Sigma xy}{N\sqrt{\frac{\Sigma x^2}{N}}\sqrt{\frac{\Sigma y^2}{N}}}$$

$$=$$

r = correlation coefficient

x = deviation from $\bar{x}$ (Arithmetic mean) of the first set of variables

y = deviation from $\bar{y}$ (Arithmetic mean) of the second set of variable

Ó = sign of summation

S_1 = standard deviation of the first set of variables

S_2 = standard deviation of the second set of variables

N = number of items in each set of variables

Example: Calculation the correlation coefficient between the following scores of history and mathematics.

Calculation the coefficient of correlation between the following scores of history and mathematics

Students	A	B	C	D	E
History (X)	65	56	69	60	75
Mathematics (Y)	60	76	74	80	85

Computation of coefficient of correlation

Student	History	Deviation		Mathematics From A.M.=65	Deviation From A.M.=65	y^2	xy
	X	x	x^2	Y	y		
A	65	0	0	60	-15	225	0
B	56	-9	81	76	+1	1	-9
C	69	+4	16	74	-1	1	-4
D	60	-5	25	80	+5	25	-25
E	75	+10	100	85	+10	100	+100
	325		Óx^2=222	375		Óy^2=352	Óxy=62
	=65			$\overline{Y} = \frac{375}{5}$ =75			

Coefficient of Correlation (r) = $\frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}} = \frac{62}{\sqrt{222 \times 352}}$

$$\frac{62}{\sqrt{78144}} = \frac{62}{\sqrt{280}} = +0.22$$

3.6.2 Regression: Concept and Meaning

In common language 'regression' means to return or to go back. In statistics, the term 'regression' is used to denote backward tendency which means going back to average or normal. The term 'regression' was first used by Sir Francis Galton.

Regression shows a relationship between the average values of two variables. So regression is average value of one variable for a given value of the other variable. It is useful for calculation of cause and effect relationship. The best average value of one variable associated with the given value of the other variable may be estimated or predicted by mean of an equation known as "Regression Equation" and also by the help of a line called as "Regression Line" which shows for a given value of other variable.

In order to estimate the best average values of the two variables, two regression equations are required and they are used separately. One equation is used for estimating the value of the first variable (X), this is called "Regression Coefficient of X on Y" or "Regression Equation of X on Y" and the second equation is used

for estimating the value of the second variable (Y) for a given value of the first variable called "Regression Coefficient of Y on X" and "Regression Equation of X on Y".

The formula for calculation of regression coefficient are:-

1) Regression Coefficient of X on Y is $b_{xy} = r \frac{S_x}{S_y}$

2) Regression Coefficient of Y on X is $b_{xy} = r$

S_x = Standard Deviation of X series

S_y = Standard Deviation of Y series

r = Correlation coefficient between X and Y

1) Regression Equation of X and Y is

$X - \bar{X} = r \quad (Y - \bar{Y})$

2) Regression Equation of Y and X is

$Y - \bar{Y} = r \quad (X - \bar{X})$

$\frac{S_y}{S_x}$

X = Value of X

Y = Value of Y

$\bar{X}$ = Arithmetic Mean of X series

$\bar{Y}$ = Arithmetic Mean of Y series

S_x = Standard Deviation of X series

S_y = Standard Deviation of Y series

r = Correlation coefficient between X and Y

Example: obtain lines of regression for the following data:

Computation of Regression Equation

$\bar{X}$	$(\bar{X} - 5)$ x	x^2	Y	(Y-12) y	y^2	xy
1	–4	16	9	–3	9	12
2	–3	9	8	–4	16	12
3	–2	4	10	–2	4	4
4	–1	1	12	0	0	0
5	0	0	11	–1	1	0
6	+1	1	13	+1	1	1
7	+2	4	14	+2	4	4

8	+3	9	16	+4	16	12
9	+4	16	15	+3	9	12
ÓX̄= 45 Óx=0		Óx²=60	ÓY=108	Óy=0	Óy²=60	Óxy=57

Regression Coefficient (b_{xy}) = r = r =

$$= \frac{9\times57-0\times0}{9\times60-0^2} = \frac{9\times57}{9\times60} = \frac{19}{20} = 0.95$$

Regression Coefficient (b_{yx}) = r = =

= = 0.95

i) The regression equation of X on Y is

$X-\bar{X} = _r \quad (Y-\bar{Y})$ is

$\bar{X} - 5 = 0.95\ (Y-12) = 0.95Y - 11.4$

$\bar{X} = 0.95Y - 11.4+5$

$\bar{X} = 0.95Y - 6.4$

ii) The regression equation of Y on X is

$Y-\bar{Y} = r \quad (\bar{X}-\bar{X})$ is

$Y - 12 = 0.95\ (\bar{X} - 5) = 0.95\bar{X} - 4.75$

$Y = 0.95\bar{X} - 4.75 + 12$

$Y = 0.95X + 7.25$

Differences between Regression and Correlation

Sl.	Correlation	Regression
1	Correlation quantifies the degree to which two variables are related. You simply are computing a correlation coefficient (r) that tells you how much one variable tends to change when the other one does.	Regression finds out the best fit line for a given set of variables.
2	With correlation you don't have to think about cause and effect. You simply quantify how well two variables relate to each other.	With regression, you do have to think about cause and effect as the regression line is determined as the best way to predict Y from X.

3	With correlation, it doesn't matter which of the two variables you call "X" and which you call "Y". You'll get the same correlation coefficient if you swap the two.	With linear regression, the decision of which variable you call "X" and which you call "Y" matters a lot, as you'll get a different best-fit line if you swap the two. The line that best predicts Y from X is not the same as the line that predicts X from Y.
4	Correlation is almost always used when you measure both variables. It rarely is appropriate when one variable is something you experimentally manipulate.	With linear regression, the X variable is often something you experimentally manipulate (time, concentration...) and the Y variable is something you measure.
5	In correlation, on the other hand, our focus is on the measurement of the strength of such a relationship.	In regression analysis, we examine the nature of the relationship between the dependent and the independent variables. In regression we try to estimate the average value of one variable h m the given
6	In correlation, all the variables are implicitly taken to be random in nature.	In regression, at our level, we take the dependent variable as random, or stochastic, and the independent variables as non-random or fixed.

In this session you read about correlation and regression, now answer the questions given in Check Your Progress-3

Check Your Progress 3

Note: a) Write your answer in about 50 words.

b) Check your answer with possible answers given at the end of the unit

1) Differentiate between correlation and regression.

.....

.....

.....

.....

.....

.....

3.7 HYPOTHESIS TESTING AND INFERENCIAL STATISTICS

3.7.1 Understanding True Difference

The analysis and interpretation of the results of our study must be related to the objectives of study. It is important to tabulate the data in univariate and/ or bi-variate or multivariate tables appropriate to the research objectives. We may find some interesting results. **For example**, in a study on nutrition, we find that 30% of the women included in the sample are anaemic as compared to only 20% of the men. How should we interpret this result?

- The observed difference of 10% might be a **true difference**, which also exists in the total population from which the sample was drawn.
- The difference might also be **due to the chance**; in reality there is no difference between men and women, but the sample of men just happened to differ from the sample of women. One can also say that the observed difference is due to sampling variation.
- A third possibility is that the observed difference of 10% is due to defects in the study design (also referred to as **Bias**). For example, we only used male interviewers, or omitted a pre-test, so we did not discover that anemia is a very important topic for women which require a female investigator.

If we feel confident that an observed difference between two groups cannot be explained by bias, we would like to find out whether this difference can be considered as a true difference. We can only conclude that this is the case if we can **rule out chance** (sampling variation) as an explanation. We accomplish this by applying a test of significance. A **test of significance** estimates the likelihood that the observed result (e.g., a difference between two groups) is due to chance or real. In other words, a significance test is used to find out whether a study result, which is observed in a sample, can be considered as a result which indeed exists in the study population from which the sample was drawn.

3.7.2 Tests of Significance

Different sets of data require different tests of significance. Throughout this module, two major sets of data will be distinguished.

- Two (or more) **groups**, which will be compared to detect **differences**. (e.g., men and women, compared to detect differences in anemia.)
- Two (or more) **variables**, which will be measured in order to detect if there is an **association** between them. (e.g., between anemia and income.)

In order to help you choose the right test, a flowchart and matrices will be presented for different sets of data. We will discuss how significance tests work. Please keep in mind that independent groups are treated as independent populations.

i) How to state Null (H_0) and Alternative (H_1) Hypothesis:

In statistical terms the assumption that **no real difference exists between groups** in the total study (target) population (or, that **no real association exists** between variables) is called the **Null Hypothesis (H_0)**. The **Alternative Hypothesis (H_1)** is that there **exists a difference between groups** or that a **real association exists** between variables. Examples of null hypotheses are

- There is no difference in the incidence of measles between vaccinated and non-vaccinated children.
- Males do not drink more alcohol than females.
- There is no association between families' income and malnutrition in their children.

If the result is statistically significant, we reject the **Null Hypothesis (H_0)** and accept the **Alternative Hypothesis (H_1)** that there is real difference

between two groups, or a real association between two variables. Examples of alternative hypotheses ($\mathbf{H_1}$) are:

- There is a difference in the incidence of measles between vaccinated and non-vaccinated children.
- Males drink more alcohol than females.
- There is an association between families' income and malnutrition in their children.

Be aware that 'statistically significant' does not mean that a difference or an association is ***of practical importance***. The tiniest and most irrelevant difference will turn out to be statistically significant if a large enough sample is taken. On the other hand, a large and important difference may fail to reach statistical significance if too small a sample is used.

ii) The Concept of Type I and Type II Error

There are four ways in which conclusion of the test might relate to in our study (i) true positive (ii) true negative and (iii) false positive and (iv) false negative. These may be expressed in terms of error in statistical test of significance in following terms:

Type I error (á): We reject the null hypothesis when it is true, or false positive error, or type I error **'á'** (called alpha). It is the error in detecting true effect.

In the above example, type I error would mean that the effects of two drugs were found to be different by statistical analysis, when, in fact, there was no difference between them.

Type II error (â): We accept the null hypothesis when it is false or false negative error; or simply, type II error **'â'** (called beta) can be stated as failure to detect true effect. In the above example, type II error would mean that the effects of two drugs were not found different by statistical analysis, when in fact there was difference.

The definition can be summarized as given below.

Actual Situation			
		True Ho	**False Ho**
Investigator's Decision	**Accept Null hypothesis**	Correct Acceptance	Error (Type II)
	Reject Null hypothesis	Error (Type I)	Correct Rejection

Note: Alpha **(á)** and beta **(â)** are the Greek letters and are used to denote probabilities for **type I error** and **type II error** respectively.

We would like to carry our test, i.e., choose our critical region so as to minimize both types of errors simultaneously, but this not possible in a given fixed sample size. In fact decreasing one type of error may very likely increase the other type. In practice, we keep type I error (**á**) fixed at a specified value (i.e., at 1% or 5%).

3.8 STATISTICAL TESTS

Depending on the aim of your study and the type of data collected, you have to choose appropriate tests of significance. Before applying any statistical test, state the null hypothesis in relation to the data to which the test is being applied. This will enable you to interpret the results of the test. The following sections will explain how you will choose an appropriate statistical test to determine differences between groups or associations between variables. Although there are many statistical tests used in drawing inferences, here we will confine our discussion to four main types of tests:

i) ÷² test

ii) T-test

iii) Z- test

iv) F-test

3.8.1 Chi-Square Test (χ^2)

Chi-square test is termed as a non parametric test. Karl Pearson first introduced the concept of chi-square and its application in testing statistical hypothesis. The value of chi-square is determined by (i) taking the difference between each observed frequency (fo) and the corresponding expected theoretical frequency (fe) (ii) squaring each difference (iii) dividing each squared difference by the corresponding expected theoretical frequency and then (iv) adding all the quotient. The value of chi-square is represented by the symbol χ^2

Thus χ^2 =

Uses of chi-square Test

The chi-square test is very powerful tool in the hands of statisticians for testing hypothesis of a variety of statistical problems. The most important purposes served by the application of test of chi-square are follows:

1) test of goodness fit – the chi-square test is used for the comparison of observed frequencies with the expected theoretical frequencies in a sample.

2) test of Independence- the chi-square test is widely used to test the independence of attributes.

3) Test of homogeneity- the chi-square test is also used to test the homogeneity of attributes is respect to of a particular characteristic.

Formula used for computation of

$$\chi^2 = \sum\left[\frac{(f_0 - f_e)}{f_e}\right]$$

χ^2 = Chi-Square

F_o = Observed frequency

F_2 = Expected frequency

Ó = Sum total

Example: Compute the chi-square of data given in table below:

Computation of Chi-square test

		f_0	f_e	$f_0 - f_e$	$(f_0 - f_e)^2$	$\frac{(f_0 - f_e)^2}{f_e}$
1.	Favorably	20	27	-7	49	1.81
2.	Unfavourably	40	27	13	169	6.25
3.	Undecided	21	27	- 6	36	1.33
		81	81			9.39

Follows steps will be used in assessing the level of significance:

Step-1 **Determining the Degree of freedom-** The Chi-Square test depends on degree of freedom. The degree of freedom deals with rows and columns of a table. The formula to calculate degree of freedom is

df = (c-1) (r-1)

df = degree of freedom

C = columns of a table

r = rows of the table

the above table (question) has 3 rows and 2 columns.

df = (C-1)(r-1)

= (3-1) (2-1)

= 2X1

= 2

Step-2 **Determining the Critical Value-** χ^2 has pre-determined value. It requires significance level (5% or 1%) for the computed degree of freedom.

The df is 2. The critical value at 5% level is 5.991 and at 1% level is 9.210 by referring to χ^2 table.

Step-3 **Comparing the critical value of Chi-Square with Computed Value-** the computed χ^2 value is 9.39. It is higher than 5% and 1% level table value. So it is significant. Consequently null hypothesis is rejected in favour of alternative hypothesis.

3.8.2 T -Test

A t-Test is a statistical hypothesis test. The T-Statistic was introduced by W.S. Gossett under the pen name "student". Therefore, the T-test is also known as the "student T-test". The T-test is a commonly used statistical analysis for testing hypothesis, since it is straight forward and easy to use. Additionally, it is flexible and adoptable to a broad range of circumstances. The T-test is applied, if you have a limited sample, usually sample size is less than 30.

The formula used for the calculation of T-test is:

$$t = \frac{\bar{d}}{S\left(\bar{X}_1 - \bar{X}_2\right)}$$

Where,

t = t-test

$\bar{d}$ = mean difference

S = standard deviation

$\bar{X}_1$ = Mean of first set of variables

$\bar{X}_2$ = Mean of second set of variables

Calculation of $\bar{d} = \frac{\Sigma d}{N}$

$$S = \sqrt{\frac{\Sigma d^2}{n-1} - \frac{\left(\Sigma d\right)^2}{n(n-1)}}$$

Example: An IQ test was administered to 5 person before and after they were trained. The result are given below.

Candidates	**I**	**II**	**III**	**IV**	**V**
IQ before Training	110	120	123	132	125
IQ after Training	120	118	125	136	121

Test whether there is any change in IQ after training programme

Candidates	**I Q before training x_1**	**I Q before training x_2**	**Difference $(x_2 - x_1)$ d**	**d^2**
I	110	120	10	100
II	120	118	-2	4
III	123	125	2	4
IV	132	136	4	16
V	125	121	-4	12
			Ód=10	Ód²=140

Estimated standard deviation of population =ó

=

=

=

=

=

= 2

$$t = \frac{\bar{d}}{s\left(\bar{X}_1 - \bar{X}_2\right)} = \frac{2}{2.45} = 0.816$$

3) Level of significance: á=0.01

4) Decision At 0.01 level of significance for 5–1=4 degrees of freedom, the critical value of t = 4.6 (using t-table) but the computed value of t = 0.816 is less than the critical value of t = 4.6 [t = 0.816 < t = 406]. Hence the computed value of t = 0.816 falls in the acceptance region. Thus the null hypothesis ($r_1 = r$) is accepted. So it may be concluded that there is no change in IQ after training programme.

$$\sqrt{\frac{\sum d^2}{N} - \bar{X}_2^2} = \sigma / \sqrt{n} = \sqrt{\frac{30}{5}} = 2.45$$

$$\sqrt{\frac{S_1^2}{n_1} + \sqrt{\frac{S_2^2}{n_2}}}$$

3.8.3 Z -Test

Z-Test is another type of test like T-test applied to compare sample and population means to know if there is a significant difference between them. Z- Test is usually applied in large sample size, having more than 30 sample.

The formula for calculation of Z-Test is

=

Where

x_1 = mean of the first variable

x_{-2} = Mean of the Second variable

S_1 = Standard deviation first equation

S_2 = Standard deviation second equation

n_1 = Standard size of first

n_2 = Standard size of Second

Example: The score in mathematics for boys and girls is given in table below, calculate whether there is significant difference in score between them.

Scores of Boys		Scores of Girls	
40	30	22	42
35	20	33	19
25	11	26	26
26	36	33	29
24	39	44	39
20	44	20	49
45	19	41	23
43	28	33	15
28	36	37	40
33	27	27	26
29	34	18	27
31	18	19	28
41	47	44	11
49	16	32	31
21	47	22	29
34	22	36	25

$\bar{X}_1 = 31.19$

$\bar{X}_2 = 29.56$

$S_1 = 10.13$

$S_2 = 9.56$

$S_1^2 = 102.802$

$= 85.67$

$Z =$

$$= \frac{1.62}{2.42} = 0.67$$

Interpretation: The tabled value of z is 1.96. Since Z < - 1.96 (0.67<1.96), we reject H_0. It means that these is no significant difference between scores of boys and girls.

3.8.4 F -Test

The F-test was first developed by R.A. Fisher. Hence it is known as fisher's test on more commonly as F test. The f-test is used either for testing the hypothesis

about the equality of two population variances of the equality of two or more population means. The ratio of two sample variances.

The formula for calculation of f is:

F =

Where,

= variance of first set of data

= variance of second set of data

Example

The time taken by workers in performing a job by method I and method II is given below.

Method I	20	16	23	27	23	22	
Method II	27	33	42	35	32	34	38

Do the data show that the variances of time distribution in a population from which these samples are drawn do not differ significantly

Solution

Computation of Variances

Method-I		
n_1X_1	$X_1-22=d_1$	d_1^2
20	–2	4
16	–6	36
26	4	16
27	5	25
23	1	1
22	0	0
	Ód$_1$= +2	Ód$_1^2$=82

Method-II		
X_2	$X_2-34=d_2$	d_1^2
27	–7	49
33	–1	1
42	8	64
35	1	1
32	–2	4
34	0	0
38	4	16
	Ód$_2$=3	Ód$_1^2$=135

Method-I

= 13.55

Variance = 13.55

Method-II

Varience

=

= 19.28 – 0.18 = 19.10

Varience = S_2^2 = 19.10

Computation of F-test statistic

Test Statistic F = = .709

3) degrees of freedom v_1= n–1 = 6 –1 = 5

And V_2= n_2–1 = 7–1 = 6

4) Decision- at 5% level of significance the critical value of F=4.95 for v_2=6 and v1=5 degrees of freedom. But the computed value of F=.709 is less than the critical value of F=4.95. Hence the null hypothesis $ó_1^2 = ó_2^2$ is accepted. So it may be concluded that the variance of time distribution in a population from which the samples are drawn do not differ significantly.

In this session you read about different deferential statistics, now answer the questions given in Check Your Progress-4

Check Your Progress 4

Note: a) Write your answer in about 50 words.

b) Check your progress with possible answers given at the end of the unit.

1) What is t-test and where it is applied?

.....

.....

.....

.....

.....

2) What is Chi-square?

.....

.....

.....

.....

.....

3.9 LET US SUM UP

Statistics is a science that deals with the collection, organization, analysis, interpretation, and presentation of information that can be presented numerically and/or graphically to help us answer a question of interest. The information or data collected may be classified as qualitative and quantitative. It may also be classified as discrete or continuous. Frequency distribution is an improved way of presenting a data. For better and more concise presentation of the information contained in a data set, the data is subjected to various calculations. If one wants to further summarize a set of observations, it is often helpful to use a measure which can be expressed in a single number like the measures of location or measures of central tendency of the distribution. The three measures used for this purpose are the mean, median, and mode. Measures of dispersion, on the other hand, give an idea about the extent to which the values are clustered or spread out. In other words, it gives an idea of homogeneity and heterogeneity of data. Two sets of data can have similar measures of central tendency but different measures of dispersion. Therefore, measures of central tendency should be reported along with measures of dispersion. The measures of dispersion include range, percentiles, mean deviation and standard deviation.

The results we obtain by subjecting our data to analysis may actually be true or may be due to chance or sampling variation. In order to rule out chance as an explanation, we use the test of significance. In this unit we have confined our discussion to four tests i.e. $\div^2$ test, Z- test, t-test and f-test.

Correlation is relationship between the two sets of continuous data; for example relationship between height and body weight. Correlation statistics is used to determine the extent to which two independent variables are related and can be expressed by a measure called the coefficient of correlation. Regression, on the other hand, deals with the cause and effect relation between two sets of data. Simple linear regression fits a straight line through the set of *n* points in such a way that makes the sum of squared *residuals* of the model (that is, vertical distances between the points of the data set and the fitted line) as small as possible. The regression line, thus, obtained helps us to predict the value of dependent variable for a given value of independent variable.

Annex I: Table of chi-square values

Degrees of freedom	χ^2 value if $\alpha = 0.05$	χ^2 value if $\alpha = 0.01$
1	3.84	6.63
2	5.99	9.21
3	7.81	11.34
4	9.49	13.28
5	11.07	15.09
6	12.59	16.81
7	14.07	18.48
8	15.51	20.09
9	16.92	21.67
10	18.31	23.21
11	19.68	24.72
12	21.03	26.22

Annexure-II

Degrees of freedom	t-value if chosen P $\alpha = 0.05$	t-value if Chosen P $\alpha = 0.01$
1	12.71	63.66
2	4.30	9.92
3	3.18	5.84
4	2.78	4.60
5	2.57	4.03
6	2.45	3.71
7	2.36	3.50
8	2.31	3.36
9	2.26	3.25
10	2.23	3.17
11	2.20	3.11
12	2.18	3.05
13	2.16	3.01
14	2.14	2.98
15	2.13	2.95
16	2.12	2.92
17	2.11	2.90

18	2.10	2.88
19	2.09	2.86
20	2.09	2.85
21	2.08	2.83
22	2.07	2.82
23	2.07	2.81
24	2.06	2.80
25	2.06	2.79
30	2.04	2.76
40	2.02	2.70
60	2.00	2.66
120	1.98	2.62
infintie	1.96	2.58.

3.10 KEYWORDS

Independent variable : The characteristic being observed or measured which is hypothesized to influence an event or outcome (dependent variable), and is not influenced by the event or outcome, but may cause it, or contribute to its variation.

Dependent variable : A variable whose value is dependent on the effect of other variables (independent variables) in the relationship being studied.

Mean : The mean (or, arithmetic mean) is also known as the average. It is calculated by totalling the results of all the observations and dividing by the total number of observations.

Median : The median is the value that divides a distribution into two equal halves. The median is useful when some measurements are in ordinal scale, i.e., much bigger or much smaller than the rest.

Mode : The mode is the most frequently occurring value in a set of observations. The mode is not very useful for numerical data that are continuous. It is most useful for numerical data that have been grouped. The mode is usually used to find the norm among populations.

Range : This can be represented as the difference between maximum and minimum value or, simply, as maximum and minimum values.

Percentiles : Percentiles are points that divide all the measurements into 100 equal parts. The 30th percentile (P3) is the value below which 30% of the measurements lie. The 50th percentile (P50), or the median, is the value below which 50% of the measurements lie.

Mean Deviation : This is the average of deviation from arithmetic mean

Standard Deviation : This denotes (approximately) the extent of variation of values from the mean.

Parametric statistical test: Is a test whose model specifies certain conditions about the parameters of the parent population from which the sample was drawn.

Non-parametric statistical test : Is a test whose model does not specify conditions about the parameters of the parent population from which sample was drawn.

Normal Distribution : The normal distribution is symmetrical around the mean. The mean, median, and mode assume the same value if observations (data) follows a normal distribution.

Sampling Variation : Any value of a variable obtained from the randomly selected sample (e.g., a sample mean) cannot assume the true value in the population. The variation is called a sampling variation.

Test of Significance : A test of significance estimates the likelihood that an observed study result (e.g., a difference between two groups) is due to chance or real.

3.11 REFERENCES AND SELECTED READINGS

Altman, D.G. (1991), *Practical Statistics for Medical Research*, Chapman and Hall, London.

Barker, D.J.P. (1982), *Practical Epidemiology*. (3rd ed.), Churchill Livingstone Edinburgh, UK.

Bradford, H. A. (1984), *A Short Textbook of Medical Statistics* (11th ed.), Hodder and Stoughton London, UK.

Castle, W.M. and North P.M. (1995), *Statistics in Small Doses.* Churchill Livingstone Edinburgh, UK.

Bose, A (1988), *Statistics,* Calcutta Book House, Calcutta

Fletcher, R. H., S. W. Fletcher and E. H. Wagner (1996), *Clinical Epidemiology: The Essentials,* Lippincott Williams and Wilkins, 351 West Canadian Street Baltimore, Maryland, USA.

Glaser, A.N. (2000), *High-yield Biostatistics*, Lippincott Williams and Wilkins, 227 East Washington Square, Philadelphia, USA.

Greenhalgh, T.(1998), *How to Read a Paper: The Basics of Evidence Based Medicine,* BMJ publishing group, BMA House, Tavistock Square, London,UK.

Hicks, C.M. (1999), *Research Methods for Clinical Therapists. 3rd Edition*, Churchill Livingstone, Robert Stevenson House, 1-3 Baxter's Place,Leith Walk, Edinburgh, UK.

Kelsey, J.L., W.D. Thompson and A.S. Evans (1986), *Methods in Observational Epidemiology,* Oxford University Press, Oxford, UK.

Kidder, L.H. and C.M. Judd (1986), *Research Methods In Social Relations*, CBS College Publishing, New York, USA.

Kleinbaum, D.G., L.L. Kupper and H. Morgenstern (1982), *Epidemiologic Research - Principles and Quantitative Methods*, Van Nostrald Reinhold, New York, USA.

Riegelman, R.F. (1981), *Studying a Study and Testing a Test*, Little Brown and Company, Boston, MA, USA.

Schlesselman, J.J. (1982), *Case-Control Studies - Design, Conduct, Analysis*, Oxford University Press, Oxford, UK.

Siegel, S. (1956), *Nonparametric Statistics for the Behavioral Sciences*, McGraw-Hill Book Company.

Swinscow, T.D.V. and M.J. Campbell (1998), *Statistics at Square One* (11th ed.), British Medical Association, London, UK.

3.12 CHECK YOUR PROGRESS – POSSIBLE ANSWERS

Check Your Progress 1

1) What are the important types of data?

   There are two types of data: (i) qualitative data, viz., occupation, sex, marital status, religion, and; (ii) quantitative data viz., age, weight, height, income, etc. These may be further be categorized in two types viz., discrete and continuous data.

2) What do you understand by non-parametric test?

   A non-parametric statistical test is a test whose model does not specify conditions about the parameters of the parent population from which sample was drawn.

Check Your Progress 2

1) What are the different measures of central tendency?

   The three measures of central tendency are the mean, median, and mode.

2) What are the different measures of dispersion?

   The measures of dispersion are range, percentiles, mean deviation and standard deviation.

Check Your Progress 3

1) Differences between Correlation and Regresion

The main difference between correlation and regression is that the correlation quantifies the degree to which two variables are related. You simply are computing a correlation coefficient (r) that tells you how much one variable tends to change when the other one does. While regression finds out the best fit line for a given set of variables.

Check Your Progress 4

1) What is t-test and where it is applied?

A t-Test is a statistical hypothesis test. The T-Statistic was introduced by W.S. Gossett under the pen name "student". Therefore, The T-test is also known as the "student T-test". The T-test is a commonly used statistical analysis for testing hypothesis, since it is straight forward and easy to use. Additionally, it is flexible and adoptable to a broad range of circumstances. The T-test is applied, if you have a limited sample, usually sample size is less than 30.

2) What is Chi-squre?

Chi-square test is termed as a non parametric test. Karl Pearson first introduced the concept of chi-square and its application in testing statistical hypothesis. The value of chi-square is determined by (i) taking the difference between each observed frequency (fo) and the corresponding expected theoretical frequency (fe) (ii) squaring each difference (iii) dividing each squared difference by the corresponding expected theoretical frequency and then (iv) adding all the quotient . The value of chi-square is represented by the symbol χ^2.