UNIT 12 ANALYSIS OF DATA

Structure

12.1 INTRODUCTION

In Unit 10 we focused on the nature of quantitative and qualitative data, the procedures for classifying and tabulating the quantitative data and presenting the data graphically. Once the data has been processed and tabulated the researcher moves on to data analysis. What are the various methods of analysis of both the quantitative and the qualitative data? This is the focus of this unit. We shall primarily deal with descriptive analysis of quantitative and qualitative data in this unit.

Objectives

After studying this unit, you will be able to:

- enlist the various methods of data analysis,
- compute measures of central tendency, variance, standard deviation, measures of relative position and measures of relationships, and
- describe various methods used for analyzing the qualitative data.

12.2 ANALYSIS OF QUANTITATIVE DATA

Analysis of quantified data means studying the organized or tabulated data in order to discover the inherent facts. The data are studied from as many angles as possible to explore the new facts. Two types of statistical methods are used in the analysis of the tabulated data measured/expressed in quantified terms. The first category of methods pertain to 'descriptive analysis' and the second, to 'inferential analysis' of data.

Descriptive statistical analysis limits generalization to the particular observed group of individuals. This analysis describes only one single group. The computed statistical values provide valuable information about the nature of that particular group only. *Inferential statistics*, on the other hand, is used to model patterns in the data accounting for randomness and drawing inferences about the larger population. In this unit, we will be concerned with 'descriptive analysis' only.

The following methods are generally used in descriptive statistical analysis of the tabulated data:

i) Measures of central tendency

ii) Measures of variability

iii) Measures of relative positions

iv) Measures of relationships.

We shall touch upon each one of them in some details as follows:

12.2.1 Measures of Central Tendency

The three most commonly used measures of central tendency are the Mean, the Median and the Mode. Let us get to know them.

I) ***The Mean (M)***

The arithmetic average of a distribution is known as its mean. The mean of a set of observations or measures is obtained by dividing the sum of all values by the total number of values. Let us consider calculating mean for grouped and ungrouped data.

a) *Mean for an ungrouped data*

The formula for finding the mean for an ungrouped data is as follows:

$$M = \frac{\sum X}{N} \qquad (1)$$

in which

M = mean

Σ = sum of

X = observations in a distribution

N = total number of observations

To illustrate the use of formula (1) let us consider the data given herewith related to nutrition knowledge score obtained for ten subjects.

16, 14, 12, 18, 21, 22, 13, 15, 16, 18

Using the formula (1), the mean nutrition knowledge score calculated as:

$$M = \frac{\sum X}{N} = \frac{16+14+12+18+21+22+13+15+16+18}{10}$$

$$= \frac{\sum X}{N} = 16.5$$

b) *Mean for grouped data*

When the number of observations or measures is large, the data is grouped in a frequency distribution.

The mean is computed by the formula:

$$M = AM + \frac{\sum fx'}{N} \times i \qquad (2)$$

where

M = mean

AM = assumed mean

x' = [Midpoint score(x) – AM]/(length of the class-interval)

$\sum fx'$ = sum of the products of frequencies and deviation of observations from the assumed mean

i = width of the class-interval

N = total number of observations

To illustrate the use of formula (2), consider the grouped data given in Table 12.1.

Computations

Step 1: Put the class-intervals in exact limits

Step 2: Find the mid-point of each class interval and take the assumed mean (AM) at the interval which has the maximum frequency (i.e.11) as highlighted in Table 12.1.

Step 3: Find the difference between each mid-point score and the assumed mean and divide it by the length of the class-interval to get the deviation x (refer to * at the bottom of Table 12.1).

Step 4: Compute fx for each class-interval (fx is the product of the frequency and deviation of the observation from the assumed mean in a particular case.)

Step 5: Find the sum of all fx

Step 6: Apply formula (2) to compute the mean.

Table 12.1: The calculation of the mean from data grouped into a frequency distribution (Ref. Table)

Class Intervals	Exact Units of Class Interval	Mid-point (x)	Frequency f	Deviation from the AM (x)	fx
35 - 39	34.5 - 39.5	37	4	2*	8
30 - 34	29.5 - 34.5	32	8	1	8
25 - 29	24.5 - 29.5	27AM	11	0	0 (+16)
20 - 24	19.5 - 24.5	22	8	–1	–8
15 - 19	14.5 - 19.5	17	6	–2	–12
10 - 14	9.5 - 14.5	12	3	–3	–9 (–29)
			$N = 40$		(–13)

$$*2 = \frac{37-27}{5}$$

Using the formula (2), the mean calculated for grouped data is:

$$M = AM + \frac{\sum fx'}{N} \times i$$

$$= 27.0 + \frac{-13}{40} \times 5$$

$$= 27.0 - 1.625$$

$$= 25.375$$

II) ***The Median***

The median is a point in an array, above and below which one half of the observations fall. It is a measure of position rather than magnitude. In other words, this is a positional central value. This divides the total number of observations into two levels. Let us understand this concept in the context of ungrouped and grouped data.

a) *Median for ungrouped data*

If the observations are ungrouped and their number is small, the observations are arranged in the order of magnitude. The middle score is determined by counting up half the value of N if the number of observation (N) is even. When the number of observations (N) is odd, the mid-observation value is median. For example, 10 is the median of odd scores: 7,8, 9,10,11,12,13. When the number of scores (N) is even, the median is the mid-point between the two middle scores. For example, if the scores are even as given herewith:

7, 8, 9, 10, 11, 12, 13, 14.

$\frac{(10+11)}{2}$ = 10.5 is the median of scores

b) *Median for grouped data*

In the case of grouped data, cumulative frequency distribution is prepared and the median is calculated by the formula:

$$\text{Mdn} = l + \frac{N/2 - F}{f} \times i \qquad (3)$$

where

Mdn = median

l = exact lower limit of the class-interval upon which the median lies

$N/2$ = one half of the total number of observations

F = sum of all frequencies below l

f = frequency within the class-interval upon which the median lies

i = width of the class interval in which the median lies

To illustrate the use of formula (3) consider the data of Table 12.1 once again, presented in the Table 12.2.

Table 12.2: The calculation of the median from data grouped into a frequency distribution

	Class Interval	Frequency (*f*)	Cumulative Frequency (*F*)
	34.5 - 39.5	4	40
	29.5 - 34.5	8	36
Median Class→	24.5 - 29.5	11	28
	19.5 - 24.5	8	17
	14.5 - 19.5	6	9
	9.5 - 14.5	3	3
		$N = 40$	

Here $N/2 = 20$, $l = 24.5$, $F = 17$, $f = 11$ and $i = 5$

Using formula (3)

$$\text{Mdn} = l + \frac{N/2 - F}{f} \times i$$

$$= 24.5 + \frac{(20-17)}{11} \times 5$$

$$= 24.5 + \frac{15}{11} = 25.86$$

III) ***The Mode***

The mode is defined as the most frequently occurring measure of an observation in a distribution. Mode has a special significance because it indicates the peak among the observations. Let us then learn how to calculate mode for grouped/ungrouped data. If only one value occurs a maximum number of times the distribution is said to have one mode; i.e. the distribution is *unimodal*. In some distributions there may be more than one mode. A two mode distribution is *bimodal* and it is *multimodal* in a distribution, which has more than two modes.

a) *Mode for ungrouped data*

In a simple ungrouped series of measures, the crude or empirical mode is that single measure which occurs most frequently. For example, in the series 7, 8, 9, 9, 10, 11 and 12 the most often recurring measures, namely, 9 is the crude or empirical mode.

b) *Mode for grouped data*

When data are grouped into a frequency distribution, the crude or empirical mode is usually taken to be the mid-point of that interval which contains the largest frequency. In the example given in Table 12.1, the interval 25 - 29 contain the largest frequency and hence 26.5, its mid-point, is the crude mode.

The true mode, that is, the point of greatest concentration in the distribution, or the point at which more measures fall than at any other point, is calculated by the formula:

$$\text{Mode} = l + \frac{fm - f_1}{2fm - f_1 - f_2} \times i \qquad (4)$$

where

l = lower limit of the modal class i.e., the class-interval having maximum frequency

fm = frequency of the modal class

f_1 = frequency of the class-interval preceding the modal class

f_2 = frequency of the class-interval following the modal class

i = width of the modal class

To illustrate, let us make use of formula (4) for the data in Table 12.1. Here the maximum frequency is 11 which lies in class interval 24.5 - 29.5.

Therefore, the modal class is (24.5 - 29.5). Here $fm = 11$, $f_1 = 8$, $f_2 = 8$, $i = 5$ and $l = 24.5$

Using Formula (4), the mode can be calculated as:

$$\text{Mode} = l + \frac{fm - f_1}{2fm - f_1 - f_2} \times i$$

$$= 24.5 + \frac{11 - 8}{2 \times 11 - 8 - 8} \times 5$$

$$= 24.5 + \frac{3}{6} \times 5$$

$$= 24.5 + 2.5 = 27.00$$

In our discussion so far we have described the measures of central tendency. Next, let us focus on measures of variability.

12.2.2 Measures of Variability

The measures of central tendency are very useful in describing the nature of a distribution of measures, but they do not give the researcher a complete picture of the data. These measures will not tell the researcher how the scores tend to be distributed. For this, we use a different set of measures which are called measures of *'variability'* or measures of *'spread'* or *'dispersion'*. The most commonly used measures of variability include the range, the variance and standard deviation. Let us study about these measures.

I) ***The Range***

The range is defined as the difference between the two extreme measures or values in a distribution (i.e. the difference between the largest and the smallest value) – Suppose the scores of 10 learners are:

50, 40, 39, 35, 29, 28, 24, 27, 19, 18.

The range for this distribution will be (50 – 18) = 32. Although the range has the advantage of being easily calculated, it has the following serious limitations:

1) As the value of range is based on only two extreme values in the total distribution, it does not give any idea of the variation of many other values of the distribution.
2) It is not a stable statistic as its value can differ from sample to sample drawn from the same population.

A better measure that uses all the observations in the data is standard deviation which is described next.

II) ***The Variance and Standard Deviation***

The average of the squared deviations of the measures or values from their means is known as the *variance.*

The *standard deviation* is the positive square root of variance. Let us understand this concept by calculating the variance and standard deviation for grouped/ungrouped data.

a) *The Variance and Standard Deviation for the ungrouped data*

The variance for the ungrouped data is found by using the formula:

$$\sigma^2 = \frac{\sum x^2}{N} \qquad (5)$$

where

σ^2 = variance of the sample

x = deviation of raw measures or values from the mean.

N = number of values or measures

Let us consider the following data of scores for the application of formula (5):

10, 10, 9, 9, 8, 8, 7, 7, 6, 6.

At the deviation of each score from the mean is required, the first thing to do is to calculate the mean. Using formula (1)

$$M = \frac{\sum x}{N} = \frac{80}{10} = 8$$

Now, from each raw score, the mean is substracted to get the value of x as shown in Table 12.3.

Table 12.3: Distribution of the test scores of ten learners

Score (X)	Deviation (X-M) (x)	Deviation Squared (x^2)
10	2	4
10	2	4
9	1	1
9	1	1
8	0	0
8	0	0
7	–1	1
7	–1	1
6	–2	4
6	–2	4
		$x^2 = 20$

Using formula (5)

$$\sigma^2 = \frac{\sum x^2}{N}$$

$$= \frac{20}{10} = 2$$

Now to get the standard deviation, we need the positive square root of the variance, σ^2

$$\text{Standard Deviation} = \sigma = \sqrt{\frac{\sum x^2}{N}}$$

$$= \sqrt{2}$$

$$= 1.41$$

The raw scores instead of deviation scores may also be used. The raw score formulae for variance and standard deviation are given as follows:

$$\textbf{\textit{Variance}} = \sigma^2 = \frac{N\sum X^2 - (\sum X)^2}{N^2} \quad (6)$$

$$\textbf{\textit{Standard Deviation}} = \sigma = \frac{\sqrt{N\sum X^2 - (\sum X)^2}}{N} \quad (7)$$

in which

X = raw score

N = the number of scores in the distribution

Using the same set of data, we can calculate variance and standard deviation with the help of raw-score formulae. Refer to Table 12.4.

Table 12.4: The calculation of variance and standard deviation from original (raw) score when the assumed mean is taken at zero and the data is ungrouped

Score (X)	X^2
10	100
10	100
9	81
9	81
8	64
8	64
7	49
7	49
6	36
6	36
$\Sigma X = 80$	$\Sigma X^2 = 660$

Using Formula (6)

$$\text{Variance} = \frac{N\sum X^2 - (\sum X)^2}{N^2}$$

$$= \frac{10\times 660 - (80)^2}{100}$$

$$= \frac{6600 - 6400}{100} = \frac{200}{100}$$

$$= 2$$

Using formula (7)

$$\text{Standard Deviation} = \sigma = \frac{\sqrt{N\sum X^2 - (\sum X)^2}}{N}$$

$$= \frac{\sqrt{100\times 660 - (80)^2}}{N}$$

$$= \frac{\sqrt{6600 - 6400}}{10}$$

$$= 1.414$$

Next, let us learn how to calculate variance and standard deviation for grouped data.

b) *Variance and Standard Deviation for grouped data*

In the case of grouped data in a frequency distribution, the variance and standard deviation are calculated by using the formulae:

$$\textbf{\textit{Variance}} = \sigma^2 = \frac{i^2}{N^2}[N\sum fx'^2 - (\sum fx')^2] \qquad (8)$$

$$\textit{Standard Deviation} = \frac{i}{N}\left[\sqrt{N\sum fx'^2 - (\sum fx')^2}\right] \quad (9)$$

where

i = width of the class interval

N = total number of measures

F = frequency of class-interval

x' = deviation of the raw measure from the assumed mean divided by the length of class-interval

To illustrate the use of these formulae let us consider the distribution given in Table 12.5.

Table 12.5: The calculation of variance and standard deviation from data grouped in a frequency distribution

Class Inverval	x	f	X'*	fx'	fx'^2
71 - 75	73	3	3*	9	27
66 - 70	68	4	2	8	16
61 - 65	63	9	1	9	9
56 - 60	58AM	15	0	0	0
51 - 55	53	8	–1	–8	8
46 - 50	48	6	–2	–12	24
41 - 45	43	5	–3	–15	45
		$N = 50$		$\Sigma fx' = -9$	$\Sigma fx'^2 = 129$

* $3 = \frac{73-58}{5}$; AM = Assumed mean

Using formula (8)

$$\text{Variance} = \sigma^2 = \frac{i^2}{N^2}[N\sum fx'^2 - (\sum fx')^2]$$

$$= \frac{(5)^2}{(50)^2}[50+129-(-9)^2]$$

$$= 6.69$$

Using formula (9)

$$\text{Standard Deviation} = \sigma = \frac{i}{N}[\sqrt{N\sum fx'^2 - (\sum fx')^2}]$$

$$= \frac{5}{50}\sqrt{(50\times129-(-9)^2)}$$

$$= \frac{1}{10}\sqrt{6369}$$

$$= \frac{1}{10}\times79.81$$

$$= 7.98$$

The standard deviation is a very useful device for comparing characteristics that may be different or expressed in different units of measurement. It is also used in describing the status or position of an individual in a group. But before this concept is developed further, it is essential to understand the nature of the 'normal probability distribution'.

Normal Probability Distribution

The normal probability distribution is based upon the law of probability. It is not an actual distribution of measures or scores; instead, it is a mathematical model. It is represented by a curve which is called the Normal Probability Curve. Figure 12.1 represents and ideal normal probability curve. You may also recall studying about this concept in Unit 11 under reference/values.

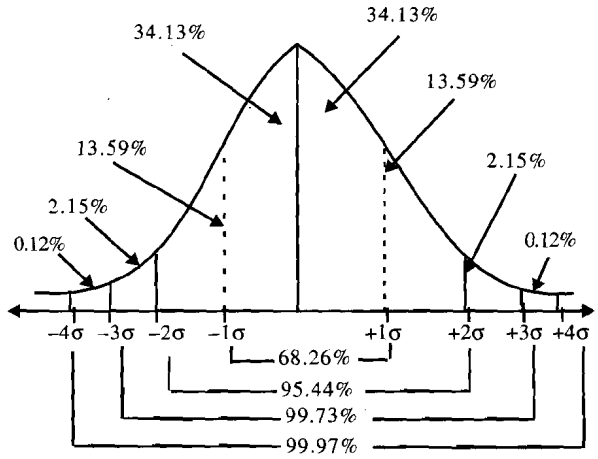

Figure 12.1: Ideal normal probability curve

The normal probability curve you may already be aware by how has the following characteristics:

1) The curve is symmetrical around its vertical axis called ordinate. It implies that the size, shape and slope of the curve on one side of the ordinate is identical to that on its other side.

2) The values of mean, mode and median computed for a distribution following this curve, are always the same.

3) The height of the vertical line called ordinate is maximum at mean and in the unit normal curve it is equal to 0.3989.

4) The curve is asymptotic. It approaches but does not meet horizontal axis and extends from $-\infty$ (minus infinity) to $+\infty$ (plus infinity).

5) The points of inflection of the curve occur at points ± 1, standard deviation ($\pm 1\sigma$), above and below the mean. Thus the curve changes from convex to concave in relation to the horizontal axis at these points.

6) About 68.26 per cent of the total area falls between the limits $M + 1\sigma$ and $M-1\sigma$; 95.44 per cent of the total area of the curve falls between the limits $M + 2\sigma$ and $M - 2\sigma$ and 99.73 per cent of the total area of the curve falls between $M + 3\sigma$ and $M - 3\sigma$.

However, these calculation are rarely necessary, as Normal Table is available from which the information about the area is readily available. For this reason it is very essential that the use of Normal Table (refer to Table I in Appendix at the end of this course) be clearly understood.

Table I give the fractional parts of the total area under the normal curve found between mean and ordinate (Y's) erected at various distances from the mean. The total area under the curve is taken arbitrarily to be 10,000, because of the greater convenience with which fractional parts of the total area may then be calculated. We know that $x = (X - M)$ measures the deviation of a raw score (X) from the mean (M). if, x is divided by σ, then this deviation is expressed in σ Units. These σ

deviation scores are called sigma scores or z-scores $\left(i.e. z = \frac{X - M}{\sigma} = \frac{x}{\sigma}\right)$. The first column of the table under $\frac{x}{\sigma}$ gives distance from the mean in the tenth of σ and distance from the mean in the hundredth of σ are given by the headings of the other columns.

To find the number of cases in the normal distribution between the mean and the ordinate erected at a distance of 1σ from the mean, we go down the $\frac{x}{\sigma}$ column unit 1.0 is reached, and in the next column under .00 we take the entry opposite to 1.0, namely 34.13. This figure means that 3413 cases in 10,000, or 34.13 per cent of the entire area of the curve lies between the mean and 1σ. Similarly, if we have to find out the percentage of the distribution between mean and 1.65σ, we go down the $\frac{x}{\sigma}$ column till 1.6, then across horizontally to the column headed .05, and take the entry 45.5. This shows that in a normal distribution, 45.05 per cent of the total area lies between the mean and 1.65σ.

Check Your Progress Exercise 1

1) Compute (i) Mean (ii) Variance and (iii) Standard Deviation for the following frequency distribution:

Class Interval	*F*
195 - 199	1
190 - 194	2
185 - 189	4
180 - 184	5
175 - 179	8
170 - 174	10
165 - 169	6
160 - 164	4
155 - 159	4
150 - 154	2
145 - 149	3
140 - 144	1

.....

.....

.....

2) Describe the characteristics of a Normal Probability Distribution.

.....

.....

.....

.....

With a basic understanding regarding measures of variability, let us next focus on measures of relative positions.

12.2.3 Measures of Relative Positions

A raw score on a test, taken by itself, has no meaning. It gets meaning only by comparison with some reference group or groups. The comparison may be done with the help of the following measures:

1) Sigma Scores
2) Standard Scores
3) Percentiles
4) Percentile Ranks.

What does each one of these measures indicate? Let us find out.

1) ***Sigma Scores***

A sigma score makes a realistic comparison of scores possible and provides a basis for equal weighting of the scores as the scores on different tests are expressed on a scale with a mean of zero and standard deviation of 1.

Let us suppose that the mean of a test is 75 and the standard deviation is 5.0. Then if A earns a score of 85 on this test, his deviation from the mean is 85–75 = 10. Dividing this deviation of 10 by the standard deviation (σ), i.e., 5.0, we give him a score of $\frac{10}{5} = 2$. If B's score on this test is 64, his deviation from the mean is 64–75 = – 11 and his score in σ units is –2.20. Deviations from the mean expressed in σ terms are called sigma scores.

Half of the scores in a distribution lie below and half above the mean, about half of σ scores are positive and half are negative.

2) ***Standard Scores***

The sigma scores, which are often small decimal fractions and half of them are negative, are somewhat inconvenient to deal with. Hence, scores are usually converted into a new distribution with mean and standard deviation so selected that it makes all scores positive and relatively easy to handle in computation. Such scores are called 'standard scores'.

The formula for the conversion of a raw score to a standard score is as follows:

$$X' = \frac{\sigma^2}{\sigma}(X - M) + M' \qquad (10)$$

in which

X' = A standard score in a new distribution
σ' and σ = *SD's* of standard and raw scores
X = A score in the original distribution
M and M' = Means of raw and standard scores

When the mean (M') and standard deviation (σ') are taken to be 50 and 10 respectively, the standard score is called a T-score.

$$\text{i.e. } T = \frac{10}{\sigma}(X - M) + 50 \qquad (11)$$

Example: To illustrate, let us consider a distribution with its mean 67 and σ = 12.5. Let us also suppose that A's score is 76 and B's score is 54. Express these scores as (i) standard scores in a distribution with a mean of 250 and σ of 50 and (ii) T-scores.

Using formula (10) the standard score can be calculated as:

$$X' = \frac{50}{12.5}(X - 67) + 250$$

Substituting A's score of 76 in the above equation we have:

$$X' = \frac{50}{12.5}(76 - 67) + 250$$

$$= \frac{50 \times 9}{12.5} + 250$$

$$= 286$$

Substituting B's score of 54 in the above equation

$$X' = \frac{50}{12.5}(54 - 67) + 250$$

$$= 198$$

Using formula (11), the T-score can be calculated as:

$$T = \frac{10}{12.5}(X - 67) + 50$$

Substituting A's score in the above equation we have:

$$T = \frac{10}{12.5}(76 - 67) + 250$$

$$= 0.8 \times 9 + 50$$

$$= 57.2$$

Substituting B's score in the above equation we have:

$$T = \frac{10}{12.5}(54 - 67) + 50$$

$$= 39.6$$

Next, let us understand the concept of quantiles

3) *Quantiles/Percentiles*

In sub-section 12.2.1 we learnt about the measures of central tendency. Apart from the central location (i.e. mean, medium, mode), there are some other locations, i.e. quantiles which, are sometimes used in research. What are quantiles? The values of variable that divide the total number of subjects into ordered groups of equal size are called *quantiles*. There are different measures for various number of divisions. Important among them are percentiles or centiles, quartiles and tertiles. These are positional values at different locations in 100 divisions, 4 divisions and 3 divisions, respectively. You may recall studying about percentiles in the context of child growth monitoring and are most commonly used for this purposes. Let us understand percentiles better.

Percentiles are the points which divide the entire scale of measurement into 100 equal parts. They are denoted by P_0, P_1, P_2, P_3, P_4, P_5 ....., P_{99}, and P_{100}.

The first percentile may be defined as that point in a frequency distribution below which lie 1 per cent of the total measures or scores. Similarly, twentieth percentile may be defined as that point in a frequency distribution below which 20 per cent of the total measures or scores fall. It is evident that the median, expressed as a percentile, is P_{50}. It should be noted that P_0 lies at the beginning of the distribution and P_{100} at the end of the distribution.

Let us understand this concept with an example. Suppose we have 300 subjects. For $n = 300$ subjects, after arranging the observations in the ascending order of magnitude, the 3rd and 95th percentiles, first quartile and second tertile are as follows.

3rd percentile = (3 × 300/100) = 9th value

95th percentile = (95 × 300/100) = 285th value

1st quartile = (1 × 300/4) = 75th value

2nd tertile = (2 × 300/3) = 200th value

The formula for calculating percentiles is as follows:

$$P_p = l + \frac{(PN - F)}{f_p} \times i \qquad (12)$$

in which

p_p = percentile of the distribution wanted

l = exact lower limit of class-interval upon which P_p lies.

PN = part of the N to be counted off in order to reach P_p

F = sum of all scores upon intervals below l.

f_p = number of scores within the interval upon which P_p falls

i = length of the class-interval.

The use of formula (12) may be illustrated by the following example. Calculate P_{25}, P_{45} and P_{95} from the following distribution given in Table 12.6.

Table 12.6: The calculation of percentiles from data grouped in a frequency distribution

Scores: Class Intervals	Frequency (f)	Cumulative Frequency (F)
81.5 - 86.5	1	80
76.5 - 81.5	4	79
71.5 - 76.5	5	75
66.5 - 71.5	10	70
61.5 - 66.5	35	60
56.5 - 61.5	12	25
51.5 - 56.5	9	13
46.5 - 51.5	2	4
41.5 - 46.5	2	2
$N = 80$		

For computing P_{25}, we have to first find PN

Here, 25 per cent of 80 is 20, $PN = 20$

Now $l = 56.5$, F = 13, $f_p = 12$ and $i = 5$

Using formula (12)

$$P_{25} = 56.5 + \frac{20-13}{12} \times 5$$

$$= 56.5 + 2.92 = 59.42$$

Similarly

$$P_{45} = 61.5 + \frac{36-25}{35} \times 5$$

$$= 61.5 + 1.57 = 63.07$$

$$P_{95} = 76.5 + \frac{76-75}{4} \times 5$$

$$= 76.5 + 1.25 = 77.25$$

Finally a word about percentile ranks.

4) ***Percentile Ranks***

The percentile rank is the point in the distribution below which a given percentage of scores falls. If the 80th percentile rank is a score of 65, then 80 per cent of the scores falls below 65. The median is the 50th percentile rank for, 50 per cent of the scores fall below it. Suppose, if the 97th percentile of weight of 2-year old boys is 14.2 kg then it means that 97% of such children have weight 14.2 kg or less. The other 3% have a higher weight.

The process of calculating percentile ranks is the reverse process of calculating percentile points. We have to calculate ranks corresponding to particular scores. If R is the rank and N is the total number of cases, then:

$$\text{Percentile Rank} = 100 - \frac{100R - 50}{N} \qquad (13)$$

Suppose A ranks 13th in the class of 80 learners, 12 learners rank above it, 67 below it. Its percentile rank is:

$$= 100 - \frac{100 \times 13 - 50}{80}$$

$$= 100 - 15.625$$

$$= 84$$

We hope the discussion presented above may have given you a good insight into the concept of measures of relative position. In the next sub-section we shall review the measures of relationship.

12.2.4 Measures of Relationship

The data in which we secure measures of two variables for each individual is called *bivariate data*. The essential feature of bivariate data is that one measure can be compared with another measure for each member of the group. When bivariate data are studied, we may like to know the degree of relationship between the variables of such data. This degree of relationship is known as *correlation*. It can be quantitatively represented by the co-efficient of correlation. Its value ranges from –1.00 to +1.00. A value of –1.00 describes a perfect negative correlation and +1.00 describes perfect positive correlation. A zero value describes complete lack of correlation between the two variables. The sign of the co-efficient indicates the direction of relationship and numerical value is its strength/magnitude. Let us get to know about the methods of correlating variables.

Methods of correlating variables

There are various methods of correlating variables. Their use is relative to the situation and type of data. *Product moment correlation* and *Rank order correlation* are mostly used for computing correlation between two variables. These are described herewith.

1) *Product-moment correlation*

In some situations, the data for two variables X and Y are expressed in interval or ratio level of measurement and the distributions of these variables have a linear relationship. Moreover, the distributions of variables are uni-modal and their variances are approximately equal. In such situations, product moment method of correlation is used generally. It is also called *Pearson's r*. The calculation of pearson's *r* from grouped/ungrouped data is explained next.

i) *Calculation of Pearson's r from ungrouped data*:

When the size of the sample is small, there is no need of grouping the data and Pearson's *r* may be calculated with the help of the following formula:

$$r_{xy} = \frac{N\sum xy - (\sum x)(\sum y)}{\sqrt{\left[N\sum x - (\sum x)^2\right]\left[N\sum y^2 - (\sum y)^2\right]}} \qquad (14)$$

in which

x = deviations of X measures from the assumed mean

y = deviations of Y measures from the assumed mean

To illustrate the use of formula (14), let us compute the product moment 'r' from the following data for the two variables X (theory marks) and Y (practical marks) for 10 learners who are enrolled in your programme study centre of IGNOU.

X : 45 54 52 58 62 46 55 49 50 54

Y : 42 50 55 46 59 41 46 48 45 48

Using formula (14) for the data in Table 12.7.

Table 12.7: The calculation of product moment correlation from ungrouped data when deviations are taken from assumed mean

X	Y	x	y	x^2	y^2	xy
45	42	–7	–6	49	36	42
54	50	2	2	4	4	4
52(AM)	55	0	7	0	49	0
59	46	6	–2	36	4	–12
62	59	10	11	100	121	110
46	41	–6	–7	36	49	42
55	46	3	–2	9	4	–6
49	48(AM)	–3	0	9	0	0
50	45	–2	–3	4	9	6
54	48	2	0	4	0	0
		$\Sigma x = 5$	$\Sigma y = 0$	$\Sigma x^2 = 251$	$\Sigma y^2 = 276$	$\Sigma xy = 186$

Using formula (14)

$$r_{xy} = \frac{N\sum xy - (\sum x)(\sum y)}{\sqrt{\left[N\sum x^2 - (\sum x)^2\right]\left[N\sum y^2 - (\sum y)^2\right]}}$$

$$r_{xy} = \frac{10\times186 - 5\times0}{\sqrt{\left[(10\times251) - (5)^2\right]\left[10\times276 - (0)^2\right]}}$$

$$= \frac{1860}{2618.89} = 0.71$$

Next, let us learn how to calculate pearson's r from ungrouped data.

ii) *Calculation of Pearson's r*from grouped data:*

When N is large or even moderate in size, and when no calculating machine is available, the best procedure is to group data in both variables X and Y and to form a scattergram. The values from the scattergram may be used in the following formula:

$$r_{xy} = \frac{N\sum xy - (\sum x)(\sum y)}{\sqrt{\left[N\sum fx^2 - (\sum fx)^2\right]\left[N\sum fy^2 - (\sum fy)^2\right]}} \quad (15)$$

To illustrate the use of the formula (15) consider the data of 50 learners enrolled with IGNOU in Course X and in Course Y in the scattergram presented in Figure 12.2.

Course y

Course x: y → / x ↓	20-21	22-23	24-25	26-27	28-29	30-31	32-33	f_y	y	fy	fy^2	fxy
51-53			$_{-2}1^4$	$_{0}2^0$				3	2	6	12	−2
48-50	$_{-6}2^{-1}$	$_{-4}2^{-2}$	$_{-3}3^4$	$_{0}3^0$				10	1	10	10	−13
45-47	$_{0}1$	$_{0}3^0$	$_{0}4^0$	$_{0}8^0$	$_{0}3^0$	$_{0}2^0$	$_{0}1^0$	22	0	0	0	0
42-44	$_{2}2^0$	$_{4}2^2$	$_{3}3^1$	$_{0}5^0$				12	−1	−12	12	13
39-41		$_{4}1^4$	$_{2}1^4$	$_{0}1^0$				3	−2	−6	12	6
f_{00}	5	8	12	19				50		−2	46	04
x	−3	−2	−1	0	1	2	3					
fx	−15	−16	−12	0	3	4	3	−33				
fx^2	45	32	12	0	3	8	9	109				
fxy	0	4	0	0	0	0	0	04				

Figure 12.2: A scattergram showing paired scores of 50 learners on the tests of course X and Course Y

The computation for the values of Σfx, Σfx^2, Σfxy etc. may be done in the following steps in the order given below:

Step 1

The distribution of Course X scores for the 50 learners is found in the f(y) column on the right of the scattergram. Assume a mean for the distribution of scores of course X (the mid-point of that interval which contains the largest frequency), and draw double lines to mark off the row in which the assumed mean falls. In the present example, the mean score for course X has been taken at 46 (mid point of internal 45 - 47) and y's (deviations from the assumed mean) have been taken from this point.

Fill in fy and then fy^2 columns.

Step 2

The distribution of the Course Y of 50 learners is found in the f(x) row at the bottom of the scattergram. Assume a mean for this distribution and draw double lines to designate the column under the assumed mean. The mean for the Course Y scores is taken at 26.5 (mid-point of interval 26 - 27), and the x's (deviations from assumed mean) are taken from this point. Fill in the fx and then fx^2 rows.

Step 3

The fxy for a cell is computed by multiplying the frequency given in the particular cell with the corresponding x and y. For example, there is a frequency 1 corresponding with Course Y score 24 - 25 and Course X score 51 - 53. The corresponding x for this cell frequency is -1 and corresponding y is $+2$. Thus fxy for this cell is (1) (-1) $(+2) = -2$. Similarly the value for fxy is computed for all the cells and their sum Σfxy is calculated row-wise, as well as, column-wise. The two sums should equal each other. In the present example, it has come to be 4.

Step 4

Substituting the values for Σfx, Σfx^2, Σfy, Σfy^2 and Σfxy in the formula (15) we get:

$$r_{xy} = \frac{50\times 4-(-33)(-2)}{\sqrt{\left[50\times 109-(-33)^2\right]\left[50\times 46-(-2)^2\right]}}$$

$$= 0.042$$

Now we hope you will be able to calculate the correlation between two variables using the learner method, which is a product moment correlation. Next, we shall learn about the rank order correlation.

2. *Rank Order Correlation*

The rank order correlation is also known as the *Spearman rank order co-efficient* of correlation and is denoted by ρ (rho). When the data are available in ordinal (rank) form of measurement rather than in interval or ratio form, this type of correlation is useful.

To find out Spearman rank order coefficient of correlation, the following formula is used.

$$\rho = 1 - \frac{6\sum D^2}{N(N^2-1)} \qquad (16)$$

in which

D = difference between the paired ranks

ΣD^2 = sum of the squared differences between ranks

N = number of paired ranks

To make use of formula (16) let us consider the following data. Two evaluators X and Y ranked 10 learners in the dissertation. The ranks given to them by the judges are given in Table 12.7.

Table 12.7: The calculation of rank difference correlation

Students	Rank Assigned by *X*	Rank Assigned by *Y*	D	D^2
A	2	3	–1	1
B	4	5	–1	1
C	5	4	1	1
D	10	9	1	1
E	8	7	1	1
F	1	2	–1	1
G	3	1	2	4
H	9	8	1	1
I	6	10	–4	16
J	7	6	1	1
				$\Sigma D^2 = 28$

Using formula (16)

$$\rho = 1 - \frac{6\sum D^2}{N(N^2-1)}$$

$$= 1 - \frac{6\times 28}{10(100-1)}$$

$$= 0.83$$

With this we end our study of measures of relationships. We hope you would use these measures as appropriate for your research.

Check Your Progress Exercise 2

1. Compute product moment correlation for the following data:

   X : 45, 55, 56, 58, 60, 65, 68, 70, 75, 80, 85

   Y : 56, 50, 48, 60, 62, 64, 65, 70, 74, 82, 90

   .....

   .....

   .....

   .....

   .....

Now that we have learnt about the different descriptive methods we can use for analyzing the quantitative data, next let us review qualitatives data analysis.

12.3 ANALYSIS OF QUALITATIVE DATA

As already discussed in Unit 10, qualitative data consists of detailed description of situations, events, interactions, programmes and observed behaviours. These data are studied from as many angles as possible either to explore the new facts or to reinterpret already known or existing facts.

Qualitative data also include all those data that are summarized in terms of *proportions, rate* or *ratio*. For example, haemoglobin estimation is quantitative but it becomes qualitative for our purpose when divided into categories such as < 7.0 g/dl (severe anaemia), 7.0 - 10.0 g/dl (moderate anaemia), > 10 g/dl but less than cut-off (mild anaemia) and ≥ 11.0 g/dl (normal) for defining the magnitude of anaemia among children. This occurs when the interest is in the proportion of subjects falling into these categories rather than in mean. Similarly, statements on risk for development of a condition in the presence of an exposure are common in nutrition/health research. For example, smoking is an important risk factor for lung cancer. The magnitude of risk or association can be measured by calculating relative risk (RR) in case of prospective studies, and by odds ratio (OR) in case of retrospective and case-control studies. Here, in this section, we shall review these commonly used methods to assess qualitative data and also methods to be used for analysis of data available in the form of detailed descriptions. Let us begin with statistical inference from proportion, relative risks and odds ratios.

12.3.1 Statistical Inference from Proportions, Relative Risk and Odds Ratio

Statistical inference from proportions where the interest is to find out that the subjects in the target group population do or do not follow a pre-specified pattern (for example whether or not the distribution (prevalence) of mild, moderate and severe anaemia among children in a community is 38%, 17% and 8%, respectively). Statistically such a problem is known as the *problem of goodness of fit* because the interest is in finding out whether or not the pattern observed in the sample fits into the specified pattern. For testing goodness of fit a statistical method called the *Chi-square* is used.

In general, the *chi-square test statistic* is of the form:

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

We will learn about the application of chi-square test in the next unit. Here, let us now focus on relative risk and odds ratio.

Relative Risk

Let us understand the concept of relative risk by considering some statements cited in epidemiological research.

Case I: Risk of lung cancer in smokers is nearly 10 times as compared to non-smokers.

Case II: In a study on infant mortality and birth weight it was found that low birth weight group has a much higher risk of early mortality as compared to normal birth weight. In this study the relative risk (RR) was = 19.8.

Having reviewed these statements you would have noticed that relative risk (RR) in simple terms is the measure of the ratio of the risk of disease or death among the exposed to the risk among the unexposed. More scientifically, *relative risk is the*

ratio of the risk of an event (or of developing a disease) in those with antecedent factors compared to those without this factor.

Relative risk can thus be calculated using the notation of Table 12.7 as follows:

RR = a/(a+c) / b/(b+d)

Table 12.7: General structure of a 2×2 contingency table

Variable - 2 (outcome)	Variable (Antecedent)		
	Present	**Absent**	**Total**
Present	a	b	(a + b)
Absent	c	d	(c + d)
	(a + c)	(b + d)	n

Thus you may have noticed that RR measures the degree of association of outcome with the antecedent factor.

In a simple comparison between an experimental group and a control group:

A relative risk of 1 means there is no difference in risk between the two groups.

A RR of < 1 means the event is less likely to occur in the experimental group than in the control group.

A RR of > 1 means the event is more likely to occur in the experimental group than in the control group.

Statisticians stress the importance of using confidence intervals (CIs). A confidence interval calculated for *a measure of treatment effect* shows a range within which the true treatment effect is likely to lie. The *confidence level* sets the boundaries of a confidence interval; this is conventionally set at 95% to coincide with the 5% convention of statistical significance in hypothesis testing. We will learn more about confidence level later in Unit 13 in sub-section 13.3.1.

The 95% of CI for RR can be obtained:

Odds Ratio

Like relative risk, odds ratio too compares the likelihood of an event between two groups.

Suppose that in a sample of 100 men, 90 are absent due to sickness in the previous week, while in a sample of 100 women only 20 are absent in the same period. Now what is the odds of men and women being absent?

Let us consider:

The odds of a man being absent are 90 to 10, or 9:1, while the odds of a woman being absent are only 20 to 80, or 1:4 = 0.25:1. Now, 9/0.25 = 36, so the odds ratio is 36, showing that men are much more likely to be absent than women.

Considering the example above, we may therefore define odds ratio as the ratio of the odds of an event occurring in one group to the odds of it occurring in another group, or to a sample-based estimate of that ratio. These groups might be men and women, an experimental group and a control group, or any other dichotomous classification as presented in Table 12.7 above.

The Odds Ratio (OR) may thus be calculated using the notations in Table 12.7 as follows:

$$OR = (a \times d)/(b \times c)$$

In case-control studies, an odd is the frequency of presence of antecedent relative to its absence. This is calculated for the cases and the control. The ratio of these two odds is called the odds ratio.

With this brief review we end our study of qualitative data based as proportions, ratio and rates. Next, let us get to know the methods used for analysis of descriptive qualitative data.

12.3.2 Analysis of Descriptive Qualitative Data

Analysis of qualitative data means studying the organized material available in the form of detailed descriptions of situations events, programmes case studies and observed behaviour in order to discover inherent facts. These data are studied from as many angles as possible either to explore the new facts or to reinterpret already known or existing facts. The following methods are generally used in the analysis of qualitative data.

i) Content Analysis

ii) Inductive Analysis

iii) Logical Analysis

A) *Content Analysis*

Content analysis is concerned with the classification, organization and comparison of the content of the document or communication. The terms, content analysis and coding are sometimes used interchangeably as both the processes involve objective, systematic, and qualitative description of any symbolic behaviour. Since content analysis involves the classification, evaluation and comparison of the content of communication or document, it is sometimes referred to as 'documentary activity' or 'information analysis'. The communication may be in the form of responses to an open-ended questionnaire, conversation as a result of an interview, or description of an observed activity. It may also be in the form of official records (census, birth, accident, crime, school, institutional and personal records), judicial decisions, laws, budget and financial records, cumulative records, courses of study, content of text books, reference works, news papers, periodicals or journals, prospectus of various educational institutions or universities, direct quotations, and notes of an interview.

There are three approaches that a researcher may adopt in content analysis. These include: (i) characteristics of content, (ii) procedures or causes of content, and (iii) audience or effects of content. In the first approach, the researcher is interested primarily in the characteristics of the content itself. He/she may focus either on the 'substantive nature' of the content or upon the 'form' of the content. In the second approach, the researcher attempts to draw valid inferences about the nature of the procedures of the content or the causes of the symbolic material from the characteristics of the material itself. In the third approach to content analysis, the researcher interprets the content so as to reveal something about the nature of its 'audience' or its 'effects'. He/she takes the content material as a basis for drawing inference about the characteristics of the 'audience' for whom the material (content) is designed or about the effects of communication. Which it brings about.

The steps involved in the process of content analysis includes (i) defining the unit of analysis, (ii) specifying variables and categories, (iii) frequency, direction and intensity of units, (iv) contingency analysis, (v) sampling of units, and (vi) constructing the content analysis outline. Defining the unit of analysis indicates whether the unit (material) is confined to single words, phrases, complete sentences, paragraphs, or to even larger amounts of materials. Once the unit is defined, the researcher conducts its analysis so as to create reproducible or objective data for scientific treatment and

generalization beyond the specific set of symbolic material analyzed. For converting symbolic material into objective data, it is necessary to specify the "variables" explicitly in terms of which descriptions are to be made. Once the unit is defined and the variables along with their categories specified, the researcher will classify units in the material to be analyzed according to: (i) the number of units (frequency), (ii) favourableness/unfavourableness of the content (direction), and (iii) the emotional impact of the units (intensity). The contingency analysis aims at considering the favourableness or unfavourableness of a single unit in the light of the reminder of the communication so that its real meaning is not lost.

B) *Inductive Analysis*

Inductive analysis means that patterns, themes, and categories of analysis emerge out of the data. In this analysis, researcher looks for natural variation in the data. The study of natural variation involves particular attention to variations in programme processes and how participants respond to and are affected by programmes. Two ways of representing the patterns emerge from the analysis of data. First, the researcher can use the categories developed and articulated in the programme studied to organize presentation of particular themes. Second, the researcher may also become aware of categories or patterns for which the people in the programme did not have labels or terms, and the analyst develops terms to describe these inductively generated categories.

C) *Logical Analysis*

Logical analysis is used for representing patterns as dimensions or categories using either participant-generated constructions or evaluator-generated constructions. It is sometimes useful to cross-classify different dimensions to generate new insights about how the data can be organized and to look for patterns that may not have been recognized in the initial induction analysis. Logical analysis aims at creating potential categories by crossing one typology with another, and then moving back and forth between the logical construction and the actual data for creating a "new typology" using cross-classification matrices.

There are other ways of analyzing qualitative data. We have not discussed all of them. The idea is to give you a feel of qualitative data analysis and show how it differs from quantitative data analysis.

Check Your Progress Exercise 3

1) Define relative risk and odds ratio.

.....

.....

.....

2) Consider the following bivariate data given herewith:

Infant Outcome	**Weight Gain During Pregnancy**		**Total**
	< 7kg	≥ 7 kg	
Dead	181	36	217
Alive	1666	651	2317
Total	**1847**	**687**	**2534**

Based on the data:

a) Calculate the relative risk of infant death in pregnancies with weight gain of less than 7 kg during pregnancy

b) Calculate the odds of dead children being born to pregnant women with weight <7 kg as compared to women with weight gain more than 7 kg.

12.4 LET US SUM UP

In this unit, we discussed the methods used in the analysis of quantitative and qualitative data. The main points are as follows:

1) The data collected through the administration of various tools on the selected samples are of (i) quantitative and (ii) qualitative nature.

2) Measures of (i) central tendency, (ii) variability, (iii) relative positions, and (iv) relationship are the four types of descriptive statistical measures.

3) Mean, median and mode are the three measures of central tendency.

4) Mean is the arithmetic average of a distribution. It is obtained by dividing the sum of all values of observation by the total number of values. The formula for finding the mean for ungrouped data is:

$$M = \frac{\sum X}{N}$$

When the number of observations is large, the data is grouped in a frequency distribution. The formula for computing the mean here is:

$$M = AM + \frac{\sum X}{N} \times i$$

5) Median is a point in an array, above and below which one half of the values or measures fall. If the values are ungrouped and their number is small, the values are arranged in order of magnitude and the middle value is determined by counting up half the value of N. When the number of values is odd, the mid-value is the median. When the number of values is even, the median is the mid-point between the two middle values.

In the case of grouped data, the median is calculated by the formula:

$$\text{Mdn} = l + \frac{\frac{N}{2} - F}{f} \times i$$

6) Mode is the most frequently occurring value in a distribution. If only one value occurs a maximum number of times, the distribution is said to have one mode (uni-modal). A two mode distribution is bi-modal, and more than a two mode distribution is called multi-modal.

In a simple ungrouped series of measures or values, the crude mode is that single measure or value which occurs most frequently.

For a group distribution, the mode is calculated by the formula:

$$\text{Mode} = l + \frac{fm - fi}{2fm - f_1 - f_2} \times i$$

7) The range, variance and standard deviation are the most commonly used measures of variability.

8) The range is the difference between the two extreme values or measures in a distribution.

9) The average of the squared deviations of the measures or values from their mean is known as variance. Standard deviation is the positive square root of variance.

Variance and standard deviation for the ungrouped data are found by the formulae:

$$\text{Variance} = \sigma^2 = \frac{N\sum X^2 - (\sum X)^2}{N^2}$$

$$\text{Standard Deviation } \sigma = \sqrt{\frac{N\sum X^2 - (\sum X)^2}{N^2}}$$

When the data are grouped in a frequency distribution, the variance and standard deviation are computed by the formulae:

$$\text{Variance} = \sigma^2 = \frac{i^2}{N^2}\left[N\sum fx'^2 - \left(\sum fx'\right)^2\right]$$

$$\text{Standard Deviation} = \sigma = \frac{i}{N}\sqrt{\left[N\sum fx'^2 - \left(\sum fx'\right)^2\right]}$$

10) The normal probability distribution is represented by a curve which has the following characteristics.

   i) The curve is symmetrical around its vertical axis called ordinate.

   ii) The mean, mode and median of the distribution have the same values.

   iii) The height of the vertical line called ordinate is maximum at the mean and in the unit normal curve, it is equal to 0.3989.

   iv) The curve is asymptotic.

   v) The points of inflections of the curve occur at points ± 1, standard deviation ($+1\sigma$), above and below the mean.

   vi) About 68.26 per cent of the total area of the curve falls between limits Mean $\pm 1\sigma$, 95.44 per cent of the total area falls between M $\pm$ 2σ, and 99.73 per cent of the total area falls between M $\pm$ 3σ.

11) Sigma scores, standard scores, percentiles and percentile ranks are the measures of relative positions.

12) A sigma score makes it possible to obtain a realistic comparison of scores and provides a basis for equal weighting of the scores as the scores on different tests are expressed on a scale with a mean of zero and standard deviation 1.

13) When the sigma scores are converted into a new distribution with mean and standard deviation so selected as to make all scores positive, the scores are called standard scores.

   The formula for the conversion of a raw score to a standard score is:

$$X = \frac{\sigma'}{\sigma}(X - M) + M'$$

   When the mean (M) and standard deviation (σ') are taken to be 50 and 10 respectively, the standard score is called a T-score. It is expressed by the formula:

$$T = \frac{10}{\sigma}(X - M) + 50$$

14) Percentiles are the points which divide the entire scale of measurement into 100 equal parts.

The formula for computing percentiles is:

$$P_p = l + \frac{(P_N - F)}{f_p} \times i$$

15) Percentile rank is the point in the distribution below which a given percentage of scores fall. If R is the rank and N is the total number of cases.

$$\text{Percentile Rank} = 100 - \frac{100R - 50}{N}$$

16) Product Moment correlation and rank-difference correlation are the commonly used measures of relationship between any two variables.

17) When the size of sample is small and the variables are measured in interval scales of measurement, the product-moment correlation is computed by the formulae:

$$r_{xy} = \frac{N\sum xy - (\sum x)(\sum y)}{\sqrt{\left[N\sum x^2 - (\sum x)^2\right]\left[N\sum y^2 - (\sum y)^2\right]}}$$

When the size of the sample is large, the product-moment correlation is found by the formulae:

$$r_{xy} = \frac{N\sum xy - (\sum fx)(\sum fy)}{\sqrt{\left[N\sum fx^2 - (\sum fx)^2\right]\left[N\sum fy^2 - (\sum fy)^2\right]}}$$

18) When the data are available in ordinal (rank) from of measurement and the size of the sample is small, the formula for computing the rank-difference correlation is:

$$\rho = 1 - \frac{6\sum D^2}{N(N^2 - 1)}$$

19) Content analysis, inductive analysis and logical analysis are some methods of qualitative analysis.

20) Content analysis pertains to the classification, quantification and comparison of the content of a document or communication under consideration. It is also called documentary or information analysis.

21) Inductive analysis leads to patterns, themes and categories emerging out of the data. In this type of analysis, the researcher looks for natural variation in the data.

22) Logical analysis is used for representing patterns as dimensions or categories, either using participant-generated constructions or evaluator-generated constructions.

23) Relative risk, odds ratio is used for statistical inference of qualitative data.

12.5 GLOSSARY

Quantitative Data : data which are expressed in nominal, ordinal, interval or ratio scales of measurement.

Qualitative Data : data which are available in the form of detailed descriptions of situations, events, people, interactions, and observed behaviour, direct quotations from people

about their experiences, attitudes, beliefs, and thoughts, and excerpts from documents, correspondence, records, and case histories.

Parametric Data : these are data which are got by applying interval or ratio scales of measurement.

Non-parametric Data : these are data which are got by applying nominal or ordinal scales of measurement. These types of data are either counted or ranked.

Central Tendency : a measure of central tendency provides a single most typical value as representative of a group of values; the 'trend' of a group of measures as indicated by some type of averages, usually the mean, median or mode.

Mean : a kind of average obtained by dividing the sum of a set of measures by their number.

Median : the middle value in a distribution or set of ranked values; the point that divides the group into two equal parts.

Mode : the value that occurs most frequently in a distribution.

Variability : the spread or dispersion of measures or values.

Range : for some specified groups, the difference between the highest and the lowest obtained measure or value on a tool. It is a rough measure of variability.

Variance : a measure of variability of a distribution. It is the average of the squared deviations of the measures or values from the mean.

Standard Deviation : the positive square root of variance.

Standard Score : a general term referring to any of the variety of 'transformed' scores, in terms of which raw scores may be expressed for reasons of convenience, comparability, ease of interpretation, etc. Sigma Scores, T-Scores etc. are the examples of standard score.

Normal Distribution : a distribution of measures that in graphic form has a distinctive bell-shaped appearance. It is symmetrical and asymptotic. The mean, mode and median for this type for distribution have equal values.

Percentile Rank : the expression of an obtained test score in terms of its position within a group of 100 scores.

Co-efficient of Correlation : a measure of the degree of relationship between two sets of measures for the same group of individuals. Its values ranges from 00, denoting a complete absence of relationship, to +1.00 and –1.00, indicating perfect positive and negative correspondence respectively.

12.6 ANSWERS TO CHECK YOUR PROGRESS EXCERSIES

Check Your Progress Exercise 1

1)

Class Interval	Mid Point	f	x'	fx'	fx'^2
195 - 199	197	1	5	5	25
190 - 194	192	2	4	8	32
185 - 189	187	4	3	12	36
180 - 184	182	5	2	10	20
175 - 179	177	8	1	8	8
170 - 174	172	10	0	0	0
165 - 169	167	6	–1	–6	6
160 - 164	162	4	–2	–8	16
155 - 159	157	4	–3	–12	36
150 - 154	152	2	–4	–8	32
145 - 149	147	3	–5	–15	75
140 - 144	142	1	–6	–6	36
		N = 50	$\sum fx' = -12$	$\sum fx'^2 = 322$	

i) Mean $= AM + \frac{\sum fx'}{N} \times i$

$= 172 + \frac{(-12)}{50} \times 5$

$= 170.80$

ii) Variance $= \sigma^2 = \frac{i^2}{N}\left[N\sum fx'^2 - \left(\sum fx'\right)^2\right]$

$= \frac{(5)^2}{(50)^2}\left[(50 \times 322 - (-12)^2\right]$

$= 159.52$

iii) Standard Deviation $= \frac{i}{N}\sqrt{\left[N\sum fx'^2 - \left(\sum fx'\right)^2\right]}$

$= \frac{5}{50}\sqrt{\left[50 \times 322 - (-12)^2\right]}$

$= 12.63$

2) Normal probability curve is symmetrical around its vertical axis i.e., ordinate.

The values of mean median and mode coincide and have the same value.

The height ordinate is maximum at the mean.

The curve is asymptotic.

The points of inflection of the curve occur at points ±1, Standard deviation (±1σ), above and below the mean.

About 68.26 per cent of the total area falls between the limits M + 1σ and M – 2σ; 95.44 per cent of the total area of the curve falls between limits M + 2σ and 99.73 per cent of the total area of the curve falls between M + 3σ and M – 3σ.

Check Your Progress Exercise 2

1)

X	Y	x	y	x^2	y^2	xy
45	56	–20	–9	400	81	180
55	50	–10	–15	100	225	150
56	48	–9	–17	81	289	153
58	60	–7	–5	49	25	35
60	–62	–5	–3	25	9	15
65(AM)	64	0	–1	0	1	0
68	65(AM)	3	0	9	0	0
70	70	5	5	25	25	25
75	74	10	9	100	81	90
80	82	15	17	225	289	255
85	90	20	25	400	625	500
	$\Sigma x = 2, \Sigma y = 6, \Sigma x^2 = 14/3, \Sigma y^2 = 1650, \Sigma xy = 1403$					

$$r_{xy} = \frac{N\sum xy - (\sum x)(\sum y)}{\sqrt{\left[N\sum x^2 - (\sum x)^2\right]\left[N\sum y^2 - (\sum y)^2\right]}}$$

$$= \frac{11\times1403 - 2\times6}{\sqrt{\left[11\times1414 - (2)^2\right]\left[11\times1650 - (6)^2\right]}}$$

$= 0.92$

Check Your Progress Exercise 3

1) Relative risk is the ratio of the risk of an event (or of developing a disease) in those with antecedent factors compared to those without this factor. Odds ratio, on the other hand, may be defined as the ratio of the odds of an event occurring in one group to the odds of it occurring in another group, or to a sample-based estimate of that ratio.

2) RR of infant death in pregnancy with weight gain < 7 kg = (181/1847) / (36/687) = 1.87.

OR = (181 × 651) / (1666 × 651) = 1.96 i.e. odds of dead children being born to pregnant women with weight <7 kg is approximately two times the odds in women with weight gain more than 7 kg.