BLOCK 4
STATISTICAL METHODS II

212 blank

UNIT 12 CORRELATION: AN INTRODUCTION*

Structure

12.0 OBJECTIVES

After reading this unit, you will be able to:

- explain the concept, direction and magnitude of correlation;
- discuss the properties, uses and limitations of correlation; and
- describe other methods of correlation.

12.1 INTRODUCTION

Let us focus on some of the statements below:

As there is rise in temperature, there is increase in sale of ice creams.

As there is increase in occupational stress experienced by employees, their performance may go down.

Sale of air purifiers can be linked to the increase in pollution.

In our day to day life, we may experience such associations, that is, when one aspect increases or decreases, another aspect also increases or decreases or when one aspect increases, the other decreases and vice versa.

To take more examples, a researcher may be interested in studying whether there exists a relationship between self esteem and achievement motivation of

* Prof. Suhas Shetgovekar, Faculty, Discipline of Psychology, School of Social Sciences, IGNOU, New Delhi

adolescents. Or a researcher may be interested in finding out if there exists any relationship between psychological wellbeing and job satisfaction of employees in an organisation. In such cases, correlation will help the researcher study the relationship and also understand its direction and magnitude.

In research studying relationship between two or more variables could be an important objective and this can be studied with the help of correlation.

In the present unit, we will focus on the concept of correlation, its direction and magnitude. The properties, uses and limitations of correlation will also be covered besides other methods of correlation.

12.2 CONCEPT OF CORRELATION, DIRECTION AND MAGNITUDE OF CORRELATION

Let us take another example, to understand the concept of correlation.

A research was carried out by a researcher on relationship between years of experience and monthly income earned by junior managers in an organisation (hypothetical data). The data given is as follows:

Table 12.1: Years of experience and monthly income of junior managers

Junior Managers	Years of Experience	Monthly income
John	1	25, 000
Ravi	2	30, 000
Maria	3	35, 000
Kuldeep	4	40, 000
Salma	5	45, 000

Looking at this data, what do you understand? You may notice that as the years of experience is increasing, the monthly income earned by the junior managers is also increasing. Thus, it can be inferred that in case of these junior managers, there is a positive relationship between the years of experience and the monthly income.

Let us take another example,

An experimenter was carrying out a study on relationship between hours of practice in a certain task and number of errors committed by the participants. The data obtained from the same is given as follows:

Table 12.2: Hours of practice in certain task and number of errors

Participants	Hours of Practice	Number of errors
Rehman	4	20
Sophia	5	12
Navjyot	7	8
Anjali	1	30
Rahul	2	24

As we look at this data, it can be seen that as the hours of practice is increasing, the number of errors committed by the participants is decreasing. Thus, it can be said that there is a negative relationship between hours of practice and number of errors committed.

With the above examples, you must have developed some idea about what is correlation. Let us now look at the concept of correlation and also focus on its direction and magnitude.

Correlations can be used to study relationship between two or more variables. It is a measure of association between two or more variables and this relationship is determined not only in terms of direction, whether negative or positive but also in terms of its magnitude, whether high or low. However, it will not provide information about any causal relationship between the variables.

Sir Francis Galton's contribution to development of correlation is noteworthy. He carried out studies on individual differences and also studies on the influence of heredity. He studied the association between the height of parents and that of their children with the help of bivariate distribution (that studies relationship between two variables) and found that the parents who are tall have children who are also tall (Veeraraghavan and Shetgovekar, 2016). Further, in 1986, Karl Pearson put forth mathematical procedure for correlation.

Correlation can be categorised in to linear and nonlinear correlation. These are discussed as follows:

Linear Correlation: Linear correlation is denoted by a single straight line in a graph that denotes linear relationship between given two variables. Such a graph indicates whether increase in one variable leads to increase in another variable and vice versa, or decrease in one variable leads to increase in another variable and vice versa. For example, if the scores on emotional intelligence increase or decrease, the scores on self esteem also increase or decrease. Linear relationship is graphically represented in figure 12.1.

Nonlinear correlation: As opposed to linear relationship, in nonlinear relationship, the relationship between two given variables is not denoted by a straight line. Thus, the relationship is curvilinear as denoted in figure12.2.

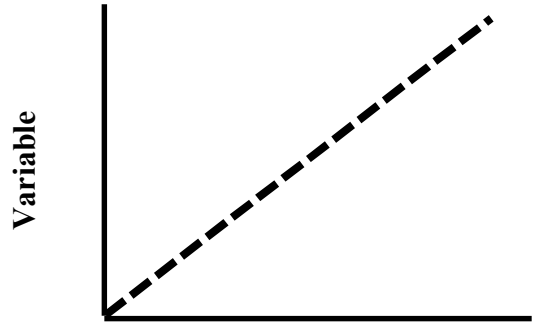

Fig. 12.1: Linear Correlation

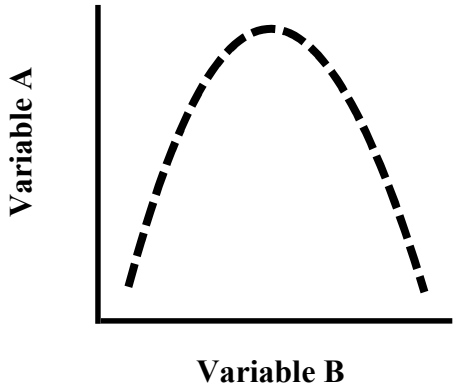

Fig. 12.2: Nonlinear Correlation

12.2.1 Direction and Magnitude of Correlation

With the help of examples that we discussed at the start of this section, it must be clear that correlation can be either positive or negative (though there could also be no relationship between the given two variables). This can be described as direction of correlation. Let us now discuss these in detail.

Positive correlation: Positive correlation denotes that increase in one variable leads to increase in another variable and decrease in one variable leads to decrease in another variable. For example, if the scores on emotional intelligence obtained by adolescents increase, then the scores obtained by them on achievement motivation will also increase or if the scores on emotional intelligence obtained by adolescents decrease then the scores obtained by them on achievement motivation will also decrease. Positive correlation indicates that both the variables are moving in same direction (refer to figure 12.3).

Fig. 12.3: Positive Correlation

Figure 12.3 is a scatter diagram denoting positive relationship between two variables, A and B. Scatter diagram can be effectively used to present a bivariate distribution that denotes relationship between the two variables.

Negative Correlation: Negative correlation denotes that increase in one variable leads to decrease in another variable or decrease in one variable leads to increase in another variable. For example, if the scores on occupational stress obtained by employees increase, then the scores obtained by them on work motivation will decrease or if the scores on occupational stress obtained by employees decrease, then the scores obtained by them on work motivation will increase. In this case, the two variables are not moving in same direction. Figure 6.4 is a diagrammatic representation of negative correlation.

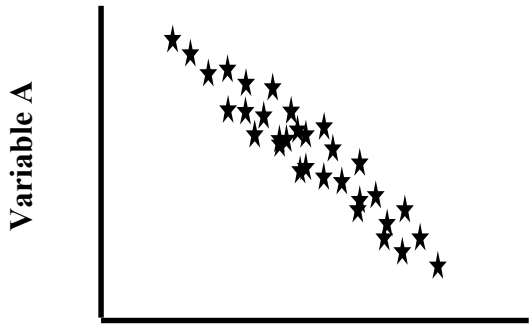

Fig. 12.4: Negative Correlation

No Correlation or Zero Correlation: It may so happen that there is no relationship between the two variables. In such a case the correlation will be zero (this will be further clear as we discuss the magnitude of correlation). Thus, in this case the relationship is neither positive or negative. There are such variables where there might be no relationship, for example, there may be no correlationship between height of persons and years of their work experience or there may exist no relationship between weight of persons and attitude towards environment. No correlation is represented in form of scatter diagram in figure 12.5.

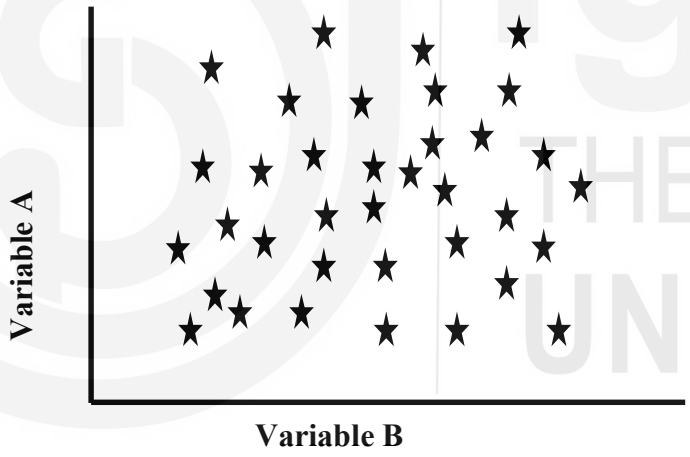

Fig. 12.5: No or Zero Correlation

Besides the direction of the correlation, it is also significant to understand the magnitude or strength of the correlation. Magnitude is denoted by the degree of linearity of the correlationship. Correlationship between any two variables is Coefficient of Correlation that is quantitatively represented. And the range for a coefficient of correlation is between -1 to +1. Thus, coefficient of correlation can be obtained as 0.28 or -0.09 or 0.75 and so on. The number will lie between -1 to +1 and the + and - signs denote the direction of correlation, whether it is positive or negative. The obtained coefficient of correlation can be interpreted with the help of the table 6.3 given as follows (Mangal, 2002, page 105):

Table 12.3: Interpretation of Coefficient of Correlation	
Coefficient of Correlation Range	**Interpretation**
+ 1 or - 1	This can be interpreted as a correlation that is perfect, though the direction could be positive or negative.
± 0.91 to 0.99	Correlation is very high
± 0.71 to 0.90	Correlation is high
± 0.41 to 0.70	Correlation is moderate
± 0.21 to 0.40	Correlation is low
0 to ± 0.20	Correlation is negligible
0	No correlation

While interpreting coefficient of correlation it is important to keep in mind the direction based on the positive and negative signs.

12.2.2 Scatter Diagram

One way in which relationship between two variables can be denoted is by using scatter diagram. Scatter diagram is also called as scatter plot or scatter gram. It is drawn by plotting the two variables in same graph, that is variable A on y axis and variable B on x axis (as shown in figures 12.3, 12.4 and 12.5).

Table 12.4: Marks obtained by students in Psychology and Sociology class test		
Students	**Marks in Psychology**	**Marks in Sociology**
Peter	25	26
Sarika	45	42
Kamaldeep	78	73
Salman	5	4
Arvind	89	84

For example, we want to study the relationship between marks obtained by five students in Psychology and in Sociology in class test. The data is given in table 12.4.

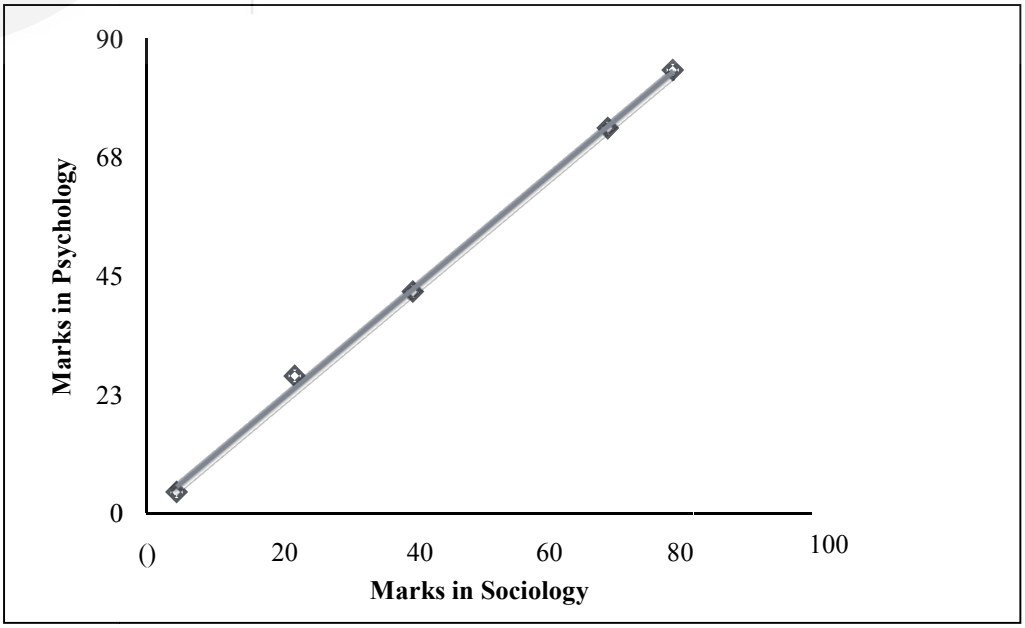

Fig. 12.6: Scatter diagram based on table 12.4

As can be seen from the graph, there is a linear relationship between the marks obtained in psychology and marks obtained in sociology.

Coefficient of Correlation can be computed with the help of Pearson's product moment correlation and Spearman's rank order correlation that will be discussed in the next unit.

Check Your Progress I

1) What is Correlation?

.....

.....

.....

.....

.....

12.3 PROPERTIES, USES AND LIMITATIONS OF CORRELATION

As the concept of correlation is now clear, we will discuss about its properties, uses and limitations.

12.3.1 Properties of Correlation

1) The variables used in correlation are quantitative in nature.

2) The value of coefficient of correlation will range from -1 to +1 and correlation can be positive or negative or there can also be zero correlation.

3) As we have discussed before, correlation will provide information about relationship between the variables, but it will not denote whether a cause and effect relationship exists between the variables. For example, we may obtain a positive correlation between organisational culture and job satisfaction of employees, but correlation will not denote if there is a cause and effect relationship between these two variables.

4) It is not possible to make any predictions based on correlation. For example, there may be positive correlation between rise in temperature and sale of ice cream or cold drinks but the sale of ice cream or cold drinks cannot be predicted based on the temperature with the help of correlation.

5) Even if variables randomly vary, correlation can be used.

6) Sampling errors can have an effect on correlation.

12.3.2 Uses of Correlation

Correlation can be used for varied purposes that have been discussed as follows (Mohanty and Misra, 2016).

1) **Validity and reliability:** Validity and reliability are important aspects of psychological testing and correlation can be used to obtain validity and

reliability of a psychological test. Validity is whether a test is measuring what it is supposed to measure and reliability provides information about consistency of a test.

2) **Verification of theory:** Correlation can also be used to verify or test certain theories by denoting whether relationship exists between the variables. For example, if a theory states that there is a relationship between parenting style and resilience, the same can be tested by computing correlation for the two variables.

3) **Putting variables in groups:** Variables that show positive correlation with each other can be grouped together and variables that show negative correlation can be grouped separately based on the coefficient of correlation obtained.

4) **Computation of further statistical analysis:** Based on the results obtained after computing correlation, various statistical techniques can be used like regression. Further, correlation is also used for multivariate statistical analysis, especially for techniques like Multivariate Analysis of Variance (MANOVA), Multivariate Analysis of Covariance (MANCOVA), Discriminant Analysis, Factor analysis and so on (Mohanty and Misra, 2016).

5) **Based on correlation, one can decide whether or not to determine prediction:** By computing correlation, it is not possible to predict one variable based on another variable, but based on the information that two or more variables are significantly related to each other, further statistical techniques can be used to make predictions. For example, if we obtain a positive correlation between family environment and adjustment of children, then further statistical techniques can be employed to find if adjustment of children can be predicted based on family environment.

12.3.3 Limitations of Correlation

Some of the limitations of correlation have been discussed as follows:

1) As was stated earlier, correlation will not provide any information about cause and effect relationship or causation.

2) The coefficient of correlation, mainly, Pearson's product moment correlation and Spearman's rank order correlation are suitable, when there is a linear relationship between the variables.

3) With regard to distributions that are discontinuous, the coefficient of correlation obtained may be overestimated or higher.

4) Sample variations can have an effect on correlation (as is also true with other statistical techniques).

5) In case of pooled sample, the correlation will be determined by relative position of the scores in X and Y dimensions or variables.

Check Your Progress II

1) List the uses of Correlation.

.....

.....

.....

.....

.....

12.4 OTHER METHODS OF CORRELATION

There are various other methods of correlation as well, that will be discussed in the present section of this unit.

1) **Partial Correlation:** In partial correlation, the relationship between two variables is studied by controlling the influence of a third variable. For example, if we are studying the relationship between emotional intelligence and self concept of adolescents, we may partial out or control the third variable, for instance, family environment .

2) **Part Correlation:** This is also called as Semipartial Correlation. This is in a way similar to partial correlation, but here as the correlationship between two variables is studied, the influence of third variable on one of the variables is controlled. Taking the example discussed under partial correlation, the influence of family environment only on emotional intelligence is controlled and not on self concept.

3) **Multiple correlation:** In multiple correlation, one variable is correlated with many other variables. For example, self concept will be correlated with various other variables like emotional intelligence, achievement motivation, quality of life and so on.

4) **Biserial Correlation:** In biserial correlation, the relationship is measured between a continuous variable and an artificially dichotomous variable. A dichotomous variable is a variable that can be categorised into two. For example, Socio-Economic Status could be high and low. An example of biserial correlation would be relationship between achievement motivation and high and low emotional intelligence. Here, achievement motivation is a continuous variable and emotional intelligence is a variable that is artificially dichotomous.

5) **Point-Biserial Correlation:** In point biserial correlation, the relationship is measured between a continuous variable and a naturally dichotomous variable. Examples of naturally dichotomous variables are gender (male and female), religion (Hindu and Muslim) and so on. For example, point biserial correlation can be used when we want to find out relationship between work motivation (continuous variable) and gender (naturally dichotomous variable).

6) **Tetrachoric Correlation:** When both the variables are artificially dichotomous, then tetrachoric correlation can be computed to study the relationship between the two variables. For example, tetrachoric

correlation can be used when we want to study the relationship between variables, emotional intelligence (that is categorised in to high and low) and adjustment (that is categorised in to well adjusted and maladjusted).

7) **Phi coefficient:** This is similar to tetrachoric correlation, but is used when both the variables are naturally dichotomous. For example, if we want to find relationship between gender, that is categorised as male and female and response to a statement in terms of agree and disagree, then phi coefficient can be computed.

Check Your Progress III

Give a brief description and nature of variables 1 and 2 for other methods of correlation

Method of Correlation	Description	Variable 1	Variable 2
Partial Correlation			
Part Correlation			
Multiple Correlation			

Biserial Correlation			
Point-Biserial			
Tetrachoric Correlation			
Phi Coefficient			

12.5 LET US SUM UP

In the present unit, we mainly discussed about the concept of correlation. Correlation can be used to study relationship between two or more variables. It is a measure of association between two or more variables and this relationship is determined not only in terms of direction, whether negative or positive, but also in terms of its magnitude, whether high or low. Correlation can be categorised in to linear and nonlinear correlation and these were discussed in the present unit with the help of figures. The concept of scatter diagram was also briefly discussed in this unit. Scatter diagram is also called as scatter plot or scatter gram and is drawn by plotting the two variables in same graph, that is, variable A on y axis and variable B on x axis. Further in the unit, we also discussed about the properties, uses and limitations of correlation that are relevant, so as to know when exactly to use correlation. In the last section of this unit, we focused on the other methods of correlation including partial correlation, part correlation, multiple correlation, biserial correlation, point biserial correlation, tetrachoric correlation and phi- coefficient. In the next unit, we will learn how to compute coefficient of correlation with the help of Pearson's product moment correlation and Spearman's rank order correlation.

12.6 KEY WORDS

Biserial Correlation: In biserial correlation, the relationship is measured between a continuous variable and an artificially dichotomous variable.

Correlation: It is a measure of association between two or more variables and this relationship is determined not only in terms of direction, whether negative or positive but also in terms of its magnitude, whether high or low.

Linear Correlation: Linear correlation is denoted by a single straight line in a graph that denotes linear relationship between given two variables.

Multiple correlation: In multiple correlation, one variable is correlated with many other variables.

Nonlinear correlation: Here the relationship is not denoted by a straight line but it is curvilinear.

Negative Correlation: The negative correlation denotes that increase in one variable leads to decrease in another variable or decrease in one variable leads to increase in another variable.

No Correlation or Zero Correlation: When there is no relationship between the two variables, the correlation will be zero. Thus in this case the relationship is neither positive or negative.

Part Correlation: This is also called as Semipartial Correlation. Here as the relationship between two variables is studied, the influence of third variable on one of the variables is controlled.

Partial Correlation: In partial correlation, the relationship between two variables is studied by controlling the influence of a third variable.

Point-Biserial Correlation: In point biserial correlation, the relationship is measured between a continuous variable and a naturally dichotomous variable.

Phi coefficient: Phi coefficient is used when both the variables are naturally dichotomous.

Positive correlation: The positive correlation denotes that increase in one variable leads to increase in another variable and decrease in one variable leads to decrease in another variable.

Scatter diagram: Scatter diagram is drawn by plotting the two variables in same graph, that is variable A on y axis and variable B on x axis.

Tetrachoric Correlation: When both the variables are artificially dichotomous, then tetrachoric correlation can be computed to study the relationship between the two variables.

12.7 REFERENCES

King, Bruce. M; Minium, Edward. W. (2008). *Statistical Reasoning in the Behavioural Sciences*. Delhi: John Wiley and Sons, Ltd.

Mangal, S. K. (2002). *Statistics in Psychology and Education*. new Delhi: Phi Learning Private Limited.

Minium, E. W., King, B. M., & Bear, G. (2001). *Statistical Reasoning in Psychology and Education*. Singapore: John-Wiley.

Mohanty, B and Misra, S. (2016). *Statistics for Behavioural and Social Sciences*. Delhi: Sage.

Veeraraghavan, V and Shetgovekar, S. (2016). *Textbook of Parametric and Nonparametric Statistics*. Delhi: Sage.

12.8 ANSWERS TO CHECK YOUR PROGRESS

Check Your Progress I

1) What is Correlation?

It is a measure of association between two or more variables and this relationship is determined not only in terms of direction, whether negative or positive but also in terms of its magnitude, whether high or low.

Check Your Progress II

1) List the uses of Correlation

Correlation can be used

- to decide to whether or not to determine prediction:
- to obtain validity and reliability of psychological tests
- for verification of theory
- for putting variables in groups
- for computation of further statistical analysis.

Check Your Progress III

Give a brief description and nature of variables 1 and 2 for other methods of correlation

Method of Correlation	Description	Variable 1	Variable 2
Partial Correlation	The correlationship between two variables is studied by controlling the influence of a third variable	Variable is continuous in nature	Variable is continuous in nature
Part Correlation	As the correlationship between two variables is studied, the influence of third variable on one of the variables is controlled	Variable is continuous in nature	Variable is continuous in nature
Multiple Correlation	One variable is correlated with many other variables (continuous)	Variable is continuous in nature	Variable is continuous in nature
Biserial Correlation	The correlationship is measured between a continuous variable and an artificially dichotomous variable	Variable is continuous in nature	Variable is artificially dichotomous
Point-Biserial	The correlationship is measured between a continuous variable and naturally dichotomous variable	Variable is continuous in nature	Variable is naturally dichotomous
Tetrachoric Correlation	It is used when both the variables are artificially dichotomous	Variable is artificially dichotomous	Variable is artificially dichotomous
Phi Coefficient	It is used when both the variables are naturally dichotomous	Variable is naturally dichotomous	Variable is naturally dichotomous

12.9 UNIT END QUESTIONS

1) Explain the concept of correlation with a focus on its direction and magnitude.
2) Discuss linear and nonlinear correlation with the help of diagrams.
3) Describe the properties of correlation.
4) Explain the uses and limitations of correlation.
5) Describe other methods of correlation.

UNIT 13 COMPUTATION OF COEFFICIENT OF CORRELATION*

Structure

13.0 OBJECTIVES

After reading this unit, you will be able to :

- learn to compute coefficient of correlation with the help of Pearson's product moment coefficient of correlation and Spearman's rank order correlation.

13.1 INTRODUCTION

In the previous unit, we discussed about the basics of correlation. We discussed that correlation indicates relationship between two or more variables. This correlation can be interpreted in terms of direction and magnitude. Thus, a relationship between given two variables can be positive, negative or there could be no correlation. Further, the correlation may range between +1 to -1.

In the present unit, we will learn about the computation of correlation with the help of Pearson's product moment correlation and Spearman's rank order correlation.

One way in which these two methods can be distinguished is that, Pearson's product moment correlation can be categorised under parametric statistics and the Spearman's rank order correlation falls under nonparametric statistics.

To distinguish between parametric and nonparametric statistics, the following table (table 13.1) can be referred to:

* Prof. Suhas Shetgovekar, Faculty, Discipline of Psychology, School of Social Sciences, IGNOU, New Delhi

Table 13.1: Difference between Parametric and Nonparametric statistics	
Parametric	**Non-parametric**
The assumed distribution is normal.	The assumed distribution may not be normal. It can be any distribution.
The variance is homogeneous.	The variance could be heterogeneous or no assumption is made with regard to the variance.
The scales of measurement used are interval or ratio.	The scales of measurement used are nominal or ordinal.
The relationship between the data needs to be independent.	There is no assumption with regard to the independence of relationship between the data.
Mean is the measure of central tendency that is used here.	Median is the measure of central tendency that is used here.
It is more complex to compute when compared to the non parametric techniques.	It is simple to calculate.
Can get affected by outliers.	Is comparatively less affected by outliers.

In the next section, we will learn how to compute Pearson's product moment correlation.

13.2 PEARSON'S PRODUCT MOMENT CORRELATION

Pearson's product moment correlation is one of the methods to compute coefficient of correlation. This is mainly used when the assumptions of parametric statistics are met. This method is named after Karl Pearson, who invented this method. It is denoted by 'r'.

13.2.1 Assumptions of Pearson's Product Moment Correlation

The assumptions of Pearson's product moment correlation are as follows:

1) The variables used to compute 'r' are continuous in nature and the scales of measurement are interval and ratio.

2) The distribution of the variables in this method is unimodal and it is close to symmetrical. The distribution need not be normal.

3) The pairs of scores involved are independent in nature and are in no way connected with other.

4) There is a linear relationship between the two variables. A scatter gram thus drawn with the help of scores in the two variables, will denote a straight line.

5) 'r' is mainly used to ascertain the sign and size of the correlation that can be positive, negative or zero correlation and will range between -1 to +1.

13.2.2 Uses of Pearson's Product Moment Correlation

1) It helps in determining the relationship between two variables quantitatively. With quantification, it is possible for us to compare.

2) Based on 'r', regression equation can be computed. Thus, after computing 'r', it is possible to compute regression and determine whether one variable can be predicted based on another variable.

3) 'r' can be used in computation of reliability and validity of psychological tests.

4) It will also assist in computation of factor analysis.

13.2.3 Computation of Pearson's Product Moment Correlation

There are two main methods that we will discuss for computing Pearson's product moment correlation. They are discussed as follows:

Method 1: The formula for the first method is give below,

$$\mathbf{r_{xy}} = \Sigma xy / N\, \sigma_x \sigma_y$$

Where,

r = Correlation

x = Deviation of any score of X from the mean of X

y = Deviation of any score of Y from the mean of Y

Σxy= Indicates the sum of all the products of deviation (that is, each x deviation is multiplied by its corresponding y deviation)

σ_x = Standard deviation of scores in X

σ_y= Standard deviation of scores in Y

N = Total number of participants (frequencies)

The formula can be simplified as follows

$\sigma_x = \sqrt{\Sigma x^2 / N}$

$\sigma_y = \sqrt{\Sigma y^2 / N}$

Thus, by substituting the values for $\sigma_{x \text{ and }} \sigma_y$, the following is obtained :

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

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

Let us understand this method and steps involved in it, with the help of an example,

A researcher wanted to study the relationship between data 1 (X) and data 2 (Y). The data is given below:

Participants (1)	Data 1 (X) (2)	Data 2 (Y) (3)	*x* (4)	*y* (5)	*xy* (6)	x^2 (7)	y^2 (8)
1	3	4	0	0	0	0	0
2	2	3	-1	-1	1	1	1
3	4	5	1	1	1	1	1
4	4	5	1	1	1	1	1
5	3	5	0	1	0	0	1
6	2	3	-1	-1	1	1	1
7	2	3	-1	-1	1	1	1
8	3	4	0	0	0	0	0
9	5	5	2	1	2	4	1
10	2	3	-1	-1	1	1	1
	$\Sigma X = 30$	$\Sigma Y = 40$			$\Sigma xy = 8$	$\Sigma x^2 = 10$	$\Sigma y^2 = 8$

Step 1: First the scores under X and and Y are totalled separately. Thus, ΣX and ΣY is obtained as can be seen above in the second and third column. N is also noted and in this case it is 10.

Step 2: Mean is now computed for the data 1 (X) and 2 (Y) as follows:

Mean for scores on X = 30 (30/10) = 3

Mean for scores on Y = 40 (40/10) = 4

Step 3: In the third step, deviation is computed for each score of X from its mean, that is, 3 in the case of this example. In a similar manner deviation is computed for each score of Y from its mean, that is, 4. These are entered in the column four and five above under headings 'x' and 'y' respectively.

Step 4: The values thus entered under 'x' and 'y' are multiplied and entered in column six and then they are also squared and entered under column seven and eight with headings x^2 and y^2. Further, the scores under each of these columns are totalled to obtain Σxy, Σx^2 and Σy^2.

Step 5: Use the formula to compute 'r'.

$$r = \Sigma xy / \sqrt{\Sigma x^2 \Sigma y^2}$$

$$= 8/ \sqrt{10 \times 8}$$

$$= 8/ \sqrt{80}$$

$$= 8/8.94$$

$$= 0.89$$

Thus, the coefficient of correlation obtained for the above data is 0.89, denoting that there is a positive and high relationship between the two data sets X and Y.

Method 2: The formula for the second method is give below,

$$r = N\Sigma XY - \Sigma X \Sigma Y / \sqrt{[N \Sigma X^2 - (\Sigma X)^2] [N \Sigma Y^2 - (\Sigma Y)^2]}$$

Where,

X and Y= the raw scores for X and Y

ΣXY = The total of the products of each X score multiplied with its corresponding Y score

N= Total number of scores.

In this method, the deviations from the mean are not computed, instead raw scores are used to compute 'r'.

Let us understand this method and steps involved in it with the help of an example used earlier for calculating 'r'.

Participants (1)	Data 1 (X) (2)	Data 2 (Y) (3)	XY (4)	X^2 (5)	Y^2 (6)
1	3	4	12	9	16
2	2	3	6	4	9
3	4	5	20	16	25
4	4	5	20	16	25
5	3	5	15	9	25
6	2	3	6	4	9
7	2	3	6	4	9
8	3	4	12	9	16
9	5	5	25	25	25
10	2	3	6	4	9
	ΣX= 30	**ΣY= 40**	**ΣXY= 128**	**ΣX^2 = 100**	**ΣY^2 = 168**

Step 1: Total scores for X and Y in column two and three are computed and denoted as ΣX and ΣY. In the case of present example, they are obtained as 30 and 40.

Step 2: In column four, XY is computed where the paired values under X and Y are multiplied. Thus, for participant 1, X value 3 and Y value 4 are multiplied and 12 is obtained. Similarly, XY is computed for all the participants and then ΣXY is also computed.

Step 3: In column five and six, X^2andY^2 are computed. These are squared values of X and Y respectively. Further, $\Sigma X^2$and ΣY^2 are computed that are summations of X^2andY^2respectively.

Step 4: Use the formula to compute 'r'.

$$r = N\Sigma XY - \Sigma X \Sigma Y / \sqrt{[N \Sigma X^2 - (\Sigma X)^2] [N\Sigma Y^2 - (\Sigma Y)^2]}$$

$$= 10 \times 128 - (30 \times 40) / \sqrt{[(10 \times 100 - (30)^2] [(10 \times 168 - (40)^2]}$$

$$= 1280 - (1200) / \sqrt{[1000 - 900] [1680 - 1600]}$$

$$= 80 / \sqrt{100 \times 80}$$

$$= 80 / \sqrt{8000}$$

$$= 80 / 89.44$$

$$= 0.89$$

Thus, 'r' obtained is 0.89, denoting positive and high correlationship between the two data sets.

Check Your Progress I

1) Pearson Product Moment Correlation is denoted as ..... .

2) The variables used to compute 'r' are continuous in nature and the scales of measurement are ..... and ..... .

3) The formula for the first method of computing Pearson's product moment correlation is ..... .

13.3 SPEARMAN'S RANK ORDER CORRELATION

Another method to compute coefficient of correlation is Spearman's rank order correlation (also know as Spearman's rho). This method is used when the assumptions of parametric statistics are not met. The method is named after Charles Spearman, who is known for his significant work on factor analysis and theory of intelligence besides Spearman's rank order correlation.

13.3.1 Assumptions for Spearman's Rank Order Correlation

The assumptions of Spearman's rank order correlation are as follows:

1) The variables are measured in terms of ordinal scale.

2) The relationship between the two variables is linear in nature.

3) The observations are independent in nature, thus denoting that the sample needs to be randomly selected.

4) The pairs of scores are independent in nature and are in no way connected with other pairs.

13.3.2 Uses of Spearman's Rank Order Correlation

1) It is used when the data is measured with the help of ordinal scale.

2) It is especially useful when the sample size is small, that is, less than 25-30 (Mohanty and Misra, 2016).

3) Many a times it is not possible to measure traits directly. Thus, they are measured in terms of ranks. Spearman's rank order correlation involves separately ranking the scores in the two data, followed by computation of correlationship between them.

4) It can be used to study the degree of relationship between two variables that are monotonic. A relationship is termed as monotonic when the variables display consistent but one directional relationship.

13.3.3 Computation of Spearman's Rank Correlation

There are two main methods that we will discuss for computing Spearman's rank order correlation, one without tied ranks and one with tied ranks. There are discussed as follows:

Method 1 (without tied ranks): The formula for the first method is give below,

$$p = 1 - [(6\Sigma d^2) / [N (N^2 - 1)]$$

where,

Σd^2= Sum of the difference squared

N = Total number of participants

Let us understand this method and steps involved in it, with the help of an example,

A researcher wanted to study the relationship between data 1 (X) and data 2 (Y). The data is given below:

Participants (c)	Data 1(X) (2)	Data 2 (Y) (3)	Rank for Data 1 (R_1) (4)	Rank for Data 2 (R_2) (5)	Difference in Ranks (R_1- R_2= |d|) (6)	Difference Squared (d^2) (7)
1	45	40	9	9	0	0
2	34	33	8	8	0	0
3	23	25	3	3	0	0
4	22	21	2	2	0	0
5	65	60	10	10	0	0
6	33	30	7	5	2	4
7	30	31	5	6	1	1
8	25	32	4	7	3	9
9	32	29	6	4	2	4
10	21	20	1	1	0	0
N= 10						$\sum d^2$= 18

Step 1: Ranks are assigned separately to scores under data 1 and those under data 2. These ranks are mentioned in column four and five respectively. Ranks can either be assigned in descending or ascending order. For instance in the present example, rank 1 is assigned to the lowest value and rank 10 to the highest value and this is followed in same way for both the data.

Step 2: Difference in Ranks are calculation and these are irrespective of their signs(R_1- R_2= |d|). These are then mentioned in column six. In the last column,

that is, column seven, difference squared(d^2) is computed and the total of this is mentioned as $\sum d^2$. In the case of present example, $\sum d^2$ is 18.

Step 3: The formula used to compute Rho is

$$p = 1 - [(6\Sigma d^2) / [N (N^2 - 1)]$$

$$= 1 - (6 \times 18) / 10 (10^2-1)$$

$$= 1 - (108)/ 10 (100-1)$$

$$= 1 - (108/ 10 \times 99)$$

$$= 1 - (108/ 990)$$

$$= 1 - 0.11$$

$$= 0.89$$

Thus, the correlation of coefficient (Rho) obtained for the above data is 0.89, thus denoting a positive and high correlationship betweem data 1 and data 2.

Method 2 (with tied ranks): The formula used for computing rho with tied ranks is the same. Only we need to understand how ranks are assigned when there are two or more similar scores in a given data.

Let us understand this method and steps involved in it, with the help of an example,

A researcher wanted to study the relationship between data 1 (X) and data 2 (Y). The data obtained is given below:

Participants (c)	Data 1 (X) (2)	Data 2 (Y) (3)	Rank for Data 1 (R_1) (4)	Rank for Data 2 (R_2) (5)	Difference in Ranks ($R_1 - R_2 = \|d\|$) (6)	Difference Squared (d^2) (7)
1	45	40	8	7	1	1
2	23@	33	2.5	6	3.5	12.25
3	23@	25	2.5	3	0.5	0.25
4	64	21	9	2	7	49
5	65	60#	10	9	1	1
6	33	60#	7	9	2	4
7	25#	31	4.5	5	0.5	0.25
8	25#	60#	4.5	9	4.5	20.25
9	32	29	6	4	2	4
10	21	20	1	1	0	0
N= 10						**$\sum d^2$= 92**

As can be seen in the above table, there are same values under data 1, that is 23, obtained by participant 2 and 3 and score of 25 obtained by participants 7 and 8. Similarly, in data 2, participants 5, 6 and 8 have obtained 60 score. In such a case ranks are assigned in a bit different manner.

As can be seen in above table, 21 is assigned with rank 1 and then there are two '23' scores that need to be equally assigned ranks 2 and 3. Thus 2 +3 = 5/ 2= 2.5. The rank 2.5 is then allotted to both these score. The next score is then allotted rank 4. But in present example, rank 4 and 5 are shared equally by score 25 obtained by 7th and 8th participants. Thus 4+5= 9/ 2 (because there are two same scores) = 4.5. Thus, 4.5 is allotted to these two scores and the next score, that is, 32 is assigned rank 6.

In data 2, the score 60 equally shares ranks 8, 9 and 10. Thus 8 + 9 +10 = 27/ 3 (because there are three same scores) = 9. Thus the score 60 is assigned rank 9.

Using the same formula Rho is computed as follows

$$p = 1 - [(6\Sigma d^2) / [N (N^2 - 1)]$$

$$= 1 - (6 \times 92) / 10 (10^2 - 1)$$

$$= 1 - (552)/ 10 (100-1)$$

$$= 1 - (552/ 10 \times 99)$$

$$= 1 - (552/ 990)$$

$$= 1 - 0.56$$

$$= 0.44$$

Thus, the correlation of coefficient (Rho) obtained for the above data is 0.44, thus denoting a positive correlation between data 1 and data 2.

Check Your Progress II

1) The variables in Spearman's Rho are measured in terms of ..... scale.

2) A relationship is termed as monotonic when

.....

.....

.....

.....

.....

3) The formula for the first method of computing Spearman's rho is

13.4 LET US SUM UP

In the present unit, we mainly discussed about the two methods of computing coefficient of correlation. The first method is Pearson's product moment correlation and the other is Spearman's rank order correlation. Pearson's product moment correlation is one of the methods to compute coefficient of correlation. This is mainly used when the assumptions of parametric statistics

are met. This method is named after Karl Pearson, who invented this method. It is denoted by 'r'. Spearman's rank order correlation is used when the assumptions of parametric statistics are not met. The method is named after Charles Spearman, who is known for his significant work on factor analysis and theory of intelligence. The assumptions and uses of these method were also discussed. The formula and computation for the two methods were discussed with the help of examples.

13.5 REFERENCES

Mangal, S. K. (2002). *Statistics in Psychology and Education*. New Delhi: Phi Learning Private Limited.

Mohanty, B and Misra, S. (2016). *Statistics for Behavioural and Social Sciences*. Delhi: Sage.

Veeraraghavan, V and Shetgovekar, S. (2016). *Textbook of Parametric and Nonparametric Statistics*. Delhi: Sage.

13.6 ANSWERS TO CHECK YOUR PROGRESS

Check Your Progress I

1) Pearson Product Moment Correlation is denoted as r.

2) The variables used to compute r are continuous in nature and the scales of measurement are interval and ratio.

3) The formula for the first method of computing Pearson's Product Moment Correlation is $\mathbf{r_{xy} = \Sigma xy / N \sigma_x \sigma_y}$

Check Your Progress II

1) The variables in Spearman's Rho are measured in terms of Ordinal scale.

2) A relationship is termed as monotonic when the variables display consistent but one directional relationship.

3) The formula for the first method of computing Spearman's rho is $p = 1 - [(6\Sigma d^2) / [N (N^2 - 1)]$

13.7 UNIT END QUESTIONS

1) Differentiate between Parametric and Nonparameric Statistics.

2) Discuss the assumptions of Pearson's product moment correlation

3) Describe the uses of Pearson's product moment correlation.

4) Discuss the assumptions of Spearman's rank order correlation.

5) Describe the steps involved in computation of Spearman's rho with the help of an example.

UNIT 14 NORMAL PROBABILITY DISTRIBUTION*

Structure

14.0 OBJECTIVES

After reading this unit, you will be able to:

- describe the concept and nature of probability;
- discuss the concept, nature, properties and relevance of normal distribution curve;
- elucidate the concept, properties and uses of standard scores; and
- explain divergence from normality.

14.1 INTRODUCTION

Let us understand the main concept of this unit, that is, normal probability curve with an example. Suppose we take variables like height or weight or psychological variables like intelligence or emotional intelligence and we

* Dr. Smita Gupta, Faculty, Discipline of Psychology, School of Social Sciences, IGNOU, New Delhi

measure the same across a large population, we are likely to obtain a graph which will indicate that the maximum scores lie in the middle of the distribution and lower scores at the extreme. With regard to height, for example, there will be more persons with average height and few persons who are taller or shorter compared to the average. And when this data is plotted on a graph, we will get a normal cure as can be seen in figure 14.1.

In the previous units, we discussed about the concept of statistics, data organisation and graphical representation, measures of central tendency and variability and we also discussed about correlation and its computation.

In the present unit, you will be introduced to the concept of probability, normal probability curve and other related aspects.

14.2 CONCEPT OF PROBABILITY

The term 'Probability' refers to chance or likelihood. For example, if you say that, "it will probably be hot tomorrow" or "probably the teacher might not come tomorrow", the sentences reveal that there are chances for the events to occur tomorrow but there is no certainty. This means that the events mentioned in the above examples are not certain to happen. Now, a question that might bother you is 'what is research and statistics to do with probability?' The probability of an event to occur due to certain reason helps the researchers to formulate hypotheses to continue with research and experiments. It serves as a baseline for the researchers to start as well as conclude their research. It is relevant in science and investigation and is also important in finance, gambling, artificial intelligence, mathematics and game theory. Researchers can draw conclusions on basis of probabilities.

In simple terms, probability refers to the chance, possibility or likelihood of an event to occur. In statistics, the term 'Probability' refers to the expected frequency (chance) of occurrence of an event among all possible similar events. The expected frequency of the occurrence of the event is based on knowledge or information about the conditions determining the occurrence of the event or phenomena. For example, if you toss a coin in air,there is an equal chance for head or tail to appear. Thus, the probability that head or tail will appear, when a coin is tossed in air, is ½ to appear. Similarly, a die (singular of dice) has six sides with dots ranging from one to six. The probability of a die to show any of its side is 1/6.

The probability is denoted in form of ratio in statistical terms, wherein a probability ratio is denoted as:

Probability Ratio = Desired Outcome/s or event(s) / Total Number of outcomes or events

To make it more clear, the probability ratio of any side of the coin to occur after each toss will be

= 1 (Either head or tail)/ 2 (Total Number of sides of coin)

And the probability ratio of any side of the dice to fall is

=1 (Any one side of the six sides)/ 6 (Total Number of sides of dice)

A probability ratio always ranges between the limit of 0.00 (impossibility of occurrence) to 1.00 (certainty of occurrence). We can say that the possibility of the Sun to set in east is 0.00 and the possibility for the Sun to set in west is 1.00. The other possible degrees of likelihood range between these limits (0.00-1.00) and are expressed in form of appropriate ratios. It is also worth mentioning here that if the probability of an event is higher, then, there are more chances that the event will occur. For example, predicting whether it will be hot the next day will have more chance to occur if on the present day the temperature might have been high enough. Similarly, if the teacher has fallen sick then it is more likely that s/he will not be coming to school on the next day.

14.2.1 Concepts Related to Probability

Probability is a statistical concept which can be measured and analysed. Due to its much scientific applications, there are certain related concepts which you need to know. These concepts are:

1) **Sample Space, Events and Power Set:** The collection or set of all possible results or outcome of an experiment or event is known as sample space. For example, in a die the all possible results to come when it is rolled, ranges between 1 to 6 (1, 2, 3, 4, 5, 6 are the number of dots assigned at each respective face of the dice). So the sample space for the outcome of die ranges from 1 to 6. The subset of the sample space which is a specified collection of possibilities from the overall possibilities is called power set. If we consider that a collection of possible results can be all even numbers (2, 4, 6) of the die, the subset (2, 4, 6) is an element of the power set of the sample space of dice rolls. These collections are called "events". In this case, {2, 4, 6} is the event that the die falls on some even number. If the results that actually occur fall in a given event, the event is said to have occurred. After considering all different collections of possible results of a sample space, a power set is formed.

2) **Mutually Exclusive/Disjoint Events:** Mutually exclusive event refers to incompatibility. Two events are said to be mutually exclusive if both cannot occur simultaneously in a single trial. In such circumstances, the occurrence of one event prevents the occurrence of other event. For example, when a single coin is tossed there is a chance of either head (H) to appear or tail (T) to appear in a single trial but both can not be up at a same time, then both H and T are called mutually exclusive events. Hence, if H an T are mutually exclusive events then, probability (HT) = 0.

3) **Exhaustive/Collective Events:** An event is said to be exhaustive when its totality includes all possible outcomes of a random experiment. Meaning thereby, that out of a set of events, atleast one event occurs in each trial. For example, if you throw a die, the outcomes will range in between 1,2,3,4,5 and 6. You can not get 7 because it is not there in the die at all. Therefore the outcomes 1, 2, 3, 4, 5, and 6 are collectively exhaustive, because they constitute the possible entire range of probable outcomes.

4) **Independent and Dependent Events:** Two or more events are said to be independent events if the outcome of one event has no effect or is not affected by the outcome of the other event. For example, the results of

tossing a coin will not be affected by the outcome of throwing a die. While dependent events are those events in which either occurrence or non occurrence of one event in any trial affects the probability of other event in other trials. For example, we have five apples and five oranges in a basket. We take out one of the fruits, which could be apple or orange. If it was apple then there are four apples and five oranges in the bag. Thus, the probability of the next frute that we take from the basket being apple is 4/9. But if the first time the fruit was orange then the probability would be 5/9.

5) **Equally Likely Events:** Events are said to be equally likely events if each event has a fair enough chance to occur approximately the same number of times. Thus, none of the event is more likely to occur more often than others. For example, if an unbiased coin is tossed, each face of the coin may be expected to be observed for the same number of times.

6) **Complementary Events:** If two events are mutually exclusive and exhaustive, they both are said to complement each other. For example, if there are two events- X and Y, then X is called the complementary event of Y and vice versa. Let us understand it by throwing a die, the occurrence of even numbers (2, 4, 6) and odd numbers (1, 3, 5) are complementary events. They are the two possible outcomes of an event, where they are the only two possible outcomes.

Check Your Progress I

1) State whether the statements are True or False

Sr. No.	Statements	True/ False
A	Events are said to be equally likely events if each event has a fair enough chance to occur approximately the same number of times.	
B	A probability ratio always ranges between the limit of -1.00 to +1.00.	
C	Two events are said to be mutually exclusive if both occur simultaneously in a single trial.	
D	Two or more events are said to be independent events if the outcome of one event does not effect as well as is not affected by the outcome of the other event.	
E	If two events are mutually exclusive and exhaustive, they both are said to complement each other.	

14.3 CONCEPT, NATURE AND PROPERTIES OF NORMAL PROBABILITY DISTRIBUTION

A continuous probability distribution for a variable is called as normal probability distribution or simply normal distribution. It is also known as Gaussian/ Gauss or LAPlace – Gauss distribution. The normal distribution is

determined by two parameters, mean and variance. The normal distributions are used to represent the real valued random variables whose distributions are unknown. They are used very frequently in the areas of natural sciences and social sciences. When the normal distribution is represented in form of a graph, it is known as normal probability distribution curve or simply normal curve. A normal curve is a bell shaped curve, bilaterally symmetrical and is continuous frequency distribution curve. Such a curve is formed as a result of plotting frequencies of scores of a continuous variable in a large sample. The curve is known as normal probability distribution curve because its y ordinates provides relative frequencies or the probabilities instead of the observed frequencies. A continuous random variable can be said to be normally distributed if the histogram of its relative frequency has shape of a normal curve (as represented in the below figure 14.1).

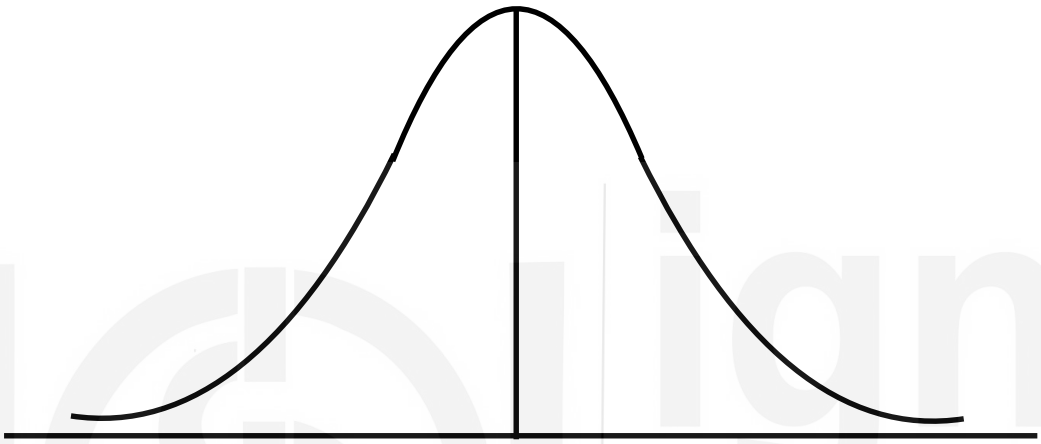

Fig. 14.1: Normal Curve

It is most important to understand the characteristics of frequency distribution of Normal Probabilty Curve (NPC) in the fields of mental measurement and experimental psychology.

14.3.1 Importance of Normal Distribution

As discussed earlier, the normal distribution plays a very significant role in the fields of natural science and other social sciences. Some of the relevance of the normal distribution are described below:

- The normal distribution is a continuous distribution and plays significant role in statistical theory and inference.
- The normal distribution has various mathematical properties which makes it convenient to express the frequency distribution in simplest form.
- It is a useful method of sampling distribution.
- Many of the variables in behavioral sciences like, weight, height, achievement, intelligence have distributions approximately like the normal curve.
- Normal distribution is a necessary component for many of the inferential statistics like z-test, t-test and F-test.

14.3.2 Properties of Normal Distribution

As discussed earlier, the representation of normal distribution of random variable in graphic form is known as Normal Probability Curve (NPC). The following are the properties of the normal curve:

- It is a bell shaped curve which is bilaterally symmetrical and has continuous frequency distribution curve.
- It is a continuous probability distribution for a random variable.
- It has two halves (right and left) and the value of mean, median and mode are equal (mean = median = mode), that is, they coincide at same point at the middle of the curve.
- The normal curve is asymptotic, that is, it approaches but never touches the x-axis, as it moves farther from mean.
- The mean lies in the middle of the curve and divides the curve in to two equal halves. The total area of the normal curve is within $z \pm 3\ \sigma$ below and above the mean.
- The area of unit under the normal curve is said to be equal to one (N=1), standard deviation is one (σ =1), variance is one (σ^2=1) and mean is zero (μ=0).
- At the points where the curve changes from curving upward to curving downward are called inflection points.
- The z-scores or the standard scores in NPC towards the right from the mean are positive and towards the left from the mean are negative.
- About 68% of the curve area falls within the limit of plus or minus one standard deviation ($\pm 1\ \sigma$) unit from the mean; about 95% of the curve area falls within the limit of plus or minus two standard deviations ($\pm 2\ \sigma$) unit from the mean and about 99.7% of the curve area falls within the limit of plus or minus three standard deviations ($\pm 3\ \sigma$) unit from the mean (refer to figure 14.2).
- The normal distribution is free from skewness, that is, it's coefficient of skewness amounts to zero.
- The fractional areas in between any two given z-scores is identical in both halves of the normal curve, for example, the fractional area between the z-scores of +1 is identical to the z-scores of –1. Further, the height of the ordinates at a particular z-score in both the halves of the normal curve is same, for example, the height of an ordinate at +1z is equal to the height of an ordinate at –1z.

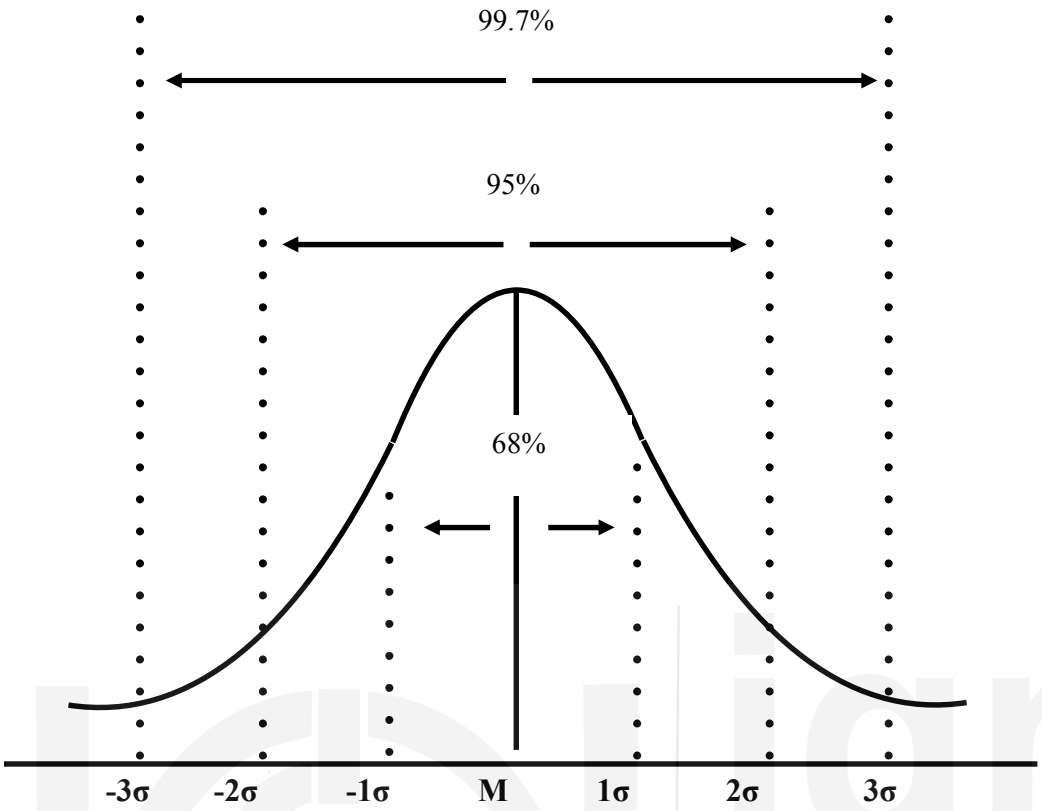

Fig. 14.2: Normal Probability Curve(NPC)

Check Your Progress II

1) Fill in the Blanks

   a) NPC is a …….shaped curve which is bilaterally symmedrical.

   b) In a normal probability curve, the value of mean, median and mode are ..... .

   c) The normal distribution is a ..... distribution and plays significant role in statistical theory and inference.

   d) The area of unit under the normal curve is said to be equal to.....

   e) The normal distribution is determined by two parameters..... and .....

14.4 STANDARD SCORES (Z-SCORES)

Standard score or z-score is a transformed score which shows the number of standard deviation units by which the value of observation (the raw score) is above or below the mean. The standard score helps in determining the probability of a score in the normal distribution. It also helps in comparing scores from different normal distributions.

14.4.1 Concept of Standard Score (z-score)

The standard score is a score that informs about the value and also where the value lies in the distribution. Typically, for example, if the value is 5 standard deviations above the mean then it refers to five times the average distance above the mean. It is a transformed score of a raw score. A raw score or sample value is the unchanged score or the direct result of measurement. A raw score (X) or sample value cannot give any information of its position within a distribution. Therefore, these raw scores are transformed in to z-scores to know the location of the original scores in the distribution. The z-scores are also used to standardise an entire distribution.

These scores (z) help compare the results of a test with the "normal" population. Results from tests or surveys have thousands of possible results and units. These results might not be meaningful without getting transformed. For example, if a result shows that height of a particular person is 6.5 feet; such findings can only be meaningful if it is compared to the average height. In such a case, the z-score can provide an idea about where the height of that person is in comparison to the average height of the population.

14.4.2 Properties of z-score

Following are some of the properties of the Standard (z) Score:

- The mean of the z-scores is always 0.
- It is also important to note that the standard deviation of the z-scores is always 1.
- Further, the graph of the z-score distribution always has the same shape as the original distribution of sample values.
- The z-scores above the value of 0 represent sample values above the mean, while z-scores below the value of 0 represent sample values below the mean.
- The shape of the distribution of the z-score will be similar or identical to the original distribution of the raw scores. Thus, if the original distribution is normal, then the distribution of the z-score will also be normal. Therefore, converting any data to z-score does not normalize the distribution of that data.

14.4.3 Uses of z-score

z-scores are useful in the following ways:

- **It helps in identifying the position of observation(s) in a population distribution:** As mentioned earlier, the z-scores helps in determining the position/distance of a value or an observation from the mean in the units of standard deviations. Further, if the distribution of the scores is like the normal distribution, then we are able to estimate the proportion of the population falling above or below a particular value. z-score has important implication in the studies related to diet and nutrition of children. It helps in estimating the values of height, weight and age of children with reference to nutrition.

- **It is used for standardising the raw data:** It helps in standardising or converting the data to enable standard measurements. For example, if you wish to compare your scores on one test with the scores achieved in another test, comparison on the basis of raw score is not possible. In such a situation, comparisons across tests can only be done when you standardise both sets of test scores.
- **It helps in comparing scores that are from different normal distributions:** As mentioned in the previous example, z-scores help in comparing scores from different normal distribution. Thus, z-scores can help in comparing the IQ scores received from two different tests.

14.4.4 Computation of z-score

As mentioned earlier, z-score refers to the distance of the sample value from the mean in the standard deviations. z-score can be computed for each value of the sample. The following formula is used to compute z-score of a sample value-

$$z = X- M/ SD \text{ or}$$

$$z= X- M/\sigma$$

where,

X = a particular raw score

M = Sample mean

SD or σ = Standard Deviation

To illustrate, suppose the following are the marks obtained by students in mathematics. The marks obtained are expressed here in terms of raw scores. The mean, SD and z-scores can be then calculated accordingly:

Students	Raw Scores (X)	X- M	z
A	50	-15	-1.24
B	60	-5	-0.41
C	66	1	0.08
D	70	5	0.41
E	80	15	1.24
N= 5			
Sum	326		
Mean	65		
SD	12.04		

The above illustration shows the z-scores of the marks obtained by each student (A,B,C,D and E). In the above example, student A is 1.24 standard deviations, or 1.24 standard deviation units below the mean. Similarly the student E is 1.24 units above the mean. The standard deviation is used as unit of measurement in standard scores. The standard score helps in normalising or

collapsing the data to a common standard based on how many standard deviations values lie from the mean.

The variation of z-scores range from -3 standard deviations (which would fall to the far left of the normal distribution curve) up to +3 standard deviations (which would fall to the far right of the normal distribution curve). Further, we need to know the values of the μ (mean) and also the σ (standard deviation) of the population.

Thus, if we want to compute z-score for X = 70, M= 65 and SD= 12.04, we will use the formula

$$\mathbf{z = X- M/ SD}$$

$$= 70- 65/ 12.04$$

$$= 5/ 12.04$$

$$= 0.42$$

Thus, the z-score is obtained as 0.42

Check Your Progress III

1) Fill in the Blanks

   a) The mean of the z-scores is always ..... .

   b) ..... is used for standardising the raw data.

   c) The variation of z-scores ranges from ..... to ..... standard deviations.

   d) The standard deviation of the z-scores is always ..... .

   e) The standard score is a score that informs about the .....

14.5 DIVERGENCE FROM NORMALITY: SKEWNESS AND KURTOSIS

Many times, the frequency curve may be more peaked or flatter than the normal probability distribution curve. In such cases, the distribution is said to have diverged from normality. Basically these divergences are of two types: Kurtosis and Skewness. Kurtosis is a measure of "tailedness" of the probability distribution of a random variable. In other words, it is a measure whether a data is heavy tailed or light tailed in relation to normal distribution. On the other hand, Skewness is a measure of asymmetry of the probability distribution of a random variable about its mean. Let us discuss both the divergences one by one.

14.5.1 Kurtosis

Kurtosis deals with the tails of distribution curve and not its peak. It does not refer to the height of the curve. Kurtosis can be quantified in various ways for a particular distribution. The kurtosis of any variable in normal distribution is 3. This value is used for comparison with respect to the other types of Kurtosis. The kurtosis is classified as follows (refer to figure 14.3):

1) **Leptokurtosis/ Leptokurtic**

In a Leptokurtic distribution, the frequency curve is narrower than the NPC and the area of the curve shifts towards the center and has longer tails at both the ends. Usually they are referred to as positive kurtosis in which the value of distribution has heavier tails than the normal distribution (refer to figure 14.3).

2) **Mesokurtosis/ Mesokurtic**

A mesokurtic curve is not too flat or not too peaked and in a way resembles normal curve (Mangal, 2002).

3) **Platykurtosis/Platykurtic**

Platykurtic curve refers to the distribution having fewer and less extreme outliers than does the normal distribution. Usually they are referred to as negative kurtosis in which the value of distribution has lesser or fewer tails or outliers than the normal distribution. The curve in platykurtic is flatter when compared with normal curve (refer to figure 14.3).

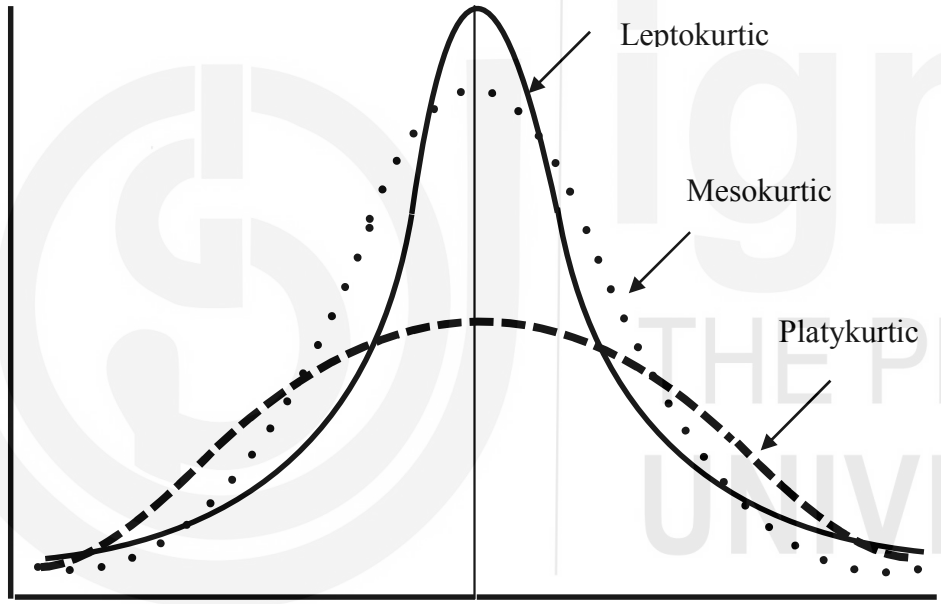

Fig. 14.3: Three Types of Kurtosis

Kurtosis can be computed with the help of the following formula:

$$K_u = Q/ P_{90} - P_{10}$$

Where,

K_u = Kurtosis

Q = Quartile Deviation

P_{90} = 90^{th} Percentile

$P1_0$ = 10^{th} Percentile

When kurtosis value is computed with the help of the above formula, the value for normal curve is obtained as 0.263. If the value obtained is below 0.263 then the curve can be termed as leptokurtic and if the value is above 0.263, then the curve can be termed as platykurtic (Mangal, 2002).

14.5.2 Skewness

As informed to you earlier that in NPC, the mean, median and mode coincide together (fall at same point) and have equal values. In a Skewed distribution, the mean, median and mode fall at different points in the distribution, and the center of gravity is shifted to one side. So, the Skewness determines the lack of symmetry in the curve and since normal probability is a symmetrical curve, it has zero skewness. Skewness can be measured in terms of Pearson's measure and percentiles. Depending upon the distribution of scores, Skewness can be classified in to two types as mentioned below.

1) **Positive Skewness**

A distribution curve is said to be positively skewed when the distribution of scores are more at the left end (refer to figure 14.4) . In a positively skewed distribution, more individuals obtian scores that are less than the mean.

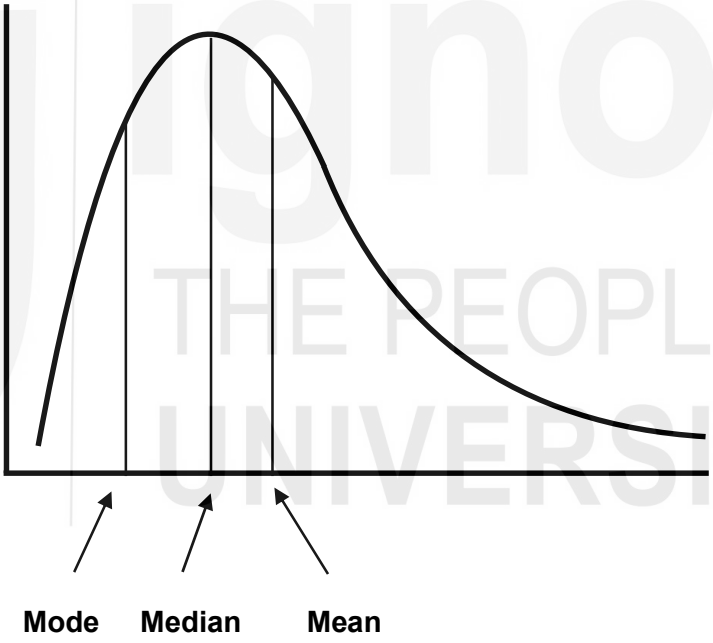

Fig. 14.4: Positive Skewness

2) **Negative Skewness**

On the other hand, if the distribution of the scores fall more towards the right side, the distribution is said to be negatively skewed (refer to figure 14.5). The median here is greater than the mean and that is why, the mean lies to the left of the median. Here, more individuals obtain scores that are higher than the mean.

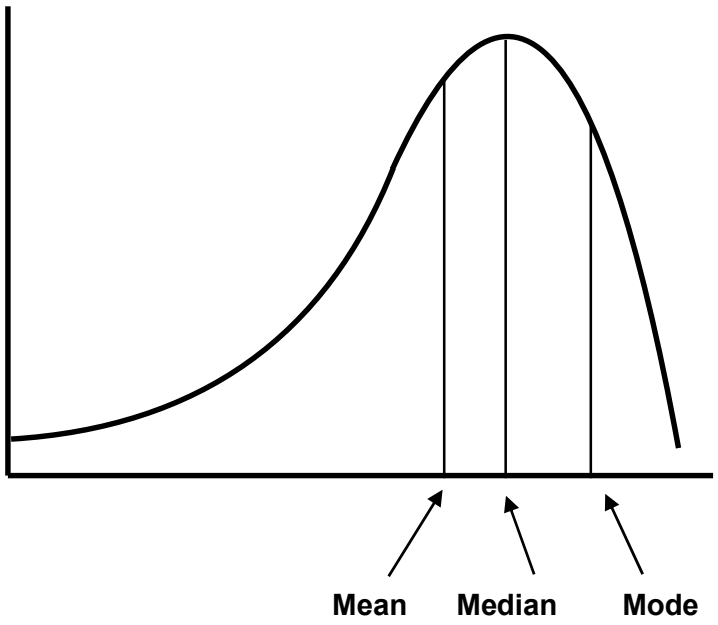

Fig. 14.5: Negative Skewness

Skewness can be computed with the help of the following formula:

$$S_k = 3 (M- M_d)/ SD$$

Where,

S_k= Skewness

M = Mean

M_d = Median

SD= Standard Deviation

There is another formula to compute skewness, that is used when the information about percentiles is available:

$$S_k = P_{90} + P_{10}/ 2 - P_{50}$$

Where,

S_k= Skewness

P_{90} = 90th Percentile

$P1_0$ = 10th Percentile

P_{50} = 50th Percentile

Check Your Progress IV

1) What is the meaning of divergence from normality?

.....

.....

.....

.....

2) What is kurtosis?

.....

.....

.....

.....

.....

3) What is skewness?

.....

.....

.....

.....

.....

4) What is the difference between kurtosis and skewness?

.....

.....

.....

.....

.....

5) What is the difference between positive skewness and negative skewness?

.....

.....

.....

.....

.....

14.6 LET US SUM UP

To sum up, in the present unit we discussed about the concept of probability. Probability refers to chance or likelihood. For example, if you say that, "it will probably be hot tomorrow" or "Probably the teacher might not come tomorrow", the sentences reveal that there are chances for the events to occur tomorrow but there is no certainty. Some of the significant concepts related to probability were also discussed like, sample space, events and power set, mutually exclusive/ disjoint events, exhaustive/ collective events, independent and dependent events, equality likely events and complementary events were also discussed. Further, the concept and nature of normal probability distribution were also highlighted with the help of a figure showing normal

curve. A normal curve is a bell shaped curve, bilaterally symmetrical and continuous frequency distribution curve. Such a curve is formed as a result of plotting frequencies of scores of a continuous variable in a large sample. The curve is known as normal probability distribution curve because its y ordinates provides relative frequencies or the probabilities instead of the observed frequencies. Standard score or z-score was also discussed in detail and the discussion covered properties, uses and computation of z-score. Lastly, the unit explained divergence from normality, where kurtosis and skewness, along with their types were discussed with help of figures.

14.7 REFERENCES

Comparison of Independent Component Analysis techniques for Acoustic Echo Cancellation during Double Talk scenario. - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/Types-of-kurtosis_fig4_275208180 [accessed 15 Oct, 2018].

A. K. Kurtz et al. (1979). Statistical Methods in Education and Psychology, New York: Springer-Verlag New York Inc.

Griffiths, D. et al. (1998). Understanding Data. Principles and Practice of Statistics. Wiley, Brisbane.

Heiman, G.W. (2001). Understanding research method and statistics. An integrated introduction for psychology. 2nd Edn.Houghton Miffin Company, Boston, NewYork.

https://faculty.elgin.edu/dkernler/statistics/ch07/7-2.html accessed on 14/11/17

http://esminfo.prenhall.com/samplechps/larson/pdfs/ch05.pdf accessed on 14/11/17

http://statistics.wikidot.com/ch7 accessed on 15/11/17

https://www.uth.tmc.edu/uth_orgs/educ_dev/oser/L1_6.HTM accessed on 4/6/18

http://influentialpoints.com/Training/z_scores-principles-properties-assumptions.htm accessed on 5/10/18

https://www.spcforexcel.com/knowledge/basic-statistics/are-skewness-and-kurtosis-useful-statistics accessed on 5/10/18

14.8 KEY WORDS

Events: A set of outcomes of an experiment. It is the sub set of the sample space.

Kurtosis: Kurtosis is a measure of "tailedness" of the probability distribution of a random variable. In other words, it is a measure whether a data is heavy tailed or light tailed in relation to normal distribution.

Normal Probability Curve: A normal curve is a bell shaped curve, bilaterally symmetrical and continuous frequency distribution curve.

Normal Probability Distribution: A continuous probability distribution for a variable is called as normal probability distribution or simply normal distribution. It is also known as Gaussian/ Gauss or LAPlace – Gauss distribution.

Probability: In statistics, the term 'Probability' refers to the expected frequency (chance) of occurrence of an event among all possible similar events.

Skewness: Skewness is a measure of asymmetry of the probability distribution of a random variable about its mean.

Standard score: Standard score or z-score is a transformed score which shows the number of standard deviation units by which the value of observation (the raw score) is above or below the mean.

14.9 ANSWERS TO CHECK YOUR PROGRESS

Check Your Progress I

1. State whether the statements are True or False		
Sr. No.	**Statements**	**True/ False**
A	Events are said to be equally likely events if each event has a fair enough chance to occur approximately the same number of times.	True
B	A probability ratio always ranges between the limit of -1.00 to +1.00.	False
C	Two events are said to be mutually exclusive if both occur simultaneously in a single trial.	False
D	Two or more events are said to be independent events if the outcome of one event does not effect as well as is not affected by the outcome of the other event.	True
E	If two events are mutually exclusive and exhaustive, they both are said to complement each other.	True

Check Your Progress II

1. Fill in the Blanks

a) NPC is a bell shaped curve which is bilaterally symmetrical.

b) In a normal probability curve, the value of mean, median and mode are equal.

c) The normal distribution is a continuous distribution and plays significant role in statistical theory and inference.

d) The area of unit under the normal curve is said to be equal to one.

e) The normal distribution is determined by two parameters, mean and variance.

Check Your Progress III

1. Fill in the Blanks

a) The mean of the z-scores is always zero.

b) z- score is used for standardising the raw data.

c) The variation of z-scores ranges from -3 standard deviations (which would fall to the far left of the normal distribution curve) up to +3 standard deviations.

d) The standard deviation of the z-scores is always one.

e) The standard score is a score that informs about the value and also where the value lies in the distribution.

Check Your Progress IV

1. What is the meaning of divergence from normality?

   When the frequency curve is more peaked or flatter than the normal probability distribution curve, the distribution is said to be diverged from normality.

2. What is kurtosis?

   Kurtosis is a measure of ”tailedness” of the probability distribution of a random variable. It is a measure of whether a data is heavy tailed or light tailed in relation to normal distribution.

3. What is skewness?

   Skewness determines the lack of symmetry in the curve and can be measured in terms of Pearson‘s measure and percentiles.

4. What is the difference between kurtosis and skewness?

   Kurtosis is a measure of ”tailedness” of the probability distribution of a random variable. In other words, it is a measure of whether a data is heavy tailed or light tailed in relation to normal distribution. On the other hand, Skewness is a measure of asymmetry of the probability distribution of a random variable about its mean.

5. What is the difference between positive skewness and negative skewness?

   A distribution curve is said to be positively skewed when the distribution of scores are more at the left end. In a positively skewed distribution the median is less than the mean which means that the mean lies to the right of the median. On the other hand, if the distribution of the scores fall more towards the right side, the distribution is said to be negatively skewed. The median here is greater than the mean and that is why, the mean lies to the left of the median.

14.10 UNIT END QUESTIONS

1) Discuss the concept and related aspects of probability.
2) Differentiate between mutual exclusive and exhaustive events.
3) Discuss the concept, nature and properties of Normal Distribution Curve.
4) Discuss the computation of z-scores and its properties.
5) Discuss the different types of divergence from normality.