BLOCK 3
STATISTICAL METHODS I

112 Blank

UNIT 7 INTRODUCTION TO STATISTICS*

Structure

7.0 OBJECTIVES

After reading this unit, you will be able to:

- explain the meaning of statistics;
- discuss the role of statistics in research;
- describe the limitations and misinterpretations of statistics;
- discuss the scales of measurement; and
- explain descriptive and inferential statistics.

7.1 INTRODUCTION

A researcher is carrying out a research on emotional intelligence and self esteem of adolescents in India. For this research, he/ she will collect the data from the adolescents (both males and females) with the help of standardised tools for emotional intelligence and self esteem. Emotional intelligence and self esteem here are the two main variables of the study. After the data collection process is over, the researcher will have to carry out statistical analysis. Based on the objectives and hypothesis(es) of his/ her research, the

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

researcher will then use varied statistical techniques to analyse the data. He/ she could use descriptive statistics or he/she may also use inferential statistics. The researcher may compute mean and standard deviation and may even graphically represent the scores. He/she may also find out the percentage of adolescents with high, moderate and low emotional intelligence and self esteem or may want to compute the mean and standard deviation for males and females with regard to the two variables. The researcher may also choose to study relationship between emotional intelligence and self esteem, or he/ she may also try to find if there exists a significant difference in emotional intelligence and self esteem with regard to gender. Thus, the researcher may choose to use varied statistical techniques based on the objectives and hypothesis(es) of his/ her research.

As it is clear from the above example, statistical methods are used in psychological research mainly to analyse data and draw inferences from them. However, before we go on to discussing various statistical techniques, we will first try to focus on the term statistics, its meaning and role.

The present unit is thus fundamental to this course and will mainly introduce the term statistics. It will also focus on certain important concepts in statistics namely, scales of measurement and descriptive and inferential statistics.

7.2 MEANING OF STATISTICS

Before we go on to develop better understanding of any subject area, we need to be clear about its basics. Keeping this in mind, in the present section of this unit, we will try to focus on the meaning of statistics and will start with defining and explaining the term statistics.

7.2.1 Definition and Nature of Statistics

What comes to your mind when the term 'Statistics' is mentioned? Well with some description in introduction section, the first thing that may come to your mind is that it is related to numbers. Some of you may also feel that it has something to do with mathematics. Others who have studied statistics before may have a better idea about the term. In the very first section of this unit, we will try to understand the meaning of the term statistics.

The term 'statistics' finds it origin in an Italian term 'Statista' that is a person who deals with State related affairs and activities. It was initially called 'state arithmetic' in which the information about the nation, for instance, tax related information and war plans, were tabulated (Aron, Aron and Coups, 2009). Thus, statistics was earlier known for its application to government related activities and data, like census. However, today it is increasingly used in various fields like economics, psychology, education, management and so on.

Statistics can be described as a branch or sub-field of mathematics that mainly deals with the organisation as well as analysis and interpretation of a group of numbers (Aron, Aron and Coups, 2009). In simple terms, statistics can be described as "the science of classifying, organising and analysing data" (King and Minium, 2008 page 3). Statistics can also be explained as science that involves use of scientific and systematic methods in order to analyse numerical data related to a phenomenon and then draw inferences and conclusions from the same. Statistics can also be defined as "a mathematical science pertaining to collection, analysis, interpretation and presentation of data" (Veeraraghavan

and Shetgovekar, (2016, page 1). It can be explained as procedures that involve not only description of data but drawing of inferences as well. In this regard, it can be mentioned that, statistics can be categorised into two main branches, descriptive statistics and inferential statistics. (These will be discussed in detail in the last section of this unit). Besides, statistics can also be categorised as parametric and nonparametric statistics, (that will be discussed in Statistical methods in psychological Research-II, that is a core course in Semester IV).

In order to understand the nature of statistics, Mohanty and Misra (2016) highlight the following points:

- Statistics can be termed as a science in which the facts related to social events are observed, recorded and computed.
- Organisation of data, its classification and analysis are the processes involved in statistics.
- Various events and phenomenon can be described, explained and compared with the help of statistics.
- A scientific enquiry can be systematically interpreted and predicted with the help of statistics. And in this regard statistics can also help in decision making.

With the above explanation, the concept of statistics must be fairly clear in your mind. But in order to understand the term further, we need to be well aware about certain basic concepts in statistics. These concepts have been described in the sub section 7.2.2.

7.2.2 Basic Concepts in Statistics

Some of the relevant basic concepts in statistics are population, sample, parameter, statistic and variable (s). These are discussed in detail as follows:

Population: The first concept relevant in statistics is population, that was discussed in unit two of this course.

Sample: Sample is yet another basic concept in statistics. This was also discussed in detail in unit two of this course.

Parameter: A parameter can be termed as a value that provides information about the population that is investigated in the research. It can be described as "a measure of the population and refers to the indices of a central value, dispersion, correlation and so on of all the individuals of the population" (Mohanty and Misra, 2016, page 3). For example, if a researcher wants to know mean weight of newly born infants in India in a given year, this can be termed as a parameter as it describes the weight of all the newly born infants in India in a given year. An exact parameter is not always easy to obtain and any parameter will have a statistic.

Statistic: As aspects of a population are measured by a parameter, aspects of a sample are measured by statistic. Thus, the researcher will measure the weight of say 500 newly born infants (a sample representing all the newborn infants) in a given year and work out a mean weight. This mean weight can be termed as a statistic.

The symbols of mean, standard deviation and variance vary for parameter and statistic, these are given in table 7.1.

Table 7.1: Symbols for Parameter and Statistic

Measure	Parameter	Statistic
Mean	μ ('mu')	$\bar{x}$ ('x-bar')
Standard Deviation	σ ('sigma')	s
Variance	σ^2 ('sigma-squared')	s^2 ("s-squared')

Variable(s): Besides the above, yet another important term that we need to discuss is variable. We discussed about variable(s) in unit 1 of this course.

Check Your Progress I

1) Define Statistics

.....

.....

.....

.....

.....

2) Identify the symbols

Symbols	Measure
μ	
σ	
σ^2	
$\bar{x}$	
s	
s^2	

7.3 ROLE OF STATISTICS IN RESEARCH

Statistics as a subject area has vast scope and application. It finds its application in fields like policy planning, management, education, marketing,

agriculture, medicine and so on, though, one of its major application is in research. Thus, our discussion in this regard will mainly focus on psychological research. But before we highlight the role of statistics in psychological research, we will try to understand the concept of research, especially in the context of Psychology.

Research in simple terms can be explained as adding to the existing fund of knowledge. The term research is derived from the French word '*recherche'* which means to travel through or survey.

Kerlinger (1995, page 10) defines scientific research as "a systematic, controlled, empirical and critical investigation of natural phenomenon guided by theory and hypotheses about the presumed relations among such phenomena".

Best and Kahn (1999) defined research as an analysis and recording of observations that is carried out in systematic and objective manner. And this analysis and recording will lead not only to generalisation but also to development of theories and predictions. Research is carried for various reasons like, investigating relationships, measuring entities, making predictions, to test hypothesis(es), make comparisons and draw conclusions about the population.

Some of the main components of research include the statement of problem, hypothesis(es), sample, research design, data collection and data analysis. These have been briefly discussed as follows:

Problem: Problem can be described as a general objective of the study. Thus, if a researcher wants to study relationship between perceived parental behaviour and self concept of adolescents, then the statement of problem will be 'To study the relationship between perceived parental behaviour and self concept of adolescents'. The statement of problem provides information about the general focus of the study. Further, there could also be specific objectives based on the statement of the problem.

Hypothesis (plural: hypotheses): Based on the statement of the problem, hypothesis(es) can be formulated. These are tentative statements that are tested with the help of scientific research. Hypothesis can be null or alternative hypothesis (these have been discussed under inferential statistics).

Sample: Any study is carried out on a sample. The nature and size of the sample will depend on the nature and purpose of the study (sample was discussed in detail under key concepts in statistics). Also, based on the requirement of the study, either probability or non-probability sampling techniques are used to derive the sample.

Research Design: Any research will also have a research design that provides information about the outline and structure of the research. Research design is important in order to not only seek solution to the research problem but also to control any variance. Thus, research design can be explained as means to allow a researcher to seek answers to research problems in an objective, valid and accurate manner, keeping in mind the economical aspect (Kerlinger, 1995). There are various research designs like experimental designs, non-experimental designs, quasi-experimental designs, factorial designs, small n designs to name a few, that can be used by the researcher while carrying out research.

Collection of data: The next component of research is collection of data. Data can be collected with the help of standardised psychological tests, interview method, observation, questionnaire and so on. Various methods can be used to collect data from the sample based on the objectives of the study.

Data analysis: Once the data collection process is over, the data can then be subjected to data analysis, qualitatively or quantitatively (or both). In the present course, we will learn about some basic statistics techniques that can be used to analyse data.

Conclusions and generalisation: Based on the results obtained in data analysis, conclusions are drawn and then the researcher is in position to generalise the results to the population.

The above discussed components are relevant to understand before we move on to understanding the role of statistics in psychological research.

Statistics plays an important role during various stages of research. For instance, while drawing a sample from population for research, statistics can be adequately used. Sample size for a research can be determined with the help of statistics. Certain formula can be used to compute sample size. Further, in test development, statistics can be used in order to ascertain the reliability and validity of the test. Techniques like factor analysis can be used effectively for reduction of data, that also finds application in the development of psychological tests. Normal distribution can be used in development of norms. Thus, statistics can play an important role in the test development process.

Statistics plays an extremely important role in analysing quantitative data collected by a researcher. The data can be organised, classified and analysed using various statistical techniques so as to draw inferences and conclusions and help in decision making. The results thus obtained can be meaningfully summarised and conclusions can be drawn and predictions can be made from the same. Both descriptive and inferential statistics can be used to analyse the data. With the help of statistics, the probability of errors while drawing inferences, can also be determined thus enhancing the degree of precision.With regard to descriptive statistics, the raw data can be classified and tabulated and then measures of central tendency and measures of variability can be used based on the objectives of the research. The data can also be graphically represented for effective presentation and easy understanding. With regard to inferential statistics, two or more sample sub groups can be compared. Further, statistics can also play a role when a researcher wants to predict one or more variable from other variable(s).

Statistics can also be categorised in to parametric and nonparametric statistics based on whether certain assumptions are met. These techniques can be effectively used in varied conditions. For instance, parametric statistics has certain requirement like, the data should be normally distributed, sample needs to be homogeneous, the variables are to be measured with interval or ratio scale and so on. Nonparametric statistics can be effectively used when a sample is heterogenous in nature, data is not normally distributed, has outliers and the variables are measured with nominal or ordinal scale and so on.

Statistics can also be univariate, bivariate or multivariate. Univariate is where there is only one variable, bivariate denotes two variables and multivariate indicates many variables. Thus, based on the objectives and nature of the

research, varied statistical techniques that could be as simple as computing mean and mode to more complex techniques like factor analysis, discriminant analysis and so on can be used.

Thus, statistics has a major and significant role in research. In the succeeding units and the units included in the course on Statistical methods for Psychological Research- II (that you will study in forth semester), you will study varied statistical techniques.

Check Your Progress II

1) Define research.

.....

.....

.....

.....

.....

.....

2) Explain the role of statistics in test development.

.....

.....

.....

.....

.....

7.4 LIMITATIONS AND MISINTERPRETATIONS OF STATISTICS

Some of limitations and misinterpretations of statistics are as follows:

1) Statistics cannot be used with single observation. To compute statistics we need a group of data or observations. For just a single observation, statistics cannot be applied.

2) Events or phenomenon that are qualitative in nature cannot be subjected to statistics. Statistics is applicable to events and phenomenon that can be measured in terms of numbers.

3) Inferences based on statistics cannot be exact as the inferences that are drawn are based on mathematical laws. Statistical laws are based on majority of the observations and may not be applicable to each and every individual.

4) In order to adequately interpret the results of statistics, knowledge about statistics is required, especially with regard to when to use what technique and how to interpret the results obtained.

5) Statistics as such has no control over the data collection process. The results obtained will provide no indication of any dishonesty or bias in data collection. Thus, it is prone to misuse and much depends in this regard on the researcher rather than statistics.

6) Statistics may not provide a complete picture about a certain phenomenon or event. There are number of factors that can have an impact on a certain phenomenon, but statistics will be able to measure only those factors that are quantitatively expressed.

7) The results obtained are also prone to be misinterpreted especially by untrained persons who lack knowledge about the statistical techniques, their computation and interpretation.

8) There is possibility of errors in statistical decisions.

Check Your Progress III

1) List any two limitations of Statistics.

.....

.....

.....

.....

.....

7.5 SCALES OF MEASUREMENT

Measurement is a process that involves assigning numbers to observations in a meaningful manner. Variations can exist in the properties of the quantification of the observations. For example, 1 kilogram of wheat is half of 2 kilograms of wheat (here wheat is measured in terms of weight). Whereas, ranks can be assigned to students based on their performance in mathematics. For example, a student who has achieved 1st rank may have obtained 95 marks, whereas a student obtaining 2nd rank may have 80 marks and a student obtaining 3rd rank may have obtained 79 marks. As can be seen, the numerical properties in both the examples are different.

In 1946 four scales of measurement were explained by S. S Stevens that can be used to measure variables (Aron, Aron and Coups, 2009). These four scales of measurement are described as follows:

1) **Nominal Scale:** Nominal scale can be used to measure variables that are qualitative as well as exclusive in nature. For example, gender, religion and so on. The term nominal is derived from latin term 'nominalis' that relates to name. Though such variables are qualitative in nature, numbers can be assigned to these variables. For example, with regard to gender, males can be assigned the number 1 and females can be assigned the number 2 or vice versa. Similarly with regard to religion, Christians can be assigned the number 1, Hindus can be assigned the number 2, Jains can be assigned the number 3, Muslims can be assigned the number 4 and Any other (belonging to any other religion besides the ones mentioned) can be assigned the number 5. These numbers in themselves have no

meaning and are purely nominal. A higher number does not indicate a higher weightage. They are mainly for the sake of identification and do not imply that a certain category is better or worse than other (s). Thus, such numbers cannot be subjected to any mathematical calculation. For example, in sports, where teams are involved, like cricket or football, the team members have numbers on their jersey that is merely for the sake of identification and does not provide information whether one player is better than the other(s).

2) **Ordinal Scale:** Ordinal scale involves ranks, that is, the data can be assigned ranks based on whether they are less or more, low or high, bad or good and so on. The data is thus ranked in terms of its magnitude. The term ordinal is derived from Latin term 'ordinalis' which indicates order. For example, based on the performance of students in mathematics, they can be ranked. Thus, a student who secures first rank has performed better than a student who has secured second rank and a student who has secured tenth rank has performed much lower when compared to the students with first and second ranks. As is with nominal scale, even in ordinal scale, the numbers cannot be subjected to any mathematical calculations. Further, in ordinal scale there is no idea about the degree of difference between the two ranks. For example, a student who has secured 75 marks in mathematics may secure 1st rank, a student who has obtained 65 may secure second rank, whereas a student who has obtained 64 marks may secure third rank. As can be see in this example, the degree of difference between the marks obtained by the 1st ranker and the second ranker is more as compared to the degree of difference between second ranker and the third ranker.

3) **Interval Scale:** Interval scale is most commonly used to measure psychological variables. These scales are similar to the ordinal scale as the categories can be ranked and are exclusive as well, but the degree of difference between two participants is same. For example, the degree of difference between individuals obtaining a score of 22 and another individual obtaining a score of 23 is same as that of an individual obtaining a score of 34 and another individual obtaining a score of 35. In interval scale there is no absolute zero, for example, their cannot be a person with zero attitude. Interval scale can be subjected to mathematical calculations.

4) **Ratio Scale:** Ratio scale has all the properties of all the scales, nominal, ordinal and interval. Besides, it also has an absolute zero, that indicates presence or absence of certain property or characteristics. Ratio scale displays equidistance between the adjacent categories. For instance, the difference between one kilogram of wheat and two kilograms of wheat and difference between five kilograms of wheat and six kilograms of wheat is same. Further 10 kilograms of wheat is half of 20 kilograms of wheat. Also zero kilogram indicates no wheat. Various mathematical calculations can be carried out with the help of ratio scale.

Refer to table 7.2 for properties and examples of the four scales of measurement

Table 7.2: Properties and Examples of Scales of Measurement				
Properties	**Nominal**	**Ordinal**	**Interval**	**Ratio**
Categories are exclusive	✓	✓	✓	✓
Categories can be arranged in an order		✓	✓	✓
Equidistance between the adjacent categories			✓	✓
Real zero				✓
Examples	Roll numbers assigned to students	Ranks obtained by students in Psychology class test	Scores obtained by individuals on an attitude scale	Errors made by Individuals on a memory test

Check Your Progress IV

1) Explain interval and ratio scales.

.....

.....

.....

.....

2) Provide examples for nominal and ordinal scales.

.....

.....

.....

.....

.....

7.6 DESCRIPTIVE AND INFERENTIAL STATISTICS

Statistics can be categorised in to descriptive and inferential statistics. In the present section of this unit, we will explain these terms in detail.

7.6.1 Descriptive Statistics

Let us try and understand descriptive statistics with the help of an example. A teacher administers a test on English writing skills of 100 marks to her students. As she receives the scores of the test, she comes to know that the average marks received by the class is 65. She also came to know that 10 percent of the students needed help with regard to English writing skills. One of her students, Tina performed very well in the test and obtained score that was better than 85% of the students in her class. From this example, it can be seen that some of the statistical techniques that this teacher used were mean or average, percentage and percentile. These and many other techniques can be categorised under the term descriptive statistics.

Descriptive statistics mainly comprises of description and organisation of the data. It can be termed as a technique that helps in summarisation of prominent characteristics of a distribution.

Based on the properties of the sample, the descriptive statistics can be categorised in to the following (Mohanty and Misra, 2016, page 7):

- **Statistics of location:** Covers techniques like measures of central tendency including mean, median and mode, frequency distribution, percentiles and so on.
- **Statistics of dispersion:** Covers techniques related to measures of dispersion including quartile deviation, standard deviation, range, average deviation and variance.
- **Statistics of correlation:** Includes coefficients of correlation like Pearson's product moment correlation, Spearman's rank order correlation and Kendall's rank correlation. Correlation mainly helps us understand the relationship between variables.

In the present course the main focus will be on descriptive statistics and the topics mentioned above will be covered in the subsequent units.

7.6.2 Inferential Statistics

Let us start our discussion with an example. A researcher was carrying out a study on emotional intelligence and self concept of adolescents in South Delhi. She selected a representative sample (N=500) from various schools in South Delhi and administered standardised tools for emotional intelligence and self concept on the adolescents. The researcher was interested in finding out if significant difference exists between the mean scores obtained by male and female adolescents with regard to both the variables. For this she used Independent t-test and found that there was a significant difference with regard to emotional intelligence. The mean scores obtained by the female students on emotional intelligence was higher than the mean scores obtained by the male students. Thus, indicating that the females had higher emotional intelligence than males. However, no significant difference was found between male and female adolescents with regard to self concept. The researcher also wanted to know if significant difference exists in emotional intelligence and self concept with regard to the phases of adolescents (early, middle and late). For this, Analysis of Variance (ANOVA) was used and the results indicated that no significant difference exists with regard to either of the variables. The independent t-test and ANOVA are techniques that can be categories under

inferential statistics (the techniques that fall under inferential statistics, which will be covered in detail in the course on Statistical Methods for Psychological Research- II that will be offered in the semester IV).

In inferential statistics, inferences are drawn about the population based on a representative sample. As stated by Veeraraghavan and Shetgovekar (2016, page 5) "Inferential statistics refers to the mathematical methods based on probability theory and helps in reasoning and inferring the characteristic features of the sample drawn from the larger population". Inferential statistics can also be effectively used to make estimations and predictions.

There are two types of procedures under inferential statistics, namely estimation and hypothesis testing. These two are discussed in details as follows:

7.6.2.1 Estimation

Estimating probability of a phenomenon is referred to as estimation (Veeraraghavan and Shetgovekar, 2016). As we know from the explanation of inferential statistics, that inferences are drawn from sample that is representative of a population and these inferences can then be generalised to the whole population. In these inferences, the researcher will make an estimation that needs to be close to the actual or true population value.

There are two types of estimation: point estimation and interval estimation.

Point estimation: This is a type of estimation in which the value is a single point. For example, the estimation for sample mean is made as 46.8 that is expected to be equal to the population mean. Point estimate comprises of sample mean and sample proportion. The population mean is 'μ', the sample mean will be '$\bar{x}$'. In similar manner, if the population proportion is '*P'* then sample proportion will be '*p'*.

Interval estimation: An interval estimate is an interval or two numbers within which the population parameter could lie. Thus, for population mean 'μ', the interval estimate will be a<x< b. The interval estimate is greater than a but lesser than b. For example, an interval estimate could be 45- 47 within which it is expected that the population mean will lie. As the researcher has an interval, he/ she is thus able to trust that the estimate is close to the population value with 95% or 99% level of confidence. Interval estimate comprises of confidence interval for mean and confidence interval for proportions.

While estimations are made there could be fluctuations and these could be due to varied reasons including chance factors and sampling error.

The inferences drawn by the researcher needs to be free of any chance factors. For example, a researcher is studying if there exists significant difference in job satisfaction of government and private bank employees. After carrying out data collection and data analysis, he/she obtains results that such a difference does exist, then such results should not be as a result of chance factors. If such a difference falls within the range ± 1.96, then the significant difference can be said to be real and not due to chance factors.

Fluctuations can also be as a result of sampling error that occur when the sample selected by the researcher is not representative of the population being studied. A sample that is not representative of the population will not possess

the same characteristics as the population and thus the results obtained from such a sample cannot be used to draw inferences for the population. Sampling errors can be avoided by being careful while selecting a sample and also by having a larger sample.

7.6.2.2 Hypothesis Testing

Besides estimation, international statistics includes hypothesis testing. We discussed about hypothesis testing in unit one of this course.

Check Your Progress V

1) List the three categories of descriptive statistics.

.....

.....

.....

.....

7.7 LET US SUM UP

To summarise, in the present unit, we mainly focused on the term Statistics. Statistics can be described as a branch or sub field of mathematics that mainly deals with the organisation as well as analysis and interpretation of a group of numbers. The term 'statistics' finds it origin on an Italian term 'Statista' that is a person who deals with State related affairs and activities. Further, the key concepts in statistics, namely, population, sample, parameter, statistics and variable(s) were also discussed. Further, the role of statistics in research was also discussed with its application from sample selection to data analysis. The limitations and misinterpretations of statistics were also discussed. The four scales of measurement, namely, nominal scale, ordinal scale, interval scale and ratio scale were discussed with examples. The last topic covered in this unit was descriptive and inferential statistics. Descriptive statistics mainly comprises description and organisation of the data. It can be termed as a technique that helps in summarisation of prominent characteristics of a distribution. In inferential statistics, inferences are drawn about the population based on a representative sample. Under inferential statistics, subtopics related to estimation and hypotheses testing were discussed.

7.8 REFERENCES

Aron and Aron (2009). Statistics for Psychology (5th ed). New Delhi: Pearson
2. Howell, D. (2009).Statistical Methods for Psychology (7th ed.). Wadsworth.

Best, J. W and Kahn, J. V. (1999). Research in Education. New Delhi: Prentice Hall of India Pvt. Ltd. for information on research designs.

Kerlinger, Fred, N. (1995). Foundations of Behavioural Research. Bangalore: Prism Books Pvt. Ltd. for information on research, research designs, types of research and methods of data collection.

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.

7.9 KEY WORDS

Interval scale: Interval scale is most commonly used to measure psychological variables. This scale is similar to the ordinal scale as the categories can be ranked and are exclusive as well, but the degree of difference between two participants is same.

Nominal Scale: Nominal scale can be used to measure variables that are qualitative as well as exclusive in nature.

Ordinal Scale: Ordinal scale involves ranks, that is, the data can be assigned ranks based on whether they are less or more, low or high, bad or good and so on. The data is thus ranked in terms of its magnitude.

Parameter: Parameter can be described as "a measure of the population and refers to the indices of a central value, dispersion, correlation and so on of all the individuals of the population" (Mohanty and Misra, 2016, page 3).

Ratio Scale: Ratio scale has all the properties of all the scales, nominal, ordinal and interval. Besides, it also has an absolute zero, that indicates presence or absence of certain property or characteristics.

Statistics: Statistics can be described as a branch or sub field of mathematics that mainly deals with the organisation as well as analysis and interpretation of a group of numbers

7.10 ANSWERS TO CHECK YOUR PROGRESS

Check Your Progress I

1) Define Statistics

Statistics can be described as a branch or sub field of mathematics that mainly deals with the organisation as well as analysis and interpretation of a group of numbers.

2) Identity the Symbols

Symbols	Measure
μ	Parameter Mean
σ	Parameter Standard Deviation
σ^2	Parameter Variance
$\bar{x}$	Statistic Mean
s	Statistic Standard Deviation
s^2	Statistic Variance

Check Your Progress II

1) Define research.

Research in simple terms can be explained as adding to the existing fund of knowledge. The term research is derived from the French word '*recherche'* which means to travel through or survey.

Kerlinger (1995, page 10) defines scientific research as "a systematic, controlled, empirical and critical investigation of natural phenomenon guided by theory and hypotheses about the presumed relations among such phenomena".

2) Explain the role of statistics in test development.

In test development statistics can be used in order to ascertain the reliability and validity of the test. Technique like factor analysis can be effectively used for reduction of data, that also find application in development of psychological tests. Normal distribution can be used in development of norms. Thus, statistics can play an important role in the test development process.

Check Your Progress III

1) List any two limitations of Statistics.

- Events or phenomenon that are qualitative in nature cannot be subjected to statistics. Statistics is applicable to events and phenomenon that can be measured in terms of numbers.
- There are number of factors that can have an impact on a certain phenomenon, but statistics will be able to measure only those that are quantitatively expressed.

Check Your Progress IV

1) Explain interval and ratio scale.

Interval scale is most commonly used to measure psychological variables. This scale is similar to the ordinal scale as the categories can be ranked and are exclusive as well, but the degree of difference between two participants is same. Ratio scale has all the properties of all the scales, nominal, ordinal and interval scale, but also has an absolute zero, that indicates presence or absence of certain property or characteristics. Ratio scale displays equidistance between the adjacent categories.

2) Provide examples for nominal and ordinal scales.

Example of nominal scale would be the jersey numbers of football players and example of ordinal scale could be the ranks obtained by students in an examination.

Check Your Progress V

1) List the three categories of descriptive statistics.

The three categories of descriptive statistics are:

- Statistics of location that includes techniques like measures of central tendency including mean, median and mode, frequency distribution, percentiles and so on.

- Statistics of dispersion, that includes techniques related to measures of dispersion including quartile deviation, standard deviation, range, average deviation and variance.
- Statistics of correlation that includes coefficients of correlation like Pearson's product moment correlation, Spearman's rank order correlation and Kendall's rank correlation. Correlation mainly helps us understand the relationship between variables.

7.11 UNIT END QUESTIONS

1) Explain the key concepts in Statistics
2) Describe the role of statistics in Research.
3) Describe the scales of measurement with suitable examples.
4) Elucidate descriptive statistics.

UNIT 8 DATA ORGANISATION AND GRAPHICAL REPRESENTATION*

Structure

8.0 OBJECTIVES

After reading this unit, you will be able to:

- discuss the classification and tabulation of statistical data;
- describe the steps in construction of a frequency distribution ;
- create a cumulative frequency distribution table;
- explain the meaning of percentile and percentile ranks; and
- discuss the graphical representation of data.

8.1 INTRODUCTION

The objective of all statistical inquiry is to describe and understand the population of interest. For example, in an exit poll survey, a news channel wants to assess the political attitude of the voters, how they are going to vote in

* Dr. Vijay Viegas, Assistant Professor, Abbé Faria Post Graduate Department of Psychology, St. Xavier's College, Goa

the upcoming election, and what are the chances of current Government to come back in power again? This information about the population of interest can be gained from a number of statistical enquiries. Exit poll surveys provide tentative information about which party will gain what percentage of votes in which state of India and so on. Such exit poll surveys make use of basic statistical techniques that can be categorised under descriptive statistics.

In the previous unit we mainly discussed about the term statistics, its definition, nature and also key terms. We also discussed about scales of measurement and the two main categories of statistics, namely descriptive and inferential statistics. In the present unit, we will mainly focus on the varied aspects of descriptive statistics, viz, classification, tabulation, organisation and graphical representation of data. One of the most basic yet important method known as frequency distribution will also be discussed in this unit. Further, we will also discuss the method of cumulative frequency distribution, percentile, percentile rank and graphical representation of data.

8.2 CLASSIFICATION AND TABULATION OF QUALITATIVE AND QUANTITATIVE DATA

Any data can be qualitative or quantitative in nature. Qualitative data are measures of types and are denoted by a name, symbol, or a number code. They are types of information that have features that can not be measured. In simple words, qualitative data are data about categorical variables. Some examples of qualitative data are the smoothness of your skin, and the colour of your eyes, the texture of your hair, the softness of your palm etc.

Whereas, quantitative data states information about quantities, that is, information that can be measured and written down with numbers. In other words, quantitative classification refers to the classification of data according to some characteristics that can be measured. Examples of quantitative data are weight, height, shoe size, and the length of fingernails, income, sales, profits, production etc.

In descriptive statistics, classification and tabulation of data, whether qualitative or quantitative, are two important functions that will help the researcher in organising the data in a better manner so that further statistical analysis (whether by computing measures of central tendency, measures of variability or inferential statistics) can be carried out.

In this context, we also need to explain the term univariate analysis. The term univariate implies that there is only one variable. And when statistical analysis is to be carried out with just one variable, descriptive statistics are used. For example, if a researcher wants to study achievement motivation of students in class tenth, the data obtained (with the help of a standardised psychological test) cannot be subjected to inferential statistics or higher level statistical techniques. The researcher will be able to classify and tabulate the data based on the students who secured higher, lower or moderate scores. He/ she may further be able to compute mean (that will be discussed in the unit on measures of central tendency) and standard deviation (that will be discussed in the unit on measures of variability).

Thus, in the context of univariate analysis, we mainly focus on the use of descriptive statistics. In the present section we will discuss classification and tabulation of data.

8.2.1 Classification

Data classification is a method of organising data into groups for its most effective and efficient use. A well-planned data classification system makes vital data easy to find and retrieve whenever required. In other words, the process of ordering data into homogenous groups or classes according to some common characteristics present in the data is called classification. For example, it is a common exercise that during the process of sorting letters in a post office, the letters are classified according to the cities and further arranged according to streets and other details, so that it becomes easier to deliver the letters to its destination.

In the context of research, the data collected by a researcher is arranged in formats that will help him/ her draw conclusions. Basically, classification involves sorting the data based on similarities. Once the data is classified, the researcher can proceed with further statistical analysis and decision making. Some of the main objectives of classification are as follows:

1) The data is presented in a concise form. A raw data as such has no meaning. But once it is classified, it will reflect some meaning.

2) Classification helps in identifying the similarities and diversities in the data. For example, based on the marks obtained in an English test, students can be grouped in to those obtaining 76-100, those obtaining marks between 51-75, those obtaining marks between 26-50 and those obtaining marks between 1-25. Each of these groups are distinct from each other in terms of marks obtained, but are grouped because of similarity of marks obtained by them (refer to table 8.1).

Table 8.1: Marks obtained by Students

Marks Obtained	Students
76- 100	28
51-75	40
26-50	12
1-25	20

3) Classification also helps in comparisons. The groups can be compared with each other and conclusions can be drawn. Computation of percentage will tell us the percentage of students falling in each of the four groups, mentioned in the above example.

4) Classification can be carried out for both qualitative as well as quantitative data. Individuals can be classified on the basis of colour of their hair or gender, that would be qualitative data. And individuals can also be categorised based on quantitative data, for example, their income, their age and so on.

One way in which quantitative data can be adequately classified is with the help of frequency distribution, that will be discussed in detail later in this unit.

8.2.2 Tabulation

Tabulation is the process of insertion of classified data into tabular form. A table is a symmetric arrangement of statistical data in rows and columns. Rows are horizontal arrangements, whereas, columns are vertical arrangements. It may be simple, double or complex depending upon the type of classification used for various purposes at any given time by an individual.

Tables are an important aspect of any research report or thesis. Any table will have some key components that are discussed as follows:

1) **Table number:** Any table needs to have a table number. In various units of this course, you will notice that all the tables are numbered. This mainly helps in identification of the table as well as provides a reference. So if you are asked to refer to say table 8.2, you know exactly where to look for it in this unit. Table numbers need to be provided in a systematic manner and in serial order, especially if you have included more than one table in your report or thesis.

2) **Title for the table:** Besides table number, a table should also have a title that should be specific in nature and should in short reflect what the table is about. Such a title also needs to be clear and self explanatory and should instantly help the reader gauge what the table is about.

3) **Captions and stubs:** Any table will then have rows and columns based on its contents. The headings given for columns are termed as captions. Whereas, stubs are the heading that are given to the rows. These again need to be concise and self explanatory. The captions and stubs will be decided by the researcher based on the research he/ she is carrying out.

4) **Body of the table:** Body of the table is the main part of the table that reflects the numerical information that is collected based on the data collection. The numerical data here will be classified based on the captions and stubs.

5) **Headnote:** Tables also have headnotes which could be written in extreme right below the title and these provide information about units of measurement.

6) **Footnote:** These are written below the table and may display crucial information about the information given in the captions and stubs.

7) **Source of data:** The source of data can then be mentioned below the table.

A table thus prepared is give below:

Table 8.2: Percentage of male and female students based on marks obtained by them in English test

Marks Obtained in English (Stub Head)	Gender (Caption head)	
	Males (N= 50) (Caption)	Females (N= 50) (Caption)
76 -100 (Stub)	20%	21%
51-75 (Stub)	12%	13%
26 to 50 (Stub)	40%	39%
1-25 (Stub)	28%	27%
Total	100%	100%

Footnote: Number of students is in terms of percentage (%).
Source: Data collected from the Term End Examination results

As discussed above classification and tabulation are significant in organising the data. Some of the merits of classification and tabulation are as follows:

1) **Clarifies the data:** The information arranged in the form of table is easily accessible and provides adequate and clear information to the user of the data.

2) **Simplification:** Classification and tabulation of data reduces the mass that is, the size of the data and present the data in simplest possible way. When the data is presented in the tables and classified, all the complexities are removed and the data is made very simple and clear for the user.

3) **Facilitates comparisons:** It enables quick comparison of the statistical data shown in rows and columns.

4) **Information can be easily referred:** When an information is tabulated, it is very easy to refer to.

Check Your Progress I

1) What is quantitative data?

.....

.....

.....

.....

.....

.....

.....

2) List the merits of classification and tabulation.

.....

.....

.....

.....

8.3 CONSTRUCTION OF FREQUENCY DISTRIBUTION

Earlier in this unit we discussed about classification and tabulation of data. And frequency distribution is a way in which raw data can be classified so as to provide a clear understanding of the data. Frequency distribution is a tabular representation, in which the raw data is organised in to class intervals.

Frequency distribution can be categorised in to three types:

1) **Relative frequency distribution:** Such a distribution denotes that the score that is allotted for each class interval is the proportion of total number of cases in a distribution. For example, in a frequency distribution of 100 employees based on years of experience, 35 employees fall in the range (class interval) 10-14 years of experience, then the relative frequency distribution will be 35/100= 0.35. Thus, it can be said that 35% of the employees fall in this class interval.

2) **Cumulative Frequency Distribution:** Such a distribution for a certain class interval is summation of the frequencies of that class interval and of the class interval below that class interval. This will be discussed in detail in the next section of this unit.

3) **Cumulative Relative Frequency Distribution:** In such a distribution, the cumulative relative frequency for a particular score is the relative frequency for that score in summation with the relative frequencies of all the scores that lie before this particular score. This will be clear from table 8.3, that provides examples of the three types of frequency distribution.

Table 8.3: Examples of relative frequency, cumulative frequency and cumulative relative frequency distributions

Scores	Frequency	Relative Frequencies	Cumulative Frequency	Cumulative Relative Frequency
34	3	10%	30	100%
23	4	13.33%	27	89.99%
22	10	33.33%	23	76.66%
21	6	20%	13	43.33%
19	7	23.33%	7	23.33%
	N= 30			

In frequency distribution there are two main methods to describe class interval.

1) **The exclusive method:** In this method, the upper limit of a certain class interval is the lower limit of the class interval next to it, thus there is a continuity between the class intervals. The score that equals the upper limit of a class interval is exclusive in the sense that it will fall in the class interval where the score is its lower limit. Thus, in exclusive method the score equal to upper limit is not included in that class interval, but a score equal to its lower limit is included in it. For example, in a distribution with class intervals using exclusive method, a score 20 will fall in class interval 20- 30 and not in 10- 20 class interval.

2) **The inclusive method:** In inclusive method there is no continuity between the class intervals and this method is especially for discrete scores. In this method, scores equal to both lower and upper limit are included in the class interval. For example, the class intervals will be 1-5, 6-10, 11- 15 and so on.

Frequency distribution can also be categorised in to ungrouped or grouped frequency distribution.

Ungrouped Frequency Distribution: An ungrouped frequency distribution is the one in which all the values are listed in an ascending or a descending order. Based on the frequency of occurrence of each score, a tally mark (/) is placed in front of the respective value and frequency (denoted by '*f*') of each score is stated in the next column. The example of ungrouped frequency distribution is given in table 8.4:

Table 8.4: Ungrouped frequency distribution

Values	Tallies	*f*
6	///	3
9	////	4
12	~~////~~	5
23	/	1
24	//	2

Grouped Frequency Distribution: Sometimes the data is too large and it is not possible to have a frequency distribution in an ungrouped form, as then the researcher will not be able to get a clear picture. In such cases a grouped frequency distribution can be used. Here the data are organised in to classes or class interval and then a tally mark is placed based on which class interval a given score falls in and then the frequency is denoted. The example is given in table 8.5.

Table 8.5: Grouped frequency distribution		
Values	**Tallies**	***f***
1-5	///	3
6-10	~~////~~	5
11-15	//	2
16-20	/	1
21-25	/	1

The concept of grouped and ungrouped frequency distribution must be clear from the above examples. We will now discuss computation of frequency distribution with the help of an example.

Suppose, in a class of forty students, following marks were obtained on a test of ten marks. The marks obtained by the forty students are given as follows:

3	8	6	5	6	4	7	6
5	3	5	6	3	5	4	4
3	6	7	8	1	10	7	6
4	5	0	7	6	5	6	7
1	7	5	4	5	8	5	7

These numbers (marks of the students) are called as *raw data*, as they are obtained from the field directly and haven't gone through any statistical analysis. Now the question is, what these numbers or raw data suggest about the target population of students? Which marks are most common? How many students got highest marks? How many students passed this test? With raw data, though, it is not possible to draw any conclusion. Thus, we need to create a frequency distribution on the basis of the raw scores. Frequency can be calculated for each of the obtained score by the students.

Frequency is the number of times a particular variable/ individual or observation (obtained marks in our context) occurs in raw data. The distribution of a variable is the pattern of frequencies of the observation. Frequency distributions are portrayed as frequency tables, histograms, or polygons. It is just the arrangement of scores and the frequency of occurrence within a group. A frequency distribution table is one way you can organise data so that it makes more sense to the reader.

As discussed earlier, there are two major types of frequency distribution, frequency distribution and ungrouped frequency distribution. The computation for both these frequency distributions are discussed as follows:

8.3.1 Computation of Ungrouped Frequency Distribution

To calculate frequency we are going to use Tally Score Method – "This method consists of making a stroke in the proper class for each observation and summing these for each class to obtain the frequency. It is customary for convenience in counting to place each fifth stroke through the preceding four . . ." (Lawal, 2014, page 13). The frequency can be tabulated as follows (based on example of marks obtained by forty students:

Table 8.6: Frequency distribution using tally method

Marks	Tallies	Frequency (*f*)
0	/	1
1	//	2
2		0
3	////	4
4	~~////~~	5
5	~~////~~ ////	9
6	~~////~~ ///	8
7	~~////~~ //	7
8	///	3
9		0
10	/	1
		∑ = 40

Please note that the total (∑) should be equal to the number of students, that is, 40. Now, we can conclude following information from frequency table:

- Only one student got full marks.
- Most common marks is five followed by six.
- Only one student scored zero on the test.

The steps involved in creating an ungrouped frequency distribution are as follows:

Step 1: Arrange your raw data in an array-ascending or descending order.

Step 2: Make a table with three columns and name them as variable (that is, marks in the case of the present example), tallies and frequency.

Step 3: Enter your variables (marks in case of this example) in the first column from lowest to highest order.

Step 4: Now, go one by one, through your raw data and make a mark (/) for each variable next to its value in the second column of your table.

Step 5: Count the tally marks for each variable and write its total in third column, that is, frequency column.

8.3.2 Computation of Grouped Frequency Distribution

One disadvantage of the ungrouped frequency distribution method is that it will be tiresome and difficult to make a table for larger values or observations. Suppose, in the above example of class test if the number of students were 250, then would it be convenient to make an ungrouped frequency distribution table for such data? Probably no! Then what can we do? We can use another statistical procedure called as grouped frequency distribution method.

To understand this method, let us take another example. Suppose, you have the scores obtained by students on class test in History:

12	7	13	14	12	23	21	14	13	23
30	12	**1**	21	23	21	23	21	5	21
11	22	30	14	4	17	**35**	24	13	17

Step 1: Range is to be found. In the case of our example, the lowest value is 1 and the highest value is 35. Range=Highest Score- Lowest Score(R=H-L)

Thus, R = 35-1 = 34.

Step 2: The class interval can be derived by dividing the range by number of categories that we need.

i = Range/ Number of categories needed

In our example, the range is obtained as 34, and total number of scores (number of students) are 30. Thus, around 6 categories would be sufficient. Thus,

i = 34/ 6= 5.7, that can be rounded off to 6.

While creating categories, ensure that not more than 10 categories are created if there are approximately 50 scores, not more than around 10 to 15 categories are created if the scores are between 50 to 100 and not more than 20 categories are created if the scores are more than 100 (Mangal, 2002). Make sure you have a few items in each category. For example, if you have 20 items, choose 5 classes (4 items per category), not 20 classes (which would give you only 1 item per category).

It is sometimes possible that the 'i' obtained is not a whole number. In such a situation, a number nearest to this obtained number can be taken. For example if 'i' is obtained as 5.8 then 6 can be taken being the nearest number.

It is also possible that the class interval or 'i' is finalised before the number of categories are decided. For convenience, the class interval of 10, 5, 2, for example, can be taken.

Thus, class interval can be derived in either way as mentioned above.

Step 3: Frequency distribution table can now be created. The following is to be done to create a frequency distribution table:

a) For this a table with three columns is to be created with variable (that is, marks in the case of the present example), tallies and frequency (this is similar to the steps followed in creating an ungrouped frequency distribution).

b) Then enter your variables in the first column.

c) Go through your raw data and make a mark (/) for each variable next to its value in the second column of your table.

d) Count the tally marks for each variable and write its total in third column, that is, frequency column.

Marks	Tallies	Frequency (*f*)
31- 36	/	1
25- 30	//	2
19- 24	~~////~~ ~~////~~ /	11
13- 18	~~////~~	5
7- 12	~~////~~ ///	8
1- 6	///	3
Total		30

Step 4: Totalling the frequencies. All the frequencies in the third column are totalled and the number thus achieved needs to be equal to the total number of scores. In case of our example, N = 30 and the total of frequencies is also 30.

Check Your Progress II

1) What is frequency distribution?

.....

.....

.....

.....

.....

2) The number of people treated in a local hospital on a daily basis is given below, construct the frequency distribution table with class interval 5.

15, 23, 12, 10, 28, 7, 12, 17, 20, 21, 18, 13, 11, 12, 26, 30, 16, 19, 22, 14, 17, 21, 28, 9, 16, 13, 11, 16, 20. 1

Class Interval	Tallies	*f*

8.4 CUMULATIVE FREQUENCY DISTRIBUTION

After understanding frequency distribution, let us now take a look at the cumulative frequency distribution. Cumulative frequency can be obtained when we successively add all the frequencies from the bottom of the distribution (Mangal, 2002). A cumulative frequency distribution table is a more meticulous table. It looks similar to a frequency distribution table but it has added column that gives the cumulative frequency.

Let us understand the cumulative frequency distribution table with an example.

In a walking race conducted by a club, all 10 of the participants had to fill out a form that gave their personal and demographic details. Participants filled in various details but here we will consider their age for constructing the cumulative frequency distribution table.

The ages (in years) of the participants were as follows:

36, 48, 54, 92, 57, 63, 66, 76, 66, 80

Now, answer the following questions based on above raw data:

- How many participants aged less than 45?
- How many participants aged more than 44?
- What is the percentage of participants who are older than 65 years?

The answer to these questions can be best given using cumulative frequency distribution method. These are called as cumulative frequencies "because they tell how many scores are accumulated up to this point on the table"(Aron, Aron and Coups, 2013, page 7).

Let us present this data in a cumulative frequency distribution table.

Step 1: Divide the values into intervals, and then count the number of values in each interval. In this case, intervals of 10 are appropriate. Since 36 is the lowest age and 92 is the highest age, start the intervals at 35 to 44 and end the intervals with 85 to 94.

Step 2: Create a table similar to the frequency distribution table but with three extra columns.

Step 3: In the first column or the lower value column, list the lower value of the intervals. For example, in the first row, you would put the number 35.

Step 4: The next column is the upper value column. Place the upper value of the intervals. For example, you would put the number 44 in the first row.

Step 5: The third column is the Frequency column. Record the number of times a value appears between the lower and upper values of the intervals. For example in the first row, place the number 1.

Step 6: The fourth column is the Cumulative frequency column. Here, we add the cumulative frequency of the previous row to the frequency of the current row. Since, this is the first row, the cumulative frequency is the same as the frequency. However, in the second row, the frequency for the 35–44 interval (i.e., 1) is added to the frequency for the 45–54 interval (i.e., 2). Thus, the cumulative frequency is 3, meaning we have 3 participants in the 34 to 54 age group.

$$1 + 2 = 3$$

Step 7 and 8 can be added to obtain cumulative percentage frequency.

Step 7: The next column is the Percentage column. In this column, list the percentage of the frequency. To do this, divide the frequency by the total number of values and multiply by 100. In this case, the frequency of the first row is 1 and the total number of values is 10. The percentage would then be 10.

$$(1 \div 10) \times 100 = 10$$

Step 8: The final column is Cumulative percentage frequency. In this column, multiply the cumulative frequency by 100 and then divide it by the total number of values. Note that the last number in this column should always equal 100.0. In this example, the cumulative frequency is 1 and the total number of values is 10, therefore the cumulative percentage frequency of the first row is 10.0.

$$1 \times 100 \div 10 = 10$$

The cumulative frequency distribution table will look like this:

Lower Value (age in years)	Upper Value (age in years)	Frequency (f)	Cumulative frequency	Percentage	Cumulative percentage frequency
85	94	1	10	10	100
75	84	2	9	20	90
65	74	2	7	20	70
55	64	2	5	20	50
45	54	2	3	20	30
35	44	1	1	10	10
		N= 10			

Based on preceding table, now following information can be obtained:

- Number of participants aged less than 45 years= 1
- Number of participants aged more than 44 years = 9
- Percentage of participants aged above 65 years = 50%

Note that cumulative frequency can easily be converted to cumulative percentage frequencies by carrying out multiplication between the cumulative frequency and 100 and dividing by N (N is the total number of frequencies in the distribution). Cumulative percentage frequencies provide information about the percentage of frequencies that lie below a certain score/ class interval (Mangal, 2002).

Check Your Progress III

1) How is cumulative frequency obtained?

.....

.....

.....

.....

.....

.....

.....

2) The number of people treated in a local hospital on a daily basis is given below, construct cumulative frequency distribution and cumulative percentage frequency with class interval 5.

15, 23, 12, 10, 28, 7, 12, 17, 20, 21, 18, 13, 11, 12, 26, 30, 16, 19, 22, 14, 17, 21, 28, 9, 16, 13, 11, 16, 20. 1

Class Interval	Tallies	*f*	Cumulative frequency	Cumulative percentage frequency

8.5 PERCENTILE AND PERCENTILE RANKS

There are two terms that are used frequently in academic and corporate world: percentile and percentile ranks. Both these statistical terms are used as indicators of performance in comparison to others in a large group. It can be said that these indicators are relative measures of one's performance. There are many tests that report scores in percentile or percentile ranks. You may have heard about Common Aptitude Test (CAT)-a common entrance exam conducted for MBA admissions in India. This exam gives result in percentile. For example, a student may obtain 90^{th} percentile in math ability and 84^{th} percentile in verbal ability.

In this section of the unit, we will discuss about the terms percentile and percentile rank and also learn how to compute them.

Percentile: A percentile can be explained as "a point on the score scale below which a given percent of cases lie" (Mangal, 2002, page 56). For example, if a student obtained 90^{th} percentile (P_{90}), it means that 90% of the students have scored below him/ her or if the student obtains 84 percentile (P_{84}) then 84% of the students lie below him/ her. Percentiles are expressed in terms of percentage of persons in the standardization sample who fall below a given raw score. A percentile will show an individual's relative position in the standardization sample. There is the difference between rank and percentile. In ranks we count from the top and the best person in the group gets Rank 1. However, in percentile we count from the bottom and lower the percentile, poorer is an individual's position in the group. The 50^{th} percentile or P_{50} is like the median. Above 50^{th} percentile denotes above average performance while below P_{50} denotes below average performance. Percentiles are different from percentage scores. Percentage scores are raw scores which are expressed in terms of percentage of correct items, while percentiles are derived scores.

Advantages of Percentile Scores

1) It is universally applicable.

2) It can be readily understood and are easy to compute even by untrained persons.

3) Is suitable for any type of test.

Drawbacks of Percentile Scores

1) Percentiles show individuals relative position in the normative score but not the individuals score compared with one another.

2) Percentile score have inequality of the unit and this is a major drawback.

Computation of percentile: Percentile can be computed as follows:

The formula for computation of percentiles is similar to that of median (Mangal, 2002).

$$P = L + [(pN/ 100- F)/ f] X i$$

Where,

L = The lower limit of the percentile class or the class where the percentile may lie.

p = Number of percentile for which calculation is to be carried out.

N = The total number of frequencies

F = Total of the frequencies that exist before the percentile class

f = Frequency of the percentile class

i = The size of the class interval

Thus, the formula for 1st percentile would be

$$P_1 = L + [(N/ 100- F)/ f] X i$$

And the formula for 10th percentile would be

$$P_{10} = L + [(10N/ 100- F)/ f] X i$$
$$= L + [(N/ 10- F)/ f] X i$$

And the formula for 75th percentile would be

$$P_{75} = L + [(75N/ 100- F)/ f] X i$$
$$= L + [(3N/ 4- F)/ f] X i$$

Let us now compute percentile with the help of an example given in table 8.7.

Table 8.7: Data for computation of Percentile

Class Interval	f
25-29	5
20-24	4
15-19	6
10-14	4
5-9	4
0-4	7
	N= 30

Now if we want to compute 30th percentile for the above data, we will compute with the help of the following steps:

Step 1: Find the class interval within which the 30th percentile will fall. P_{30} indicates that 30% of the scores lie below this point. Thus, 30% of N = 30 X 30/100 = 9. Now as we look at the data, the 9th score from below lies in the class interval 5-9.

Step 2: L, that is, the lower limit of the percentile class or the class where the percentile may fall is identified. In the case of this example, it will be 4.5 that is the lower limit of class interval 5-9.

Step 3: F, that is, total of the frequencies that exist before the percentile class is 7. In case of this example and *f*, that is, frequency of the percentile class is 4.

Step 4: Let us now substitute the values in the formula

$$\mathbf{P_{30} = L + [(30N/ 100- F)/ \mathit{f}] X i}$$

$$= 4.5 + [(30 X 30/ 100- 7)/ 4] X 5$$

$$= 4.5 + [(9-7)/4] X 5$$

$$= 4.5 + 2/4 X5$$

$$= 4.5 +2.5$$

$$= 7$$

Thus, the obtained P_{30} is 7 that falls in the class interval 5-9.

Percentile Ranks: In statistics, percentile rank refers to the percentage of scores that are identical to or less than a given score. Percentile rank can be explained as "the number representing the percentage of the total number of cases lying below the given score" (Mangal, 2002, page 60). Percentile ranks, like percentages, fall on a continuum from 0 to 100. For example, a percentile rank of 50 indicates that 50% of the scores in a distribution of scores fall at or below the score at the 50th percentile. Percentile ranks are beneficial when you want to quickly understand how a specific score compares to the other scores in a distribution. For instance, knowing someone scored 300 points in an exam doesnot tell you much. You do not know how many points were possible, and even if you did, you would not know how that person scored compared to the rest of his/her classmates. If, however, you were told that he/she scored at the 95th percentile rank, then you would know that he/she did as well or better than 95% of his/her class.

Computation of percentile rank: Percentile rank can be computed for an ungrouped data as well as grouped data. These computations have been discussed as follows with the help of examples:

Computation of Percentile rank for ungrouped data: The formula for computation of percentile rank for ungrouped data is:

$$\mathbf{PR= 100-100R- 50/ N}$$

Where,

PR= Percentile Rank

R = The rank position of the person for whom the percentile rank is to be computed.

N= Total number of persons in the group.

We will now compute percentile rank with the help of the following data:

The marks obtained by 10 students in a psychology test are given as follows:

34, 45, 23, 67, 43, 78, 87, 56, 88, 46

We will now find percentile rank for the marks 67.

Step 1: The marks are to arranged in descending order as follows:

Marks	Rank order
88	1
87	2
78	3
67	**4**
56	5
46	6
45	7
43	8
34	9
23	10

Step 2: Rank for the marks are identified. As can be seen above, the Rank for marks 67 is 4 and N is 10.

Step 3: Let us now substitute the values in the formula

PR= 100-(100R- 50/ N)

= 100- (100 X 4- 50/ 10)

= 100- (400-50/ 10)

= 100- 350/10)

= 100- 35

= 65

Thus, the percentile rank obtained for rank 67 is 65.

Computation of Percentile rank for grouped data: There are two methods for computing percentile rank for grouped data. One is where as such formula is not required and the other where formula is required.

We will now compute percentile rank with the help of the following data:

Marks	***f***
90-99	1
80-89	3
70-79	2
60-69	10
50-59	9
40-49	3
30-39	**6**
20-29	7
10-19	8
0-9	1
	N= 50

We will compute percentile for marks 35.

Method 1: Without formula

The steps in this computation are discussed as follows:

Step 1: We know that the marks 35 fall in the class interval 30-39. If we add the frequencies that are below the upper limit of class interval 20- 29, that is 29. 5, there are (7 +8 +1)= 16 cases.

Step 2: We need to find out the number of cases that lie below 35. Thus, 35-29.5= 5.5.

Step 3: The frequency distribution for class intervals 30- 39 is 6. Thus, these 10 marks (30-39) are shared by 6 individuals. The interval shared by each of the 6 individuals is 5.5. 6/10 x 5.5 = 3.3

Step 4: Thus, up to marks 35, there are 16 +3.3 = 19. 3 or 19 cases.

Step 5: To present these cases on a scale of 100. we multiply these cases with 100/N. N= 50.

$$19.3 \times 100/ 50 = 1930/50 = 38.6$$

Thus, the percentile rank is 38.6 or 39 for marks 35.

Method 2: With formula

The formula for computation of percentile rank for grouped data is:

$$\mathbf{PR= 100/ N [F + (X-L/ i) x f]}$$

Where,

PR= Percentile Rank

F= The cumulative frequency that lies below the class interval that consists of X

X= The marks for which the percentile rank is to be computed.

L= The lower limit of the class interval that consists of X

i= Size of the class interval

f = Frequency of the class interval that consists of X

N= Total number of cases in the distribution

We will take the same example discussed above and compute the percentile rank for marks 35 with the help of the formula.

Step 1: The cumulative frequency below the class interval (30-39) that consists of X (35) is 16 (7 +8 +1). Thus F is 16.

Step 2: L, that is, the lower limit of the class interval that consists of X, is 29.5, i = 10 and f = 6.

Step 3: Let us now substitute the values in the formula

$$\mathbf{PR = 100/\ N\ [F + (X\text{-}L/\ i)\ x\ f]}$$

$$= 100/\ 50\ [16 + (35\text{-}29.5/10)\ x\ 6]$$

$$= 2\ [16 + 5.5/\ 10\ x\ 6$$

$$= 2\ [16+3.3]$$

$$= 2\ x\ 19.3$$

$$= 38.6$$

Thus, the percentile rank is 38.6 or 39 for marks 35.

Percentile and percentile rank can be termed as important in statistics as they not only provide information about the comparative position of an individual in a particular group based on certain characteristics, but they also help in comparing individuals in two or more groups or under two or more circumstances or conditions. For example, if a learner from one college obtained 55 marks in psychology and another learner from another college obtained 65 marks, these cannot be compared, but if these marks are converted in to percentile rank and then it is stated that both have 60th percentile rank, then a comparison is possible. Percentiles also play an important role in standardisation of psychological tests where the raw data can be converted to percentiles and interpreted.

Check Your Progress IV

1) What is percentile?

.....

.....

.....

.....

.....

2) Compute percentile rank for 22 in the following data:

23, 34, 22, 33, 45, 55, 32, 43, 46, 21

8.6 GRAPHICAL REPRESENTATION OF DATA

All the available numerical data can be represented graphically. A graph is the representation of data by using graphical symbols such as lines, bars, pie diagrams, dots etc. A graph represents a numerical data in the form of a structure and provides important information to the user of the data.

When an organised data is graphically represented it not only looks attractive but it is easier to understand. A large amount of data can be presented in a very concise and attractive manner. Graphs are effective and economical as well. They are also easy to interpret and adequately reflect any comparison between two sets of data.

There are various types of graphs like bar graph, histogram, frequency polygon etc. that can be effectively used to graphically represent data. However, one must know when to use which graphs.

Let us now discuss various types of graphs.

8.6.1 Bar Graph or Bar Diagram

A bar graph is also called as bar diagram. It is the most frequently used graph in statistics. A bar graph is a type of graph, which contains rectangles or rectangular bars. The lengths of these bars should be proportional to the numerical values represented by them. In bar graph, the bars may be plotted either horizontally or vertically depending on the interest of the plotter.

Bar graph or diagram can be easily drawn for raw scores, frequencies, percentages and mean (Mangal, 2002).

The following needs to be taken care of while drawing bar graphs (Mangal, 2002):

1) Rules need to be followed with regard to the length of the bars, though no rules are applicable to the width, all the bars need to be of equal width. The lengths or heights of the bars in the bar graph need to in proportion with the amount of variables.

2) The space between two bars could be around half of the width of a bar and the space between any two bars should be same.

The steps followed while drawing a vertical bar graph are as follows:

Step 1: On a graph paper draw the vertical (y axis) and horizontal (x axis) lines. These lines should be perpendicular to each other and need to intersect at 0.

Step 2: Provide adequate labels to the y axis and x axis.

Step 3: A scale needs to be selected for the length of the bars that is usually written on the extreme right at the top of the bar graph.

Step 4: On x axis, we need to select a width for the bars as well as the gap between the bars that needs to be uniform.

Step 5: Based on your data you may then draw the graph.

An example of bar graph or diagram is given in figure 8.1, which is based on the table 8.1 that reflects the marks obtained by students in a class test in Psychology of 100 marks. There are 20 students who scored marks between 1-25, 12 who secured marks between 26 and 50, 40 students who secured marks between 51 and 75 and 28 students secured between 76-100 marks:

The bar graphs based on table 8.1 will look as follows:

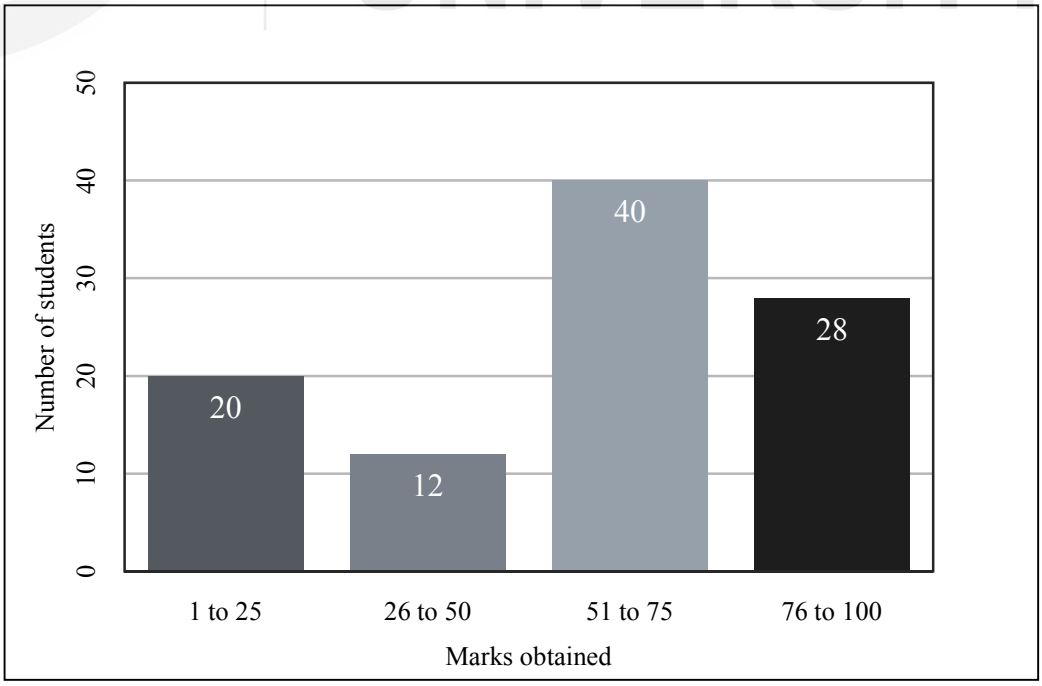

Fig. 8.1: Bar Graph

8.6.2 Histogram

A Histogram is a bar diagram that can be drawn based on frequency distribution. The following steps are to be taken while drawing a histogram.

Step 1: Histogram is based on frequency distribution and a grouped frequency distribution has class intervals, therefore, before drawing a histogram, two more class intervals are added, one below and one above. As can be seen in table 8.8. The frequency distribution originally had 5 class interval, but two more, one below and one above have been added.

Step 2: Further for histogram, the class intervals are changed as can be seen in figure 8.2. where class interval 10-19 has changed to 9.5-19.5 and so on.

Step 3: On x axis, the actual lower limits of all the class intervals are then plotted. And frequencies are plotted on the y axis.

Step 4: A single rectangle will then represent each frequency.

Ensure that the height of the graph is around 75% of its width.

Table 8.8: Data for Histogram

Class Intervals (10)	Class Intervals taken for Histogram	Frequencies
70-79	69.5- 79.5	0
60-69	**59.5- 69.5**	**5**
50-59	**49.5- 59.5**	**4**
40-49	**39.5- 49.5**	**13**
30-39	**29.5- 39.5**	**12**
20- 29	**19.5- 29.5**	**10**
10-19	9.5- 19.5	0

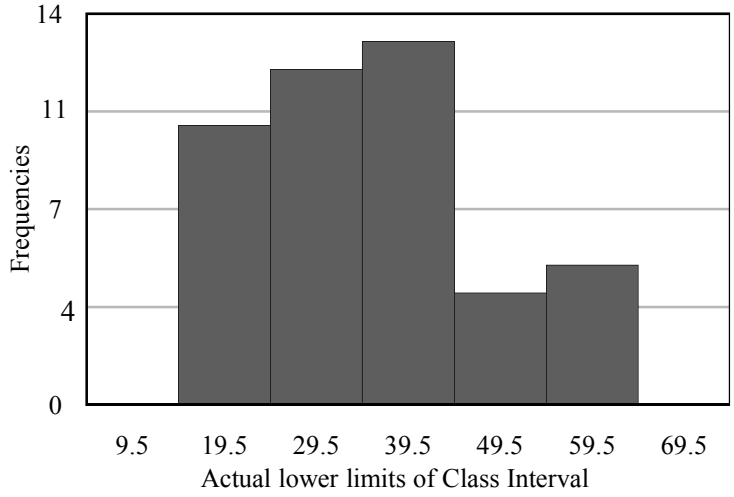

Fig. 8.2: Histogram

8.6.3 Frequency Polygon

A line graph used for plotting frequency distribution is called frequency polygon. Frequency polygon can either be constructed directly or it can also be constructed by drawing a straight line through the midpoints of the upper base of the histogram (Mangal, 2002), that is shown in figure 8.4.

Steps followed while drawing a frequency polygon are as follows:

Step 1: As we know that the frequency polygon is based on frequency distribution. In case of frequency polygon as well before drawing a frequency polygon, two more class interval are added, one below and one above. As can be in table 2.9.

Step 2: For all the class intervals, midpoints are computed.

Step 3: Like every graph, frequency polygon also has x axis and y axis. On x axis, the midpoints are to be plotted and the frequencies will be represented on the y axis.

Step 4: The corresponding frequencies of the class intervals are then plotted based on the midpoints given on x axis.

Step 5: These points are then joined to form a line.

Ensure that the height of the graph is around 75% of its width.

Once plotted, the frequency polygon will look as given in figure 8.3.

Table 8.9: Data for Frequency Polygon

Class Intervals (10)	Midpoints of Class Intervals	Frequencies
70-79	74.5	0
60-69	**64.5**	**5**
50-59	**54.5**	**4**
40-49	**44.5**	**13**
30-39	**34.5**	**12**
20- 29	**24.5**	**10**
10-19	14.5	0

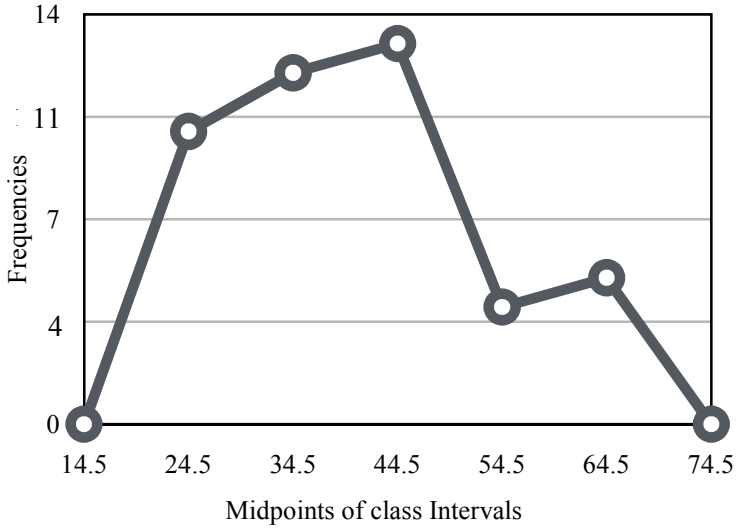

Fig. 8.3: Frequency Polygon

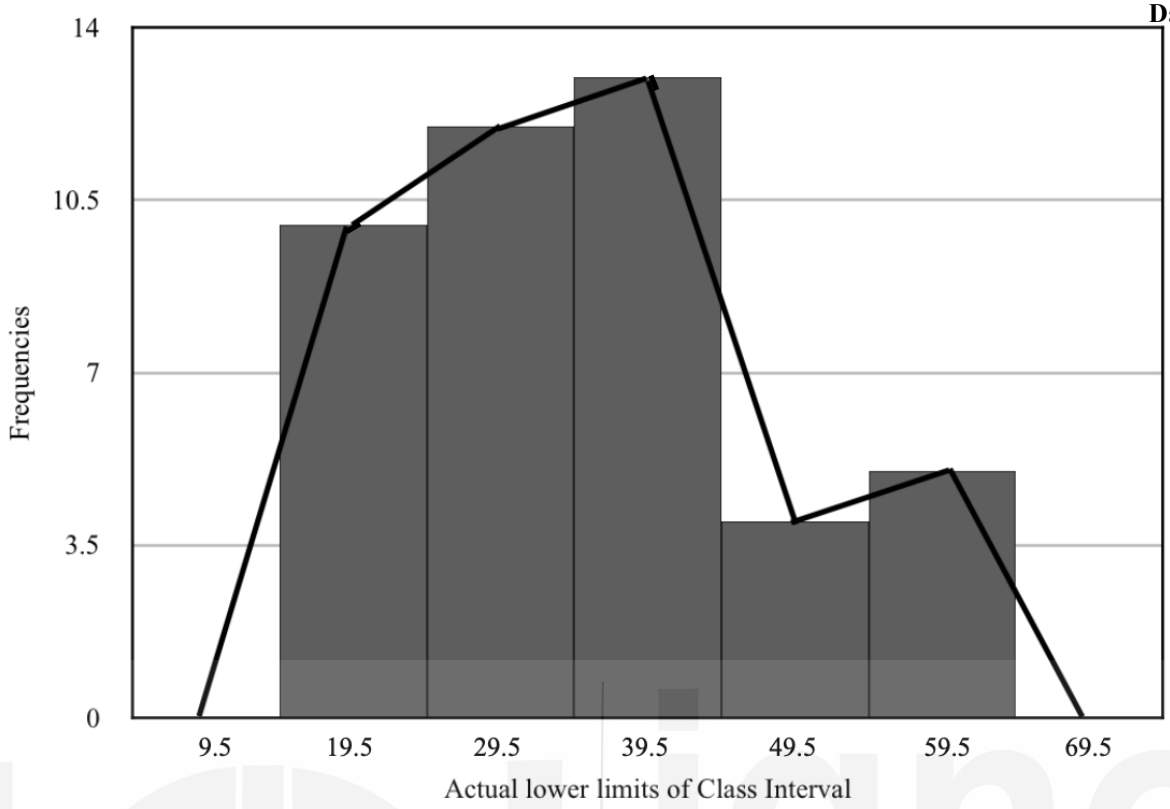

Fig. 8.4 : Frequency Polygon drawn with the help of Histogram

8.6.4 Cumulative Frequency Percentage Curve or Ogive

Cumulative frequency percentage can be plotted in form of a graph and this graph is called as cumulative frequency percentage curve or ogive. Such a graph is a line graph. On y axis the cumulative frequency percentages are plotted and on x axis, the upper limit of the class intervals are plotted. This graph lacks a negative slope and when a certain class interval has zero frequency then the line or curve will remain horizontal.

As was discussed under section on cumulative frequency distribution, cumulative frequency percentage is computed by multiplying the cumulative frequency by 100/ N, where N stands for total number of frequencies.

The steps to draw a cumulative frequency percentage curve or ogive are as follows:

Step 1: The frequency distribution table should be ready with computation of cumulative frequency percentages.

Step 2: Plot the cumulative frequency percentage on y axis and the upper limits of class interval on x axis.

Step 3: Plot the points representing the cumulative frequency percentage for each class interval.

Step 4: Join the points with the help of a line.

Table 8.10 : Data for Cumulative frequency and cumulative frequency percentage				
Class Intervals (10)	**Upper Limit of Class Intervals**	**Frequencies**	**Cumulative frequencies**	**Cumulative frequency percentage**
60-69	69.5	10	44	100
50-59	59.5	12	34	77.27
40-49	49.5	13	22	50
30-39	39.5	4	9	20.45
20- 29	29.5	5	5	11.36
10-19	19.5	0	0	0

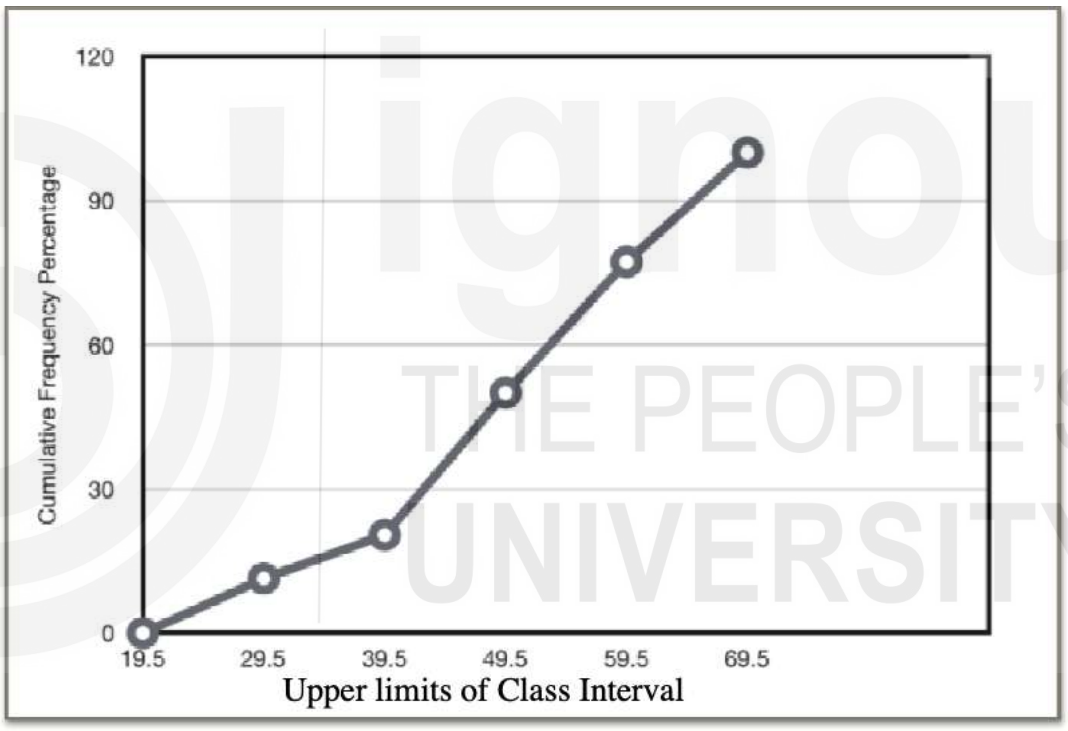

Fig. 8.5: Cumulative Frequency Percentage Curve or Ogive

8.6.5 Circle Graph or Pie Chart

A pie chart is also known as a circle graph. A pie chart is defined as a graph, which contains a circle which is divided into sectors. These sectors illustrate the numerical proportion of the data. Each portion of the circle represents the data. This circle graph is called as pie chart because 'pie' (π) is a quantity that is considered when the circumference of a circle is determined (Mangal, 2002).

Steps in construction of a pie chart:

Step 1: The data represented here is presented through 360º because the surface area of the circle covers 2π or 360º.

Step 2: The total frequency is considered equal to 360º and then angle for each component part is computed. This is done by using the formula:

(Frequency of the component/ Total frequency) X360º.

If the components are presented in percentages then the formula used is

(Percentage value of a particular component/ 100) X360º

Step 3: The sections are then drawn after the angles are determined.

Table 8.11: Data for Pie Chart

Occupation	Number of Individuals	Angle of the circle
Lawyer	5	5/30 x360º= 60º
Accountant	6	6/30 x360º= 72º
Psychologist	4	4/30 x360º= 48º
Engineer	7	7/30 x360º= 84º
Doctor	8	8/30 x360º= 96º
Total	30	360º

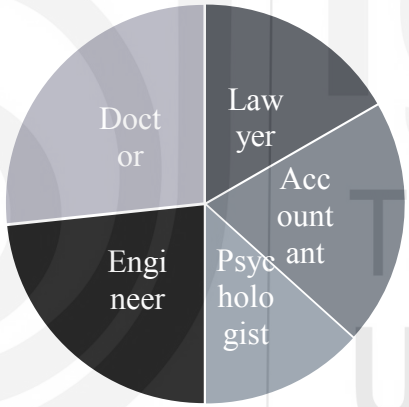

Fig. 8.6: Circle Graph pr Pie Chart

Check Your Progress V

1) What care needs to be taken while drawing a bar graph?

.....

.....

.....

.....

2) What is a pie chart?

.....

.....

.....

8.7 LET US SUM UP

In this unit we initially discussed about classification and tabulation of qualitative and quantitative data. In descriptive statistics classification and tabulation of data, whether qualitative or quantitative, are two important functions that help researchers in organising the data in a better manner and then to subject it to further statistical analysis. Data classification is the method of organising data into groups for its most effective and efficient use. Well-planned data classification system makes vital data easy to find and retrieve whenever required. Tabulation, on the other hand, is the process of insertion of classified data into tabular form. A table is a symmetric arrangement of statistical data in rows and columns. We also discussed about the key components of tabulation. The significance of classification and tabulation was also highlighted.

Further in this unit, we discussed about frequency distribution. Frequency distribution is arranged in a tabular form in which the raw data is organised in to class intervals. Frequency distribution can be categorised as relative frequency distribution, cumulative frequency distribution and cumulative relative frequency distribution, which were discussed in the unit with the help of examples. Besides the two main methods, namely, the exclusive and inclusive methods, of describing class interval in frequency distribution were also discussed. The unit then focused on computing frequency distribution for both ungrouped and grouped data. The steps involved in creating a cumulative frequency distribution were also highlighted. Cumulative frequency percentage was also explained in the unit.

Further, the unit focused on the concepts and computation of percentile and percentile rank with the help of examples. A percentile can be explained as a point on the score scale below which a given percent of cases lie and percentile rank refers to the percentage of scores that are identical to or less than a given score.

The last section of the unit explained the graphical representation of data. A graph is the representation of data that uses graphical symbols such as lines, bars, pie diagrams, dots etc. When an organised data is graphically represented, it not only looks attractive but it is easier to understand. A large amount of data can be presented in a very concise and attractive manner. Graphs are effective and economical as well. In the present unit, bar graph, histogram, frequency polygon, cumulative frequency percentage curve or ogive and piechart were discussed in detail with the help of examples and figures.

8.8 REFERENCES

Kurtz, A. K., & Mayo, S. T. (2012). *Statistical Methods in Education and Psychology*. Springer Science & Business Media.

Kurtz A.K., Mayo S.T. (1979) Percentiles and Percentile Ranks. In: Statistical Methods in Education and Psychology. Springer, New York, NY

Miles, J. N. V., & Banyard, P. (2007). *Understanding and Using Statistics in Psychology: A Practical Introduction*. London: Sage.

Wright, D. B., & London, K. (2009). *First Steps in Statistics* (2nd ed.). London: Sage.

Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics*. Sage.

Rosnow, R. L., & Rosenthal, R. (2005). *Beginning Behavioural Research: A Conceptual Primer* (5th ed.). Englewood Cliffs, NJ: Pearson/Prentice Hall.

Aron, A., Coups, E. J. & Aron, E. N. (2013). *Statistics for Psychology* (6th ed.). Pearson Education

8.9 KEY WORDS

Classification: It is the process of ordering data into homogenous groups or classes according to some common characteristics present in the data is called classification.

Tabulation: It is the process of insertion of classified data into tabular form.

Frequency: It is the number of times a particular variable/ individual or observation (obtained marks in our context) occurs in raw data.

Percentiles: These are expressed in terms of percentage of persons in the standardisation sample who fall below a given raw score. A percentile will show an individual's relative position in the standardisation sample.

Percentile ranks: refers to the percentage of scores that are identical to or less than a given score. Percentile ranks, like percentages, fall on a continuum from 0 to 100.

8.10 ANSWERS TO CHECK YOUR PROGRESS

Check Your Progress I

1) What is quantitative data?

   Quantitative data states information about quantities, that is, information that can be measured and written down with numbers.

2) List the merits of classification and tabulation.

The merits of classification and tabulation are as follows:

a) It helps in clarifying the data

b) The data is presented in simple form

c) Comparison is possible between the data

d) Information can be easily referred to

Check Your Progress II

1) What is frequency distribution?

Frequency distribution is a way in which raw data can be classified so as to provide a clearer understanding of the data.

2) The number of people treated in a local hospital on a daily basis is given below, construct the frequency distribution table with class interval 5.

15, 23, 12, 10, 28, 7, 12, 17, 20, 21, 18, 13, 11, 12, 26, 30, 16, 19, 22, 14, 17, 21, 28, 9, 16, 13, 11, 16, 20, 1

Class Interval	Tallies	*f*
30-34	/	1
25-29	///	3
20-24	~~///~~ /	6
15-19	~~///~~ ///	8
10-14	~~///~~ ////	9
5-9	//	2
0-4	/	1
		N= 30

Check Your Progress III

1) How is cumulative frequency obtained?

Cumulative frequency can be obtained when we successively add all the frequencies from the bottom of the distribution

2) The number of people treated in a local hospital on a daily basis is given below, construct the cumulative frequency distribution table with class interval 5.

15, 23, 12, 10, 28, 7, 12, 17, 20, 21, 18, 13, 11, 12, 26, 30, 16, 19, 22, 14, 17, 21, 28, 9, 16, 13, 11, 16, 20. 1

Class Interval	*f*	Cumulative Frequency	Cumulative Percentage Frequency
30-34	1	30	100
25-29	3	29	96.67
20-24	6	26	86.67
15-19	8	20	66.67
10-14	9	12	40
5-9	2	3	10
0-4	1	1	3.33
	N= 30		

Check Your Progress IV

1) What is percentile?

Percentile can be described as a point on the score scale below which a given percent of cases lie.

2) Compute percentile rank for 22 in the following data:

23, 34, 22, 33, 45, 55, 32, 43, 46, 21

Data	Rank order
55	1
46	**2**
45	3
43	4
34	5
33	6
32	7
23	8
22	9
21	10

The percentile rank for 22 is 15.

Check Your Progress V

1) What care needs to be taken while drawing a bar graph?

The lengths or heights of the bars in the bar graph need to in proportion with the amount of variables. The space between two bars could be around half of the width of a bar and the space between any two bars should be same.

2) What is a piechart?

A pie chart is defined as a circular graph, which contains a circle which is divided into sectors.

8.11 UNIT END QUESTIONS

1) Explain classification of data with a focus on its objective.
2) Describe the key components of a table.
3) Elucidate percentile and percentile ranks with suitable examples.
4) Describe bar diagram with suitable diagram
5) Discuss the steps involved in drawing a cumulative frequency percentage curve or ogive.

UNIT 9 INTRODUCTION TO MEASURES OF CENTRAL TENDENCY*

Structure

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

9.0 OBJECTIVES

After reading this unit, you will be able to:

- explain the concept of central tendency of data;
- describe the different measures of central tendency;
- discuss the properties, advantages and limitations of mean, median and mode; and
- compute measures of central tendency for ungrouped and grouped data.

9.1 INTRODUCTION

Suppose you have data, for instance, marks in psychology obtained by students in 12th standard and you want to analyse it statistically, what statistical techniques will you employ? You can of course organise the data with the help of classification and tabulation that we discussed in the previous Unit and the data can also be graphically represented. But if you want to further analyse the data then you can compute the average marks obtained by the whole class or find the midpoint for marks above and below which will lie half of the students or you can also find out most frequent marks obtained by the students. The techniques you are employing here are mean, median and mode. These are called measures of central tendency and can be categorised under descriptive statistics.

In the previous unit, we discussed about classification, tabulation and also graphical representations of data. In the present unit, we will discuss the measures of central tendency, viz., mean, median and mode. We will not only understand what these techniques are, but will also focus on their properties, advantages and limitations. Further, we will also learn how to compute mean, median and mode for ungrouped and grouped data.

9.2 CONCEPT OF CENTRAL TENDENCY OF DATA

Measures of central tendency provides a single value that indicates the general magnitude of the data and this single value provides information about the characteristics of the data by identifying the value at or near the central location of the data (Bordens and Abbott, 2011). King and Minium (2013) described measures of central tendency as a summary figure that helps in describing a central location for a certain group of scores. Tate (1955, page 78) defined measures of central tendency as "a sort of average or typical value of the items in the series and its function is to summarise the series in terms of this average value".

The main functions of measures of central tendency are as follows:

1) They provide a summary figure with the help of which the central location of the whole data can be explained. When we compute an average of a certain group we get an idea about the whole data.

2) Large amount of data can be easily reduced to a single figure. Mean, median and mode can be computed for a large data and a single figure can be derived.

3) When mean is computed for a certain sample, it will help gauge the population mean.

4) The results obtained from computing measures of central tendency will help in making certain decisions. This holds true not only to decisions with regard to research but could have applications in varied areas like policy making, marketing and sales and so on.

5) Comparison can be carried out based on single figures computed with the help of measures of central tendency. For example, with regard to performance of students in mathematics test, the mean marks obtained by girls and the mean marks obtained by boys can be compared.

A good measure of central tendency needs to have the following characteristics:

1) The definition of the central tendency needs to be adequately specified and should be clear. It should not be subject to varied interpretations and needs to be unaffected by any individual bias. The definition should be rigid so that a stable value is obtained that represents the data.

2) The measure of central tendency should be easy to understand and easy to compute. It should not involve elaborate mathematical calculations.

3) For the value obtained from the computation of measures of central tendency to be representative of the data, the whole data needs to be computed.

4) The data needs to be collected from a sample that truly represents the population. The sample thus needs to be randomly selected.

5) The measure of central tendency needs to display sampling stability and should not be affected by any fluctuations in the sample. For example, if two different researchers obtain a representative sample from a same population, the means computed by them for their respective sample should display least variation.

6) The measure of central tendency should not be affected by outliers. Outliers are extreme values in data or distribution.

7) The measure of central tendency should render itself to further mathematical computations.

Check Your Progress I

1) Define measures of central tendency.

.....

.....

.....

.....

.....

.....

2) List the functions of measures of central tendency.

.....

.....

.....

.....

.....

.....

9.3 DIFFERENT MEASURES OF CENTRAL TENDENCY

As the concept of central tendency is now clear, we will now proceed to discuss the three measures of central tendency. The three measures of central tendency that we will be discussing are:

1) Mean or Arithmetic mean
2) Median
3) Mode

In this section of the Unit, we will try to understand these concepts and then in the next section we will be focusing on the properties, advantages and limitations of each of these measures.

9.3.1 Mean or Arithmetic Mean

Mean for sample is denoted by symbol 'M or $\bar{x}$ ('x-bar')' and mean for population is denoted by 'μ' (mu). It is one of the most commonly used measures of central tendency and is often referred to as average. It can also be termed as one of the most sensitive measure of central tendency as all the scores in a data are taken in to consideration when it is computed (Bordens and Abbott, 2011). Further statistical techniques can be computed based on mean, thus, making it even more useful.

Mean is a total of all the scores in data divided by the total number of scores. For example, if there are 100 students in a class and we want to find mean or average marks obtained by them in a psychology test, we will add all their marks and divide by 100, (that is the number of students) to obtain mean.

9.3.2 Median

Median is a point in any distribution below and above which lie half of the scores. Median is also referred to as P_{50} (King and Minium, 2008). The symbol for median is 'M_d'. As stated by Bordens and Abbott (2011, page 411), 'median is the middle score in an ordered distribution'. If we take the example discussed earlier of the marks obtained by 100 students in a psychology test, these marks are to be arranged in an order, either ascending or descending. The middle score in this distribution is then identified as median. Though this would seem easy for an odd number of scores, in case of even number of scores a certain procedure is followed that will be discussed when we learn how to compute median later in this unit.

9.3.3 Mode

Mode is denoted by symbol 'M_o'. Mode is the score in a distribution that occurs most frequently. Taking the example of the marks obtained by a group of 100 students in psychology test discussed earlier, if out of these 100 students, 10 students obtained 35 marks. 35 is thus, most frequently occurring value and will be termed as mode. Certain distributions can be bimodal as well, where there are two modes. For instance if there were other 10 students in this group of 100 students, who secured 47 marks, 47 is the value that is occurring as frequently as 35 and thus, will be termed as mode along with 35. In a similar way, when there are three modes, the term used is trimodal and when there are four or more modes, we use the term multimodal.

Though if the scores in a distribution greatly vary then it is possible that there is no mode. Mode as such does not provide an adequate characterisation of the distribution because it just takes in to consideration the most frequent scores and other scores are not considered.

How to choose a measure of central tendency?

The choice of a measure of central tendency will depend on first of all, the scales of measurement that we discussed in the first unit. For nominal scales one can compute mode but not mean or median. For example, in case of males and females, the males can be coded as 1 and females can be coded as 2 (or vice versa) in such a case, we can compute frequently occurring score, that will provide us information whether there are more males or more females. However, it is not possible to compute mean or median. With regard to ordinal scale median or mode can be used. For example, if we rank the students based on their performance in mathematics test, it is possible to find median below and above which lie half of the ranks. Mode can also be computed if more than one student gets same rank. With regard to interval scale and ratio scale mean can be computed.

Yet another aspect that is important while making a choice with regard to which measure of central tendency to use is, whether the data is normally distributed or not. If the data is normally distributed we will compute mean and if it is not normally distributed, we will compute median or mode. This is because mean may not adequately represent the data when the data is not normally distributed. We will discuss normal distribution in detail in the last unit (unit 8) of this course.

Check Your Progress II

1) Describe mean, median and mode.

Measure	Description	Example
Mean		

Median		
Mode		

2) How to choose a measure of central tendency?

.....

.....

.....

.....

.....

9.4 PROPERTIES, ADVANTAGES AND LIMITATIONS OF MEAN, MEDIAN AND MODE

Let us now discuss the properties, advantages and limitations of mean, median and mode.

9.4.1 Properties of Mean

1) Mean is sensitive to the actual position of each and every score in a distribution and if another score is included in the distribution, then the mean or average of that distribution will change. For example, mean of the scores 5, 4 ,6 ,3, 2 is 4 [We got the value 4 by adding 5+4+6+3+2= 20 and then dividing it by 5, that is the total number of scores (N)]. But if we change the scores to 5, 4, 6, 3, 2, 8, the mean will be 4.67 [We got the value 4.67 by adding 5+4+6+3+2+8= 28 and then dividing it by 6, that is the total number of scores (N)]

2) Mean denotes a balance point of any distribution and the total of positive deviations from the mean is equal to the negative deviations from the mean (King and Minium, 2008).

3) Mean is especially effective when we want the measure of central tendency to reflect the sum of the scores.

9.4.2 Advantages of Mean

1) The definition of mean is rigid which is a quality of a good measure of central tendency.

2) It is not only easy to understand but also easy to calculate.

3) All the scores in the distribution are considered when mean is computed.

4) Further mathematical calculations can be carried out on the basis of mean.

5) Fluctuations in sampling are least likely to affect mean.

9.4.3 Limitations of Mean

1) Outliers or extreme values can have an impact on mean.

2) When there are open ended classes, such as 10 and above or below 5, mean cannot be computed. In such cases median and mode can be computed. This is mainly because in such distributions mid point cannot be determined to carry out calculations.

3) If a score in the data is missing or lost or not clear, then mean cannot be computed unless mean is computed for rest of the data by not considering the lost score and dropping it all together.

4) It is not possible to determine mean through inspection. Further, it cannot be determined based on a graph.

5) It is not suitable for data that is skewed or is very asymmetrical as then in such cases mean will not adequately represent the data.

9.4.4 Properties of Median

1) When compared to mean, median is less sensitive to extreme scores or outliers.

2) When a distribution is skewed or is asymmetrical median can be adequately used.

3) When a distribution is open ended, that is, actual score at one end of the distribution is not known, then median can be computed.

9.4.5 Advantages of Median

1) The definition of median is rigid which is a quality of a good measure of central tendency.

2) It is easy to understand and calculate.

3) It is not affected by outliers or extreme scores in data.

4) Unless the median falls in an open ended class, it can be computed for grouped data with open ended classes.

5) In certain cases it is possible to identify median through inspection as well as graphically.

9.4.6 Limitations of Median

1) Some statistical procedures using median are quite complex. Computation of median can be time consuming when large data is involved because the data needs to be arranged in an order before median is computed.

2) Median cannot be computed exactly when an ungrouped data is even. In such cases, median is estimated as mean of the scores in the middle of the distribution.

3) It is not based on each and every score in the distribution.

4) It can be affected by sampling fluctuations and thus can be termed as less stable than mean.

9.4.7 Properties of Mode

1) Mode can be used with variables that can be measured on nominal scale.

2) Mode is easier to compute than mean and median. But it is not used often because of lack of stability from one sample to another and also because a single set of data may possibly have more than one mode. Also, when there are more than one mode, then the modes cannot be termed to adequately measure central location.

3) Mode is not affected by outliers or extreme scores.

9.4.8 Advantages of Mode

1) It is not only easy to comprehend and calculate but it can also be determined by mere inspection.

2) It can be used with quantitative as well as qualitative data.

3) It is not affected by outliers or extreme scores.

4) Even if a distribution has one or more than one open ended classe(s), mode can easily be computed.

9.4. 9 Limitations of Mode

1) It is sometimes possible that the scores in the data vary from each other and in such cases the data may have no mode.

2) Mode cannot be rigidly defined.

3) In case of bimodal, trimodal or multimodal distribution, interpretation and comparison becomes difficult.

4) Mode is not based on the whole distribution.

5) It may not be possible to compute further mathematical procedures based on mode.

6) Sampling fluctuations can have an impact on mode.

Check Your Progress II

1) List the properties of mean.

.....

.....

.....

.....

.....

.....

2) List the advantages of median.

.....

.....

.....

.....

.....

.....

3) List the limitations of mode.

.....

.....

.....

.....

.....

.....

9.5 COMPUTATION OF MEASURES OF CENTRAL TENDENCY IN UNGROUPED AND GROUPED DATA

Now as we have developed a fair idea about the three measures of central tendency, we will move on to learn how to compute them. While computing each of these measures, we will do so for ungrouped and grouped data. Ungrouped and grouped data are explained as follows:

Ungrouped data: Any data that has not been categorised in any way is termed as an ungrouped data. For example, we have an individual who is 25 years old, another who is 30 years old and yet another individual who is 50 years old. These are independent figures and not organised in any way, thus they are ungrouped data.

Grouped data: A data that is categories or organised is termed as grouped data. Mainly such data is organised in frequency distribution. For example, we can have age range 26- 30 years, 31- 35 years, 36- 40 years and so on. Grouped data are convenient especially when the data is large.

9.5.1 Computation of Mean for Ungrouped Data

The formula for computing mean for ungrouped data is

$M = \sum X / N$

Where,

M = Mean

$\sum X$= Summation of scores in the distribution

N = Total number of scores.

Let us now compute mean with the help of an example

The scores obtained by 10 students on psychology test are as follows:

58 34 32 47 74 67 35 34 30 39

Step 1: In order to obtain mean for the above data we will first add the marks to obtain $\sum X$:

$58+ 34+ 32+ 47+ 74+ 67+ 35+ 34+ 30+ 39 = 450$

Step 2: Now using the formula, we will compute mean

$M = \sum X / N$

$\Sigma X = 450$, N= 10 (Total number of students)

Thus,

$M = 450 / 10 = 45$

Thus, the mean obtained for the above data is 45

9.5.2 Computation of Mean for Grouped Data

The formula for computing mean for grouped data is

$M = \sum fX / N$

Where,

M= Mean

$\sum$= Summation

X= Midpoint of the distribution

f = The respective frequency

N = Total number of scores.

Let us now compute mean with the help of an example.

A class of 30 students were given a psychology test and the marks obtained by them were categorised in to six categories. The lowest marks obtained were 10 and highest marks obtained were 35. A class interval of 5 was employed. The data is given as follows:

Marks	Frequencies (*f*)	Midpoint (X)	*f*X
35- 39	5	37	185
30-34	7	32	224
25-29	5	27	135
20-24	6	22	132
15-19	4	17	68
10- 14	3	12	36
	N= 30		**•fX = 780**

The steps followed for computation of mean with grouped data are as follows:

Step 1: The data is arranged in a tabular form with marks grouped in categories with class interval of 5.

Step 2: Once the categories are created, the marks are entered under frequency column based on which category they fall under.

Step 3: The midpoints of the categories are computed and entered under X.

Step 4: *f*X is obtained by multiplying the frequencies and midpoints for each category.

Step 5: *f*X for all the categories are added to obtain ΣfX, in case of our example it is obtained as 780

Step 6: The formula M = Σ*f*X/ N is used, N is equal to 30.

M = Σ*f*X/ N

M = 780/ 30= 26

Thus, the mean obtained is 26

9.5.3 Computation of Mean by Shortcut Method (with Assumed mean)

In certain cases data is very large and it is not possible to compute each *f*X. In such situations, a short cut method with the help of assumed mean can be computed. A real mean can thus be computed with application of correction.

The formula is

M =AM+(Σ*f*x′/N×i)

Where,

AM= Assumed mean,

Σ = Summation

i = Class interval

x′ = {(X − AM)/ i}, X the midpoint of the class intervals

f = the respective frequency of the midpoint

N = The total number of frequencies or students.

Let us discuss the steps followed for computation of mean with the help of an example given below:

Class Intervals (Marks)	**Frequencies (*f*)**	**Midpoint (X)**	**x′ = {(X − AM)/ i}**	***f x′***
35- 39	5	37	3	15
30-34	7	32	2	14
25-29	5	27	1	5
20-24	6	22	0	0
15-19	4	17	-1	-4
10- 14	3	12	-2	-6
	N= 30			**• *f*x′= 24**

Step 1: We will assume mean (AM) as 22.

Step 2: Difference is obtained between each of the midpoints and the assumed mean and then the same is divided by 'i' that is the class interval (5 in this case), these are then entered under column with heading x′= {(X − AM)/ i}. The x′ for 22 will be 0.

Step 3: Frequency (*f*) is then multiplied with x′ to obtain *f x′*.

Step 4: All *fx′* are added to obtain ∑*f*x′, in the present example it is 24.

Step 5: The formula for mean is now applied

M =AM+(∑*f*x′/N×i)

M= 22 + (24/30 x 5)

= 22 + 4 = 26

Thus, mean is obtained as 26.

And if you refer to the mean obtained by the direct method and mean obtained with the shortcut method, the mean is the same, that is 26.

9.5.4 Computation of Median for Ungrouped Data

With regard to computation of median for ungrouped data, different procedures are followed for data that is odd and data that is even.

9.5.4.1 Odd Data

When the data is odd the median is computed in the following manner:

Data: 58 34 32 47 74 67 35 34 30 (N= 9)

Step 1: First the data is to be arranged in either ascending or descending order. We will arrange the data in ascending order and it will look like this:

30 32 34 34 35 47 58 67 74

Step 2: The following formula is then used to compute Median:

M_d = (N +1)/ 2th score

Thus (9 +1)/2= 10/2 =5th item

In our data the 5th item is 35, that is the median of this data.

9.5.4.2 Even Data

When the data is even, the median is computed in the following manner:

58 34 32 47 74 67 35 34 30 39 (N= 10)

Step 1: First the data is to be arranged in either ascending or descending order. We will arrange the data in ascending order and it will look like this:

30 32 34 34 35 39 47 58 67 74

Step 2: The following formula is used to compute median:

M_d = (N/2)th score + [(N/2)th score + 1]/ 2

The (N/2)th score is the 5th score, that is 35.

The score of (N/2)th score + 1 is the 6th score, that is 39.

Thus 35 +39/ 2= 37

The median thus obtained is 37.

9.5.5 Computation of Median for Grouped Data

The formula used for computation of median for grouped data is as follows:

$$M_d = L + \left[(N/2) - F/f_m\right] \times i$$

Where,

L = The lower limit of the median class

N = Total of all the frequencies

F = Sum of frequencies before the median class

f_m = frequency within the interval upon which the median falls

i = class interval.

Let us discuss the steps followed for computation of median with the help of the example given below:

Class Intervals (Marks)	Frequencies (*f*)
35- 39	5
30-34	7
25-29	**5**
20-24	6
15-19	4
10- 14	3
	N= 30

The steps in computing median for grouped data are as follows:

Step 1: The first step is to compute N/2, that is 30/2 so that we obtain one half of the scores in the data (15 in this case).

Step 2: As the scores are even in number (N= 30), the median should fall between 15th and 16th score. Whether we add the frequencies from above (5+7+5= 17) or from below (3+4+6+5= 18), the median will fall in the class interval 25-29. Further L that is the lower limit of the median class can also be mentioned. As the median class is 25-29, its lower limit will be 24.5.

Step 3: Compute F, that is sum of frequencies before the median class. In our example it would be 3 +4 +6 = 13

Step 4: f_m is computed. It is the frequency within the interval upon which the median falls. In the present example the median class interval is 25-29 and the frequency for this class interval is 5. So f_m is 5.

Step 5: The values can now be put in the formula to obtain the median

$$\mathbf{M_d = L + [(N/2) - F/f_m] \times i}$$

$$= 24.5 + [(30/2) - 13/5] \times 5$$

$$= 24.5 + [15 - 13/5] \times 5$$

$$= 24.5 + [2/5] \times 5$$

$$= 24.5 + 10/5$$

$$= 24.5 + 2$$

$$= 26.5$$

Thus, the median obtained is 26. 5. And it falls in the median class interval 25-29.

9.5.6 Computation of Mode for Ungrouped Data

Let us now learn how to compute mode for an ungrouped data with the help of the following example:

58 34 32 47 74 67 35 34 30 39

The mode can be calculated in simple manner by just counting the scores that appears maximum number of times in the data. In our example, the score occurring maximum number of times is 34, that occurs twice. Thus the mode is 34.

9.5.7 Computation of Mode for Grouped Data

There are two methods by which mode for grouped data can be computed:

9.5.7.1 First Method

The first method is by using the following formula

Mode=3Mdn−2M

Where,

Mdn = Median

M= Mean

Let us now compute mode with the help of the following example:

Class Intervals (Marks)	Frequencies (*f*)	Midpoint (X)	*f*X
50- 59	5	54.5	272.5
40- 49	7	44.5	**311.5**
30- 39	8	34.5	276
20- 29	10	24.5	245
10- 19	15	14.5	217.5
0- 9	5	4.5	54.522.5
	N= 50		**•fX = 1345**

The formula M = ∑*f*X/ N is used, N is equal to 50.

Step 1: Compute mean

M = ∑*f*X/ N

M = 1370/ 50= 26.9

Step 2: Compute median

$$\mathbf{M_d = L + [(N/2) - F/f_m] \times i}$$

$$= 19.5 + [(50/2) - 20/10] \times 10$$

$$= 19.5 + [25 - 20/10] \times 10$$

$$= 19.5 + [5/10] \times 10$$

$$= 19.5 + 5$$

$$= 24.5$$

Step 3: Let us now use these values in our formula and compute mode

$$\mathbf{M_o = 3Mdn - 2M}$$

$$M_o = 3 \times 24.5 - 2 \times 26.9$$

$$= 73.5 - 53.8$$

$$= 19.7$$

Thus the mode computed is 19.7

Also we can make one observation here that the mean obtained for our example is 26.9 the median is 24.5 and the mode is 19.7. All the three values are not close to each other indicating that the distribution of the data may not be normal as the values do not fall in the central area of the distribution. If the values of mean, median and mode were similar, then we could have said that the data is normally distributed.

9.5.7.2 Second Method

In the second method of computing mode for grouped data the following formula is used:

$$\mathbf{M_o = L + /d_1 / d_1 + d_2] \times i}$$

Where,

L= Lower limit of the class interval in which the mode may lie, called as modal class

i = Class interval

d_1 = difference between frequencies of modal class and class interval below it.

d_2 = difference between frequencies of modal class and class interval above it.

Let us discuss the steps followed for computation of mode with the help of the example given below:

Class Intervals (Marks)	**Frequencies (*f*)**
35- 39	5
30-34	**7**
25-29	5
20-24	6
15-19	4
10- 14	3
	N= 30

Step 1: The mode is most likely to fall in the the class intervals 30-34 as that has the highest frequencies (7). Thus this is our modal class and the lower limit of the same (L) will be 29.5.

Step 2: The class interval (i) for this example is 5.

Step 3: Compute d_1, that is, difference between frequencies of modal class and class interval below it and d_2, that is, difference between frequencies of modal class and class interval above it.

$d_1 = f_m - f_{m-1}$

$d_2 = f_m - f_{m+1}$

Where,

f_m = the frequency of the modal class (7 in case of our example).

f_{m-1} = the frequency of the class interval below the modal class (5 in case of our example).

f_{m+1} = the frequency of the class interval above the modal class (5 in case of our example).

Thus, d_1= 7-5 = 2 and d_2= 7-5 = 2 in case of our example.

Step 4: Now let us compute mode with the help of the formula

$$\mathbf{M_o = L + /d_1/ d_1 + d_2] \ x \ i}$$

$$= 29.5 + [2/2+2] x \ 5$$

$$= 29.5 + 2/4 \ x \ 5$$

$$= 29.5 + 10 /4$$

$$= 29.5 + 2.5$$

$$= 32$$

Thus, the mode obtained is 32.

Check Your Progress IV

1) Compute mean, median and mode for the following data:

23, 34, 43, 65, 67, 67, 78, 65, 43, 34, 45, 33, 23, 67, 60 (N= 15)

2) Compute mean for the following data:

Class Intervals (Marks)	Frequencies (*f*)
50- 59	4
40- 49	5
30- 39	6
20- 29	5
10- 19	5
1- 9	5
	N= 30

9.6 LET US SUM UP

In the present unit, we discussed the concept of central tendency. The measures of central tendency was explained as summary figures that help in describing a central location for a certain group of scores. It was further explained as providing information about the characteristics of the data by identifying the value at or near the central location of the data. The functions of measures of central tendency besides the characteristics of good measures of central tendency were also discussed. Further, the unit focused on the three measures of central central tendency, namely, mean, median and mode. Mean is a total of all the scores in data divided by the total number of scores. It is one of the most frequently used measure of central tendency and is often referred to as an average. It can also be termed as one of the most sensitive measure of central

tendency as all the scores in a data are taken in to consideration when it is computed. Median is the middle score in an ordered distribution. Median is a point in any distribution below and above which lie half of the scores. Mode is the score in a distribution that occurs most frequently. Certain distributions are bimodal, where there are two modes.When there are three modes, the term used is trimodal and when there are four or more modes, we use the term multimodal. Though, if the scores in a distribution greatly vary, then it is possible that there is no mode. The properties, advantages and limitations of mean, median and mode were also discussed in detail. Further, the computation of each of these measures of central tendency was also discussed for both ungrouped and grouped data with stepwise explanation.

9.7 REFERENCES

Bordens, K. S. and Abbott, B. B. (2011). *Research Design and Methods: A Process Approach*. New Dekhi:McGraw Hill Education(India) Private Limited.

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.

Tate, M. W.(1955). *Statistics in Education*. New York: Macmillion Co.

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

9.8 KEY WORDS

Measures of Central Tendency: Measures of central tendency can be explained as a summary figure that helps in describing a central location for a certain group of scores.

Mean: Mean is a total of all the scores in data divided by the total number of scores.

Median: Median is a point in any distribution below and above which lie half of the scores.

Mode: Mode is the score in a distribution that occurs most frequently.

9.9 ANSWERS TO CHECK YOUR PROGRESS

Check Your Progress I

1) Define measures of central tendency

Measures of central tendency can be defined as a summary figure that helps in describing a central location for a certain group of scores. It is a value that determines the general magnitude of a distribution.

1) List the functions of measures of central tendency.

   a) They provide a summary figure with the help of which the central location of the whole data can be explained.

   b) The large amount of data can be easily reduced to a single figure.

   c) When mean is computed for a certain sample, it will help us gain idea about the population mean.

   d) The results obtained from computing measures of central tendency will help a researcher make certain decisions.

   e) Comparison can be carried out with the help of the single figures computed with the help of measures of central tendency.

Check Your Progress II

1) Describe mean, median and mode with suitable examples.

Measure	Description	Example
Mean	Mean is a total of all the scores in data divided by the total number of scores. It is one of the most often used measures of central tendency and is often referred to as average. It can also be termed as one of the most sensitive measures of central tendency as all the scores in a data are taken in to consideration when it is computed.	Scores on Job Satisfaction obtained by 5 employees 23, 34, 54, 34, 22 (N= 5) Thus Mean would be 23 +34 +54 + 34 +22= 167 Thus 167/5= 33.4
Median	Median is the middle score in an ordered distribution. Median is a point in any distribution below and above which lie half of the scores.	In above example, the data is arranged in ascending order 22, 23, 34, 34, 54 Median thus is 34
Mode	Mode is the score in a distribution that occurs most frequently. Certain distributions can be bimodal as well, where there are two modes.When there are three modes, the term used is trimodal and when there are four or more modes, we use the term multimodal. Though if the scores in a distributions greatly vary then it is possible that there is no mode.	In above example, 23, 34, 54, 34, 22 Mode is 34 that occurs twice

2) How to choose which measure of central tendency to use?

Choice of measure of central tendency will depend on the scales of measurement and also whether the data is normally distributed or not.

Check Your Progress III

1) List the properties of mean

   a) Mean is sensitive to the actual position of each and every score in a distribution and if another score is included in the distribution, then the mean or average of that distribution will change.

   b) Mean denotes a balance point of any distribution and the total of positive deviations from the mean is equal to the negative deviations from the mean.

   c) Mean is especially effective when we want the measure of central tendency to reflect the sum of the scores.

2) List the advantages of median.

   a) The definition of median is rigid which is a quality of a good measure of central tendency.

   b) It is easy to understand and calculate.

   c) It is not affected by outliers or extreme scores in data.

   d) Unless the median falls in an open ended class, it can be computed for grouped data with open ended classes.

   e) In certain cases it is possible to identify median through inspection as well as graphically.

3) List the limitations of mode.

   a) It is sometimes possible that the scores in the data vary from each other and in such cases the data may have no mode.

   b) Mode cannot be rigidly defined.

   c) In case of bimodal, trimodal or multimodal distribution, interpretation and comparison becomes difficult.

   d) Mode is not based on the whole distribution.

   e) It may not be possible to compute further mathematical procedures based on mode.

   f) Sampling fluctuations can have an impact on mode.

Check Your Progress IV

1) Compute mean, median and mode for the following data:

   23, 34, 43, 65, 67, 67, 78, 65, 43, 34, 45, 33, 23, 67, 60 (N= 15)

   Mean = 49.8, Median = 45, Mode: 67

2) Compute mean for the following data:

Class Intervals (Marks)	Frequencies (*f*)
50- 59	4
40- 49	5
30- 39	6
20- 29	5
10- 19	5
1- 9	5
	N= 30

Mean = 28.83

9.10 UNIT END QUESTIONS

1) Discuss the concept of measures of central tendency with a focus on characteristics of a good measure of central tendency.

2) Explain the properties of mean, median and mode.

3) Discuss the limitations of mean, median and mode.

4) Compute mean, median and mode for the following data:

   44, 32, 34, 34, 45,54, 56, 54, 55, 58, 45, 56, 54, 55, 56, 67, 79, 77, 88, 66, 89, 65, 43, 45, 54

5) Compute mean, median and mode for the following data:

Class Intervals (Marks)	Frequencies (*f*)
50- 59	12
40- 49	10
30- 39	9
20- 29	11
10- 19	8
1- 9	10
	N= 60

UNIT 10 INTRODUCTION TO MEASURES OF VARIABILITY*

Structure

10.0 OBJECTIVES

After reading this unit, you will be able to:

- explain the concept of variability in data;
- describe the main properties, limitation and uses of the range, quartile deviation, average deviation and standard deviation; and
- explain variance and coefficient of variance.

10.1 INTRODUCTION

Look at the two data given below:

Data A: 8, 2, 6, 4, 8, 2, 10, 5, 5, 10 (N= 10, Total = 60, Mean = 6)

Data B: 7, 7, 7, 6, 7, 5, 5, 6, 5, 5 (N = 10, Total = 60, Mean = 6)

* Dr. Usha Kulshreshtha, Faculty, Psychology, University of Rajasthan, Jaipur.

A single glance at the data A and B given above tell us that data A is more homogeneous when compared to the data B that seems to display more variability. Though to further understand the variance in the data, various measures of variability need to be computed.

In the previous unit, we discussed the measures of central tendency, viz, mean, median and mode. These measures give us an average of a set of observations or data. However, the average cannot be a true representation of data because of variations in the distribution. As can be seen in the above example, the mean is same for data A and data B, but the data vary in terms of their deviation from the mean. Thus, it is very important to consider the variations in the data or set of observations. In this unit, we will be explaining the concept of variability (also known as dispersion) in data. Dispersion actually refers to the variations that exist within and amongst the scores obtained by a group. In average there is a convergence of scores towards a mid-point in a normal distribution. In dispersion, we try and see how each score in the group varies from the mean or the average score. The larger the dispersion, less is the homogeneity of the group concerned and if the dispersion is less it means that the group is homogeneous. Dispersion is an important statistic which helps us to know how far the sample population varies from the universe population. It tells us about the standard error of the mean.

In the present unit, we will discuss the meaning and significance of variability. The main properties and limitation of the range, quartile deviation, average deviation and standard deviation that are the measures of variability will also be discussed. Further, the concept of variance and coefficient of variance will also be highlighted.

10.2 CONCEPT OF VARIABILITY IN DATA

Variability in statistics means deviation of scores in a group or series, from their mean scores. It actually refers to the spread of scores in the group in relation to the mean. It is also known as dispersion. For instance, in a group of 10 participants who have scored differently on a mathematics test, each individual varies from the other in terms of the marks that he/she has scored. These variations can be measured with the help of measure of variability, that measure the dispersion of different values for the average value or average score. Variability or dispersion also means the scatter of the values in a group. High variability in the distribution means that scores are widely spread and are not homogeneous. Low variability means that the scores are similar and homogeneous and are concentrated in the middle.

According to Minium, King and Bear (2001), measures of variability express quantitatively the extent to which the score in a distribution scatter around or cluster together. They describe the spread of an entire set of scores, they do not specify how far a particular score diverges from the centre of the group. These measures of variability do not provide information about the shape of a distribution or the level of performance of a group.

Measures of variability fall under descriptive statistics that describe how similar a set of scores are to each other. The greater the similarity of the scores to each other, lower would be the measure of variability or dispersion. The less the similarity of the scores to each other, higher will be the measure of variability or dispersion. In general, the more the spread of a distribution,

larger will be the measure of dispersion. To state it succinctly, the variation between the data values in a sample is called dispersion. The most commonly used measures of dispersion are the range, and standard deviation.

In the previous unit, measures of central tendency were discussed. While measures of central tendencies are indeed very valuable, their usefulness is rather limited. Although through these measures we can compare the two or more groups, a measure of central tendency is not sufficient for the comparison of two or more groups. They do not show how the individual scores are spread out. Let us take another example, similar to the one that we discussed under the section on introduction. A math teacher is interested to know the performance of two groups (A and B) of his /her students. He/she gives them a test of 40 points. The marks obtained by the students of groups A and B in the test are as follows:

Marks of Group A: 5,4,38,38,20,36,17,19,18,5 (N = 10, Total = 200, Mean = 20)

Marks of Group B: 22,18,19,21,20,23,17,20,18,22 (N = 10, Total = 200, Mean = 20)

The mean scores of both the groups is 20, as far as mean goes there is no difference in the performance of the two groups. But there is a difference in the performance of the two groups in terms of how each individual student varies in marks from that of the other. For instance, the test scores of group A are found to range from 5 to 38 and the test scores of group B range from 18 to 23.

It means that some of the students of group A are doing very well, some are doing very poorly and performance of some of the students is falling at the average level. On the other hand, the performance of all the students of the second group is falling within and near about the average (mean) that is 20. It is evident from this that the measures of central tendency provide us incomplete picture of a set of data. It gives insufficient base for the comparison of two or more sets of scores. Thus, in addition to a measure of central tendency, we need an index of how the scores are scattered around the center of the distribution. In other words, we need a measure of dispersion or variability. A measure of central tendency is a summary of scores, and a measure of dispersion is summary of the spread of scores. Information about variability is often as important as that about the central tendency.

The term variability or dispersion is also known as the *average of the second degree*, because here we consider the arithmetic mean of the deviations from the mean of the values of the individual items. To describe a distribution adequately, therefore, we usually must provide a measure of central tendency and a measure of variability. Measures of variability are important in statistical inference. With the help of measures of dispersion, we can know about fluctuation in random sampling. How much fluctuation will occur in random sampling? This question in fundamental to every problem in statistical inference, it is a question about variability.

The measures of variability are important for the following purposes:

- Measures of variability are used to test the extent to which an average represents the characteristics of a data. If the variation is small then it indicates high uniformity of values in the distribution and the average represents the characteristics of the data. On the other hand, if variation is large then it indicates lower degree of uniformity and unreliable average.

- Measures of variability help in identifying the nature and cause of variation. Such information can be useful to control the variation.
- Measures of variability help in the comparison of the spread in two or more sets of data with respect to their uniformity or consistency.
- Measures of variability facilitate the use of other statistical techniques such as correlation, regression analysis, and so on.

10.2.1 Functions of Variability

The major functions of dispersion or variability are as follows:

- It is used for calculating other statistics such as analysis of variance, degree of correlation, regression etc.
- It is also used for comparing the variability in the data obtained as in the case of Socio-Economic Status, income, education etc.
- To find out if the average or the mean/median/mode worked out is reliable. If the variation is small then we could state that the average calculated is reliable, but if variation is too large, then the average may be erroneous.
- Dispersion gives us an idea if the variability is adversely affecting the data and thus helps in controlling the variability.

10.2.2 Absolute Dispersion and Relative Dispersion

Measures of dispersion give an estimate and express quantitatively the deviation of individual scores in a given sample from the mean and median. Thus, the numerical measures of variability spread or scatter around a central value.

In measuring dispersion, it is imperative to know the amount of variation (absolute measure) and the degree of variation (relative measure). In the former case, we consider the range, quartile deviation, mean deviation, standard deviation etc. In the latter case, we consider the coefficient of range, the coefficient of mean deviation, the coefficient of variation etc. Thus, there are two broad classes of the measures of dispersion or variability. They are absolute measure of dispersion and relative measure of dispersion.

Absolute dispersion usually refers to the standard deviation, a measure of variation from the mean. The units of standard deviation are the same as for the data. In other words, absolute measure is expressed in terms of the original units of a distribution. Therefore, absolute dispersion is not suitable for comparing the variability of two distributions since the two variables are expressed and measured in two different units. For instance, the variability in body height (cm) and body weight (kg) cannot be compared because the absolute measure (standard deviation) is expressed in cm and kg. The absolute measure is also not appropriate for two sets of scores expressed in the same units with wide divergence in means (central value). Nevertheless, absolute measures are widely used, except in the exceptional cases like above. The absolute measures include range, mean deviation, standard deviation, and variance.

Relative dispersion, sometimes called the coefficient of variation, is the result of dividing the standard deviation by the mean and it may be presented as a quotient or as a percentage. Thus, relative measures are computed from the

absolute measures of dispersion and its corresponding central values. A low value of relative dispersion usually implies that the standard deviation is small in comparison to the magnitude of the mean. To give an example, if standard deviation for mean of 30 marks is 6.0, then the coefficient of variation will be

$$6.0/ 30 = 0.2(\text{about } 20\%)$$

If the mean is 60 marks and the standard deviation remains the same as 6.0, the coefficient of variation will be

$$6.0 / 60 = 0.1 \ (10\%).$$

However, with measurements on either side of zero and a mean being close to zero the relative dispersion could be greater than 1. At the same time, we must remember that the two distributions in quite a few cases can have the same variability. Sometimes the distributions may be skewed and not normal with mean, mode and median at different points in the continuum. These distributions are called skewed distributions (Skewness will be discussed in detail in the unit 8). It is also possible to have two distributions that have equal variability but unequal means or different shapes. Thus, the relative measure is derived from a ratio of an absolute measure like standard deviation and mean (measure of central value) and is expressed in percentage of the mean. So, the relative measure is suitable for comparing the variabilities of two sets of scores given in different units. They are also preferred in comparing two sets of scores given in the same unit, when the mean widely diverges. The relative measures include the coefficient of variation, the coefficient of quartile deviation, and the coefficient of mean deviation.

Check Your Progress I

1) State any one function of variability.

.....

.....

.....

.....

.....

2) List the two broad classes of the measures of dispersion.

.....

.....

.....

.....

.....

10.3 TYPES OF MEASURES OF DISPERSION OR VARIABILITY

The measures of variability most commonly used in psychological statistics are as follow:

1) Range

2) Quartile Deviation

3) Average Deviation or Mean Deviation

4) Standard Deviation

5) Variance

Range and quartile deviation measure dispersion by computing the spread within which the values fall, while as average deviation and standard deviation compute the extent to which the values differ from the average. We will introduce each and discuss their properties in detail.

10.3.1 The Range (R)

Range can be defined as the difference between the highest and lowest score in the distribution. This is calculated by subtracting the lowest score from the highest score in the distribution. The equation is as follows:

Range = Highest Score – Lowest Score(R=H-L)

The range is a rough measure of dispersion because it tells about the spread of the extreme scores and not the spread of any of the scores in between. For instance, the range for the distribution 4,10,12,20, 25, 50 will be 50 - 4 = 46.

10.3.1.1 Merits and Limitations of the Range

Some of the merits of range as a measure of variability are explained in this section.

1) It is easiest to compute when compared with other measures of variability and its meaning is direct.

2) The range is ideal for preliminary work or in other circumstances where precision is not an important requirement (Minium et. al., 2001).

3) It is quite useful in case where the purpose is only to find out the extent of extreme variation, such as temperature, rainfall etc.

4) Range is effectively used in the application of tests of significance with small samples.

The following are the limitations of the range as a measure of variability:

1) The calculation of range is based only on two extreme values in the data set and does not consider other values of the data set. Sometimes, the extreme values of the two different data sets may be same or similar, but the two data sets may be differ in dispersion.

2) Its value is sensitive to change in sampling. The range varies more with sampling fluctuation. That is different sample of the same size from the same population may have different range.

3) Its value is influenced by large samples. In many types of distribution, including normal distribution, the range is dependent on sample size. The sampling variance increases rapidly with increase in sample size.

4) Range cannot be used for open-ended class intervals since the highest and the lowest scores of the distribution are not available and thus the range cannot be computed.

5) Further mathematical calculations are not possible for range.

6) Range indicates two extreme scores, thus the magnitude or frequency of intermediate scores is missing.

7) It does not indicate the form of distribution, like skewness, kurtosis, or modal distribution of scores.

8) A single extreme score may also increase the range disproportionately.

10.3.1.2 Uses of the Range

Range is applied in diverse areas discussed as follows:

- Range is used in areas where there are small fluctuations, such as stock market, rate of exchange, etc.
- Range may be used in day-to-day activities like, daily sales in a grocery store, monthly wages in a factory, etc.
- Range is used in weather forecasts, like variation in temperature in a day.
- When the researcher is only interested in the extreme scores or total spread of the scores, range is the most useful measure of variability.
- Range can also be used when the data are too scant or too scattered to justify the use of most appropriate measure of variability.

10.3.2 The Quartile Deviation (QD)

Since a large number of values in the data lie in the middle of the frequencies distribution and range depends on the extreme (outliers) of a distribution, we need another measure of variability. The Quartile deviation, is a measure that depends on the relatively stable central portion of a distribution. According to Garret (1966), the Quartile deviation is half the scale distance between 75^{th} and 25^{th} per cent in a frequency distribution. The entire data is divided into four equal parts and each part contains 25% of the values. According to Guilford (1963) the Semi-Interquartile range is the one half the range of the middle 50 percent of the cases.

On the basis of above definitions, it can be said that quartile deviation is half the distance between Q_1 and Q_3.

Inter Quartile Range (IQR): The range computed for the middle 50% of the distribution is the Inter Quartile Range. The upper quartile (Q_3) and lower quartile (Q_1) is used to compute IQR. This is $Q_3 - Q_1$. IQR is not affected by extreme values.

Semi-Inter Quartile Range (SIQR) or Quartile Deviation (QD): Half of the IQR is called as Semi-Inter Quartile Range. SIQR is also called as quartile deviation or QD. Thus, QD is computed as;

$$QD = Q_3 - Q_1/2$$

Thus, quartile deviation is obtained by dividing IQR by 2. Quartile deviation is an absolute measure of dispersion and is expressed in the same unit as the scores.

Quartile deviation is closely related to the median because median is responsive to the number of scores lying below it rather than to their exact positions and Q_1 and Q_3 are defined in a same manner. The median and quartile deviation have common properties. Both median and quartile deviation are not affected by extreme values. In a symmetrical distribution, the two quartiles Q_1 and Q_3 are at equal distance from the median or $Q_1 = Q_3$- Median. Thus, like median, quartile deviation covers exactly 50 per cent of observed values in the data. In normal distribution, quartile deviation is called the Probable Error or PE. If the distribution is open-class, then quartile deviation is the only measure of variability that is reasonable to compute.

In an asymmetric or skewed distribution, Q_1 and Q_3 are not equidistant from Q_2 or median. In such a distribution, the median of the IQR moves towards the skewed tail. The degree and direction of skewness can be assessed from quartile deviation and the relative distance between Q_1, Q_2 and Q_3.

Kurtosis is proportional to quartile deviation. Smaller the quartile deviation, greater is the concentration of scores in the middle of the distribution, thus making the distribution with high peak and narrow body. The scores that are widely dispersed indicate a large quartile deviation and thus, long IQR. This distribution has a low peak and broad body.

10.3.2.1 Merits and Limitations of Quartile Deviation

From the explanation in the above section, it becomes clear that quartile deviation is easy to understand and compute.

1) Quartile deviation is a better measure of dispersion than range because it takes into account 50 per cent of the data, unlike the range which is based on two values of the data, that is highest value and the lowest value.
2) Secondly, quartile deviation is not affected by extreme scores since it does not consider 25 per cent data from the beginning and 25 per cent from the end of the data.
3) Lastly, quartile deviation is the only measure of dispersion which can be computed from the frequency distribution with open-end class.

Despite the major merits of quartile deviation, there are limitations to it as well.

1) The value of quartile deviation is based on the middle 50 percent values, it is not based on all the observations. Thus, it is not regarded as a stable measure of variability
2) The value of quartile deviation is affected by sampling fluctuation.
3) The value of quartile deviation is not affected by the distribution of the individual values within the intervals of middle 50 percent observed values.

10.3.2.2 Uses of Quartile Deviation

1) The distribution contains few and very extreme scores.
2) When the median is the measure of central tendency.
3) When our primary interest is to determine the concentration around the median.

10.3.3 The Average Deviation (AD) or Mean Deviation (MD)

The two measures of variation, range and quartile deviation, which we discussed in the earlier subsections, do not show how values of the data are scattered about a central value. R and QD attempt to compute spread of values and not compute how far the values are from their average. To measure the variation, as a degree to which values within a data deviate from their mean, we use average deviation.

Before discussing average deviation, first we should know about the meaning of deviation. Deviation score express the location of the scores by indicating how many score points it lies above or below the mean of the distribution. Deviation score may be defined as (X-Mean), that is, when we subtract the means from each of the raw scores the resulting deviation scores states the position of the scores, relative to the mean.

According to Garrett (1971, as cited in Mangal 2002, page 70) "The average deviation is the mean of the deviation of all of the separate scores is a series taken from their mean". According to Guilford (1963) average deviation can be described as an average or mean of all the deviations when the algebraic signs are not taken in to the account.

Average is a central value and thus, some deviations will be positive (+) and some may be negative (-). Mean deviation ignores the signs of the deviations, and it considers all the deviations to be positive. This is so because the algebraic sum of all the deviations from the mean equals to zero. AD or MD is arithmetic mean of the difference of the values from the average. The average is either the arithmetic mean or the median. It is a measure of variability that takes into account the variations of all the scores in the data. It is an absolute measure of dispersion and is expressed in the same unit as the raw scores.

The calculation of average deviation is easy therefore it is a popular measure. When we calculate average deviation, equal weight is given to each observed value and thus it indicates how far each observation lies from the mean. AD or MD can be obtained from any of the measures of central tendency, that is mean, median, or mode. Mode is ill defined, hence, AD or MD is computed about the mean or median. AD or MD calculated about the median will be less than the AD or MD about the mean or mode. For a symmetrical distribution, MD about mean and MD about median covers 57.5 per cent of the observations of the data. Thus, a small value of MD will indicate less variability. AD is thus somewhat larger (57.5 per cent of the cases) than QD (50 per cent of the cases).

10.3.3.1 Merits and Limitations of the Average Deviation

The main merits of AD are as follows:

1) AD or MD is easy to understand and compute.

2) It is based on all observations, unlike R or QD.

3) It is an accurate measure of variability since it averages the absolute deviations.

4) It is less affected by extreme observations.

5) It is based on average thus, it is a better measure to compare about the formations of different distributions.

The main limitations of average deviation are as follows:

1) While calculating average deviation we ignore the plus minus sign and consider all values as plus. Because of this mathematical property, it is not used in inferential statistics.

2) AD cannot be computed for open-end classes.

3) It tends to increase with the size of the sample.

10.3.3.2 Use of Average Deviation

Despite the limitations, AD or MD is used by economists and business statisticians. It is also used in computing the distribution of personal wealth in a community or a nation. According to National Bureau of Economic Research, MD is the most practical measure of dispersion to be used for this purpose (Mohanty and Misra, 2016, pg. 133).

1) When it is desired to weight all deviation from the mean according to their size.

2) When the standard deviation in unduly influenced by the presence of extreme scores.

3) Distribution of the score is not near to normal.

10.3.4 The Standard Deviation (SD)

The term standard deviation was first used in writing by Karl Pearson in 1894. The standard deviation of population is denoted by 'σ' (Greek letter sigma) and that for a sample is 's'**.** A useful property of SD is that unlike variance it is expressed in the same unit as the data. This is most widely used method of variability. The standard deviation indicates the average of distance of all the scores around the mean. It is the positive square root of the mean of squared deviations of all the scores from the mean. It is the positive square root of variance. It is also called as 'root mean square deviation'. Mangal (2002, page 71) defined standard deviation as "as the square root of the average of the squares of the deviations of each score from the mean". SD is an absolute measure of dispersion and it is the most stable and reliable measure of variability.

Standard deviation shows how much variation there is, from the mean. SD is calculated from the mean only. If standard deviation is low it means that the data is close to the mean. A high standard deviation indicates that the data is spread out over a large range of values. Standard deviation may serve as a measure of uncertainty. If you want to test the theory or in other word, want to decide whether measurements agree with a theoretical prediction, the standard deviation provides the information. If the difference between mean and standard deviation is very large then the theory being tested probably needs to be revised. The mean with smaller standard deviation is more reliable than mean with large standard deviation. A smaller SD shows the homogeneity of the data. The value of standard deviation is based on every observation in a set of data. It is the only measure of dispersion capable of algebraic treatment therefore, SD is used in further statistical analysis.

10.3.4.1 Merits and Limitations of the Standard Deviation

The main merits of using standard deviation are as follows:

1) It is widely used because it is the best measure of variation by virtue of its mathematical characteristics.

2) It is based on all the observations of the data.

3) It gives an accurate estimate of population parameter when compared with other measures of variation.

4) SD is least affected by sample fluctuations

5) It is also possible to calculate combined SD, that is not possible with other measures.

6) Further statistics can be applied on the basis of SD like, correlation, regression, tests of significance, etc.

7) Coefficient of variation is based on mean and SD. It is the most appropriate method to compare variability of two or more distributions.

The limitations of SD are as follows:

1) While calculating standard deviation more weight is given to extreme values and less to those, near the means. When we calculate SD, we take deviation from mean (X-M) and square these obtained deviations. Therefore, large deviations, when squared are proportionally more than small deviations. For example, the deviations 2 and 10 are in the ratio of 1:5 but their square 4 and l00 are in the ratio 1:25.

2) It is difficult to compute as compared to other measures of dispersion.

10.3.4.2 Uses of Standard Deviation

The uses of standard deviation are as follows:

1) SD is used when one requires a more reliable and accurate measure of variability but it is recommended when the distribution is normal or near to normal.

2) It is used when further statistics like, correlation, regression, tests of significance, etc. have to be computed.

10.3.5 Variance

The term variance was used to describe the square of the standard deviation by R.A. Fisher in 1913. The concept of variance is of great importance in advanced work where it is possible to split the total into several parts, each attributable to one of the factors causing variations in their original series. Variance is a measure of the dispersion of a set of data points around their mean value. It is a mathematical expectation of the average squared deviations from the mean. The variance (s^2) or mean square (MS) is the arithmetic mean of the squared deviations of individual scores from their means. In other words, it is the mean of the squared deviation of scores.Variance is expressed as $V = SD^2$.

The variance and the closely related standard deviation are measures that indicate how the scores are spread out in a distribution. In other words, they are

measures of variability. The variance is computed as the average squared deviation of each number from its mean.

Calculating the variance is an important part of many statistical applications and analysis. It is a good absolute measure of variability and is useful in computation of Analysis of Variance (ANOVA) to find out the significance of differences between sample means.

10.3.5.1 Merits and Demerits of Variance

The main merits of variance are listed as follows:

1) It is rigidly defined and based on all observations.

2) It is amenable to further algebraic treatment.

3) It is not affected by sampling fluctuations.

4) It is less erratic.

The main demerits of variance are listed as follows:

1) It is difficult to understand and calculate.

2) It gives greater weight to extreme values.

10.3.5.2 Co-efficient of Variation (CV)

The relative measure corresponding to SD is the coefficient of variation. It is a relative measure of dispersion developed by Karl Pearson. When we want to compare the variations (dispersion) of two different series, relative measures of standard deviation must be calculated. This is known as coefficient of variation or the coefficient of SD. It is defined as the SD expressed as a percentage of the mean.The coefficient of variation represents the ratio of the standard deviation to the mean, and it is a useful statistic for comparing the degree of variation from one data series to another, even if the means are drastically different from each other. Thus, it is more suitable than SD or variance. It is given as a percentage and is used to compare the consistency or variability of two or more data series.

The formula for computing coefficient of variation is as follows:

$V = 100 \times \sigma / M$

Where,

V = Variance

σ = Standard deviation

M = Mean

To understand the computation with the help of an example,

If the standard deviation of marks obtained by 10 students in a class test in English is 10 and Mean is 79, then,

$V = 100 \times 10 / 79$

$= 1000 / 79$

$= 12.65$

Check Your Progress II

1) What is range?

.....

.....

.....

.....

.....

.....

2) List the merits of quartile deviation.

.....

.....

.....

.....

.....

.....

3) What is variance?

.....

.....

.....

.....

.....

.....

10.4 LET US SUM UP

To summarise, the measures of central tendency are not sufficient to describe data. Thus, to describe distribution adequately, we must provide a measure of variability or dispersion.The measures of variability are summary figures that express quantitatively, the extent to which, scores in a distribution scatter around or cluster together. The measures of variability are range, quartile deviation, average deviation, standard deviation and variance. Range is easy to calculate and useful for preliminary work. But this is based on extreme items only, and does not consider intermediate scores. Thus, it is not useful as a descriptive measure. Quartile deviation is related to the median in its

properties. It takes into consideration the number of scores lying above or below the outer quartile point but not to their magnitude. This is useful with open ended distribution.The average deviation takes into account the exact position of each score in the distribution. The means deviation gives a more precise measure of the spread of scores but is mathematically inadequate. The average deviation is less affected by sampling fluctuation.The standard deviation is the most stable measure of variability. Standard deviation shows how much the score departs from the mean. It is expressed in original scores unit. Thus, it is most widely used measure of variability in descriptive statistics.The variance (s^2) or mean square (MS) is the arithmetic mean of the squared deviations of individual scores from their means. In other words, it the mean of the squared deviation of scores.The relative measure corresponding to SD is the coefficient of variation. It is a useful measure of relative variation.

10.5 REFERENCES

Garrett, H.E. (1981), *Statistics in Psychology and Education*, (Tenth edition), Bombay, Vakils Feffer and Simons Ltd.

McBride, Dawn M. (2018). *The Process of Statistical Analysis in Psychology.* Sage. USA

Minium, E.W., King, B.M. & Bear. G (2001). *Statistical Reasoning in Psychology and Education* (3rd edition), Singapore, John Wiley & Sons, Inc.

Mohanty, B. & Misra, Santa (2016). *Statistics for Behavioural and Social Sciences.* Sage. New Delhi.

10.6 KEY WORDS

Average Deviation or Mean Deviation: A measure of dispersion that gives the average difference (ignoring plus and minus sign) between each item and the mean.

Dispersion: The spread or variability is a set of data.

Deviation: The difference between raw score and mean.

Quartile Deviation: A measure of dispersion that can be obtained by dividing the difference between Q_3 and Q_1 by two.

Range: Difference between the largest and smallest value in a data.

Standard deviation: The square root of the variance in a series.

Variance: Variance is a measure of the dispersion of a set of data points around their mean value. It is a mathematical expectation of the average squared deviations from the mean.

10.7 ANSWERS TO CHECK YOUR PROGRESS

Check Your Progress I

1) State any one function of variability

   Variability is used for calculating other statistics such as analysis of variance, degree of correlation, regression etc.

2) List the two broad classes of the measures of dispersion.

Absolute dispersion

Relative dispersion.

Check Your progress II

1) What is range?

Range can be defined as the difference between the highest and lowest score in the distribution.

2) List the merits of quartile deviation.

The merits of quartile deviation are as follows:

- Quartile deviation is a better measure of dispersion than range because it takes into account 50 percent of the data, unlike the range which is based on two values of the data, that is highest value and the lowest value.
- Secondly, quartile deviation is not affected by extreme scores since it does not consider 25 percent data from the beginning and 25 percent from the end of the data.
- Lastly, quartile deviation is the only measure of dispersion which can be computed from the frequency distribution with open-end class.

3) What is variance?

Variance is a measure of the dispersion of a set of data points around their mean value. It is a mathematical expectation of the average squared deviations from the mean. The variance (s^2) or mean square (MS) is the arithmetic mean of the squared deviations of individual scores from their means. In other words, it is the mean of the squared deviation of scores.

10.8 UNIT END QUESTIONS

1) Explain the concept and significance of variability.

2) Discuss the merits and limitation of range and quartile deviation.

3) List the merits and limitations of standard deviation.

4) Elucidate average deviation or mean deviation.

5) Explain coefficient of variance with example.

UNIT 11 COMPUTATION OF MEASURES OF VARIABILITY*

Structure

11.0 OBJECTIVES

After reading this unit, you will be able to:

- compute range;
- compute quartile deviation, for ungrouped and grouped data;
- compute average deviation, for ungrouped and grouped data; and
- compute standard deviation, for ungrouped and grouped data and with the help of short method.

11.1 INTRODUCTION

In the previous two units, we discussed about measures of central tendency and measures of variability. We discussed that average like mean, median and mode condense the series into a single figure. These measures of central tendency tell us something about the general level of magnitude of the distribution but they fail to show anything further about the distribution. It is not fully representative of a population unless we know the manner in which the individual items scatter around it. A further description of the series is necessary if we are to gauge how representative the average is. To cite an

* Dr. Usha Kulshreshtha, Faculty, Psychology, University of Rajasthan, Jaipur.

example, in a country the average income may be very high, yet there may be great disparity in its distribution among people. As a result, a majority of the people may be living below the poverty line. When we want to make comparison between two groups, it is seen that at times the value of the means is the same in both the groups but there is a large difference between individual participants in the groups. This difference amongst the participants within the same group is termed as variation, that is within the groups the participants vary a great deal even though they have the same means. Therefore to make accurate and meaningful comparisons between groups, we should use variability along with central tendency. In the last unit we have learned about the concept of variability, different measures of variability, their merits, limitations and uses. In this unit, we will learn how to compute the range, quartile deviation, average deviation and standard deviation.

11.2 COMPUTING DIFFERENT MEASURES OF DISPERSION (VARIABILITY)

As explained in the last unit that there are four measures of computing variability or dispersion within a set of scores:

1) Range(R)
2) Quartile Deviation (QD)
3) Average Deviation (AD) or Mean Deviation (MD)
4) Standard Deviation (SD)

Each of the above measures of variability give us the degree of variability or dispersion by the use of a single number and tells us how the individual scores are spread throughout the distribution. In the following sections, we will discuss the methods of computation of the above measures of dispersion.

11.2.1 Range(R)

Range is the difference between the highest and the lowest score for a group of participants whose scores are given.

The formula for Range is as follows:

R=H-L

Where,

H=Highest scores in the distribution

L=Lowest score in the distribution

Let us understand the steps in computation of range with the help of an example,

For example, if there are 10 students who have obtained marks in history as mentioned below:

50, 45, 42, 46, 55, 54, 59, 60, 62, 64

Step 1: Arrange the scores in ascending order.

42, 45, 46, 50, 54, 55, 59, 60, 62, 64

Step 2: Identify the lowest and the highest score in the data

In the above data, the lowest score is 42 and the highest score is 64.

Step 3: Compute range with the help of the following formula:

R=H-L

64–42=22.

Thus, the range obtained is 22.

11.2.2 Quartile Deviation (QD)

The inter quartile range is a measure of dispersion and is equal to the difference between the third and first quartiles. Half of the inter quartile range is called semi interquartile range or Quartile Deviation. The formula for computation of QD is

$QD = Q_3 – Q_1/2$

Where,

Q_1 = first quartile of the data

Q_3= third quartile of the data

Quartile is an additional way of disaggregating data. Each quartile represents one-fourth of an entire population or the group. The quartile deviation has an attractive feature that the range "median+QD" contains approximately 50% of the data. The quartile deviation is also an absolute measure of dispersion. Its relative measure is called coefficient of quartiled deviation or semi-inter quartile range. It is defined by the relation;

Coefficient of quartile deviation = $(Q_3-Q_1) / (Q_3+Q_1)$

The quartile deviation is one half the scale distance between the 75^{th} and 25^{th} percentile in a frequency distribution. The 25^{th} percentile or Q_1 is the first quartile on the score scale, the point below which lies 25% of the scores.The 75^{th} percentile or Q_3 is the third quartile on the score scale, the point below which lie 75% of the scores. To find quartile deviation, we must first compute Q_3 and Q_1.

There are grouped data and ungrouped data in all cases and thus to compute quartile deviation, we have to find out first if it is a grouped data ore ungrouped data. We will first see how the quartile deviation is calculated from ungrouped data.

11.2.2.1 Calculation of Quartile Deviation for Ungrouped Data

Let us understand steps in computation of quartile deviation for ungrouped data with the help of the following example:

The scores obtained by students in Psychology class test are :

24, 25, 23, 26, 29, 30, 27, 35, 34, 36, 28

Step 1: Arrange the data in ascending order.

23, 24, 25, 26, 27, 28, 29, 30, 34, 35, 36

Step 2: Compute Q_1

$Q_1 = (N+1)/4^{th}$ position

N = 11

$Q_1 = 11+1/4^{th}$ position = 3rd position = 25

Step 3: Compute Q_3

$Q_3 = 3(N+1)/4^{th}$ position

N = 11

$Q_3 = 3(11+1)/4 = 9^{th}$ position = 34

Step 4: Compute QD with help of the following formula

$QD = Q_3 - Q_1/2$

In the case of our data Q_3 is 34 and $Q_{1/2}$ is 25

QD =34–25/2

= 9/ 2

= 4.5

11.2.2.2 Calculation of Quartile Deviation for Grouped Data

From the grouped data, Quartile Deviation can be computed with the following formula:

$QD = Q_3 - Q_1/2$

Further,

$Q_1 = l+i[(N/4\text{-cum } fi)]/fq$

$Q_3 = l+i[(3N/4\text{-cum } fi)]/fq$

Where,

l = the exact lower limit of the interval in which the quartile falls.

i = the length of the interval

cum fi = cumulative f up to interval which contains the quartile

fq = the f on the interval containing the quartile.

Let us understand steps in computation of quartile deviation for grouped data with the help of the following example:

Class intervals	Frequencies (*f*)	Cumulative frequencies
195-199	1	50
190-194	2	49
185-189	4	47
180-184	5	43
175-179	8	38
170-174	10	30
165-169	6	20
160-164	4	14
155-159	4	10(1+3+2+4)
150-154	2	6 (1+3+2)
145-149	3	4 (1+3)
140-144	1	1

Step 1: Q_1 is calculated using the formula given below:

$Q_1 = l+i[(N/4–cum fi)]/fq$

To locate the Q_1, we take N/4. In the above example N/4=12.5

l = 159.5 (50/4 = 12.5 th item from down below) (falls in 160-164)

fi = 10 cumulated scores upto interval containing Q_1

fq = 4, the f on the interval on which Q_1 falls

i = 5 (Class Interval)

Substituting in formula we have,

$Q_1=159.5+5\{(12.5-10)\}/4=162.62$

Step 2: To calculate the third quartile, that is Q_3.

$Q_3 = l+i\,[(3N/4–cum\,fi)]/fq$

To locate the Q_3 we take 3xN/4 of our scores. In the above example, 3N/4 is 3×50/4 = 37.5

3/4N = 37.5 (37.5th item that falls in 175-179)

l = 174.5 is the exact lower limit of interval which contains Q_3

Cum fi = 30, sum of scores upto interval which contains Q_3

i=5

fq=8

$Q_3=174.5+5(37.5-30)/8 =179.19$

Step 3: Finally, substituting in formula, we have the quartile deviation as given below in the formula;

QD = $Q_3 - Q_2 / 2$

Q = {(179.19)–-(162.62)}/2

= 8.28

Thus, quartile deviation for the above data is 8.28.

11.2.3 Average Deviation (AD)

Average deviation as well can be computed for ungrouped and grouped data.

11.2.3.1 Computation of Average Deviation for Ungrouped Data

In case of ungrouped data, the average deviation is calculated by the following formula,

AD=∑ | x | /N

Where,

∑ | x | = Total of deviation from mean

N= Total number of observations

Let us understand the steps in computation of average deviation for ungrouped data with the help of example given below:

Following are the scores obtained by five students on a test:

Students	Scores	Deviation from the Mean | x |
1	6	-4
2	8	-2
3	10	0
4	12	2
5	14	4
	Total=50	**Total=12(Ignore signs)**
	Mean=50/5=10	

Step 1: Compute mean with the help of formula M=∑X/N

M=50/5 =10

Step 2: Compute deviation from the mean as has been computed in the third column in above example.

It is seen from the table, that the deviations (x) are =0,-2,-4, +2, +4

Step 3: Compute total for these deviations without considering the + and – signs. the total is obtained as 12.

Step 4: Use the formula to compute average deviation.

AD = ∑ | x | /N

= 12/5

=2.4.

Thus, the average deviation is obtained as 2.4.

11.2.3.2 Calculation of Average Deviation for Grouped Data

The average deviation for grouped data can be computed by the following formula,

AD= $\sum |fx| /N$

Where,

$\sum |fx|$ =Add all the fx without considering the + and – sign

N= Number of observations

The above formula and calculation of AD can be illustrated by the example given below.

Let us understand the steps in computation of average deviation for grouped data with the help of example given below:

Class Interval	**Frequency (*f*)**	**Mid Point (X)**	***f*X**	**x=M-X**	***f*x**
(1)	(2)	(3)	(4)	(5)	(6)
110-114	3	112	336 (112×3)	20.33 (112-91.67=20.33)	60.99 (20.33×3)
105-109	4	107	428	15.33	61.32
100-104	6	102	612	10.33	61.98
95-99	8	97	776	5.33	42.64
90-94	15	92	1380	.33	4.95
85-89	10	87	870	–4.67	–46.67
80-84	7	82	574	–9.67	–67.69
75-79	4	77	308	–14.67	–58.68
70-74	3	72	216	–19.67	–59.01
	Total= 60		**Total= 5500**		**Total = 463.93**
			Mean=91.67		

Step 1: Identify midpoints of the class interval and mention them in column three, as can be seen above.

Step 2: Multiply respective frequencies and mid-points as shown in column 4.

Step 3: For obtained *f*X, compute mean with the help of formula M=$\sum$*f*X/N.

M=5500/60

=91.67

Step 4: x is then computed with the formula x=M–X.

Step 5: Average deviation is then computed with the help of following formula:

$AD=\sum |fx| /N$

$=463.93/60$

$=7.73$

Thus, the average deviation obtained is 7.73.

11.2.4 The Standard Deviation (SO)

Standard Deviation is the most stable measure of variability. Therefore, it is most commonly used in research studies. Standard deviation can be computed for ungrouped and grouped data.

11.2.4.1 Calculation of Standard Deviation (SD) for Ungrouped Data

Standard deviation for ungrouped data can be computed by the following formula:

$\mathbf{SD} = \sqrt{\sum x^2/N}$

The above formula can be explained by the following example.

Let us understand the steps in computation of standard deviation for ungrouped data with the help of example given below:

Scores (X)	Deviation from the mean (x)	Deviation square (x^2)
52	-8	64
50	-10	100
56	-4	16
68	8	64
65	5	25
62	2	4
57	-3	9
70	10	100
Total = 480		$\sum x^{2=}$ **382**
Mean= 60		

Step 1: Add all the scores ($\sum X$) and divide this sum by the number of scores (N) and find out mean. Mean of the given scores is

$M = \sum X/N$

$=480/8$

$=60.$

Step 2: Find out deviation x by computing X – x, as given in second column above.

Step 3: Square all the deviation to get x^2.

Step 4: Add all the squared deviation to get $\sum x^2$.

Step 5: Compute standard deviation with the help of the formula:

SD= $\sqrt{(\sum x^2)}$ **/N**

$= \sqrt{382/8}$

$= \sqrt{47.7}$

=6.91

Thus, the standard deviation for this data is 6.91.

11.2.4.2 Computations of SD for Grouped Data by Long Method

Standard deviation of grouped data can be computed by the formula,

SD= $\sqrt{\sum fx^2}$**/N**

Where,

$\sum fx^2$= when frequencies (*f*) are multiplied with their respective devation squared (x^2), fx^2 is obtained total of all fx^2 is fx^2

N= Total number of scores

Let us understand the steps in computation of standard deviation for grouped data with the help of example given below:

Class interval	Frequency (*f*)	Midpoint (X)	*f*X	Deviation of midpoint from the mean (x= X- M)	Deviation squared (x^2)	fx^2
(1)	(2)	(3)	(4)	(5)	(6)	(7)
127-129	1	128	128	17.6	309.76	309.76
124-126	2	125	250	14.6	213.16	426.32
121-124	2	122	244	11.6	134.56	269.12
118-120	2	119	238	8.6	73.96	147.92
115-117	4	116	464	5.6	31.36	125.44
112-114	4	113	452	2.6	6.76	27.04
109-111	4	110	440	–0.4	0.16	0.64
106-108	2	107	214	–3.4	11.56	23.12
103-106	2	104	208	–6.4	40.96	81.92
100-102	2	101	202	–9.4	88.36	176.72
	Total= 25		**$\sum fx$= 2760**			**$\sum fx^2$= 1588**

Step 1: Midpoint is computed for respective class intervals and entered in column 3 as shown in the above table.

Step 2: *f*X is then computed by multiplying the frequencies and the midpoint, the values thus obtained are entered in column 4.

Step 3: Add all the scores under *f*X and divide this sum by the number of scores (N) and find out mean. Thus, M= $\sum f$X/ N, that is, 2760/ 25= 110.4. The mean obtained is 110.4.

Step 4: x is now computed by subtracting the M from X (midpoint). The values thus obtained are entered in column 5.

Step 5: The x is then squared to obtain x^2 (column 6).

Step 6: fx^2 is then computed by multiplying f and x^2. The values thus obtained are entered in column 7.

Step 7: Add the fx^2 to obtain $\sum fx^2$. In the present example it is obtained as 1588.

Step 8: Compute standard deviation with help of the formula:

$$\textbf{SD} = \sqrt{\sum fx^2/N}$$

$$= \sqrt{1588/25}$$

$$= \sqrt{63.52}$$

$$= 7.97$$

The standard deviation thus obtained is 7.97.

11.2.4.3 Calculation of SD for Grouped Data by Short Method

Standard deviation from grouped data can also be computed by the following formula.

$$\textbf{SD} = i \sqrt{\sum fx'^2/N - (\sum (fx')/N)^2}$$

where,

i= The size of the class interval

$\sum fx'^2$= when frequencies are multiplied with their respective deviations of midpoint from the mean fx' is obtained square of fx' is fx'^2 and total of all the fx'^2 is fx'^2

Let us understand the steps in computation of standard deviation for grouped data by short method with the help of example given below:

Class interval	Frequency (f)	Midpoint (X)	Deviation of midpoint from the mean ($x' = X - AM/i$)	(fx')	fx'^2
(1)	(2)	(3)	(5)	(6)	(7)
127-129	1	128	4	4	16
124-126	2	125	3	6	36
121-124	2	122	2	4	16
118-120	2	119	1	2	4
115-117	4	116	0	0	0
112-114	4	113	-1	-4	16
109-111	4	110	-2	-8	64
106-108	2	107	-3	-6	36
103-106	2	104	-4	-8	64
100-102	2	101	-5	-10	100
	Total= 25			**$\sum fx' = -20$**	**$\sum fx'^2 = 352$**

Step 1: Find out midpoint of each class interval, that is entered in column 3.

Step 2: Assume one value as mean. In this example, assumed mean is taken as 116.

Step 3: Find out the difference between midpoint and assumed mean and divide it by class intervals to get x′, (x′= X- AM/i) and enter the obtained value in column 5.

Step 4: Multiply each x′ by respective frequency and get fx′ (column 6).

Step 5: Multiply *f*x′ is squared to get *f*x′^{2}. Add all the *f*x′^{2} to obtain ∑*f*x′.

Step 6: Compute standard deviation with the help of the formula

SD= i $\sqrt{\sum fx'^2/N-(\sum(fx')/N)^2}$

$= 3\sqrt{352/25- (-20/25)^2}$

$= 3\sqrt{14.08 - (-0.8)^2}$

$= 3\sqrt{14.08 - 0.64}$

$= 3\sqrt{3.44}$

$= 3 \times 1.85$

$= 5.55$

The standard deviation thus obtained is 5.55.

Check Your Progress 1

1) What is the formula for range?

2) List the steps in computation of standard deviation for ungrouped data.

.....

.....

.....

.....

.....

11.3 LET US SUM UP

In this unit, we covered the computation of different measures of variability, like the range, average deviation, quartile deviation and standard deviation. Computation of each of the measures of dispersion was discussed with the help of steps and examples.

11.4 REFERENCES

Garrett, H.E. (1981), *Statistics in Psychology and Education*, (Tenth edition), Bombay, Vakils Feffer and Simons Ltd.

McBride, Dawn M. (2018). *The Process of Statistical Analysis in Psychology.*

Sage. USA

Minium, E.W., King, B.M. & Bear. G (2001). *Statistical Reasoning in Psychology and Education* (3rd edition), Singapore, John Wiley &Sons, Inc.

Mohanty, B. & Misra, Santa (2016). *Statistics for Behavioural and Social Sciences.* Sage. New Delhi.

11.5 ANSWERS TO CHECK YOUR PROGRESS

Check Your Progress 1

1) What is the formula for range?

R=H-L

2) List the steps in computation of standard deviation for ungrouped data.

The steps in computation of standard deviation for ungrouped data are

Step 1: Add all the scores ($\sum x$) and divide this sum by the number of scores (N) and find out mean.

Step 2: Find out deviation x by computing X- x.

Step 3: Square all the deviation to get x^2.

Step 4: Add all the squared deviation to get $\sum x^2$.

Step 5: Compute standard deviation with the help of the formula.

11.6 UNIT END QUESTIONS

1) Compute the range, average deviation and standard deviation from the following ungroupeddata:

a) 30, 35, 36, 39, 42, 46, 38, 34, 35

b) 52, 50, 56, 68, 65, 62, 57, 70

2) Calculate quartile deviationforthefollowingscores:

6, 3, 9, 9, 5, 7, 9, 6, 8, 4, 8, 5, 7, 9, 3, 2, 9, 5, 7

3) Calculate Average deviation of the following scores:

Class Interval	Frequency
40-44	3
35-39	4
30-34	6
25-29	12
20-24	7
15-19	5
10-14	1

4) Calculate the Quartile deviation and Standard Deviation for the following frequency distribution:

Scores	Frequency
70-71	2
68-69	2
66-67	3
64-65	4
62-63	6
60-61	7
58-59	5
56-57	1
54-55	2
52-53	3
50-51	1

210 blank