BPCC-104

STATISTICAL METHODS FOR PSYCHOLOGICAL RESEARCH-I

School of Social Sciences
Indira Gandhi National Open University

EXPERT COMMITTEE

COURSE PREPARATION TEAM

	Block	Unit Writer
Block 1	**Introduction**	
Unit 1	Introduction to Statistics	Prof. Suhas Shetgovekar
Unit 2	Data Organisation and Graphical Representation	Dr. Vijay Viegas (Edited by Dr. Arti Singh, Academic Associate, Discipline of Psychology, SOSS)
Block 2	**Measures of Central Tendency and Variability**	
Unit 3	Introduction to Measures of Central Tendency	Prof. Suhas Shetgovekar
Unit 4	Introduction to Measures of Variability	Dr. Usha Kulshreshta adapted by Dr. Monika Misra
Unit 5	Computation of Measures of Variability	Dr. Usha Kulshreshta adapted by Dr. Monika Misra
Block 3	**Correlation**	
Unit 6	Correlation: An Introduction	Prof. Suhas Shetgovekar
Unit 7	Computation of Coefficient of Correlation	Prof. Suhas Shetgovekar
Block 4	**Normal Probability Distribution**	
Unit 8	Normal Probability Distribution	Dr. Smita Gupta

Course Coordinator : Prof. Suhas Shetgovekar, Faculty, Discipline of Psychology, School of Social Sciences, IGNOU, New Delhi

General Editors : Prof. Suhas Shetgovekar and Dr. Monika Misra, Discipline of Psychology, SOSS, IGNOU.

Acknowledgement: This is to acknowledge the contribution of Ms. Shivani Arora and Ms. Komal Beri, research scholars of Ph. D in Psychology programme, July 2018 admission cycle, to the preparation of the workbook and to the editing process of this course.

Cover Page: Ms. Arvinder Chawla

Print Production

Mr. Manjit Singh
Section Officer (Publication)
IGNOU, New Delhi

Secretarial Assistance/Graphics

Ms. Suresh Kumar
JAT/CO, SOSS
IGNOU, New Delhi

January, 2020

ISBN :

Further information on Indira Gandhi National Open University courses may be obtained from the University's office at Maidan Garhi, New Delhi-110 068 or visit University's Website http://www.ignou.ac.in.

Printed and published on behalf of the Indira Gandhi National Open University, New Delhi by Director, School of Social Sciences.

Lasertypesetted at Graphic Printers, 204, Pankaj Tower, Mayur Vihar, Phase-I, Delhi-110091.

Printed at :

BPCC-104 STATISTICAL METHODS FOR PSYCHOLOGICAL RESEARCH- I: HOW TO PROCEED IN THE COURSE?

The course on Statistical Methods for Psychological Research- I (BPCC-104) is the second core course offered in IInd Semester of BA Psychology Honours Programme, under Bachelors Degree Programme of IGNOU. The course is of 06 Credits, comprising Theory (04 Credits) and Tutorials (02 Credits). The course is divided into four blocks. Each of these blocks represents a specific theme which is discussed in terms of units. The units are arranged in a logical sequence so as to cover the main aspects of each theme.

Before proceeding to read the units, you are advised to go through instructions about how to read the course material. Given below is the explanation of the organization and sequencing of the unit.

Organization and Sequencing of a Unit

The following is the structure of each unit:

1.0 Objectives

1.1 Introduction

1.2 Section (Theme of the section)

    1.2.1 Subsection

……………….

Check Your Progress I

1.3 Section (Theme of the section)

    1.3.1 Subsection

……………….

Check Your Progress II

Let Us Sum Up

References

Key Words

Answers to Check Your Progress

Unit End Questions

As the scheme suggests, each unit is divided into sections for easy reading and better comprehension. The numbering and length of each section and subsection may vary from one unit to the other, depending upon the depth of information in each unit. Each section is indicated by **BOLD CAPITALS** and each sub-section by a **relatively smaller but bold typeface**. Divisions within the sub-sections are in **relatively smaller bold typeface** so as to make it easy for you to understand.

Let us now discuss each section of a unit.

Objectives

We begin each unit with the section Objectives. It tells you briefly about the objectives of the unit and what you will learn after you study the unit.

Introduction

The section Introduction will mainly focus on introducing the theme of the unit.

Illustration

There are several illustrations in each unit in the form of figures and diagrams. The main purpose of these illustrations is to make the study comprehensive and interesting.

Check Your Progress

We have given self-check exercises under the caption Check Your Progress at the end of main sections. You can provide your answers in the space give below each question/ exercise.You will be tempted to have a glance of the main text as soon as you come across anexercise. But we do hope that you will resist this temptation and turn to the main textonly after completing the answers.You should read each unit and note the important points in the margin provided in thecourse material. This will help in your study. It will also help you to answer the self-check exercises and the assignment questions, as well as help in revising your coursebefore appearing for Term End Examination (TEE).

Let Us Sum Up

Let Us Sum Up summarises the whole unit for the purpose of ready reference and recapitulation.

References

We have given a list of references at the end of each unit. This is a list of books and articles used by the course writers to prepare the units. This reflects that your course material is based on a wide spectrum of literature available on a particular theme, related to your course. This also informs you of the wide literature available in the particular area of study. If interested in widening your knowledge, you may look for the mentioned references. Each reference mentions the name of the author, year of publication, title of the book/article, name of publisher and place of publication.Suggested readings help you to increase your level of understanding of a particular themein each unit.

Key words

The key words at the end of the unit explain the basic ideas, technical terms and difficult words.

Answers to Check Your Progress

Answers to the exercises under Check Your Progress are given under this.

Unit End Questions

Besides Check Your Progress, we have given Unit End Questions in each unit. Practicing these questions will help you in answering assignments and Term End Examination. Though the pattern and style of questions asked may not be similar.

Audio and Video Aids

Some units have been selected for the audio and video programmes to supplement the printed material. These will help you to understand the units with greater clarity.

Apart from this, you may also access IGNOU's FM radio channel, Gyanvani (105.6 FM), which is available across many cities in India, for regular programmes, related to themes on Psychology. You can listen to the live discussions by faculty and experts on the topic of the day and interact with them through telephone, email, and through chat mode.

You may also watch Gyandarshan TV channel (free to air educational channel), for programmes related to topics on Psychology. The schedule of Gyanvani and Gyandarshan is displayed on www.ignou.ac.in. The radio and TV channels may also be accessed on Gyandhara, webcast facility for Gyanvani and Gyandarshan, provided by the University.

Tutorial

The course on Statistical Methods for Psychological Research- I will include tutorial, that is compulsory. It will be in form of activities and evaluated by your academic counsellor. You should carefully read the course material and apply the information while answering the tutorial. These activities are meant to develop your ability to relate your knowledge to day-to day life experiences. The tutorial will be made available to you along with the assignments and they have to be submitted at your study centre.

Assignment

You will receive a set of assignments for the whole programme. These are Tutor Marked Assignments, which are to be submitted to the respective Study Centre after completion. These assignments will be evaluated by academic counsellor(s) from your Study Centre. Ensure that you complete all your assignments because the grades that you get in each of these assignments are included in the final evaluation of your degree. Before answering the assignments, read all the units and additional material (if available).

Guidelines for tutorial and assignment

While working on the tutorial and assignments, kindly ensure the following points:

1) Clearly write your Enrollment number.
2) Answer them in your handwriting and in your own words **(do not copy the sentences from the course material or any other source)**.
3) Write clearly and neatly so that it is easy to read your answers.
4) Leave margins on one side of your answer-sheets so that evaluator may write his/her comments on your performance.
5) Organise your answers well based on the question asked.
6) You will submit the tutorial/ assignments at your Study Centre on or before the date mentioned as per the admission cycle. Kindly check the dates from www.ignou.ac.in or your Regional Centre website.

Term End Examination (TEE)

Consider the following points while answering TEE.

1) Questions need to be answered in **one's own words** and they need to be focused based on the questions asked.

2) Answer the questions keeping in mind the word limit.

3) Organise answers well based on the questions asked and also keep in mind any bifurcation given in the marks.

4) Ensure that you mention correct question numbers for respective answers.

Preparation of Course Material

The syllabus of course material BPCC-104 is designed by an Expert Committee (see page 2 of this course) and prepared by Course Preparation Team which comprises the author(s) of units, content editor(s), language editor, and the course coordinator. The expert committee selected the themes and sub-themes of the blocks and units, keeping in view the prescribed syllabi of UGC (CBCS model).The authors of units have provided their expertise in elaborating them in the form of the main text of each unit. The content editor has carefully examined the course contents and has made an attempt to make the material clear and comprehendible.

For any query or feedback related to the course, you may kindly contact the course coordinator at:

Prof. Suhas Shetgovekar
Room No. 121, Block-F,
School of Social Sciences
IGNOU, New Delhi
Email: sshetgovekar@ignou.ac.in

COURSE CONTENT

COURSE INTRODUCTION

Dear Learner,

In the first semester of this programme you must have developed a fair idea about basics and principles of psychology as a subject area. In the second semester, one of the courses that you will study is Statistical Methods for Psychological Research- I. In psychology both quantitative and qualitative research is carried out. When quantitative research is carried out, statistical analysis is involved and thus, one needs to have fair idea about the statistical methods, that can be used to statistically analyse the data collected.

The main objective of this course is to introduce the descriptive statistics used in psychology and to acquaint the learners with statistical computations and applications.

The present course will also be helpful in the sixth semester of this programme when you take up the course on project (BPCE144).

To provide an overview of this course, the course structure is divided in to four blocks.

Block 1: The first block is titled **'Introduction'** and it is further divided in to two units, unit 1 and unit 2. In the very first unit you will learn about what statistics is and its role in research. Limitations and misinterpretations of statistics will also be covered in this unit. Further, this unit will also introduce you to the scales of measurement and will focus on two significant terms, descriptive and inferential statistics. The second unit will focus on the data organisation and graphical representation, which is important when you have raw data with you and want to make it meaningful. The unit will cover subtopics like classification and tabulation of qualitative and quantitative data, construction of frequency distribution, cumulative frequency distribution, percentile and percentile ranks and graphical representation of data.

Block 2: The second block is titled **'Measures of Central Tendency and Variability'** and has three units, unit 3, 4 and 5. The third unit is about the measures of central tendency and will cover, concept of central tendency of data, the different measures of central tendency, namely, mean, median and mode, their properties, advantages, limitations and their computation. Unit four is titled measures of variability and deals with the concept of variability in data, the different measures of variability, viz., range, quartile deviation, average deviation, standard deviation and variance,their properties, advantages and limitations. Unit five will deal with the computation of various measures of variability.

Block 3: The next block, that is Block 3, is titled **'Correlation'** and includes two units, unit 6 and unit 7. Unit six will focus on the concept of correlation, its direction and magnitude, its properties, uses and limitations and unit seven will discuss the computation of coefficient of correlation with the help of Pearson's product moment coefficient of correlation and Spearman's rank order correlation.

Block 4: The last block is titled '**Normal Probability Distribution'**, that consists of unit 8, which will cover the concept of probability, the concept, nature and properties of normal probability distribution, concept, properties and computation of standard scores and divergence from normality, namely, kurtosis and skewness.

It is very important that you read all the units very carefully and try to understand the subject matter. Knowledge of statistics is important in Psychology as the same can be applied to data analyse data while carrying out quantitative research.

Some suggestions and tips to enhance your study of this course are as follows:

1) Ensure that your basic concepts, given in this course, are clear, whether it is the term statistics or terms mean, median, mode, normal probability curve etc. If you don't understand the terms, read again. The first unit is a foundation to rest of the units, so ensure that you read and learn this unit first.

2) Another important aspect is to understand the formulae and also the steps in computing various statistical techniques. These have been discussed with the help of examples. Practicing these steps along with the formulae will help you learn them in a better way.

3) Besides the above, you also need to develop an understanding about when to use which technique. For example, when to compute mean, when to compute median, when to compute Pearson's product moment correlation and when to compute Spearman's rank order correlation. While reading the units, try to focus on when a certain statistics technique is computed.

4) Answering the exercises given in Check Your Progress (given in each unit) as well as in the Workbook (given at the end of the booklet) will help you practice and learn the subject matter and the method for computing various statistical techniques.

10 Blank

BLOCK 1
INTRODUCTION

BLOCK INTRODUCTION

This is the very first block of this course that will provide introduction to statistics in psychological research. This block is divided into two units. **Unit 1** covers introduction to statistics. The main topics that will be discussed are meaning of statistics, thats includes the definition and also the basic concepts in statistics. The unit will also cover the role of statistics in research. Further, the unit also deals with the limitations and misinterpretations of statistics. The unit also focuses on the scales of measurement, namely, nominal, ordinal, interval and ratio that are explained with the help of examples. Lastly, the unit discusses about descriptive and inferential statistics. Discussion on descriptive statistics is especially important as it serves as a foundation to the successive units.

Unit 2 in this block is titled, Data organisation and graphical representation and these are two relevant aspects of descriptive statistics. The unit starts with discussion on classification and tabulation of quantitative and qualitative data. The unit then moves on to explaining frequency distribution and its computation. Cumulative frequency distribution will also be discussed in this unit. The unit further focuses on yet another significant topic that is percentile and percentile ranks. Lastly, the graphical representation of data will be discussed in detail with a focus on various types of graphs and the steps involved in constructing them.

Check Your Progress and Unit End Questions are also included that will help you understand the subject matter in a better way.

UNIT 1 INTRODUCTION TO STATISTICS*

Structure

1.0 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.

1.1 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.

1.2 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.

1.2.1 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 1.2.2.

1.2.2 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: This term can be used to describe the persons, objects, elements, animals or even reactions that display a pattern of characteristics that is unique. It can also be explained as set of persons, objects, elements animals, reactions that the researcher wants to study. If a researcher wants to carry out a study on adolescents in New Delhi, then his/ her population will be all the adolescents in New Delhi. Population can be finite or infinite in nature (Mohanty and Misra, 2016, page 3). An example of finite population is number of students in a school who have failed in mathematics. And an example of infinite populations would be number of stars in the sky.

Sample: In simple terms sample can be defined as the group of individuals who participate in the research. If we take the example of adolescents in New Delhi, discussed above, it is not feasible for the researcher to contact and collect data from all the adolescents in New Delhi. Thus, the researcher will take a sample (preferably representative) from that population. To take another example, in a cold drink factory, if the quality inspector wants to find out if the quality and taste of the cold drink is adequate, he/she will not test all the bottles of cold drink, but take a sample and test the same and preferably this sample is taken randomly.

Thus, a sample is a smaller group from the population that participates in the research. It is important that the sample is representative of the population, that is, it is as much as possible similar to the population or possesses the same characteristics or elements as the population (Mohanty and Misra, 2016). Thus, sampling techniques are relevant in research, which help in selection of sample. Sampling techniques can be categorised in to probability sampling and non probability sampling (refer to table 1.1).

Table 1.1: Sampling Techniques

Probability Sampling	Non-probability Sampling
Simple random sampling	Convenient sampling
Systematic random sampling	Voluntary sampling
Stratified random sampling	Judgement sampling
Cluster sampling	Quota sampling
Multistage random sampling	Snowball sampling

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 1.2.

Table 1.2: 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. Variable means something that varies. It can also be explained as quantity or a number that will vary or will have different values. In the introduction section, a study on emotional intelligence and self esteem of adolescents in India was mentioned. In this study, emotional intelligence and self esteem can be termed as variables. Emotional intelligence can be high or low as can self esteem be. Both these variables can possess varied values. Even gender can be termed as a variable because it will vary in terms of males or females. Different types of variables have been discussed in table 1.3.

Table 1.3: Types of Variables

Type	Description	Example
Independent Variable (IV)	Variable that is manipulated by the researcher is independent variable.	In a study on effect of light on performance of individuals, a researcher can manipulate light to bright, dim or normal. Light can be an example of independent variable.
Dependent Variable (DV)	In a research, variable that is measured for any changes when independent variable is manipulated is dependent variable.	In the above example, performance is an example of dependent variable.
Extraneous Variable (EV)	Variable that may impede or interfere in the relationship between independent variable and dependent variable is called extraneous variable.	In the above example, noise can interfere in relationship between IV and DV and it is possible that changes in DV, that is performance is due to EV, that is, noise rather than IV, that is, light.
Quantitative Variable	This is a variable that is numerically represented.	Intelligence Quotient (IQ), weight, height.
Qualitative Variable	This has measurable characteristics, that are not numerical but categorical.	Gender (Male and Female), Socio Economic Status (High and Low), Religion (Christian Hindu, Muslim).
Continuous Variable	Such a variable has any value and is continuous in nature.	Weight: 56.98 kg, Age: 2 years 5 months.
Discontinuous or discrete Variable	This is a set of integers that is distinct.	Number of children, number of two wheelers.

Check Your Progress I

1) Define Statistics

....

....

....

....

....

2) Explain population and sample with suitable examples.

Population	Sample
Example:	Example:

3) Discuss independent variable and dependent variable with suitable examples.

....

....

....

....

4) Identify the symbols

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

1.3 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.

....

....

....

....

....

1.4 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.

....

....

....

....

....

1.5 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 1^{st} rank may have obtained 95 marks, whereas a student obtaining 2^{nd} rank may have 80 marks and a student obtaining 3^{rd} 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 1.4 for properties and examples of the four scales of measurement

Table 1.4: 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.

....

....

....

....

....

1.6 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.

1.6.1 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.

1.6.2 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 adolescents on emotional intelligence was higher than the mean scores obtained by the male adolescents. Thus, indicating that the female adolescents had higher emotional intelligence than male adolescents. 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:

1.6.2.1 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 '$\overline{x}$. 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.

1.6.2.2 Hypothesis Testing

One of the significant aspects of research is hypothesis testing. Hypothesis is a tentative statement that is investigated during the process of research. Hypothesis pertains to certain phenomenon and is based on a theory. Data is then collected by the researcher in order to validate the hypothesis. Thus, the hypothesis is rejected or accepted by a researcher based on the results obtained. Hypothesis testing is also referred to as "a process involving statistical decision making with regard to population value that is based on the sample value" (Veeraraghavan and Shetgovekar, 2016, page 9). Though, while drawing inferences one needs to ensure that hypothesis is not incorrectly accepted or rejected due to results that are influenced by chance factor or confounding factors.

Types of Hypothesis: There are two types of hypothesis:

Null hypothesis: Once a researcher finalises his/her research topic and variables after reviewing literature, he/ she will have to formulate hypothesis based on the problem statement. In a research attempts are made by the researcher to reject the null hypothesis(es). Null hypothesis is also referred to as hypothesis of 'no difference' as null hypothesis implies that there is no difference between two groups. For example, if a researcher wants to study if there is significant difference in job satisfaction of government and private bank employees, he/ she will formulate a null hypothesis as 'No significant difference exists in job satisfaction of public and private bank employees'. And then based on the results obtained the hypothesis will be rejected or accepted.

Alternative hypothesis: Alternative hypothesis can be termed as a counter proposition to the null hypothesis (Veeraraghavan and Shetgovekar, 2016). It is formulated by a researcher based on a theory. The hypothesis states that there is a significant difference between two groups. For example, the researcher can formulate the alternative hypothesis as 'There will be a significant difference in job satisfaction of public and private bank employees'. This is also a neutral alternative hypothesis that does not provide any direction and thus is termed as nondirectional hypothesis. Alternative hypothesis can also be directional. For example, 'Public bank employee have higher (or lower) job satisfaction when compared with private bank employees'. Such hypothese are formulated based on the review of literature.

Steps in Hypothesis testing: The steps involved in hypothesis testing are as follows:

Step 1: Null hypothesis(es)/ alternative hypothesis(es) are specified by the researcher.

Step 2: A level of significance is selected. This level of significance could be 0.05 level or 0.01 level. The term significance in the context of statistics implies 'probably true' that indicates that results are free from chance factor at the specified level of significance. For example, if there exists a significant difference in job satisfaction of public and private bank employees then the researcher could either be 95% (0.05 level of significance, $P<0.05$) confident or 99% (0.01 level of significance, $P<0.01$) confident about the results obtained. ['P' stands for probability value]. This is because when carrying out research with human participants, 100% accuracy cannot be achieved. Thus, there could be 5% or 1% chance that the results are due to chance or confounding factors. Whether the null hypothesis is accepted or rejected will depend on whether the statistical value obtained after data analysis is more or less than the table value (tables for various statistical techniques are provided at the end of any book on Statistics) specified at 0.05 or 0.01 level of significance. If the obtained value is higher than the table value, then the null hypothesis is rejected, and if the obtained value is less than the table value then the null hypothesis is accepted.

Type I and Type II Errors: In the process of hypothesis testing, two errors could occur, type I and type II errors. These will be more clear from table 1.5.

Table 1.5: Type I and Type II Errors

	Null Hypothesis is true	**Null Hypothesis is false**
Null Hypothesis is Rejected	Type I error	Decision is correct
Null Hypothesis is Accepted	Decision is correct	Type II

As can be seen in table 1.5 a researcher will be making a correct decision when a false null hypothesis is rejected and when a null hypothesis that is true is accepted. However, it may so happen that a null hypothesis is rejected even when it is true and this is termed as type I error. On the other hand, when a false null hypothesis is accepted then it is termed as type II error.

Step 3: Based on the parameter specified in the null hypothesis (es), the statistic is calculated. A sample is taken by the researcher and data is collected. Statistic that is thus obtained from the sample (representative) is then used in order to make an estimation about the population parameters.

Step 4: Decision is made whether to accept or reject the null hypothesis (es). In this regard, the P value or probability level is selected as has been discussed under step 2 and accordingly decision is taken by the researcher. Whether to accept or reject the null hypothesis.

Check Your Progress V

1) List the three categories of descriptive statistics.

....

....

....

....

....

2) What is null hypothesis?

....

....

....

....

....

3) Explain Type I and Type II Errors.

Type I Error	Type II Error

1.7 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.

1.8 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.

1.9 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).

Population: This term can be used to describe the persons, objects, elements, animals or even reactions that display a pattern of characteristics that is unique.

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.

Sample: Sample can be defined as the group of individuals who participate in the research.

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

1.10 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) Explain population and sample with suitable example.

Population	Sample
It can be explained as set of persons, objects, elements animals, reactions that the researcher wants to study.	Sample can be defined as the group of individuals who participate in the research.
Example: Working Women in Mumbai	Example: The working women (N= 200) included in a study.

3) Discuss independent variable and dependent variable with suitable example.

Variable that is manipulated by the researcher is independent variable. In a research, variable that is measured for any changes when independent variable is manipulated is dependent variable. For example, in a study on effect of temperature (high and low) on performance, temperature is the independent variable and performance is the dependent variable.

4) 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.

1) 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.

2) 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.

2) What is null hypothesis?

Null hypothesis indicates that there is no significant difference between two groups.

3) Explain Type I and Type II Errors.

Type I Error	Type II Error
Occurs when the null hypothesis is true but we reject it.	Occurs when the null hypothesis is false but we accept it.

1.11 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.
5) Explain hypothesis testing with a focus on steps involved in hypothesis testing.

UNIT 2 DATA ORGANISATION AND GRAPHICAL REPRESENTATION*

Structure

2.0 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.

2.1 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.

2.2 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 aboutcategorical variables. Some examples of qualitative data are the smoothness ofyour skin, and the colour of your eyes, the texture of your hair, the softness ofyour 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.

2.2.1 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 2.1).

Table 2.1: 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.

2.2.2 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 2.2, 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 2.2: 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%
50-75 (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.

....

....

……………………………………………………………………………………………

……………………………………………………………………………………………

……………………………………………………………………………………………

2.3 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 for that class interval and for 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 2.3, that provides examples of the three types of frequency distribution.

Table 2.3: 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 2.4:

Table 2.4: 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 2.5.

Table 2.5: 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, grouped frequency distribution and ungrouped frequency distribution. The computation for both these frequency distributions are discussed as follows:

2.3.1 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 2.6: 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
		$\sum = 40$

Please note that the total ($\sum$) 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.

2.3.2 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 statisticalprocedure 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*

2.4 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 35 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.

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

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 frequencyof 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 frequencies can easily be converted to cumulative percentage frequencies by carrying out multiplication between the cumulative frequencies 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

2.5 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 90th percentile in math ability and 84th 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 90th 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 50th percentile or P_{50} is like the median. Above 50th 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).

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

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

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

And the formula for 10th percentile would be

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

And the formula for 75th percentile would be

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

Let us now compute percentile with the help of an example given in table 2.7.

Table 2.7: 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

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

$$= 100- (100 \text{ X } 4- 50/ 10)$$

$$= 100- (400-50/ 10)$$

$$= 100- 350/10)$$

$$= 100- 35$$

$$= 65$$

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 eachof 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:

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

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

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

$$= 100/\ 50\ [16 + (35-29.5/10) \times 6]$$

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

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

$$= 2 \times 19.3$$

$$= 38.6$$

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

2.6 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.

2.6.1 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 2.1, which is based on the table 2.1 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 2.1 will look as follows:

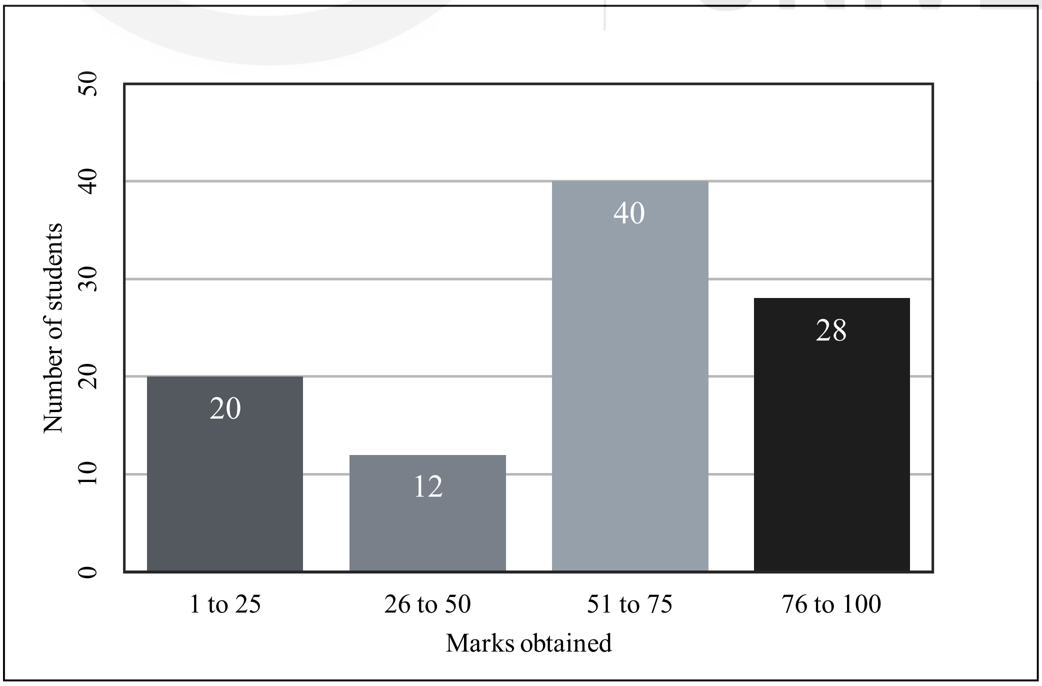

Fig. 2.1: Bar Graph

2.6.2 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 2.8. 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 2.2. 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 2.8: 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

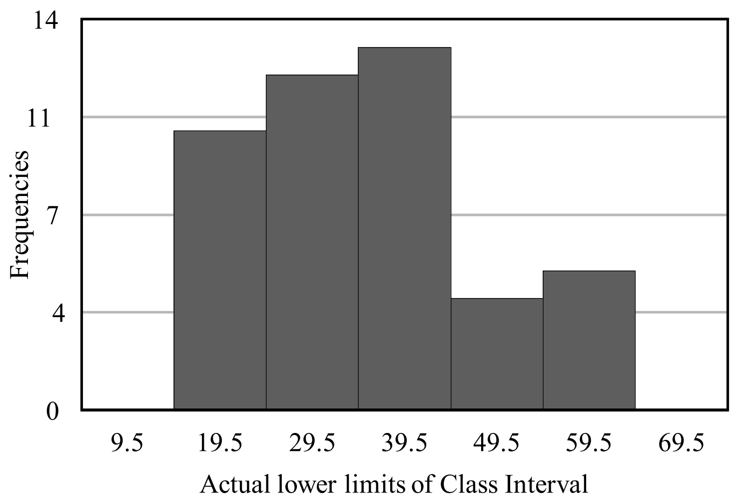

Fig. 2.2: Histogram

2.6.3 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 2.4.

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 2.3.

Table 2.9: 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

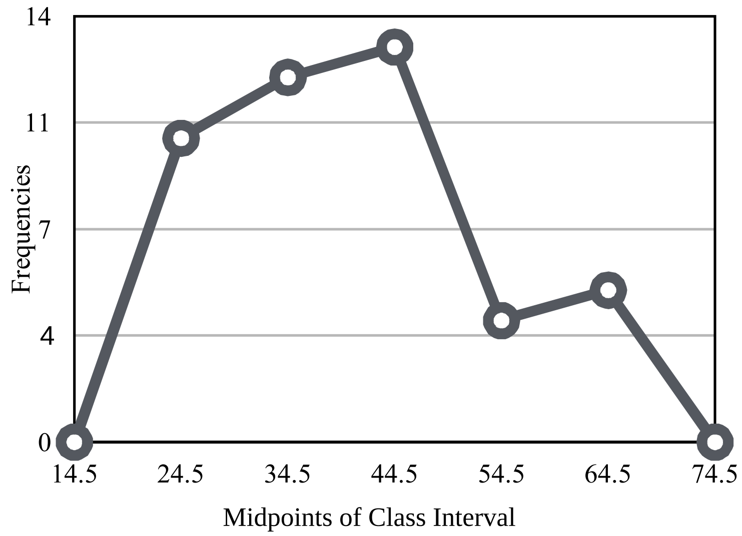

Fig. 2.3: Frequency Polygon

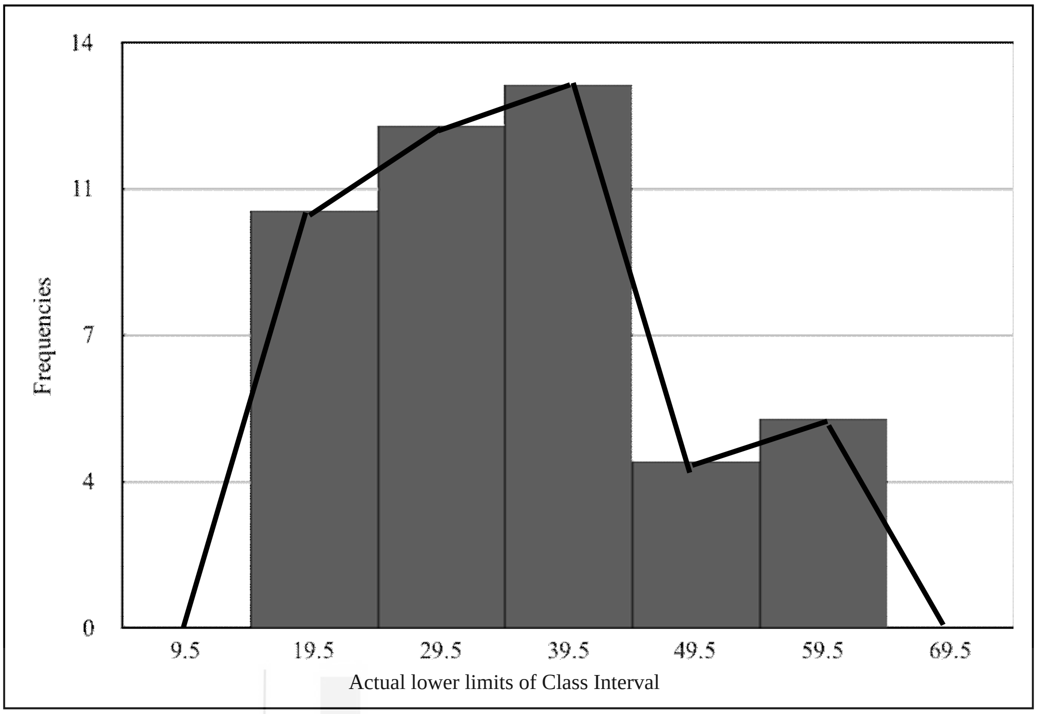

Fig. 2.4 : Frequency Polygon drawn with the help of Histogram

2.6.4 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 2.10 : 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

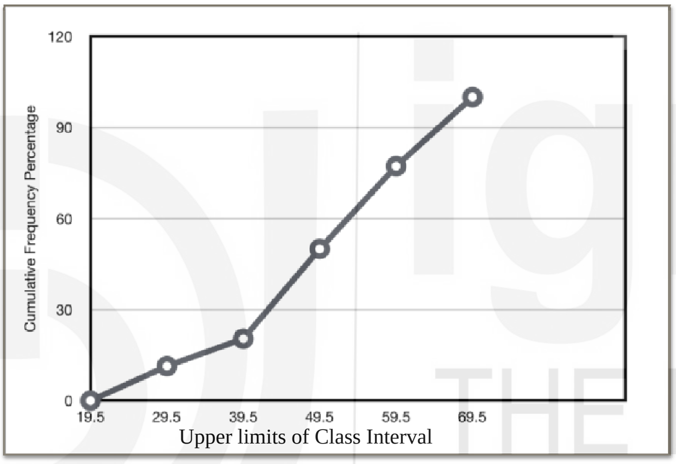

Fig. 2.5: Cumulative Frequency Percentage Curve or Ogive

2.6.5 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 2.11: 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

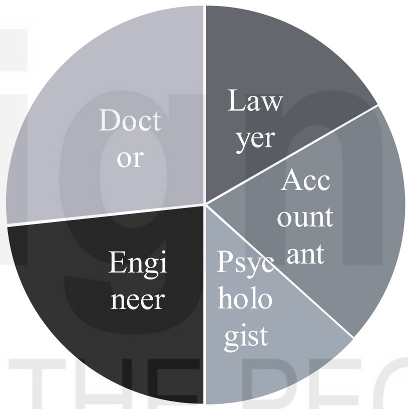

Fig. 2.6: Circle Graph pr Pie Chart

Check Your Progress V

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

....

....

....

....

....

2) What is a pie chart?

....

....

....

....

....

2.7 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.

2.8 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* (6^{th} ed.). Pearson Education

2.9 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.

2.10 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.

2.11 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.